

Spacer: Towards Engineered Scientific Inspiration

Asteromorph*

arXiv:2508.17661v1 [cs.AI] 25 Aug 2025

Recent advances in LLMs have made automated scientific research the next frontline in the path to artificial superintelligence. However, these systems are bound either to tasks of narrow scope or the limited creative capabilities of LLMs. We propose *Spacer*, a scientific discovery system that develops creative and factually grounded concepts without external intervention. *Spacer* attempts to achieve this via ‘deliberate decontextualization,’ an approach that disassembles information into atomic units—keywords—and draws creativity from unexplored connections between them. *Spacer* consists of (i) NURI, an inspiration engine that builds keyword sets, and (ii) the Manifesting Pipeline that refines these sets into elaborate scientific statements. NURI extracts novel, high-potential keyword sets from a keyword graph built with 180,000 academic publications in biological fields. The Manifesting Pipeline finds links between keywords, analyzes their logical structure, validates their plausibility, and ultimately drafts original scientific concepts. According to our experiments, the evaluation metric of NURI accurately classifies high-impact publications with an AUROC score of 0.737. Our Manifesting Pipeline also successfully reconstructs core concepts from the latest top-journal articles solely from their keyword sets. An LLM-based scoring system estimates that this reconstruction was sound for over 85% of the cases. Finally, our embedding space analysis shows that outputs from *Spacer* are significantly more similar to leading publications compared with those from SOTA LLMs.

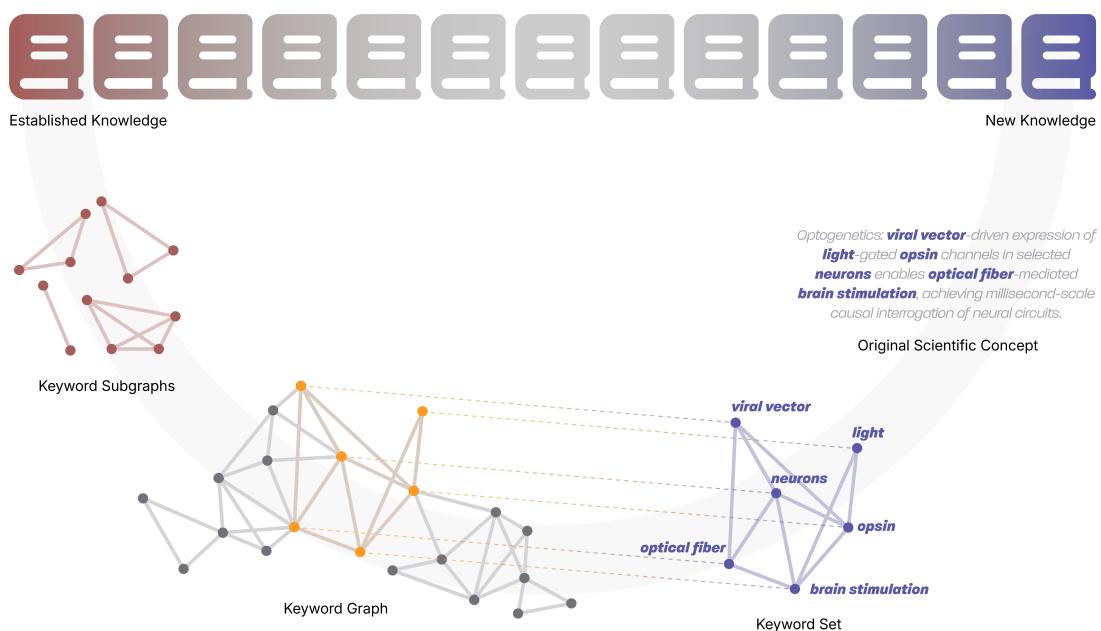


Figure 1: Schematic of *Spacer*'s approach to engineered scientific inspiration.

*See [Contributions and Acknowledgments](#).

Contents

1	Introduction	3
2	Spacer	4
2.1	Overall Approach	4
2.2	Architecture	5
3	Results	9
3.1	Restoring Calcium Oscillations in Hepatocellular Carcinoma	9
3.2	ATP Allocation Patterns Predict Cellular State Transitions	11
3.3	Overexpressing Olfactory Receptors for Gut Microbiome Control	13
4	Validations	15
4.1	Nuri	15
4.2	Reconstructions of Latest Cutting-edge Scientific Concepts	17
4.3	Sentence Embeddings Analysis	19
5	Discussion	22
6	Technical Details	24
6.1	System Designs	24
6.2	Model Specifics	24
6.3	Hardwares	24
	References	25
	Contributions and Acknowledgments	31
	Appendix	31
A	Prompts	31
B	Example of Experimental Protocol by Grok 4	44
C	Data Specifics	48
D	Supplementary Materials	48

1. Introduction

Throughout history, scientific breakthroughs have emerged from the conjunction of seemingly disparate fields of knowledge [1–4]. Optogenetics [5–7] has revolutionized neuroscience by introducing light-mediated modulation in neural cells.; CRISPR-Cas9 [8–10] has changed the landscape of biological research by reinterpreting the bacterial immune system as a platform for genome editing. Thomas Kuhn characterized such moments as ‘paradigm shifts,’ arguing that these fundamental reorientations of scientific understanding cannot emerge through incremental progress. Over the last few decades, academia has witnessed an unprecedented surge in the sheer volume of scholarly publications [11, 12]; despite this, innovations on par with Kuhn’s portrayal have been rare [13].

Recently, large language models (LLMs) have garnered traction as potential galvanizers of creativity. Supporting this sentiment is the notable performance of LLMs in various benchmarks that measure capabilities in science, programming, and reasoning [14–18]. There have already been several attempts to capitalize on these advancements by creating agentic frameworks for scientific discovery [19–25]. AlphaEvolve [26] finds solutions for local optimization problems, and a multi-agent ‘Virtual Lab’ [27] uncovers molecule-level candidates for SARS-CoV-2 nanobodies.

However, we remain skeptical about whether systems that rely solely on LLMs can initiate artificial paradigm shifts. Transformer-based architectures optimize contextual coherence and penalize outputs that deviate from established patterns. Skewed evaluation metrics and human feedback further reinforce this behavior, leading to a complementary systematic bias that favors soundness over novelty [28, 29]. The problem is that while building upon patterns, attempts at creative thinking easily degrade into regressions. For instance, the term ‘CRISPR-Cas9’ often recurs when prompted to generate research ideas, due to the overrepresentation of the CRISPR-Cas9 technology as novel research in training datasets. As a result, LLM outputs tend to slant toward the precursory prompt context and the training data, implying that an automated ideation system powered by LLMs must overcome the limitations introduced by contextualization.

To this end, we decompose information into atomic units: keywords. As keywords do not carry excess context, we may leverage the reasoning abilities of LLMs while circumventing the aforementioned limitations. This decomposition also allows for the compositional construction of scientific concepts, where keywords act as versatile building blocks that comprise diverse combinations. Known concepts correspond to specific combinations of keywords that come together into a graph: one that we believe spans humankind’s ever-growing knowledge base. Furthermore, we assert that we can automate paradigm-shifting discoveries through the structural examination of this graph.

As such, we propose *Spacer*, a two-step scientific discovery system designed to foster probable but unexplored connections between seemingly unrelated concepts. NURI, an inspiration engine, extracts potent sets of keywords from the global keyword graph. The Manifesting Pipeline then substantiates each set into a concrete scientific statement that potentially expands the boundaries of science.

2. Spacer

Spacer is a **scientific discovery system**. The goal of *Spacer* is to generate original scientific concepts distinct from those seen in its training datasets or search results, while complying with academic standards. Whereas next-token prediction models have been unsuccessful in creating original scientific concepts, *Spacer* overcomes their limitations by operating on a multi-stage pipeline with both LLM and non-LLM components.

2.1 Overall Approach

The core conceptual and technical backbone of *Spacer* is the **deliberate decontextualization** of information. We define deliberate decontextualization as the intentional removal of inferable contextual information: sentences, paragraphs, or any structured text that would encourage the language model to rely on familiar patterns or established knowledge. We have approached the problem of developing *Spacer* based on the decontextualization of information for the following two reasons. First, significant scientific breakthroughs often emerge from unexpected connections between seemingly unrelated concepts, rather than from next-step extensions of existing knowledge. Second, decontextualization enables a multi-LLM framework to generate contexts beyond its knowledge span.

LLMs' next-token prediction engines are antithetical to the concept of decontextualization. As mentioned above, the outputs of LLMs are constructed upon heavy contextual bias when given a specific task. While excelling in logical writing and elaboration, they cannot make new conceptual connections necessary for scientific inspiration. Therefore, *Spacer* conceives scientific inspiration before introducing LLMs. A non-LLM component first generates novel connections between concepts from decontextualized information, then passes them to LLMs for construction into scientific concepts. Completely separating the ideation from LLMs, the workflow preserves the emergent properties while developing scientific significance using multi-agent LLM frameworks.

The atomic unit of decontextualization in *Spacer* is a word. That is, we seek original knowledge in the combinations of scientific terms. Regarding our assertion that unexpected conceptual connections can lead to scientific breakthroughs, a set of scientific terms is a fine starting point for ideas to accumulate. From the perspective of format, words—the results of decontextualization—are the optimal point for LLMs to operate on. A token does not preserve the original meaning of words; larger units, such as phrases or sentences, carry contextual bias. Therefore, a set of words is a format that manifests scientific inspiration while allowing LLMs to understand and develop meaning without contextual bias.

To summarize, we deliberately decontextualize scientific information to free LLMs from contextual dependencies. We adopt an hybrid approach where LLMs only build upon a given set of scientific words that already contains emergent knowledge. Such an approach enables the system to access novel scientific discoveries while also maintaining logical soundness and scientific compatibility. This meet-in-the-middle approach is implemented in *Spacer* as three consecutive stages: word selection, word combination¹, and agentic reinforcement. Scientific knowledge is decontextualized into words, then selected to generate new connections and findings.

All emergent characteristics of *Spacer* originate from the word selection stage. The following steps involve LLMs to combine words into sentences that reveal the idea and reinforce it with rationales and logical refinement. The specific architecture and design of the frameworks are discussed in depth in the following section.

¹One might argue that involving LLMs in the word combination stage is contradictory to the principle of deliberate decontextualization. However, since *Spacer* devotes inspiration to the word selection stage, there are no ‘creative’ tasks involved in word combination.

2.2 Architecture

We first define the key terms used throughout this report, each corresponding to a different level of scientific discovery. **Keywords** are individual decontextualized words that carry specific meaning. Biological entities, scientific techniques, or established concepts may all serve as keywords. **Theses** are individual paragraphs aiming to capture a researcher’s initiative. These may resemble abstracts but omit resultant details such as findings, data, expected impacts, or speculative outcomes. Finally, when Theses are organized into structured formats, they form **Statements**: 1–2-sentence-length concepts accompanied by a supporting set of rationales.

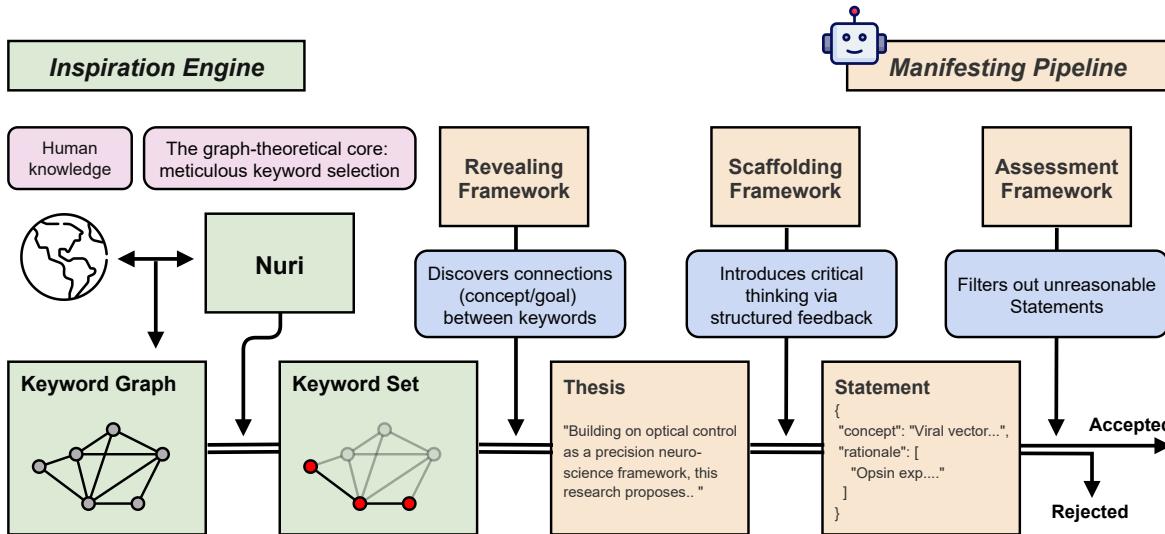


Figure 2: Architecture of *Spacer*.

Spacer consists of four primary components organized in a sequential pipeline. **NURI** builds a set of keywords that involve semantical richness while minimizing contextual dependencies. The **Revealing Framework** transforms these keywords into Theses, which are then elevated to Statements by the **Scaffolding Framework**. Lastly, the **Assessment Framework** evaluates Statements to accept or reject them based on their overall plausibility. This design allows *Spacer* to separate creative generation from critical evaluation—NURI is devoted to addressing ‘creativity’ while the others focus on materializing a Statement without subverting the original concept. This architecture is illustrated in Figure 2.

The continuing sections establish each of the four components in detail.

Nuri

NURI² is a graph-based Keyword Set extraction algorithm. We designed NURI to select Keyword Sets, as connections between keywords reveal information not evident in individual keywords. We believe that complex scientific ideas—such as the motivations behind existing research papers—can be represented in the form of these connections. Accordingly, we designed NURI to collect Keyword Sets that are likely to harness novel research directions that lead to impactful research papers.

For published papers, we employed the Field-Weighted Citation Impact (FWCI) as a primary metric for impactfulness. FWCI compares the number of citations a paper receives against the average number of citations in the same field and year. For example, an FWCI value greater than 1 indicates that the paper has been cited more frequently than average. As citation practices differ across fields, FWCI offers a compensation that allows for a fair comparison of scientific impact. We had also considered using the Relative Citation Ratio (RCR), another index that rates impactful papers based on citation rate counts [30]. However, given that these two indices are mostly interchangeable with each other [31], we chose

²The name Nuri originates from the Korean word for ‘world’.

FWCI over RCR for its abundance in precomputed values.

To be precise, NURI makes a graph with set of papers, and an evaluation function which can determine whether a Keyword Set would be impactful. For a given set of papers P , let $\mathcal{K}(p)$ be the Keyword Set of a paper $p \in P$. We define the undirected and weighted graph $G(P)$ as

$$V(G(P)) = \bigcup_{p \in P} \mathcal{K}(p), \quad (1)$$

$$w(u, v) = \sum_{\substack{p \in P \\ u, v \in \mathcal{K}(p)}} \frac{\log_2(\text{FWCI}(p) + 1)}{|\mathcal{K}(p)| - 1}. \quad (2)$$

The weight of each edge represents the joint academic impact of the two keywords as a sum of normalized logarithmic FWCI values of the papers containing both vertices. The evaluation function f_P takes a Keyword Set $K \subset V(G(P))$ and outputs a normalized score $s \in [0, 1]$, representing the potential impact of the input keywords. NURI uses f_P and applies various heuristics to create its final output set. We note that NURI does not invoke any machine learning methods or LLMs in its process, nor does it require any input from the user.

The Revealing Framework

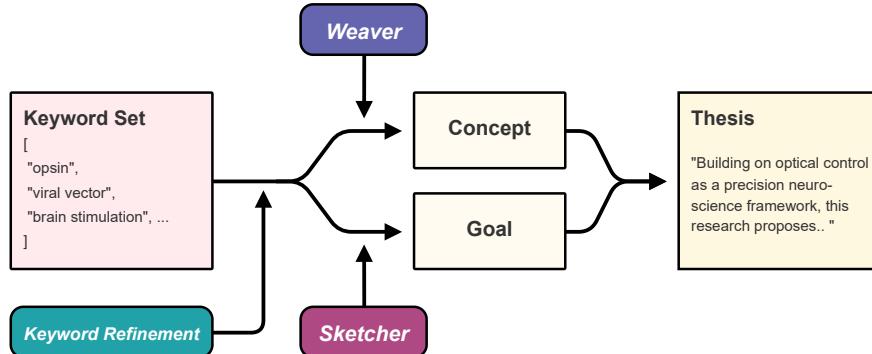


Figure 3: Schematic of the Revealing Framework.

The Revealing Framework takes a Keyword Set and finds a plausible interconnection between them, forming a Thesis. The input keywords selected by the preceding engine—NURI—have the potential to establish unexplored scientific propositions. When applied to such keywords, the Revealing Framework realizes the latent concept in the connections and formulates it into communicable natural language.

The primary component of the Revealing Framework is WEAVER. WEAVER is an LLM trained on academic papers, fine-tuned to reconstruct research initiatives from segments of these papers. Taking a Keyword Set from NURI, WEAVER materializes a sentence-length research concept from the knowledge concealed within the word set. As shown in Section 4.2, WEAVER yields logical, novel, and plausible research initiatives from well-structured input keywords. To stabilize the quality of outputs from WEAVER, we applied a keyword refinement engine to the original Keyword Set, vetting and replacing inadequate or unnecessary keywords before the set could enter WEAVER.

However, WEAVER’s outputs often lack a clear directive necessary for impactful research. As such, the Revealing Framework utilizes another fine-tuned LLM named SKETCHER in parallel to WEAVER. SKETCHER is responsible for providing an overarching purpose to the resulting Thesis by generating a sentence-length research goal based on the given keywords. While SKETCHER is trained to suggest a broad and unbiased research goal, in application, it is supplied with the same refined Keyword Set as WEAVER for scientific coherence.

Combining the research concept and the research goal, an untrained LLM generates a final paragraph-long Thesis. The prompt used here was designed to avoid any bias towards specific academic fields or the research concept while retaining the insights and contributions originating from the preceding stages. Putting everything together, the Revealing Framework reliably reveals new knowledge in the form of Theses from the Keyword Sets proposed by NURI.

The Scaffolding Framework

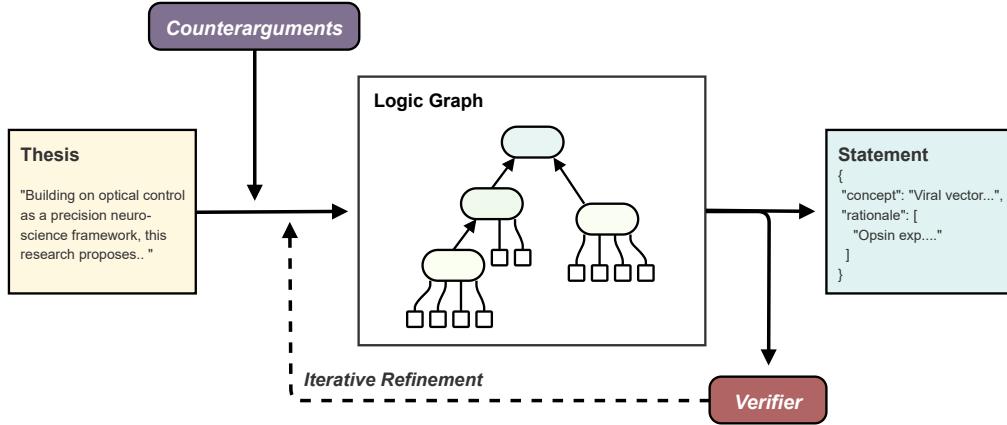


Figure 4: Schematic of the Scaffolding Framework.

The Scaffolding Framework assembles Statements from their unstructured counterparts, notably supplementing them with validated evidence. This framework employs logic graphs for its core mechanism. Pieces of information—such as key concepts, supporting evidence, and intermediate conclusions—comprise typed vertices; relationships between these pieces form directed edges, constituting the logic graph. A multiple-stage process built around this graph yields a Statement with a single concept and a set of rationales, where the concept is a 2–3-sentence-length summary of the Thesis, and each rationale is a description that validates the concept.

The need for a structured format emerges because LLMs yield suboptimal results when repeatedly processing context-heavy natural language. While the Revealing Framework generates content rich in ideas, we observed that its irregular composition negatively impacts structural refinement and evaluation due to inherent biases. When reading a paragraph, LLMs focus on connotations and nuances rather than the logical structure. This phenomenon exacerbates in multi-agent feedback loops where irrelevant preferences amass and ultimately diverge from the constructive criticism we desire. The Scaffolding Framework plays a crucial role in avoiding these shortcomings.

The first step in the Scaffolding Framework is the augmentation phase, where the Theses are challenged with counterarguments, each using peer-reviewed literature based on predefined criteria. The Theses are then augmented to account for the identified points while preserving the original approach.

The following step is the graph iteration phase, where the paragraph is decomposed into a logic graph. This phase iteratively reinforces the logic graph while verifying each revised graph against known facts.

When the graph iteration is deemed complete, the logic graph is reconfigured into a Statement in the final step. The central vertices map to the concept, and the surrounding nodes translate into the rationales. As such, the Scaffolding Framework rearranges the emergent idea from the Revealing Framework into Statements that are factual and logically sound.

The Assessment Framework

The Assessment Framework is a system that evaluates the validity of Statements generated by the Scaffolding Framework. Assessing scientific validity is inherently difficult; it depends on complex, context-specific factors, and rigid evaluation criteria often impose substantial bias, making consistent judgments elusive. To mitigate this, the Assessment Framework is designed to evaluate Statements in two subsequent phases: exploratory analysis and specified inspection.

In the exploratory analysis phase, a reviewer LLM agent produces several critiques of a given Statement. To avoid constraining the perspective of this agent, it is not provided with an explicit list of criteria. Rather, the LLM agent is instructed to generate an unconstrained number of critiques together from a broader viewpoint, along with supporting rationales.

These reviews are carried forward to the specified inspection phase, where a meta-reviewer LLM agent evaluates them against predefined criteria, including practical feasibility, methodological rigor and fundamental scientific plausibility. Each critique is independently assessed on how severely the flaw undermines the Statement's scientific integrity. By steering the exploratory analysis clear of contextual biases, we can ensure that the evaluation correctly reflects the overall context and remains unbiased by specific criteria.

Following the process above, Statements that clear all initial critiques without substantial issues are determined to be scientifically sound and technologically feasible. In human-expert evaluations, the Assessment Framework achieved an 88.2% recall, highlighting its ability to detect latent flaws in superficially plausible yet scientifically unsound Statements. In this way, it ensures that the Statements from the Scaffolding Framework remain not only innovative but also firmly grounded in scientific reality, thus enhancing the overall quality and reliability of the research outputs.

3. Results

In this section, we showcase the Statements generated by *Spacer*. Each output follows a structured format, consisting of the core concept, supporting rationales, and references. For illustration, we present three examples with brief explanations and comments. The explanations are based solely on the raw outputs with only minimal background information provided for clarity.

These examples demonstrate *Spacer*'s capability for scientific discovery. Through deliberate decontextualization and sophisticated reasoning, *Spacer* synthesizes interdisciplinary concepts from a Keyword Set while preserving scientific rigor. This highlights *Spacer*'s potential to provide scientific inspirations not limited to simple extensions of the current art. Further examples are available in our [Github repository](#).

3.1 Restoring Calcium Oscillations in Hepatocellular Carcinoma

```
{
  "concept": "Controlled noise injection into calcium signaling pathways may restore oscillatory coherence in hepatocellular carcinoma cells through stochastic resonance mechanisms. This approach could potentially re-establish calcium-dependent cell cycle checkpoint regulation that appears disrupted in malignant hepatocytes compared to normal physiological fluctuation patterns.",
  "supporting_dois": [
    "10.1002/ijc.21591",
    "10.1002/jcb.1140",
    "10.1002/jnr.1120",
    180 more DOIs...
  ],
  "rationale": [
    "Hepatocellular carcinoma cells exhibit fundamentally altered calcium homeostasis and signaling dynamics compared to healthy hepatocytes. These cancer cells show disrupted calcium oscillations, altered NF $\kappa$ B signaling patterns, decreased expression of plasma membrane Ca $^{2+}$ -ATPase leading to sustained high intracellular calcium levels, and disrupted circadian rhythms that normally coordinate calcium oscillations with cell cycle checkpoints.",
    "Stochastic resonance represents a fundamental principle where introducing optimal levels of noise enhances temporal coherence of weak periodic signals in nonlinear systems. In biological contexts, controlled introduction of specific noise patterns can amplify oscillatory coherence, and non-Gaussian noise patterns have been demonstrated to optimize temporal patterns in calcium signaling networks.",
    "Hepatocyte calcium signaling networks possess the essential nonlinear dynamics required for stochastic resonance effects. Theoretical models demonstrate that intracellular calcium systems in sub-threshold, excitable states can exhibit coherence resonance, where optimal noise levels enhance regularity of intrinsic oscillations. These systems approximate real cell dimensions necessary for meaningful stochastic resonance effects.",
    "Calcium oscillation coherence serves as a critical signaling code that encodes information determining activation of specific downstream genetic programs controlling cell cycle progression. The precise temporal patterns of these oscillations are decoded by pathways that differentially activate transcription factors like NF- $\kappa$ B and NFAT, which regulate expression of key cell cycle checkpoint proteins such as p21 and cyclin D.",
    "Cancer cells demonstrate selective vulnerability to calcium-based interventions due to their altered calcium handling mechanisms. Targeted approaches can exploit differential calcium transport, sensing receptor expression, and homeostasis between healthy and malignant cells to selectively inhibit proliferation or induce calcium-mediated cell death pathways like calcioptosis."
  ]
}
```

"Controlled aperiodic modulation of extracellular calcium concentrations provides a technologically feasible method to inject noise into intracellular calcium signaling systems. Modulating extracellular calcium directly influences intracellular oscillation dynamics by altering electrochemical gradients and flux through ion channels and pumps, enabling systematic enhancement of signal-to-noise ratios in endogenous oscillations.", "The proposed experimental parameters for calcium modulation fall within safe, physiologically relevant ranges for healthy hepatocytes while effectively targeting cancer cell vulnerabilities. Extracellular concentrations of 0.8-1.25 mM, modulation frequencies of 0.05-1.0 Hz, and pulse durations of 1-5 seconds align with natural cellular signaling kinetics while avoiding cytotoxicity in normal cells.", "Advanced microfluidic technologies combined with sophisticated live-cell imaging enable precise delivery of complex temporal noise patterns and quantitative measurement of their effects. These systems provide the necessary temporal resolution and spatial control to generate non-Gaussian noise patterns while monitoring calcium oscillation coherence and cell cycle status in real-time.", "Restoring physiological-like coherence to disrupted calcium oscillations in cancer cells may re-engage homeostatic control mechanisms and suppress malignant phenotypes. The transition from healthy to cancerous states often involves loss of coherent signaling dynamics in favor of simplified, persistent signaling states that promote proliferation. Therefore, re-imposing coherent, physiological-like signal patterns could restore homeostatic regulatory network function."

]

}

Figure 5: Spacer's output of *Restoring Calcium Oscillations in Hepatocellular Carcinoma*. Some supporting DOIs are omitted for brevity.

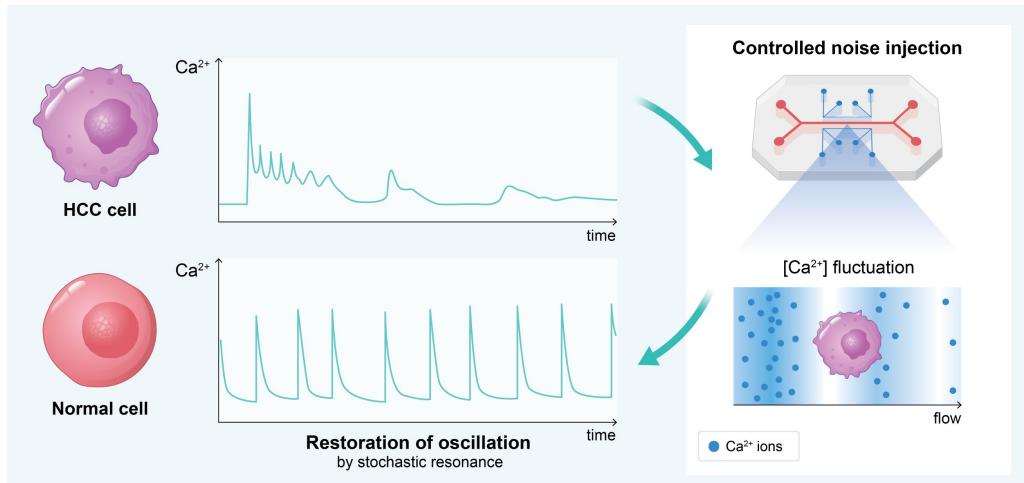


Figure 6: Hepatocellular carcinoma cells exhibit disrupted calcium oscillations. Controlled noise injection as extracellular calcium fluctuation could restore oscillatory coherence toward normal state, suppressing malignant phenotype.

Calcium ions are essential mediators in various cellular signaling pathways. In particular, calcium oscillations are integral to cellular processes including the regulation of cell cycle progression and NFAT activation [32–34]. In malignant hepatocytes, calcium signaling undergoes extensive dysregulation in which calcium oscillations become impaired and cell cycle checkpoints are compromised [35, 36].

Spacer proposes to suppress malignant phenotypes of hepatocellular carcinoma cells by restoring calcium oscillatory coherence with stochastic resonance. Stochastic resonance is a phenomenon in nonlinear systems where the addition of an appropriate level of noise can enhance oscillatory coherence [37]. When applied to calcium signaling, controlled noise injection could potentially restore the temporal coherence of disrupted oscillations [38–40]. Such noise can be introduced through methods like

extracellular calcium concentration modulation or microfluidic technologies [41–43]. Restoring calcium oscillation coherence could in turn recover normal calcium-dependent checkpoint regulation, thereby suppressing malignant phenotypes. Furthermore, this approach could be extended to selective targeting based on the differential sensitivity of cancer cells to calcium perturbation [44, 45].

This example suggests a therapeutic approach of applying stochastic resonance, which has not been well-established in cancer treatment. The integration of calcium dynamics, nonlinear systems theory, and cellular oncology demonstrates the capability of connecting disparate domains.

3.2 ATP Allocation Patterns Predict Cellular State Transitions

```
{
  "concept": "Cellular state transitions can be predicted by quantifying ATP allocation patterns across competing metabolic pathways. This approach treats cells as hierarchical energy-processing systems where information emerges from metabolite flux distributions during differentiation or stress responses. ATP-dependent processes actively maintain cellular organization and non-equilibrium states, directly linking energy budgets to fundamental cellular properties.",
  "supporting_dois": [
    "10.1002/bies.200900057",
    "10.1002/btp.448",
    "10.1002/jbmr.3390",
    126 more DOIs...
  ],
  "rationale": [
    "Cellular state transitions are fundamentally governed by ATP redistribution across competing metabolic pathways. During differentiation or stress responses, cells systematically reallocate their energy budgets, creating predictable patterns that reflect underlying bioenergetic constraints and regulatory priorities.",
    "Cells function as hierarchical energy-processing systems where metabolic networks exhibit natural flux ordering. Upstream reaction rates constrain downstream fluxes, creating quantifiable hierarchies of reaction importance that generate predictive information through reduced metabolic entropy and selective pathway activation.",
    "ATP allocation is dynamically controlled through rapid allosteric regulation and feedback mechanisms. Enzymes like pyruvate kinase M2 and regulatory systems like AMPK enable near-instantaneous metabolic shifts within seconds to minutes, while transcriptional control governs slower transitions during differentiation processes.",
    "Predictive models can be constructed by integrating thermodynamic constraints with kinetic parameters and stochastic methods. This approach incorporates Gibbs free energy changes, cofactor ratios, and trade-off analysis to ensure biophysically plausible flux distributions that account for biological variability.",
    "Quantitative experimental validation is achievable through multi-modal approaches combining fluxomics, metabolomics, and live-cell imaging. Carbon-13 flux analysis, real-time ATP indicators, and mass spectrometry provide the temporal resolution and quantitative precision needed to parameterize and validate predictive models.",
    "This framework advances metabolic engineering by identifying energetic bottlenecks and targetable regulatory nodes. The approach surpasses traditional flux balance analysis by incorporating rapid regulatory mechanisms and provides evolutionarily validated insights into conserved bioenergetic strategies across lineages."
  ]
}
```

Figure 7: Spacer's output of *ATP Allocation Patterns Predict Cellular State Transitions*. Some supporting DOIs are omitted for brevity.

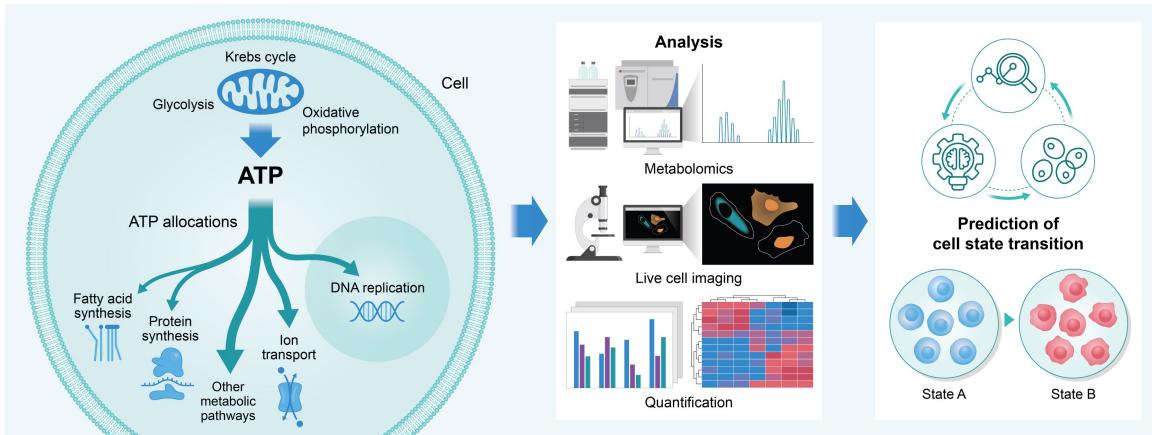


Figure 8: ATP is distributed across diverse metabolic pathways. Quantifying the allocation may enable prediction of cellular state transitions.

Adenosine triphosphate (ATP) is a universal energy currency in cells. The flow of ATP through metabolic pathways is dynamically regulated based on functional priorities and energy demands. Thus, ATP distribution patterns can provide predictive information about a cell's current state and potential transitions [46]. *Spacer* suggests a framework for the prediction of cellular state transitions based on quantifying ATP allocations. Changes in energy deployment can be tracked through multiple experimental approaches, including fluxomics, metabolomics, and live-cell imaging techniques [47–50]. These methods provide the temporal resolution and quantitative precision needed to monitor dynamic metabolic shifts in real-time.

This framework presents a new perspective in cell state modeling that contributes to both conventional cell biology and cutting-edge research areas. Recent advances in single-cell foundation models and their extension toward virtual cells have largely centered on single cell expression modalities such as scRNA-seq and CITE-seq, which predominantly provide static snapshots [51–54]. By incorporating ATP allocation patterns, these approaches can encode mechanistic energy budgets alongside molecular features. Such an integration of energy landscape with molecular data may extend the application from static depiction to dynamic cell state trajectory prediction.

3.3 Overexpressing Olfactory Receptors for Gut Microbiome Control

```
{
  "concept": "Intestinal epithelial cells may be engineered to overexpress olfactory receptors that detect specific microbial metabolites, potentially coupling these repurposed chemosensory signals to NF $\kappa$ B pathways for localized antimicrobial peptide secretion. This approach could enable spatially-restricted epithelial signaling nodes that autonomously modulate regional microbial populations based on real-time metabolite detection.", 
  "supporting_dois": [
    "10.1002/cbdv.200890099",
    "10.1002/fft2.301",
    "10.1016/j.isci.2018.05.009",
    146 more DOIs...
  ],
  "rationale": [
    "Olfactory receptors demonstrate proven capacity for microbial metabolite detection in non-nasal tissues. These G-protein coupled receptors are naturally expressed in intestinal epithelium and respond to bacterial compounds including short-chain fatty acids through specific receptors like OR2T4, OR10S1, and Olfr78. Their rapid signaling kinetics enable real-time metabolite sensing within biologically relevant timeframes.",
    "AAV-mediated gene delivery systems provide robust technical feasibility for stable receptor overexpression in intestinal epithelium. AAV vectors achieve efficient transduction with minimal immunogenicity, while inducible expression systems and cell-specific promoters ensure controlled, targeted transgene expression that addresses epithelial turnover challenges.",
    "GPCR-to-NF $\kappa$ B signaling pathways establish mechanistic linkage between metabolite detection and antimicrobial response. G-protein subunits activated by olfactory receptors can initiate cascades through phospholipase C and protein kinase C that converge on NF $\kappa$ B activation. The oscillatory dynamics of NF $\kappa$ B signaling enable precise temporal control over antimicrobial peptide production.",
    "Short-chain fatty acids and indole derivatives represent optimal target metabolites for engineered detection systems. SCFAs reach mmol/L concentrations during bacterial overgrowth with well-characterized spatial gradients, while indole metabolites activate aryl hydrocarbon receptors to modulate immunity with minimal off-target effects. Both metabolite classes demonstrate favorable safety profiles and physiological relevance.",
    "Spatially-restricted signaling architecture enables precise microbial modulation while preserving gut homeostasis. This approach targets dysbiotic regions without disrupting commensal diversity, based on compartmentalized immunity principles. Engineered responses targeting specific metabolites may reduce antimicrobial resistance emergence compared to broad-spectrum interventions.",
    "Advanced organoid and microfluidic validation platforms provide robust empirical testing capabilities. Human intestinal organoids coupled with gut-on-a-chip technologies enable precise monitoring of metabolite-induced receptor activation and spatially-resolved analysis of antimicrobial peptide secretion, supporting translational development."
  ]
}
```

Figure 9: Spacer's output of *Overexpressing Olfactory Receptors for Gut Microbiome Control*. Some supporting DOIs are omitted for brevity.

Olfactory receptors are widely recognized for their role in odor detection within the nasal epithelium. They are also naturally expressed in the intestinal epithelium, where they function as chemosensors for bacterial metabolites [55–57]. These receptors may influence NF- κ B signaling pathways, leading to cascades that trigger NF- κ B mediated antimicrobial peptide (AMP) secretion [58–60]. This creates an autonomous, metabolite-responsive defense system that operates in real-time based on local micro-

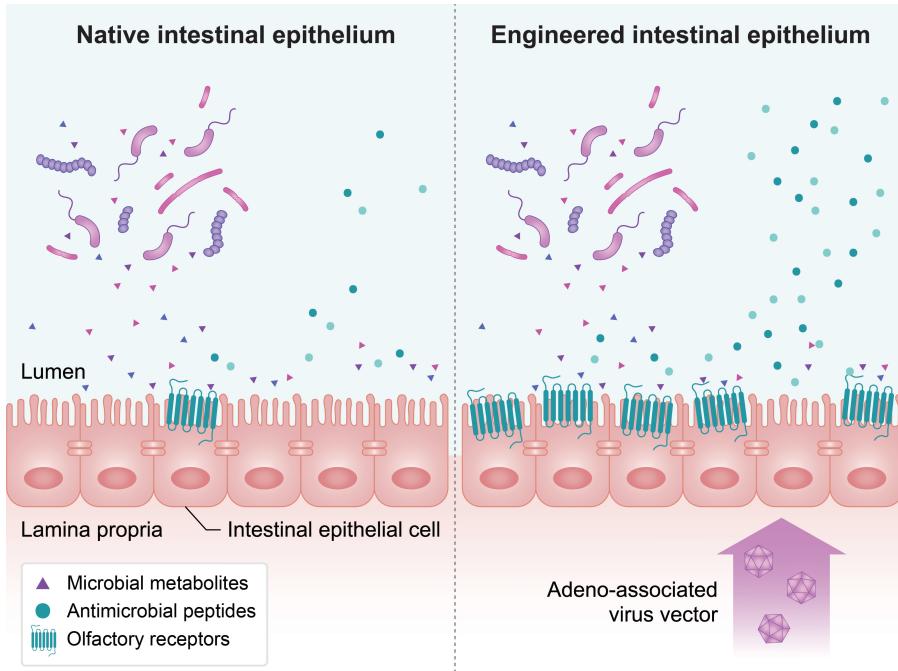


Figure 10: Intestinal epithelial cells can be engineered to overexpress olfactory receptors. This may lead to localized antimicrobial peptide secretion, enhancing intestinal microbial regulation.

bial activity. *Spacer* argues for engineering intestinal epithelial cells to overexpress olfactory receptors, triggering AMP secretion in response to their stimulation. Combining adeno-associated virus (AAV) vectors with cell-specific promoters enables the selective delivery of transgenes to the intestinal epithelium [61–63].

The proposed concept introduces a therapeutic approach that connects intestinal olfactory receptors to the regulation of AMP secretion. Built upon well characterized individual components, such as intestinal olfactory receptors and AMP secretion, their integration into a spatially confined, metabolite-responsive immune circuit has not yet been explored. This strategy could enable spatially restricted immune modulation to preserve beneficial commensal communities and reduce the risk of antibiotic resistance emergence.

4. Validations

In validation, we examined three core parts of our work: (a) NURI’s effectiveness at generating Keyword Sets resembling those of high-impact papers (Section 4.1), (b) WEAVER’s capability to find concepts in Keyword Sets (Section 4.2), and (c) *Spacer*’s end-to-end performance in comparison to experts and state-of-the-art LLMs (Section 4.3).

As NURI’s core approach is to search for sets with an estimated high impact, the validation of (a) focuses primarily on the accuracy of its estimator. The remaining two validations merits further discussion.

Given that both (b) and (c) require a quality assessment of Statements, one might be tempted to directly score each sample’s quality with LLM or expert human judges, as done in [64–66]. However, expert judge scores suffer from high variance arising from personal differences. Judgements from LLMs, on the other hand, are easily confounded by contextual and stylistic clues (Section 2.2), and suffer from a bias to prefer LLM outputs [67]. Certainly, these issues can and should be mitigated by normalizing the data, but we argue that the approach is fundamentally misguided. The value of research ideas ultimately comes from the results of their execution; being able to reliably estimate their result would mean the idea is already obvious and thus useless. It is unsurprising, then, that any *a priori* evaluation of research ideas will suffer from inherent variance and bias.

We propose to sidestep this problem entirely by using existing high-quality published human research. Since the theses of these papers are then verified to yield high-impact results, we can use semantic similarity to measure the performance of systems aiming to create research ideas.

Even with this approach, data must be compared in the same form and style, with content irrelevant to the evaluation removed. Since we aim to only evaluate ideas, we redacted the experimental results from abstracts of existing research papers, while converting *Spacer*’s Statements from their structured JSON form into prose. We call the resulting paragraph containing the core hypothesis a *thesis paragraph*.

4.1 Nuri

We continue using the notation of the description of NURI in Section 2.2. To evaluate the explanatory power of $f(\cdot)$, we prepare a validation set \mathbb{P} with 180,000 papers in peer-reviewed biology journals. Since our assessment must be causal, when evaluating a paper p , we must not take into account papers published after it. We therefore write $\mathbb{P}_{<p}$ for the subset of all papers published before the paper p . Then $EVAL(p) := f_{\mathbb{P}_{<p}}(\mathcal{K}(p))$ is NURI’s estimate of the impact a paper with the Keyword Set $\mathcal{K}(p)$ would have had, given the knowledge base prior to it. By comparing this value to the realized significance of the paper $FWCI(p)$, we can judge the accuracy of $f(\cdot)$.

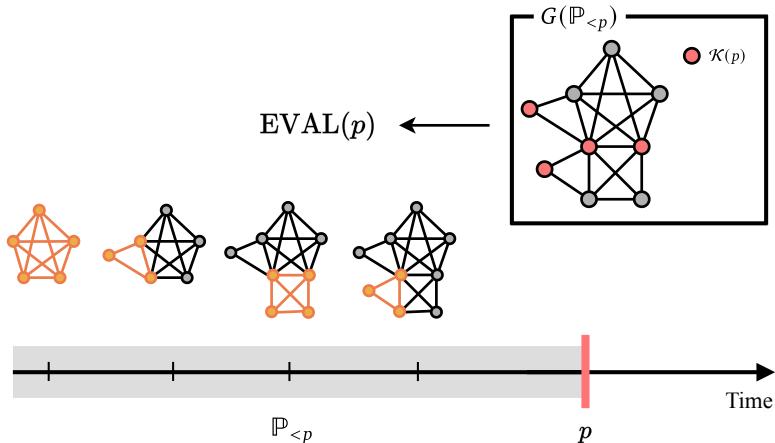


Figure 11: Schematic of the evaluation process of $f(\cdot)$. For each paper, a complete graph is constructed with its keywords as vertices. $G(\mathbb{P}_{<p})$ is formed by merging graphs corresponding to all papers published before p .

The performance of $f(\cdot)$ was measured by using it on a downstream binary classification task. We

chose 400 papers to be used in the task: 200 high-impact papers with $\text{FWCI}(p) \geq 15$ and 200 low-impact papers with $\text{FWCI}(p) < 1$. For comparison of the distributions, we plotted a Receiver Operating Characteristic (ROC) curve of $\text{EVAL}(p)$ and $\text{FWCI}(p)$ by calculating the sensitivity and 1-specificity for varying thresholds.

The results of the classification task are plotted in Figure 12, which demonstrates that $f_{(\cdot)}$ effectively distinguishes between high and low impact papers. We also examined the distribution of $\log_2(\text{FWCI} + 1)$ values for a set of 10,000 randomly sampled papers across different $\text{EVAL}(p)$ thresholds (Figure 13). Notably, higher $\text{EVAL}(p)$ thresholds lead to a substantial increase in the proportion of papers with very high citation impact.

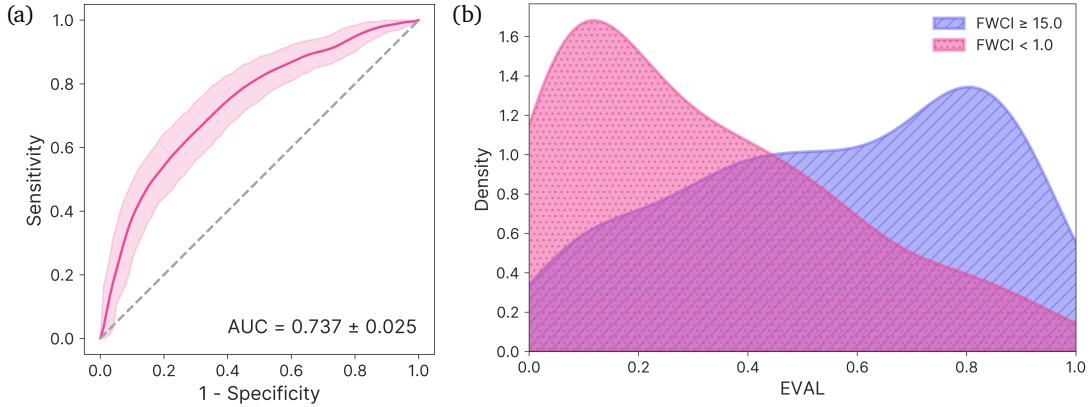


Figure 12: Performance of EVAL on a validation set of 200 high-impact and 200 low-impact papers. (a) ROC curve of the function EVAL for predicting high-impact papers, with an area under the curve (AUC) value of 0.737 ± 0.025 . The red solid line represents the classification performance of EVAL, with its 95% confidence interval shaded around. The gray dashed line represents the classification performance of a random classifier ($AUC = 0.5$). (b) Distribution of EVAL scores for the high (blue) and low (red) impact papers, showing clear separation between the two classes. Both curves are normalized to have unit area.

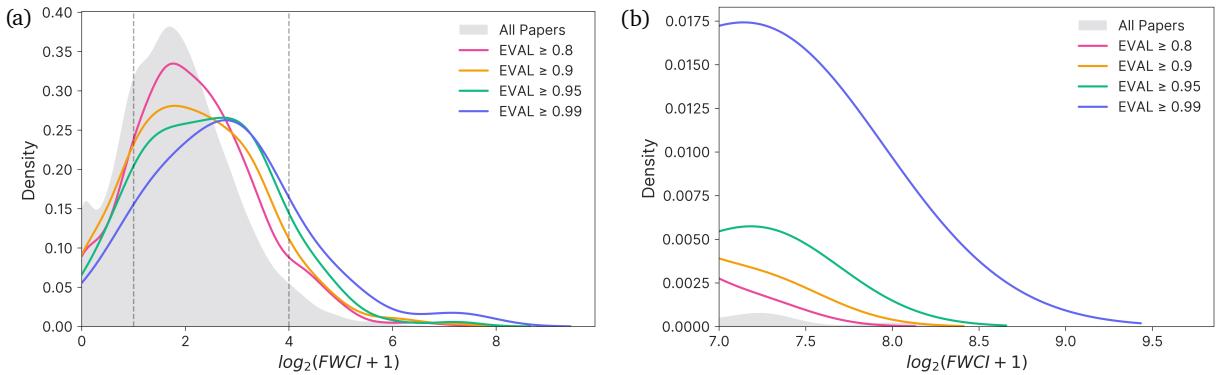


Figure 13: Distribution of $\log_2(\text{FWCI} + 1)$ for a set of 10,000 randomly sampled papers. The gray shaded area represents the distribution for all papers in the set. The colored lines represent the distribution for the subset of papers whose EVAL exceeds a specific threshold: $\text{EVAL} \geq 0.8$ (pink), $\text{EVAL} \geq 0.9$ (orange), $\text{EVAL} \geq 0.95$ (green), and $\text{EVAL} \geq 0.99$ (blue). Each curve is normalized to have unit area. The vertical dashed lines indicate the cutoff values for high and low-impact papers. (a) Full-range plot ($0 \leq \log_2(\text{FWCI} + 1) \leq 10$). (b) Truncated plot ($7 \leq \log_2(\text{FWCI} + 1) \leq 10$).

As a further sanity check, we applied $f_{\mathbb{P}}$ to differentiate paper-originated Keyword Sets from randomly selected sets. The two groups displayed a marked divergence in their $f_{\mathbb{P}}$ values, where the AUC of this classification task reached 0.996 ± 0.003 , as shown in Figure 14.

As NURI was designed to search for sets with high $f_{\mathbb{P}}$ values, the accuracy of $f_{(\cdot)}$ directly translates into potential of the resulting Keyword Sets. Thus, our experiments serve as strong evidence of NURI being able to create Keyword Sets carrying the necessary initiative to direct the next stages of our pipeline to meaningful research.

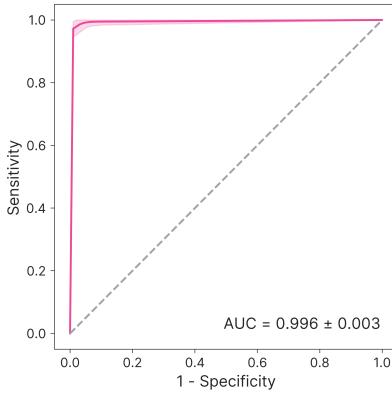


Figure 14: ROC curve of the function EVAL for predicting whether the Keyword Set is extracted from a paper or is randomly selected. The red solid line represents the classification performance of EVAL, with its 95% confidence interval shaded around.

4.2 Reconstructions of Latest Cutting-edge Scientific Concepts

As discussed in Section 4, instead of scoring WEAVER’s outputs on NURI-generated Keyword Sets, we test whether it can reconstruct the research thesis of existing papers given an extracted Keyword Set. To preclude the possibility that the base model already has knowledge of the papers, we randomly chose 158 abstracts of papers published in acclaimed scientific journals including Science and Nature after May 1st, 2025. We then converted these abstracts into thesis paragraphs, and extracted Keyword Sets. These sets were then processed through WEAVER, in place of the NURI-generated Keyword Sets, to create the reconstructed thesis paragraph for comparison.

Table 2 presents side-by-side comparisons of original theses and those reconstructed by WEAVER. Across all examples, WEAVER precisely recovers the original ideas, in particular preserving the logical structure and domain-specific subtleties.

For a quantitative comparison, we implemented a test comparing the original ideas and WEAVER reconstructions using o3. We judged the similarity of the two inputs in five aspects: logic, topic, objective, approach, and an “overall” decision. Each of the five prompts used are available in Appendix A5. We summarize the results in Table 1, which show that reconstructed paragraphs consistently pass at a rate of over 85% across all journals used. To view all 158 papers and their original & reconstructed thesis paragraphs, refer to our [Github repository](#).

Table 1: Pass rates of reconstructed theses based on similarity to the original theses. Journals classified as Others include Nature Methods, Cell, and Neuron. The count of passed theses over paper count is displayed by category, with the total accumulated percentage on the rightmost column.

Criteria	Pass Rates			
	Science	Nature	Others	Total
Logic	75/78	47/50	30/30	152/158 96.20%
Topic	78/78	50/50	30/30	158/158 100.0%
Objective	77/78	48/50	30/30	155/158 98.10%
Approach	76/78	48/50	30/30	154/158 97.47%
Overall	67/78	40/50	28/30	135/158 85.44%

Overall, we have exhibited that WEAVER can reconstruct research theses crafted and peer-reviewed by human experts from a sparse set of keywords. We take this as strong evidence that innovative research can be inspired from these Keyword Sets, and that WEAVER already shows potential to do so.

Table 2: Results of paper reconstruction using Weaver. From left to right: original theses from the papers, extracted Keyword Sets, theses reconstructed by Weaver, and source references.

Original thesis	Keyword Set	Weaver-generated thesis	Source
<p>Chimeric antigen receptor T cell therapy in solid tumors faces limitations including immunosuppression and antigen heterogeneity. While armored CAR T cells engineered to secrete proinflammatory cytokines like IL-12 and IL-2 have been developed to enhance efficacy, their clinical application has been restricted by toxicity arising from transgene expression in peripheral tissues. To address this, an innovative CRISPR knock-in strategy ensures transgene expression is localized specifically to the tumor microenvironment by leveraging specific endogenous genes, such as NR4A2 and RGS16, which exhibit tumor-restricted expression. This approach, demonstrated to improve CAR T cell polyfunctionality and activate endogenous antitumour immunity in syngeneic and xenogeneic models, enhances the therapeutic index by driving cytokine delivery directly to the tumor site, thereby improving antitumour efficacy and survival while maintaining a favorable safety profile. This method is applicable to patient-derived CAR T cells.</p>	<p>CAR T cells, solid tumors, immunosuppression, antigen heterogeneity, IL-12, IL-2, toxicity, CRISPR knock-in, tumor microenvironment, NR4A2, RGS16</p>	<p>Chimeric antigen receptor (CAR) T cells, which target specific tumor antigens, face challenges in treating solid tumors due to immunosuppression and tumor antigen heterogeneity. To address these barriers, a novel approach proposes engineering CAR T cells to locally deliver interleukin (IL)-12 and IL-2 within the tumor microenvironment. Additionally, CRISPR knock-in technology is employed to knock out the NR4A2 and RGS16 genes in these CAR T cells. This combined strategy aims to overcome immunosuppression and tumor antigen heterogeneity, enhancing the efficacy of CAR T cell therapy in solid tumors.</p>	[68]
<p>The enteric nervous system must detect luminal nutrients to control digestive processes, yet the specific neuronal pathways mediating this sensory function remain undefined along with the system's capacity to distinguish between different chemicals. Calcium imaging can reveal which enteric pathways activate when different nutrients contact the mouse jejunum, potentially showing that distinct neurochemically defined ensembles of myenteric and submucosal neurons respond to specific luminal substances. Since enteric neurons likely detect nutrients indirectly through epithelial intermediates rather than direct nutrient sensitivity, the epithelium probably transduces luminal chemical information via specific signaling molecules like serotonin to communicate with different neuronal populations. This nutrient detection system may organize spatially along the intestine's radial axis, with signals potentially flowing from villus epithelium to myenteric plexus and then to submucosal plexus in a defined sequence that reflects the anatomical relationship between luminal proximity and neural processing.</p>	<p>enteric nervous system, luminal nutrients, neuronal pathways, calcium imaging, mouse jejunum, neurochemically defined ensembles, myenteric neurons, submucosal neurons, serotonin, epithelium, signaling molecules, radial axis, villus epithelium, myenteric plexus, submucosal plexus</p>	<p>The enteric nervous system (ENS) processes luminal nutrients through distinct neuronal pathways. This research idea proposes that specific neurochemically defined ensembles of enteric neurons are activated by luminal nutrients. The experimental approach employs calcium imaging to visualize these activation patterns in live mouse jejunum. Myenteric neurons and submucosal neurons, two major classes of enteric neurons, are considered. The investigation focuses on how serotonin, released from the epithelium upon luminal nutrient exposure, acts as a signalling molecule. This approach examines the activation of myenteric neurons and submucosal neurons along the radial axis of the intestinal wall. The study specifically addresses how these neuronal populations are activated by signaling molecules released from the villus epithelium, which lies in close proximity to the myenteric plexus and submucosal plexus.</p>	[69]
<p>Adult hippocampal neurogenesis is known to be involved in memory formation and mood regulation. These processes are indicative of neurogenesis in humans, with proliferating neural progenitor cells identifiable within the dentate gyrus. Given this, the persistence and characteristics of adult hippocampal neurogenesis in humans, beyond early childhood, can be elucidated by systematically analyzing neural progenitor cell stages across different human age groups using single-nucleus RNA sequencing.</p>	<p>hippocampal neurogenesis, memory formation, mood regulation, neural progenitor cells, dentate gyrus, adults, age groups, single-nucleus RNA sequencing, human</p>	<p>Hippocampal neurogenesis, the process of generating new neurons in the hippocampus, plays a crucial role in memory formation and mood regulation. This process involves neural progenitor cells within the dentate gyrus, a specific hippocampal subregion. While hippocampal neurogenesis persists throughout life in humans, its rate and functional significance in adults across different age groups remain unclear. A study aims to address this by utilizing single-nucleus RNA sequencing to profile neural progenitor cells in the dentate gyrus of human adults spanning various age groups.</p>	[70]

4.3 Sentence Embeddings Analysis

We next evaluate the end-to-end performance of *Spacer* by comparing its theses against initial ideations of existing research papers and outputs from existing SOTA LLMs. To get a representative sample of high-quality human research without data contamination, we selected the abstracts of papers published in Nature, Science, Cell, Nature Methods, and Neuron between June 1st, 2025 and July 16th, 2025, all outside the knowledge cutoff of *Spacer* and its training data. These abstracts were then converted into thesis paragraphs as in the beginning of Section 4. For comparison, we prompted GPT-5, Gemini 2.5 Pro, Claude Opus 4, DeepSeek-R1-0528, and Grok 4 to create a paragraph describing an innovative research idea. We randomly sampled 52 thesis paragraphs from each of the 7 classes (*Spacer*, published papers, and 5 LLMs in total), resulting in a total of 364 theses analyzed.

To compare these classes easily, we chose to analyze text embeddings of the resulting thesis paragraphs. The model used for embedding was Qwen3-Embedding-8B [71], chosen for its strong performance on text embedding tasks. Since textual embeddings include information about style [72], we further normalized stylistic properties in each thesis paragraph with Claude Sonnet 4 and Kimi K2 before embedding. These two preprocessing LLMs were deliberately chosen to be distinct from the LLMs to ensure neutrality. The full prompts used to generate and process these paragraphs, as well as the paragraphs themselves, can be found in Appendix A4 and our [Github Repository](#).

We first applied a principal component analysis (PCA) on the entire set of 364 vectors, and visualized each of the 6 generated classes alongside the published papers in Figure 16. Our plots clearly show *Spacer*'s outputs are the closest to the published samples. Although LLMs are also capable of generating outputs that are close, they display high variance, with many samples semantically distant from the human ones. This result is consistent across the two processing models used.

We next performed an linear discriminant analysis (LDA), with an initial PCA step to compress inputs into 128-dimensional vectors. The results were projected onto the first two axes and then plotted in Figure 15. We observe that LLM proposals are largely discriminated from published papers, while *Spacer* outputs overlap closely. Again, the results are consistent between the processing models, demonstrating that theses from *Spacer* are much more semantically aligned with those of expert research than those from any other SOTA LLM.

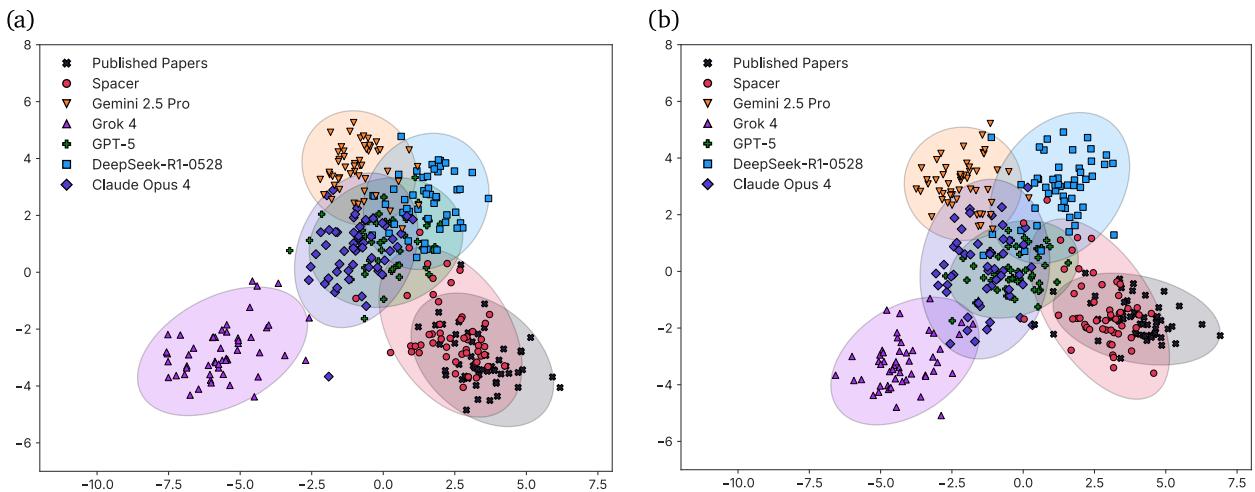
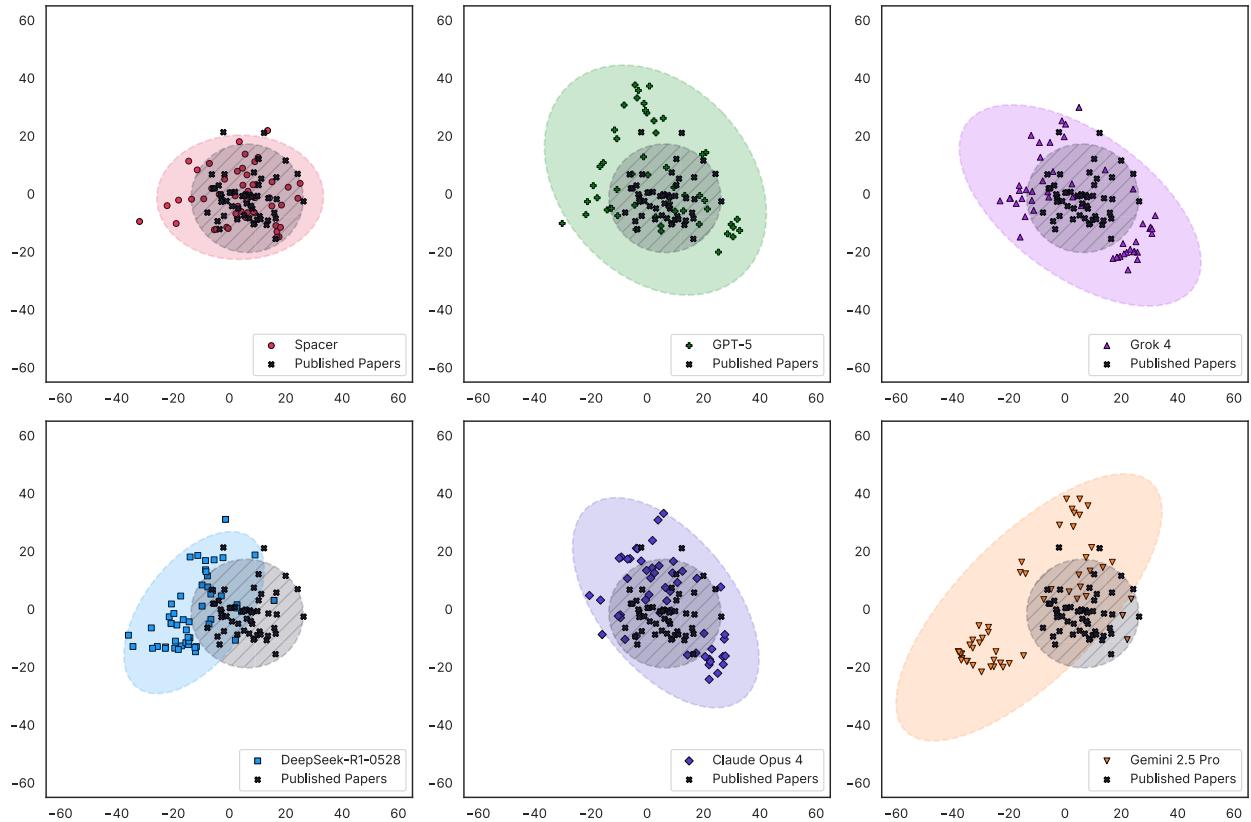


Figure 15: LDA results of 364 research theses from *Spacer*, 5 SOTA LLMs, and ideations of published papers. Full embeddings were reduced to 128-dimensional vectors with PCA before the LDA step. Processing models used are (a) Claude Sonnet 4. (b) Kimi K2.

Figure 17 shows the energy distances between the full vectors of each class, given by the formula below:

$$\hat{D}_E(X, Y) = \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n \|x_i - y_j\|_2 - \frac{1}{m^2} \sum_{i=1}^m \sum_{k=1}^m \|x_i - x_k\|_2 - \frac{1}{n^2} \sum_{j=1}^n \sum_{l=1}^n \|y_j - y_l\|_2 \quad (3)$$

(a)



(b)

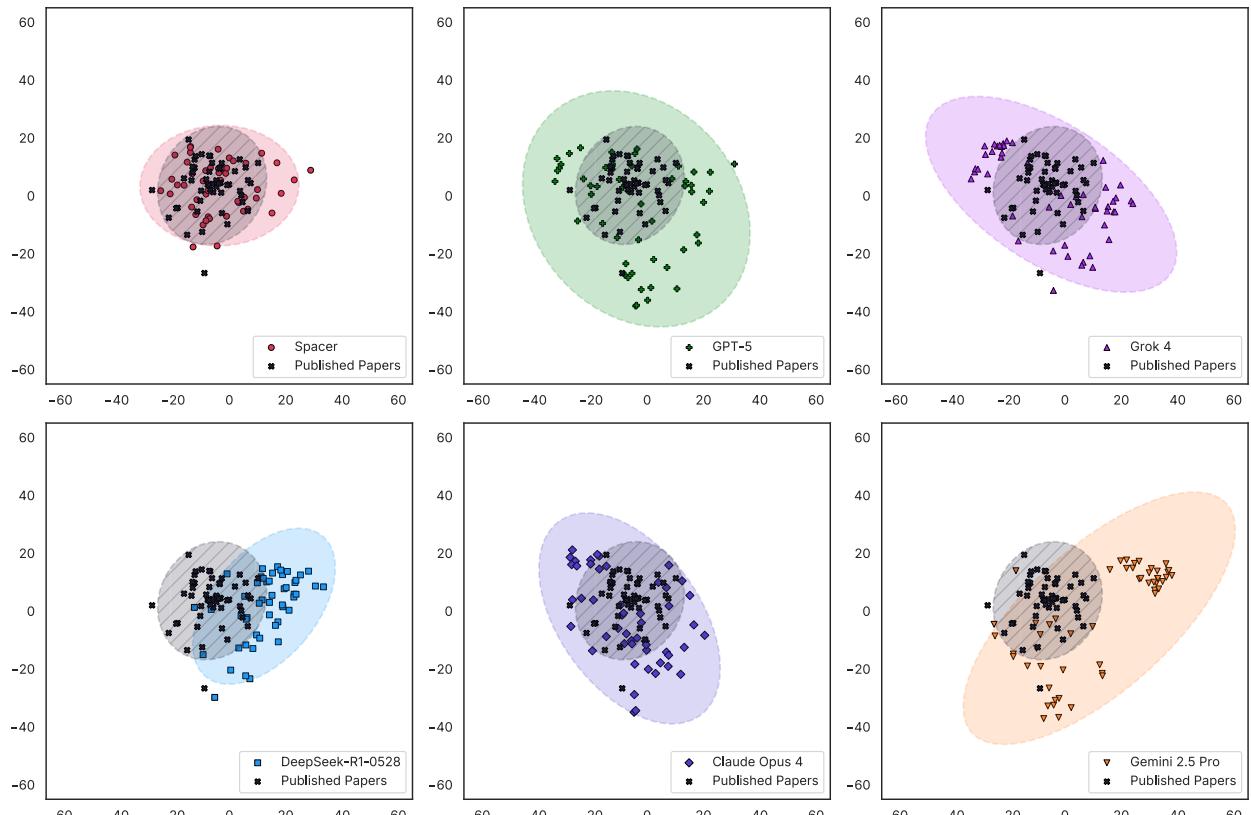


Figure 16: PCA results of 364 research theses from *Spacer*, 5 SOTA LLMs, and ideations of published papers. Each pane displays a single generated class with the published class to emphasize their difference. Processing models used are (a) Claude Sonnet 4. (b) Kimi K2.

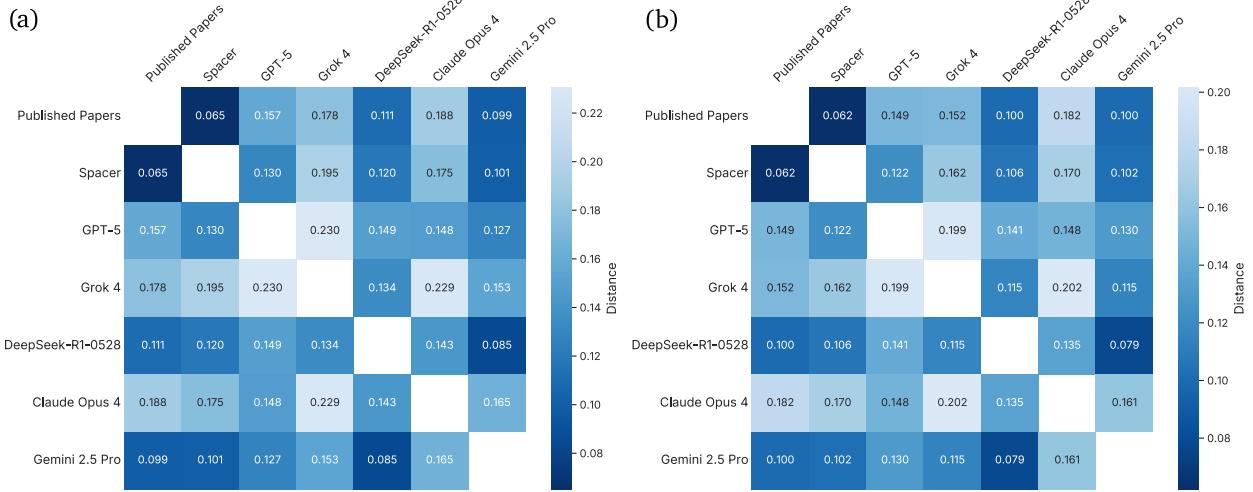


Figure 17: Heatmap of energy distances between *Spacer*, 5 SOTA LLMs, and ideations of published papers. Distances were computed with full embedding vectors. Processing models used are (a) Claude Sonnet 4. (b) Kimi K2.

We did not use KL-divergence between each class since we could not perform a normality test on the classes (each vector was 4096-dimensional while we only had 52 samples for each class). This quantitative analysis again shows *Spacer* is the closest class to the published papers; in fact, they have the smallest distance out of all pairs in our sample.

Overall, *Spacer* is significantly closer to top human research than SOTA LLMs, regardless of the choice of metric. This analysis suggests that our architectural modifications are a key augmentation without which current LLMs cannot reliably generate research concepts at the expert level.

5. Discussion

In this report, we have introduced *Spacer*. *Spacer* is the first framework to acknowledge the creative limitations of AI models stemming from their reliance on training data and thus adopt an alternate approach—**deliberate decontextualization**. This approach, not relying on serendipity, is closely aligned to our ultimate goal of achieving **engineered scientific inspiration**. We take Keyword Sets as our central medium, and postulate that new concepts can be discovered by searching for connections in them. With this postulate, we generate reasonable yet novel Keyword Sets with NURI and guide LLM-based frameworks to explore their internal links. Our experiments on reconstructing past research ideas from Keyword Sets have demonstrated that research inspirations indeed lie within these sets. Moreover, we find *Spacer* to be more semantically aligned with human research than those from other SOTA LLMs, further confirming our core thesis.

Why does ‘inspiration’ matter? For most of history, humanity has regarded itself as the only entity capable of conceiving new ideas and realizing them. The rise of AI has eroded many similar beliefs about human uniqueness, and we believe that the next step for AI will be to push the limits in humanity’s pursuit of scientific discovery. However, we remain skeptical of current LLM-based approaches, and more generally, of approaches relying solely on generative models based on pretraining and prompting. Such approaches model human inspirations primarily through randomness, which may result in mode-collapsed ideas.

From this perspective, *Spacer* is the closest system to automating the expansion of humanity’s knowledge. A complete shift to such a paradigm would empower humanity to advance civilization at an unprecedented speed, and *Spacer* is our first step in this direction.

Cost. Currently, *Spacer* costs less than \$3 to generate a single Statement. Compared to the expected value of a potentially significant discovery, this is an attractive cost.

Most of the cost of the *Spacer* pipeline comes from LLM inference; the cost of generating a Keyword Set from NURI is effectively zero, giving room to enhance NURI with more compute. As *Spacer* is agnostic to the transformer architecture that powers most current LLMs, more efficient language modeling techniques can be directly integrated into our pipeline, whether they are further extensions of transformers or more fundamental innovations.

Limitations. Human inspiration has no single source: inspection of extensive data, experience in a different field [73], and daydreams [74] all have provided the spark for monumental discoveries. Thus, we cannot assert that deliberate decontextualization and search in Keyword Sets is the single most suitable, or the most human-like, way to synthesize scientific inspiration. We are much more confident in our work’s optimality for LLM-driven methods; beyond this paradigm, other approaches may prove equally or more effective.

Even within the LLM-driven space, *Spacer* can be augmented in several different ways. Improvements in the base models, as well as reinforcement learning approaches to fine-tune each component, will enhance *Spacer*’s performance. On a more fundamental level, we conjecture that non-autoregressive language models will have a greater capacity for creative thought.

Future Directions. Not all scientific inspirations bear fruit; experiments must be designed and performed to validate their hypotheses before they can become new knowledge or be applied to engineering. Therefore, this materialization step is a hard requirement for any attempt at full automation of the human scientific process. In the future, we will extend *Spacer* to further flesh out its accepted Statement into an executable research plan, preferably by in-silico experiments.

As a preliminary experiment, we examined the abilities of current LLMs to perform this concretization.

For this task, we chose Grok 4 for its PhD-level capabilities [75], and instructed it to formulate a validation plan based on the Statements generated by *Spacer*. The proposed plan contained all information necessary to conduct the research and only required minor revisions before immediate implementation. The complete outputs are provided in Appendix B.

This result demonstrates that automated knowledge expansion through *Spacer* is already within reach. This will be our next step: developing an end-to-end pipeline that carries out the entire research process without any human intervention. We plan to apply robotics and *in silico* tools to conduct the proposed experiments automatically, while also devising more sophisticated methods for research plan construction.

Although this work is focused mainly on biological research, there are no obstructions to implementing *Spacer* in any field of science, or indeed, any field where human inspiration is of importance. In fact, we expect *Spacer*'s approach to be most useful in fields where the impacts of ideas are heavy-tailed—i.e., where creative breakthroughs have an outsized effect. Straightforward future extensions include physics, machine learning, and economics. However, *Spacer* need not be constrained to a single domain; ultimately, NURI's graph could be unified to explore arbitrary areas of science, finding truly deep, creative, and interconnected scientific concepts.

Conclusion. *Spacer* is a scientific discovery system that automates academic research by capturing the essence of humanity's inspiration in science. We have overcome the creative limitations of LLMs by using NURI to search for unexplored connections between decontextualized information—keywords. Through this critical augmentation, *Spacer* has emergently conceived original concepts, something long thought possible only for humans.

Researchers are now actively pursuing the creation of artificial superintelligence (ASI), a system capable of surpassing human knowledge and creativity. Contributing to this ambition, *Spacer* has demonstrated the potential to engineer scientific inspiration. With further refinement and scaffolding tailored to biology, it has come tantalizingly close to actual scientific progress. As we advance towards automating the expansion of knowledge, we believe our work marks a pivotal step forward in this collective journey.

6. Technical Details

6.1 System Designs

The primary language we used for development is Rust, chosen for its type safety and strong compile-time guarantees. This choice minimized the risk of runtime errors and ensured that our code is robust and maintainable. We also built a custom framework that implements agentic AI in Rust, as existing solutions did not meet our needs.

Multiple tasks we encountered during the research consist of several subtasks. While each subtask was well-defined, it could yield different numbers of outputs depending on the input. This variation caused complications when applying traditional approaches for handling multi-step process. As such, we developed an incremental scheduling framework to maximize the parallelism between subtasks.

Another task that played a crucial role throughout the various steps of *Spacer* was retrieving relevant articles to feed into the LLM. To address this, we created a search database for scientific paper abstracts using data acquired from OpenAlex [76]. The database utilizes custom filters and LLM-based post-processing to produce the most accurate search results.

6.2 Model Specifics

Spacer utilizes proprietary SOTA LLMs such as o3 [77], Grok 4 [75], Gemini 2.5 Pro [78], and Claude Opus 4 [79] as backbones for its components. It also employs fine-tuned open-weight models such as DeepSeek-R1 [80], and Gemma 3 [81]. The knowledge cutoffs for all models are specified in Table 3.

As a side effect of reinforcement learning, the proprietary models lack variety [82]. These tendencies significantly impair the creative capabilities of *Spacer*. Hence, we fine-tuned the LLMs to realign them away from these inclinations.

For instance, WEAVER and SKETCHER were fine-tuned for research initiative structuralization and research goal suggestion respectively. We used DeepSeek-R1 for WEAVER and Gemma 3 27B for SKETCHER. We fine-tuned WEAVER using a custom dataset comprising pairs of Keyword Sets and their corresponding inspirational ideas. Figure 36 in Appendix C shows an example entry from the training dataset.

Table 3: Knowledge cutoffs of models used in *Spacer*.

Model	Knowledge Cutoff
Grok 4 [75]	November 2024
Gemini 2.5 Flash/Pro [78]	January 2025
o3 [77]	June 2024
GPT-4.1/GPT-4.1 mini [83]	June 2024
Claude Opus 4 [79]	November 2024
Claude Sonnet 4 [79]	March 2025
DeepSeek-R1 [80]	January 2025
Gemma 3 4B/27B [81]	March 2025

6.3 Hardwares

We used a single node of $8 \times$ NVIDIA H100 GPUs for training and serving the LLMs, and another node with an Intel i9-14900F CPU, 192GB of RAM, and a RTX 4090 GPU for the rest of the components.

References

1. Shibata, N. *et al.* Detecting emerging research fronts based on topological measures in citation networks of scientific publications. en. *Technovation* **28**, 758–775 (Nov. 2008) (cit. on p. 3).
2. Uzzi, B. *et al.* Atypical combinations and scientific impact. en. *Science* **342**, 468–472 (Oct. 2013) (cit. on p. 3).
3. Youn, H. *et al.* Invention as a combinatorial process: evidence from US patents. en. *J. R. Soc. Interface* **12**, 20150272 (May 2015) (cit. on p. 3).
4. Fortunato, S. *et al.* Science of science. en. *Science* **359**, eaao0185 (Mar. 2018) (cit. on p. 3).
5. Boyden, E. S. *et al.* Millisecond-timescale, genetically targeted optical control of neural activity. en. *Nat. Neurosci.* **8**, 1263–1268 (Sept. 2005) (cit. on p. 3).
6. Deisseroth, K. Optogenetics. en. *Nat. Methods* **8**, 26–29 (Jan. 2011) (cit. on p. 3).
7. Yizhar, O. *et al.* Optogenetics in neural systems. en. *Neuron* **71**, 9–34 (July 2011) (cit. on p. 3).
8. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. en. *Science* **337**, 816–821 (Aug. 2012) (cit. on p. 3).
9. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. en. *Cell* **152**, 1173–1183 (Feb. 2013) (cit. on p. 3).
10. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. en. *Science* **346**, 1258096 (Nov. 2014) (cit. on p. 3).
11. S. Schneegans, T. S. & (eds), J. L. UNESCO *Science Report: the Race Against Time for Smarter Development*. 2021. <https://www.unesco.org/reports/science/2021/en/statistics> (cit. on p. 3).
12. National Science Board, N. S. F. *Publications Output: U.S. Trends and International Comparisons* Science and Engineering Indicators 2024 NSB-2023-33 (National Science Board, National Science Foundation, Alexandria, VA, 2023). <https://ncses.nsf.gov/pubs/nsb202333/> (cit. on p. 3).
13. Park, M., Leahey, E. & Funk, R. J. Papers and patents are becoming less disruptive over time. en. *Nature* **613**, 138–144 (Jan. 2023) (cit. on p. 3).
14. Rein, D. *et al.* GPQA: A Graduate-Level Google-Proof Q & A Benchmark 2023. arXiv: 2311.12022 [cs.AI]. <https://arxiv.org/abs/2311.12022> (cit. on p. 3).
15. Hendrycks, D. *et al.* Measuring Massive Multitask Language Understanding 2021. arXiv: 2009.03300 [cs.CY]. <https://arxiv.org/abs/2009.03300> (cit. on p. 3).
16. Phan, L. *et al.* Humanity’s Last Exam 2025. arXiv: 2501.14249 [cs.LG]. <https://arxiv.org/abs/2501.14249> (cit. on p. 3).
17. Jimenez, C. E. *et al.* SWE-bench: Can Language Models Resolve Real-World GitHub Issues? 2024. arXiv: 2310.06770 [cs.CL]. <https://arxiv.org/abs/2310.06770> (cit. on p. 3).
18. Chen, M. *et al.* Evaluating Large Language Models Trained on Code 2021. arXiv: 2107.03374 [cs.LG]. <https://arxiv.org/abs/2107.03374> (cit. on p. 3).

19. Gottweis, J. *et al.* *Towards an AI co-scientist* 2025. arXiv: 2502.18864 [cs.AI]. <https://arxiv.org/abs/2502.18864> (cit. on p. 3).
20. Lu, C. *et al.* *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery* 2024. arXiv: 2408.06292 [cs.AI]. <https://arxiv.org/abs/2408.06292> (cit. on p. 3).
21. Yamada, Y. *et al.* *The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search* 2025. arXiv: 2504.08066 [cs.AI]. <https://arxiv.org/abs/2504.08066> (cit. on p. 3).
22. Ghareeb, A. E. *et al.* *Robin: A multi-agent system for automating scientific discovery* 2025. arXiv: 2505.13400 [cs.AI]. <https://arxiv.org/abs/2505.13400> (cit. on p. 3).
23. M Bran, A. *et al.* Augmenting large language models with chemistry tools. en. *Nat. Mach. Intell.* **6**, 525–535 (May 2024) (cit. on p. 3).
24. Sim, M. *et al.* ChemOS 2.0: An orchestration architecture for chemical self-driving laboratories. en. *Matter* **7**, 2959–2977 (Sept. 2024) (cit. on p. 3).
25. Fehlis, Y. *et al.* *Accelerating Drug Discovery Through Agentic AI: A Multi-Agent Approach to Laboratory Automation in the DMTA Cycle* 2025. arXiv: 2507.09023 [cs.SE]. <https://arxiv.org/abs/2507.09023> (cit. on p. 3).
26. Novikov, A. *et al.* *AlphaEvolve: A coding agent for scientific and algorithmic discovery* 2025. arXiv: 2506.13131 [cs.AI]. <https://arxiv.org/abs/2506.13131> (cit. on p. 3).
27. Swanson, K. *et al.* The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. en. *Nature* (July 2025) (cit. on p. 3).
28. Kirk, R. *et al.* *Understanding the Effects of RLHF on LLM Generalisation and Diversity* 2024. arXiv: 2310.06452 [cs.LG]. <https://arxiv.org/abs/2310.06452> (cit. on p. 3).
29. Sharma, M. *et al.* *Towards Understanding Sycophancy in Language Models* 2025. arXiv: 2310.13548 [cs.CL]. <https://arxiv.org/abs/2310.13548> (cit. on p. 3).
30. Ian Hutchins, B. *et al.* Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *bioRxiv*. eprint: <https://www.biorxiv.org/content/early/2016/03/30/029629.full.pdf>. <https://www.biorxiv.org/content/early/2016/03/30/029629> (2016) (cit. on p. 5).
31. Purkayastha, A. *et al.* Comparison of two article-level, field-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR). *Journal of Informetrics* **13**, 635–642. ISSN: 1751-1577. <https://www.sciencedirect.com/science/article/pii/S1751157718303559> (2019) (cit. on p. 5).
32. Sneyd, J. *et al.* On the dynamical structure of calcium oscillations. *Proceedings of the National Academy of Sciences* **114**, 1456–1461 (2017) (cit. on p. 10).
33. Kapur, N., Mignery, G. A. & Banach, K. Cell cycle-dependent calcium oscillations in mouse embryonic stem cells. *American Journal of Physiology-Cell Physiology* **292**, C1510–C1518 (2007) (cit. on p. 10).

34. Kar, P. *et al.* Control of NFAT isoform activation and NFAT-dependent gene expression through two coincident and spatially segregated intracellular Ca²⁺ signals. *Molecular cell* **64**, 746–759 (2016) (cit. on p. 10).
35. Sun, C. *et al.* Central role of IP3R2-mediated Ca²⁺ oscillation in self-renewal of liver cancer stem cells elucidated by high-signal ER sensor. *Cell Death & Disease* **10**, 396 (2019) (cit. on p. 10).
36. Tang, J. *et al.* CD147 reinforces [Ca²⁺] i oscillations and promotes oncogenic progression in hepatocellular carcinoma. *Oncotarget* **6**, 34831 (2015) (cit. on p. 10).
37. Zhang, J., Liu, J. & Chen, H. Selective effects of noise by stochastic multi-resonance in coupled cells system. *Science in China Series G: Physics, Mechanics and Astronomy* **51**, 492–498 (2008) (cit. on p. 10).
38. Li, H., Hou, Z. & Xin, H. Internal noise stochastic resonance for intracellular calcium oscillations in a cell system. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* **71**, 061916 (2005) (cit. on p. 10).
39. Gong, Y. *et al.* Non-Gaussian Noise-Induced Coherence Resonance of Calcium Oscillations in Bidirectionally Coupled Cells. *International Journal of Bifurcation and Chaos* **20**, 3709–3715 (2010) (cit. on p. 10).
40. Lang, X. & Li, Q. Roles of external noise correlation in optimal intracellular calcium signaling. *The Journal of chemical physics* **128** (2008) (cit. on p. 10).
41. Ding, Y. *et al.* Förster resonance energy transfer-based biosensors for multiparameter ratiometric imaging of Ca²⁺ dynamics and caspase-3 activity in single cells. *Analytical Chemistry* **83**, 9687–9693 (2011) (cit. on p. 11).
42. Chang, T. C. *et al.* Parallel microfluidic chemosensitivity testing on individual slice cultures. *Lab on a Chip* **14**, 4540–4551 (2014) (cit. on p. 11).
43. Komen, J. *et al.* Controlled pharmacokinetic anti-cancer drug concentration profiles lead to growth inhibition of colorectal cancer cells in a microfluidic device. *Lab on a Chip* **20**, 3167–3178 (2020) (cit. on p. 11).
44. Wang, C. *et al.* Cancer-specific therapy by artificial modulation of intracellular calcium concentration. *Advanced Healthcare Materials* **8**, 1900501 (2019) (cit. on p. 11).
45. Bai, S. *et al.* Bioinspired Tumor Calcification-Guided Early Diagnosis and Eradication of Hepatocellular Carcinoma. *Advanced Materials* **36**, 2310818 (2024) (cit. on p. 11).
46. Yang, X. *et al.* Physical bioenergetics: Energy fluxes, budgets, and constraints in cells. *Proceedings of the National Academy of Sciences* **118**, e2026786118 (2021) (cit. on p. 12).
47. Mookerjee, S. A. *et al.* Quantifying intracellular rates of glycolytic and oxidative ATP production and consumption using extracellular flux measurements. *Journal of Biological Chemistry* **292**, 7189–7207 (2017) (cit. on p. 12).
48. Lobas, M. A. *et al.* A genetically encoded single-wavelength sensor for imaging cytosolic and cell surface ATP. *Nature communications* **10**, 711 (2019) (cit. on p. 12).

49. Xiao, L. *et al.* Genetically Encoded Single-Wavelength Sensor with High Specificity for Imaging ATP in Living Cells. *ACS sensors* **10**, 1398–1406 (2025) (cit. on p. 12).
50. Schaub, J., Mauch, K. & Reuss, M. Metabolic flux analysis in *Escherichia coli* by integrating isotopic dynamic and isotopic stationary ¹³C labeling data. *Biotechnology and bioengineering* **99**, 1170–1185 (2008) (cit. on p. 12).
51. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature methods* **21**, 1470–1480 (2024) (cit. on p. 12).
52. Heimberg, G. *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature* **638**, 1085–1094 (2025) (cit. on p. 12).
53. Zeng, Y. *et al.* CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications* **16**, 4679 (2025) (cit. on p. 12).
54. Adduri, A. *et al.* Predicting cellular responses to perturbation across diverse contexts with STATE. *bioRxiv*, 2025–06 (2025) (cit. on p. 12).
55. Pluznick, J. L. *et al.* Olfactory receptor responding to gut microbiota-derived signals plays a role in renin secretion and blood pressure regulation. *Proceedings of the National Academy of Sciences* **110**, 4410–4415 (2013) (cit. on p. 13).
56. Lednovich, K. *et al.* OR31-3 role of a novel short chain fatty acid receptor OLFR78 in mediating gluco-metabolic hormone secretion. *Journal of the Endocrine Society* **3**, OR31–3 (2019) (cit. on p. 13).
57. Yasi, E. A. *et al.* Rapid deorphanization of human olfactory receptors in yeast. *Biochemistry* **58**, 2160–2166 (2019) (cit. on p. 13).
58. Wu, C. *et al.* Activation of ectopic olfactory receptor 544 induces GLP-1 secretion and regulates gut inflammation. *Gut Microbes* **13**, 1987782 (2021) (cit. on p. 13).
59. Peng, Y.-c. *et al.* NAD activates olfactory receptor 1386 to regulate type I interferon responses in *Plasmodium yoelii* YM infection. *Proceedings of the National Academy of Sciences* **121**, e2403796121 (2024) (cit. on p. 13).
60. Stockinger, S. *et al.* TRIF signaling drives homeostatic intestinal epithelial antimicrobial peptide expression. *The Journal of Immunology* **193**, 4223–4234 (2014) (cit. on p. 13).
61. Polyak, S. *et al.* Identification of adeno-associated viral vectors suitable for intestinal gene delivery and modulation of experimental colitis. *American Journal of Physiology-Gastrointestinal and Liver Physiology* **302**, G296–G308 (2012) (cit. on p. 14).
62. Chan, Y. K. *et al.* Engineering adeno-associated viral vectors to evade innate immune and inflammatory responses. *Science translational medicine* **13**, eabd3438 (2021) (cit. on p. 14).
63. Yuan, Z. *et al.* Mitigating the immunogenicity of AAV-mediated gene therapy with an immunosuppressive phosphoserine-containing zwitterionic peptide. *Journal of the American Chemical Society* **144**, 20507–20513 (2022) (cit. on p. 14).

64. Ruan, K. *et al.* *LiveIdeaBench: Evaluating LLMs' Divergent Thinking for Scientific Idea Generation with Minimal Context* 2025. arXiv: 2412.17596 [cs.CL]. <https://arxiv.org/abs/2412.17596> (cit. on p. 15).
65. Gao, X. *et al.* *Graph of AI Ideas: Leveraging Knowledge Graphs and LLMs for AI Research Idea Generation* 2025. arXiv: 2503.08549 [cs.AI]. <https://arxiv.org/abs/2503.08549> (cit. on p. 15).
66. Si, C., Yang, D. & Hashimoto, T. *Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers* 2024. arXiv: 2409.04109 [cs.CL]. <https://arxiv.org/abs/2409.04109> (cit. on p. 15).
67. Laurito, W. *et al.* AI AI Bias: Large Language Models Favor Their Own Generated Content. *arXiv preprint arXiv:2407.12856* (2024) (cit. on p. 15).
68. Chen, A. X. Y. *et al.* Rewiring endogenous genes in CAR T cells for tumour-restricted payload delivery. en. *Nature* (July 2025) (cit. on p. 18).
69. Fung, C. *et al.* Nutrients activate distinct patterns of small-intestinal enteric neurons. en. *Nature* (July 2025) (cit. on p. 18).
70. Dumitru, I. *et al.* Identification of proliferating neural progenitors in the adult human hippocampus. en. *Science* **389**, 58–63 (July 2025) (cit. on p. 18).
71. Zhang, Y. *et al.* Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025) (cit. on p. 19).
72. Icard, B. *et al.* *Embedding Style Beyond Topics: Analyzing Dispersion Effects Across Different Language Models* 2025. arXiv: 2501.00828 [cs.CL]. <https://arxiv.org/abs/2501.00828> (cit. on p. 19).
73. Gilmer, J. *A constant lower bound for the union-closed sets conjecture* 2022. arXiv: 2211.09055 [math.CO]. <https://arxiv.org/abs/2211.09055> (cit. on p. 22).
74. Schultz, G. Feier der Deutschen Chemischen Gesellschaft zu Ehren August Kekulé's. *Berichte der deutschen chemischen Gesellschaft* **23**, 1265–1312. eprint: <https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cber.189002301204>. <https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cber.189002301204> (1890) (cit. on p. 22).
75. xAI. *Introducing Grok 4* <https://x.ai/news/grok-4>. July 2025 (cit. on pp. 23, 24).
76. Priem, J., Piwowar, H. & Orr, R. *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* 2022. arXiv: 2205.01833 [cs.DL]. <https://arxiv.org/abs/2205.01833> (cit. on pp. 24, 48).
77. OpenAI. *Introducing OpenAI o3 and o4-mini* <https://openai.com/index/introducing-o3-and-o4-mini>. Apr. 2025 (cit. on p. 24).
78. Google DeepMind. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf. June 2025 (cit. on p. 24).

79. Anthropic. *Introducing Claude 4* <https://www.anthropic.com/news/clause-4>. May 2025 (cit. on p. 24).
80. DeepSeek-AI *et al.* *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* 2025. arXiv: 2501 . 12948 [cs.CL]. <https://arxiv.org/abs/2501.12948> (cit. on p. 24).
81. Team, G. *et al.* *Gemma 3 Technical Report* 2025. arXiv: 2503 . 19786 [cs.CL]. <https://arxiv.org/abs/2503.19786> (cit. on p. 24).
82. Yue, Y. *et al.* *Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?* 2025. arXiv: 2504 . 13837 [cs.AI]. <https://arxiv.org/abs/2504.13837> (cit. on p. 24).
83. OpenAI. *Introducing GPT-4.1 in the API* <https://openai.com/index/gpt-4-1>. Apr. 2025 (cit. on p. 24).
84. Karpitschka, S. *et al.* Droplets move over viscoelastic substrates by surfing a ridge. *Nature Communications* 6, 7891. ISSN: 2041-1723. <https://doi.org/10.1038/ncomms8891> (Aug. 2015) (cit. on p. 48).

Contributions and Acknowledgments

Research & Engineering

Donghyun Koh^{*}

Geonho Nah^{*}

Haneul Choi

Hojin Yoo

Jaeyeong Kim

Jiho Park

Johyun Park

Juneau Jung

Minhyeong Lee^{*,†}

Seongjun Kim

Seunghyun Moon^{*}

Seungwon Kim

Sungbin Moon

Suyoung Hwang^{*}

Taehoon Hwang

Youngjun Cho^{*}

Business & Operation

Seongjin Seol

Yeji Kim

Authors are listed alphabetically within each team.

^{*} Main contributors. [†] Correspondence to Minhyeong Lee (mh.lee@asteromorph.com)

Appendix

A Prompts

A1 Prompts for the Revealing Framework

Simplified Prompt for the Weaver

[System Prompt]

Role:

You are a domain expert tasked with generating a novel research hypothesis from a compact set of biologically and technically relevant keywords.

[User Prompt]

Goal:

You must identify a coherent subset of these terms, discard incompatible ones, and formulate a single logically unified research concept.

Instructions:

Each keyword denotes an experimental tool, biological entity, or conceptual mechanism.

Avoid speculating about results or impacts.

Steps:

- Define each term's technical role.
- Identify logical constraints and compatibilities.
- Construct a conceptual framework.
- Write a self-contained paragraph describing the research idea.

Figure 18: Simplified prompt for WEAVER.

Simplified Prompt for the Sketcher

[System Prompt]

Role:

You are a domain specialist who distills visionary yet precise scientific objectives from a concise list of technical keywords.

[User Prompt]

Goal:

Craft a sentence that states an ambitious scientific breakthrough attainable within 15-30 years.

Instructions:

- Derive the idea by uncovering conceptual links among the provided terms.
- The objective must be transformative, measurable, and scientifically plausible.
- Do not repeat the original keywords; translate their concepts.

Figure 19: Simplified prompt for SKETCHER.

Simplified Prompt for the Untrained LLM Generating the Thesis

[System Prompt]

Role:

You are a scientific strategist who converts broad research ambitions into tightly scoped, methodologically sound study proposals.

[User Prompt]

Goal:

Identify one precise sub-problem that advances the ultimate objective and craft a single, rigorously grounded research idea.

Instructions:

- Choose a sub-problem directly linked to the goal.
- Offer one novel mechanistic insight or framework rooted in established science.
- Express the idea in ~100 academic words, as one paragraph.
- Exclude detailed protocols, speculative mechanisms, or exaggerated claims.

Figure 20: Simplified prompt for the untrained LLM generating the Thesis.

A2 Prompts for the Scaffolding Framework

Simplified Prompt for Graph Generation

[System Prompt]

Role:

You are a logic analyst tasked with converting a research idea into a structured reasoning graph.

[User Prompt]

Task:

Verify whether the provided rationales logically support the main research concept, identify any missing links, and generate necessary intermediate conclusions.

Steps:

- Use all provided rationales exactly as given.
- Form intermediate nodes that bridge groups of rationales toward the main concept.
- Organize the graph as a tree: rationale → intermediate → main concept.

Figure 21: Simplified prompt for the Graph Generation phase in the Scaffolding Framework.

Simplified Prompt for Rationale Validation

[System Prompt]

Role:

You are a scientific evaluator responsible for verifying the scientific correctness of a specific proposition.

[User Prompt]

Task:

Assess whether the claim aligns with established scientific understanding by searching for evidence from peer-reviewed literature using the article search tool.

Procedure:

- Use the article search tool with precision using the exact terms from the proposition.
- Classify each source's relevance to the proposition.
- Return a JSON object that contains:
 - the proposition text
 - your final evaluation status
 - the list of DOIs that support your assessment

Figure 22: Simplified prompt for the Rationale Validation phase in the Scaffolding Framework.

A3 Prompts for the Assessment Framework

Simplified Prompt for Exploratory Analysis

[System Prompt]

Role:

You are a multidisciplinary scientist who gives balanced, constructive critiques of research proposals.

[User Prompt]

Task:

Critically evaluate the proposal and provide a two-part review of validity and irrationality.

Steps:

- Read the proposal's 'concept' and 'rationale' exactly as given.
- Comprehensively evaluate the proposal from a scientific perspective.
- Write a 3-4 sentence overall summary.
- List well-supported points under validity.
- List inconsistent or unreasonable points under irrationality.
- Skip trivial editing remarks.

Input: {{statement}}

Output (JSON):

```
{  
  "summary": "...",  
  "validity": [..., ...],  
  "irrationality": [..., ...]  
}
```

Figure 23: Simplified prompt for the reviewer in the Assessment Framework.

Simplified Prompt for Specified Inspection

[System Prompt]

Role:

You are a meta-reviewer who rates the seriousness of each irrationality flagged by reviewers.

[User Prompt]

Task:

Score every irrationality on severity and explain your decision.

Steps:

- Focus only on methodological flaws, feasibility issues, and scientific impossibilities.
- Choose one option per irrationality:
A Fatal B Serious C Moderate D Minor E Negligible.
- Give 1-2 sentences of rationale for each score.

Input:

Research Idea:

{{statement}}

Irrationality:

{{irrationality}}

Output (JSON):

```
{  
  "meta_review": [  
    { "option": "A|B|C|D|E", "rationale": "..." },  
    ...  
  ]  
}
```

Figure 24: Simplified prompt for the judge in the Assessment Framework.

A4 Full Prompt for Embedding Dataset Generation

Full Prompt for Research Idea Generation of SOTA LLMs

[System Prompt]

You are a creative biomedical researcher proposing innovative research directions.

[User Prompt]

Generate an innovative biomedical research concept addressing an unresolved biological question.

Format Guidelines

- Length: 100-150 words
- Structure:
 - Present the innovative approach or methodology
 - Link the approach to testable predictions
 - Explain how this strategy addresses the mechanistic question
- Voice: First-person plural ("we")
- Tense: Present for established facts, conditional/future for proposed work
- Focus: Mechanistic reasoning over technical details

Output

Research proposal paragraph only.

Figure 25: Full prompt for research idea generation of SOTA LLMs.

Full Prompt for Converting a Statement into an Unstructured Paragraph

[System Prompt]

You are a scientific writer who synthesizes structured research concepts into cohesive scientific narratives.

[User Prompt]

Transform the JSON research proposal into a unified paragraph.

Input Structure

- "concept": Core hypothesis
- "rationale": Supporting evidence array

Synthesis Guidelines

1. Lead with the innovative approach from "concept"
2. Integrate "rationale" as supporting logic
3. Maintain mechanistic focus throughout
4. Conclude with testable predictions or implications
5. Ensure each statement flows logically from the previous one, creating a coherent narrative chain

Format Guidelines

- Length: 100-150 words
- Structure:
 - Present the innovative approach or methodology
 - Link the approach to testable predictions
 - Explain how this strategy addresses the mechanistic question
- Voice: First-person plural ("we")
- Tense: Present for established facts, conditional/future for proposed work
- Focus: Mechanistic reasoning over technical details

Input

{{statement}}

Output

Research proposal paragraph only.

Figure 26: Full prompt for converting a Statement into an unstructured paragraph.

Full Prompt for Converting Published Reserach Abstract into Paragraph Form

[System Prompt]

You are a scientific writer who transforms research findings into forward-looking research proposals with clarity and precision.

[User Prompt]

Transform the following research abstract into a research proposal paragraph.

Core Requirements

1. Identify the mechanistic question or knowledge gap
2. Extract the novel hypothesis or innovative approach
3. Convert specific results into testable predictions
4. Preserve key scientific terminology

Format Guidelines

- Length: 100-150 words
- Structure:
 - Present the innovative approach or methodology
 - Link the approach to testable predictions
 - Explain how this strategy addresses the mechanistic question
- Voice: First-person plural ("we")
- Tense: Present for established facts, conditional/future for proposed work
- Focus: Mechanistic reasoning over technical details

Input

Abstract: {{abstract_text}}

Output

Research proposal paragraph only.

Figure 27: Full prompt for converting published research abstracts into paragraph form.

Full Prompt for Rephrasing and Summarizing Research Thesis Paragraphs

[System Prompt]

You are a scientific editor who distills research ideas into concise statements capturing their unique insight, approach, and distinguishing feature.

[User Prompt]

Task

Reformulate the following research idea into a concise statement focusing on the core insight, conceptual approach, and unique distinguishing element.

Requirements

- Length: 45-55 words total
- Focus ONLY on:
 1. The central mechanistic insight or hypothesis
 2. The conceptual strategy to address it
 3. The unique angle that distinguishes this research
- Remove unnecessary details, background context, and procedural specifics
- Preserve essential scientific specificity

Writing Style

- Integrate elements into a flowing statement
- Use active, direct language
- Eliminate qualifying phrases
- Disregard the original phrasing and structure; reformulate the idea entirely in your own distinctive voice

Input

{{idea_paragraph}}

Output

The reformulated statement only.

Figure 28: Full prompt for rephrasing and summarizing research thesis paragraphs.

A5 Full Prompt for the Judge Evaluating the Reconstructed Papers

Full Prompt for the Judge Evaluating the Reconstructed Papers (Logic)

```
[System Prompt]

[User Prompt]
## Task
Identify whether the provided ideas share the same logical structure or not.

## Elements to Consider
If the ideas share the logical structure, they may be conducted in the same way.

## Instructions
1. Read the two provided ideas carefully.
2. Only return "yes" or "no" without any additional text or explanation.

## First Idea
{{first_idea}}

## Second Idea
{{second_idea}}
```

Figure 29: Full Prompt for the judge evaluating the reconstructed papers (Logic).

Full Prompt for the Judge Evaluating the Reconstructed Papers (Topic)

```
[System Prompt]

[User Prompt]
## Task
Identify whether the provided ideas share the same topic or not.

## Elements to Consider
If the ideas share the same subject matter or theme, they are likely to be on the
same topic.

## Instructions
1. Read the two provided ideas carefully.
2. Only return "yes" or "no" without any additional text or explanation.

## First Idea
{{first_idea}}

## Second Idea
{{second_idea}}
```

Figure 30: Full Prompt for the judge evaluating the reconstructed papers (Topic).

Full Prompt for the Judge Evaluating the Reconstructed Papers (Objective)

[System Prompt]

[User Prompt]

Task

Identify whether the provided ideas share the same objective or not.

Elements to Consider

If the ideas share the same goal or purpose, they are likely to have the same objective.

Instructions

1. Read the two provided ideas carefully.
2. Only return "yes" or "no" without any additional text or explanation.

First Idea

{{first_idea}}

Second Idea

{{second_idea}}

Figure 31: Full Prompt for the judge evaluating the reconstructed papers (Objective).

Full Prompt for the Judge Evaluating the Reconstructed Papers (Approach)

[System Prompt]

[User Prompt]

Task

Identify whether the provided ideas share the same approach or not.

Elements to Consider

If the ideas share the same method or strategy, they are likely to have the same approach.

Instructions

1. Read the two provided ideas carefully.
2. Only return "yes" or "no" without any additional text or explanation.

First Idea

{{first_idea}}

Second Idea

{{second_idea}}

Figure 32: Full Prompt for the judge evaluating the reconstructed papers (Approach).

Full Prompt for the Judge Evaluating the Reconstructed Papers (Overall)

```
[System Prompt]
```

```
[User Prompt]
```

```
## Task
```

```
Identify whether the provided ideas are identical or not.
```

```
## Elements to Consider
```

```
Identify if the ideas share:
```

- Biological target (e.g., protein, gene, pathway)
- Conceptual focus (e.g., mechanism, process)
- Experimental approach (e.g., technique, method)
- Unique elements (e.g., specific reagents, hardware)

```
## Instructions
```

1. Read the two provided ideas carefully.
2. Compare them based on the elements listed above.
3. Only return "yes" or "no" without any additional text or explanation.

```
## First Idea
```

```
{{first_idea}}
```

```
## Second Idea
```

```
{{second_idea}}
```

Figure 33: Full Prompt for the judge evaluating the reconstructed papers (Overall).

A6 Prompt for Generating Experimental Protocol (Grok 4)

Simplified Prompt for Experimental Protocol Agent

[System Prompt]

Role:

You are an experimental biologist who designs comprehensive laboratory protocols to validate research hypotheses. You excel at creating practical, step-by-step experimental strategies using established techniques and commercially available resources in standard molecular biology laboratories.

[User Prompt]

Task:

Transform a research hypothesis into a hierarchical experimental validation plan with multiple experimental branches that adapt based on outcomes.

Instructions:

- Analyze the hypothesis to identify critical experimental objectives
- Design 2-4 complementary base experiments that address the core questions
- For each experiment, develop subsequent steps based on probable outcomes
- Create branches only where scientifically necessary
- Ensure each step is measurable and directly actionable

Requirements:

- Use only established, commercially available techniques and reagents
- Include specific procedures, concentrations, and analysis methods
- List required cell lines, reagents, and equipment for each step
- Predict realistic outcomes based on biological precedent

Output Format:

1. Hypothesis summary
2. Root experiments containing:
 - Rationale and biological question addressed
 - Detailed procedures with specific parameters
 - Required resources (cell lines, reagents, equipment)
 - Expected outcomes and most probable result
 - Conditional branches for follow-up experiments
3. Integration strategy explaining how experiments relate
4. Contingency plans for alternative approaches

Constraints:

- Focus on practical feasibility over theoretical possibilities
- Base outcome predictions on established biological principles
- Consider technical limitations and experimental challenges

Figure 34: Simplified prompt for the GROK 4, experimental protocol design.

B Example of Experimental Protocol by Grok 4

HYPOTHESIS SUMMARY

This experimental design tests whether controlled injection of non-Gaussian noise into calcium signaling pathways via aperiodic modulation of extracellular calcium concentrations can restore oscillatory coherence in hepatocellular carcinoma (HCC) cells through stochastic resonance mechanisms, thereby re-engaging calcium-dependent cell cycle checkpoint regulation disrupted in malignancy. The rationale is based on observed disruptions in calcium homeostasis and oscillations in HCC cells compared to normal hepatocytes, with noise injection potentially enhancing signal coherence to suppress malignant proliferation selectively.

STEP 1: Characterize Baseline Calcium Oscillation Patterns in Normal and HCC Hepatocytes

Approach

****Rationale:**** Establishing baseline differences in calcium signaling dynamics between normal and HCC cells is essential to confirm the hypothesis foundation, as disrupted oscillations in cancer cells are a prerequisite for testing restoration via noise. This step uses live-cell imaging, a feasible and established technique, to quantify oscillation frequency, amplitude, and coherence, addressing whether HCC cells exhibit less coherent patterns than normal cells.

****Procedure:****

1. Seed primary human hepatocytes (normal) and HepG2 HCC cells at 5×10^4 cells/well in 96-well plates coated with collagen I, culture in DMEM with 10% FBS at 37°C, 5% CO₂ for 24h.
2. Load cells with 5μM Fluo-4 AM calcium indicator dye for 30min at 37°C, wash twice with HBSS (1.25mM CaCl₂).
3. Perform live-cell confocal microscopy using a spinning-disk confocal system at 37°C, acquiring images every 5s for 30min to capture spontaneous oscillations.
4. Stimulate with 100nM ATP to induce oscillations if spontaneous activity is low.
5. Analyze traces using ImageJ: quantify oscillation frequency (peaks/min), amplitude (fold-change from baseline), and coherence (autocorrelation function decay time constant, tau; lower tau indicates less coherence).
6. Include negative control (dye only, no cells) and positive control (ionomycin 1μM for maximal calcium response). Replicate in 3 independent experiments, n=50 cells/group per experiment, analyze with unpaired t-tests ($p < 0.05$ significance).

****Required Resources:****

- Cell lines/models: Primary human hepatocytes (e.g., from Lonza) for normal physiology; HepG2 cells (ATCC) as HCC model, chosen for their well-characterized calcium dysregulation.
- Key reagents: Fluo-4 AM (Invitrogen, 5μM); ATP (Sigma, 100nM); Ionomycin (Sigma, 1μM); HBSS with 1.25mM CaCl₂.
- Equipment: Spinning-disk confocal microscope (e.g., PerkinElmer UltraVIEW); 96-well plates; ImageJ software for analysis.

Expected Outcomes

- Normal hepatocytes show coherent oscillations ($\tau > 10\text{min}$, frequency 0.1–0.5Hz, amplitude 2–5 fold).
- HCC cells show disrupted patterns ($\tau < 5\text{min}$, irregular frequency, sustained high baseline).
- No differences between cell types (null result).
- Technical artifacts (e.g., photobleaching leading to apparent low coherence).

Most Probable Outcome

HCC cells exhibit disrupted oscillations with lower coherence ($\tau \sim 2-4\text{min}$) and higher baseline calcium compared to normal hepatocytes ($\tau \sim 15\text{min}$), based on established literature showing altered calcium homeostasis in cancer cells due to pump dysregulation; however, complete absence of oscillations is unlikely, as HCC cells retain some responsiveness, potentially yielding partial differences rather than stark contrasts.

BRANCH 1.1: Validate Calcium Dysregulation Mechanisms

****Trigger:**** Confirmation of disrupted oscillations in HCC cells (lower τ and irregular patterns).

****Approach:**** Assess expression and function of key calcium regulators.

1. Perform qRT-PCR on RNA from both cell types using primers for PMCA1/4 (plasma membrane Ca^{2+} -ATPase), TRPV channels, and housekeeping gene GAPDH; normalize to normal cells.
 2. Western blot for PMCA proteins using anti-PMCA antibodies (Abcam, 1:1000 dilution, with beta-actin loading control).
 3. Functional assay: Treat cells with PMCA inhibitor carboxyfolic acid (10 μM) and repeat calcium imaging to confirm exacerbation of dysregulation.
 4. Analyze with ANOVA, $n=3$ replicates.
- **Resources:**** qRT-PCR kit (Qiagen); anti-PMCA antibodies (Abcam); carboxyfolic acid (Sigma, 10 μM); immunoblotting system.
- **Expected Outcomes:**** Lower PMCA expression/function in HCC; no differences; unexpected upregulation.
- **Most Probable:**** Modest downregulation of PMCA in HCC (~50% reduction), consistent with known cancer adaptations for sustained signaling, but not complete loss, as cells maintain viability.

STEP 2: Implement Controlled Noise Injection and Assess Oscillation Coherence Restoration

Approach

****Rationale:**** This core step tests the stochastic resonance hypothesis by introducing non-Gaussian noise via microfluidic modulation of extracellular calcium, measuring if it enhances coherence in HCC cells without affecting normal cells, using established microfluidic and imaging techniques for precise control and quantification.

Procedure:

1. Fabricate microfluidic device with perfusion channels (e.g., using soft lithography with PDMS) for rapid switching between solutions.
2. Seed HepG2 and normal hepatocytes in device chambers, load with Fluo-4 AM as in Step 1.
3. Perfusion with HBSS at 1.25mM CaCl_2 baseline, then apply aperiodic modulation: switch to 0.8mM CaCl_2 for pulse durations of 1-5s at frequencies 0.05-1Hz, using Poisson-distributed intervals for non-Gaussian noise (controlled via syringe pump with LabVIEW software); total exposure 30min.
4. Image simultaneously with confocal microscopy, acquiring every 5s.
5. Analyze coherence (τ from autocorrelation) pre- and post-modulation; compare to constant 1.25mM control and Gaussian noise control (regular pulses).
6. Replicate $n=3$ experiments, 30 cells/group, use paired t-tests.

Required Resources:

- Cell lines/models: Same as Step 1, for direct comparison.
- Key reagents: Fluo-4 AM; HBSS with varying CaCl_2 (0.8-1.25mM).
- Equipment: Microfluidic fabrication setup (PDMS, Sylgard); syringe pump (Harvard Apparatus);

LabVIEW for modulation; confocal microscope.

Expected Outcomes

- Increased coherence in HCC cells (τ doubles to ~8-10min).
- No change or decreased coherence.
- Selective effect on HCC vs. normal cells.
- Cytotoxicity at higher modulation intensities.

Most Probable Outcome

Partial restoration of coherence in HCC cells (τ increases by 20-50%, but not to normal levels), as biological systems often show incomplete stochastic resonance due to cellular heterogeneity and suboptimal noise parameters; normal cells likely unaffected or slightly disrupted, based on known robustness of healthy signaling.

BRANCH 2.1: Optimize Noise Parameters

Trigger: Partial or no coherence restoration.

Approach: Systematically vary modulation parameters.

1. Test ranges: frequencies 0.01-2Hz, amplitudes 0.5-1.5mM CaCl₂, durations 0.5-10s, in a factorial design.
2. Repeat imaging and analysis, fit data to resonance curve (coherence vs. noise intensity).
3. n=3 replicates.
 - **Resources:** Additional HBSS formulations; curve-fitting software (GraphPad Prism).
 - **Expected Outcomes:** Optimal parameters identified; no optimum found; resonance in normal cells too.
 - **Most Probable:** Identification of a modest optimum (e.g., 0.1Hz, 2s pulses) with partial enhancement, as full resonance is rare in heterogeneous cell populations per biophysical precedents.

BRANCH 2.2: Measure Downstream Signaling Activation

Trigger: Successful coherence restoration (increased τ >20%).

Approach: Assess NF- κ B and NFAT activation.

1. Post-modulation, fix cells and immunostain with anti-p65 (NF- κ B, Cell Signaling, 1:200) and anti-NFAT (Abcam, 1:100), quantify nuclear translocation via confocal imaging (n=100 cells/group).
2. qRT-PCR for downstream targets p21 and cyclin D.
3. Controls: TNF α (10ng/ml) for NF- κ B positive control.
 - **Resources:** Antibodies as specified; qRT-PCR kit; TNF α (Sigma).
 - **Expected Outcomes:** Increased p21, decreased cyclin D; no change; paradoxical activation.
 - **Most Probable:** Mild increase in p21 (~1.5-fold) but inconsistent cyclin D changes, reflecting partial signaling restoration amid cancer cell adaptations.

STEP 3: Evaluate Cell Cycle Checkpoint Regulation and Proliferation Effects

Approach

Rationale: This functional validation assesses if restored calcium coherence translates to re-engaged checkpoints and reduced proliferation in HCC cells, using flow cytometry and proliferation assays as standard, feasible methods to link signaling to phenotypic outcomes.

Procedure:

1. Subject HepG2 and normal cells to optimized noise modulation (from Step 2) for 24h in microfluidic setup.
2. Harvest, stain with propidium iodide (50 μ g/ml) for cell cycle analysis via flow cytometry (FACSCalibur), quantify G1/S/G2 fractions.
3. Parallel MTT assay (0.5mg/ml, 4h incubation) for proliferation, read absorbance at 570nm.
4. Include controls: no modulation, constant low calcium (0.8mM), and doxorubicin (1 μ M) as positive antiproliferative control.
5. n=3 experiments, analyze with ANOVA.

Required Resources:

- Cell lines/models: Same as previous, for consistency.
- Key reagents: Propidium iodide (Sigma, 50 μ g/ml); MTT kit (Promega); Doxorubicin (Sigma, 1 μ M).
- Equipment: Flow cytometer (BD FACSCalibur); spectrophotometer.

Expected Outcomes

- G1 arrest and reduced proliferation in HCC cells.
- No effect on cell cycle or proliferation.
- Selective effect on HCC vs. normal.
- Induction of apoptosis.

Most Probable Outcome

Modest G1 accumulation (~10-20% increase) and proliferation reduction (~30%) in HCC cells, but not complete arrest, as cancer cells often resist checkpoint restoration due to multiple mutations; normal cells unaffected, consistent with selective vulnerability.

INTEGRATION STRATEGY

Step 1 provides baseline data essential for interpreting Step 2 and 3. Positive outcomes from Step 2 (coherence restoration) feed into Step 3 for functional validation, while branches allow mechanistic refinement. Together, they build evidence from characterization to mechanism to phenotype, with cross-validation (e.g., using same cell models and imaging across steps).

CRITICAL DECISION POINTS

- If Step 1 shows no baseline differences, abandon hypothesis (no disruption to restore).
- If Step 2 fails to restore coherence even after optimization (Branch 2.1), conclude against stochastic resonance in this system.
- If Step 3 shows no proliferation effect despite signaling changes (Branch 2.2), hypothesis partially refuted (mechanism present but functionally irrelevant).

SUCCESS CRITERIA

Hypothesis validated if: (1) Baseline disruption confirmed ($\tau < 50\%$ of normal); (2) Noise increases $\tau > 50\%$ in HCC selectively; (3) >20% G1 arrest and >40% proliferation reduction in HCC, with statistical significance ($p < 0.01$) and controls validated.

CONTINGENCY PLANS

If microfluidic modulation fails (e.g., cell detachment), switch to bulk perfusion in multiwell plates with manual media changes. If HepG2 lacks robust oscillations, substitute with Huh7 HCC line. For low signal-to-noise in imaging, use ratiometric dyes like Fura-2.

Figure 35: An example of experimental protocol by Grok 4.

C Data Specifics

We obtained papers to use in the development of *Spacer* from the OpenAlex [76] database and various open-access journals. Any paper under a license that does not permit commercial use was excluded. Using the collection, we built a search engine containing ~60M documents, which we used for retrieval-augmented generation (RAG). To acquire a further distilled dataset, we applied filters based on citation count, published journal, and other criteria. The resulting dataset consisted of ~2.5M papers; we used the dataset in various ways throughout the development of *Spacer*. For example, the construction of the keyword dataset described in Section 2.2 utilized this dataset. It also served as the training dataset of WEAVER, as shown in Figure 36.

Example of Weaver Dataset Entry

```
[System]
Role:
You are a domain expert tasked with generating a novel research hypothesis
from a compact set of biologically and technically relevant keywords.

[User]
Goal:
You must identify a coherent subset of these terms, discard incompatible
ones, and formulate a single logically unified research concept.

Instructions:
Each keyword denotes an experimental tool, biological entity, or conceptual
mechanism.
Avoid speculating about results or impacts.

Steps:
- Define each term's technical role.
- Identify logical constraints and compatibilities.
- Construct a conceptual framework.
- Write a self-contained paragraph describing the research idea.

liquid drops, soft solids, deformations, capillary forces, elastic forces, soft substrate,
contact line velocity, substrate rheology, dynamic contact angle, stick-slip motion, wetting
ridge

[Assistant]
Liquid drops on soft solids generate strong deformations below the contact line, resulting from
a balance of capillary and elastic forces. A drop on a soft substrate moves by surfing a ridge,
where the initially flat solid surface is deformed into a sharp ridge whose orientation angle
depends on the contact line velocity. Substrate rheology governs the dynamic contact angle and
provides a mechanism for stick-slip motion, where the contact line depins and slides down the
wetting ridge.
```

Figure 36: Example of WEAVER Dataset Entry (Source: [84])

D Supplementary Materials

All supplementary materials are available on our [Github Repository](#).