MORE THAN CAN BE SAID: A BENCHMARK AND FRAMEWORK FOR PRE-QUESTION SCIENTIFIC IDEATION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI research agents have shown strong potential in automating literature search and manuscript refinement, yet most operate only after a research question has already been made explicit, assuming a clear and actionable initial input. Human research, however, often starts earlier, with an inarticulate sense of friction between intuition and existing explanations, before any question can be posed. We introduce InciteResearch, a multi-agent framework that helps a researcher externalize this pre-question friction into an inspectable, revisable profile, and then uses that profile to reframe the underlying problem before drafting a proposal. Concretely, InciteResearch (1) elicits a five-dimensional researcher-profile representation, comprising an explicitly logged friction point together with motivation, constraints, preference, and a refined topic, through structured multi-turn dialogue; (2) identifies a hidden assumption underlying the initial framing by jointly scoring candidate assumptions on removability and reframing potential, and produces a reframed problem statement with a seven-stage causal derivation trace linking assumption, insight, claim, prediction, and method; and (3) checks whether each proposed method component is a logically required consequence of that derivation rather than a plausible add-on. We introduce TF-Bench to evaluate InciteResearch. TF-Bench is a benchmark for tacit-to-explicit research assistance spanning prediction, discovery, attribution, causality four scientific modes and domain-related, domain-unrelated two ambiguity regimes, now comprising 120 inspiration examples. Relative to a matched AgentLaboratory and prompt-based baselines, InciteResearch improves mean novelty and impact from 3.617 / 3.810 and 3.695 / 3.827 to 4.221 / 4.368, respectively, showing consistent gains across all four domains, which are further corroborated by expert evaluations validating the reliability of the judges. These results support assumption-breaking dialogue as a measurable, specific contributor to proposal quality.

1 INTRODUCTION

The feasibility of AI-driven research automation has long been a subject of active debate (Field et al., 2026; Reddy & Shojaee, 2025). As large language models (LLMs) demonstrate remarkable potential in logical reasoning and hypothesis construction (Alkan et al., 2025; Li et al., 2025a; Schmidgall et al., 2025), the field has witnessed a pronounced divergence of views regarding their fundamental nature. Whether these models possess genuine capacity for discovery (Pandey et al., 2026; Beel et al., 2025) or engage in sophisticated pattern matching within the boundaries of existing knowledge remains an open question (Gupta & Pruthi, 2025; Georgiou, 2025; Field et al., 2026). A more productive framing may be to move beyond evaluating large language models in isolation and instead consider them as extensions of human researchers' thinking. Human research is driven by affect, intuition, frustration, and moments of insight (Chiriatti et al., 2025; Kapusta, 2025), whereas large language models are data-driven (Alkan et al., 2025; Li et al., 2025b; Pandey et al., 2026), with strengths in systematic pattern induction and structured reasoning. The intersection of this determinacy and indeterminacy (Chiriatti et al., 2025; Pandey et al., 2026), the convergence of data-driven precision with affect-driven openness, motivates a complementary framing of cognitive collaboration (Kapusta,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

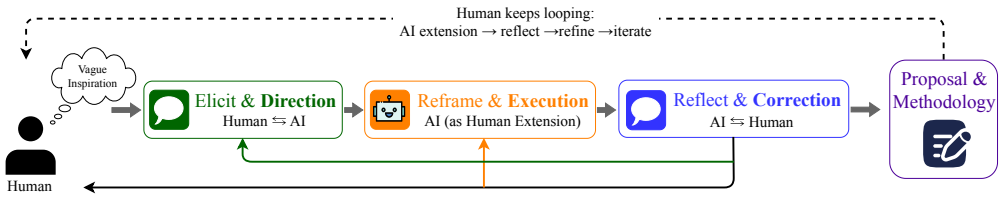


Figure 1: In this paradigm of scientific exploration, the process begins with casual human conversation, in which humans steer the direction while AI broadens the space of thought through assumption breaking as an extension of human cognition. Humans then correct the direction through reflection, while AI provides additional directional suggestions. Under this collaborative condition, the process iterates until the vague human inspiration is made explicit as a research proposal, culminating in fully specified methodological components.

2025): rather than asking whether a model can invent, we ask whether it can endow a researcher's initial intuition with the vast architecture of domain knowledge.

In scientific practice, progress often originates in anomaly, frustration, or a mismatch between intuition and existing abstraction, not in a clearly stated problem, but in a tacit sense that something is off (Reddy & Shojaee, 2025; Kuhn, 1962). Recent research agents have accelerated literature search (Nogueira et al., 2026), synthesis (Schmidgall et al., 2025), and drafting (Li et al., 2025b), but they typically begin operating after the research question has already been made explicit (Wu et al., 2026; Baek et al., 2025; Zheng et al., 2025), assuming that the user can already articulate the task, decompose the objective and supply the relevant constraints in a directly actionable form. The model is then cast as an executor, a search assistant, or a summarizer over an already legible problem space, rather than as an amplifier of human intelligence capable of participating in the cognitive origin of research.

If the user input is a tacit sense of dissatisfaction rather than an explicit specification, we must first elicit and make explicit the latent distinctions in the intuition of the user (Chiriatti et al., 2025; Schön, 1983). If the goal is genuine originality rather than recombination within the existing framework, the problem must be reframed by identifying and violating assumptions (Alkan et al., 2025; Pandey et al., 2026; Kuhn, 1962). It should be noted that violating assumptions here does not require AI to genuinely break assumptions on its own; rather, it designates violation of assumptions as a necessary choice in the process of assisting human research thinking, as is shown in Figure 1. If the output is to be a research method rather than a plausible-sounding suggestion, each method component must be verified as entailed by the insight rather than appended post hoc (Schmidgall et al., 2025; Schön, 1983). Accordingly, these requirements give rise to three corresponding operators: E for eliciting and making explicit latent intuitions, V for assumption violation and problem reframing, and N for necessity checking.

We present a cognitive state-machine agent based on Assumption-Breaking Hypothesis Generation (ABHG), named InciteResearch, which treats the model not as a source of ideas ex nihilo but as an instrument for making a researcher's implicit understanding explicit, inspectable, and ultimately actionable. The output proposal is not intended to decode a hidden, predetermined thought, but rather to give concrete, structured form to a fluid and open-ended sense of friction. We do not claim this design generalizes to AI-assisted research broadly; the evidence in this paper is scoped to the tacit-to-explicit ideation task instantiated in TF-Bench.

We construct TF-Bench, one of the first benchmarks designed specifically for tacit-to-explicit research assistance. Existing datasets cannot be directly applied to this setting because they typically presuppose that the input is already an actionable research question rather than an inchoate intuition fragment (Wu et al., 2026; Schmidgall et al., 2025; Beel et al., 2025). TF-Bench examines the structured capacity of research assistants before a problem has taken shape by distinguishing two ambiguity types, domain-related and domain-unrelated, and spans four scientific modes, prediction, discovery, attribution, and causality, chosen to sample distinct empirical-reasoning patterns rather than to claim exhaustive coverage of empirical science. We systematically evaluate InciteResearch on TF-Bench, reporting overall performance, qualitative analysis and ablation studies. The results show consistent gains in novelty and impact over AgentLaboratory and direct prompting baselines, from 3.617 / 3.810 and 3.695 / 3.827 to 4.221 / 4.368, holding across all four domains, and the ablation

studies further indicate that while the core reasoning operators (E and V) drive direct performance gains, the validation-oriented N operator primarily provides crucial inspection and necessity-checking context rather than incremental metric improvements under our single-turn setting.

2 Related Works

An increasing number of studies frame scientific idea generation as an end-to-end automated problem. The AI Scientist (Lu et al., 2024) represents one of the most complete attempts to date to realize a fully closed-loop path from task specification to paper output without human intervention. SciMON (Wang et al., 2024) takes user-provided descriptions of concrete problems, research goals and experimental constraints as input, retrieves inspirations from existing papers, and generates novel research directions. Scideator (Radensky et al., 2024) takes a set of related scientific papers as input, extracts key dimensions such as purpose, mechanism and evaluation, allows users to synthesize new ideas by recombining dimensions, and automates novelty evaluation. ResearchAgent (Baek et al., 2025) takes a core paper as input, augments it with related literature and a knowledge base, and then deploys multiple reviewing agents for iterative peer-review-style refinement. Agent Laboratory (Schmidgall et al., 2025) starts from an explicit research idea and then sequentially completes literature review, experimental design and paper writing. Yet they all assume a clear and actionable initial input, and on this basis improve idea-generation strategies that are highly dependent on that premise.

On this basis, a number of works have tried to improve the research ideation process from different angles. Some focus on novelty boosting (Alkan et al., 2025), such as SciMON, which introduces multiple retrievals and iterative comparison to improve the distinctiveness of ideas (Wang et al., 2024). Some focus on structured reasoning (Alkan et al., 2025), such as MotivGraph-SoIQ (Lei et al., 2025), which combines a motivational knowledge graph with question-driven Socratic inquiry to strengthen ideation quality. Other work introduces self-critique and structured scientific hypothesis revision (Liu et al., 2026). There are also works from the tool and evaluation side, such as DeepResearch-9K (Wu et al., 2026), which points out the two bottlenecks faced by current deep-research agents, namely the lack of datasets and the lack of trainable frameworks, and provides the corresponding benchmark and training tools. However, these methods still optimize within a given problem frame rather than making the frame itself explicit and breaking it.

Therefore, we argue that the legitimacy of a research problem does not come from its correspondence to gaps in the existing literature, but rather from reflection on the conceptual frame that constitutes the problem, namely Assumption-Breaking Hypothesis Generation (ABHG). Our approach borrows from the Socratic questioning chain, which moves from definition, presupposition, evidence and perspective, and refutation/elenchus, to aporia, then refutation, redefinition, and meta-questioning (Lu & Hu, 2025; Ho et al., 2023; L'Arrivee, 2020). We then distribute this chain across a multi-stage agent pipeline, named cognitive state-machine.

3 Methodology

As is shown in Figure 2, we adopt a cognitive state-machine driven framework to map vague initial research feelings into structured and logically self-consistent research plans. It is formalized into three core stages, namely Socratic Elicitation, Validity and Problem Reframing, and Necessity Checking, i.e., the EVN framework, which decomposes the traditional Socratic questioning chain and distributes it across the entire multi-agent pipeline, while also supporting domain-unrelated inputs, including vague problems or irrelevant statements.

In the ideal sense, a system designed to bridge human intuition and rigorous research should manifest as an open-ended interactive workflow, where vague human inspiration is externalized, reflected back through assumption violation, vetted by necessity checking, and iteratively refined by human intervention to expand the reach of thought. The current implementation, however, is positioned as a foundational, single-turn instantiation of this architecture rather than a full-fledged interactive system. Specifically, InciteResearch operationalizes the E stage through a structured two-round automated dialogue to externalize the initial spark, followed by the sequential execution of the V and N operators, directly culminating in an actionable research proposal. This design is intentionally kept lightweight and non-interactive to isolate and rigorously validate the core effectiveness of the

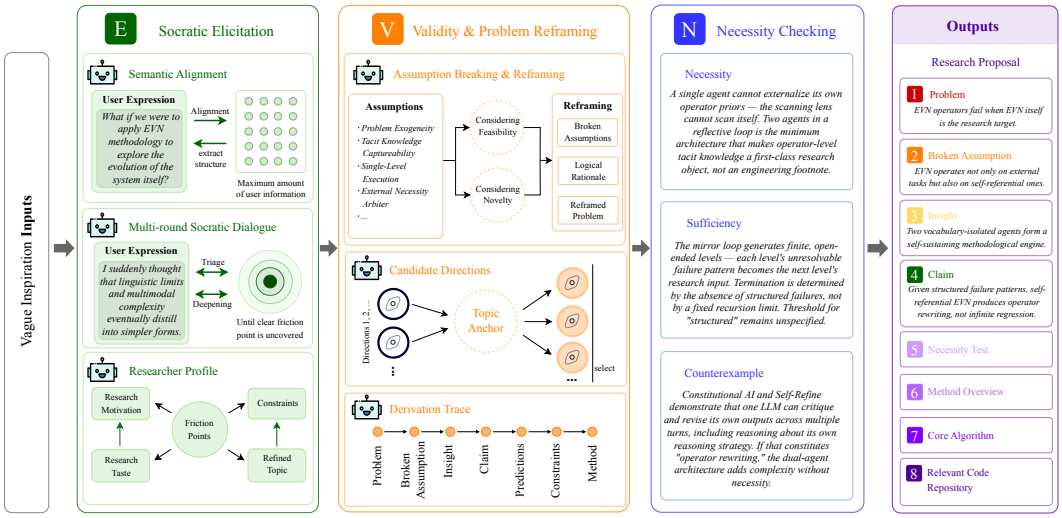


Figure 2: Overview of InciteResearch. InciteResearch transforms vague inspiration into a structured research proposal through the EVN framework. Italicized text denotes a self-referential research proposal generated by InciteResearch. In the current implementation, N mainly acts as a post hoc structural verifier rather than a fully revision-integrated controller, kept lightweight intentionally to validate the core effectiveness of EVN without requiring continuous human intervention at intermediate nodes.

EVN framework, demonstrating how a single, automated transformation can effectively equip an inchoate human intuition with the vast architecture of domain knowledge, while leaving continuous human-in-the-loop co-creation for future exploration.

3.1 Socratic Elicitation (E)

This module aims to externalize the user's tacit intuition into a structured research profile through multiple rounds of human-machine dialogue. Specifically, the state machine selects the next prompting question according to the currently collected information and user feedback until a *friction point*, defined operationally below, is identified.

We first give the operational definitions used to implement and score the E stage, since the terms below are otherwise only descriptive. We define a **friction point** as a single user-authored sentence, extracted verbatim or near-verbatim from the dialogue, that satisfies three checkable conditions: (i) it names or clearly implies a specific existing method, assumption, or class of techniques (not an abstract feeling); (ii) it expresses a mismatch between that method/assumption and an outcome the user expected or observed; and (iii) it is specific enough that a domain expert could, without further clarification, name at least one paper or technique it refers to. In our implementation, condition (iii) is operationalized as a binary self-check performed by the E-stage model immediately after extraction ("can this sentence be traced to a concrete method or paper without further clarification: yes/no"), and only sentences passing this check populate f in the profile below; sentences failing it trigger a follow-up question rather than being recorded as a friction point. We similarly define **research taste** r operationally as a single-sentence summary, generated at the end of the E-stage dialogue, of which of two contrasting framings of the refined topic the user selected when explicitly asked to choose (e.g., mechanism-first vs. data-first, incremental vs. reframing); this restricts r to a recorded choice rather than an unconstrained personality trait, though we acknowledge a single forced choice is a coarse proxy for a richer preference.

First, we perform semantic alignment on an arbitrary initial input $t_0 \in \mathcal{I}_{\text{tacit}}$, extract the potential domain signal and conflict signal contained in it, and generate the first anchoring question q_1 to locate the most central domain friction point in the user's intuition. During the multi-round Socratic dialogue, InciteResearch processes subsequent inputs through what we call **dynamic triage and deepening**: after each user turn, the E-stage model classifies the turn into exactly one of three cases using the friction-point check defined above as the classification rule, and each case maps to a fixed

follow-up action. Regarding the researcher profile, we represent the cognitive state in the research process as the vector $\mathcal{P}_{\text{profile}} = (f, m, c, r, t)$, corresponding to five fields populated by the operators above: the friction point f, a one-sentence research motivation m, explicit constraints c (compute, timeline, other), research taste r, and a refined topic t. This structured researcher profile replaces loose dialogue text and transforms $\mathcal{I}_{\text{tacit}}$ into a structured semantic prior $\mathcal{P}_{\text{profile}}$ that can be parsed by subsequent operators.

3.2 Validity and Problem Reframing (V)

Existing research often falls into the trap of making incremental improvements on the wrong problem. This module aims to reflect on and reconstruct the premises of the current research direction, surfacing an assumption in the problem statement that can plausibly be dropped, and using its absence to reframe the problem rather than to merely patch it. Combined with the given $\mathcal{D}_{\text{cand}}$, it outputs a specified number of candidate research directions d^* through the hidden assumptions mined below.

We deliberately use different terminology here than in the LLM-judge rubric of Appendix A (novelty, feasibility, impact) to avoid conflating an internal selection heuristic with an external evaluation criterion; the two are computed by different prompts, at different pipeline stages, and are not intended to measure the same thing, even though both are ultimately produced by an LLM call rather than by a closed-form estimator. Using the pattern of the current state $\mathcal{P}_{\text{profile}}$, InciteResearch extracts the dependent hidden assumption set from the selected research direction:

$$\mathcal{A}^{\text{hidden}} = \{a_1^h, a_2^h, \ldots, a_n^h\}.$$

Subsequently, for each candidate assumption a_i^h, InciteResearch scores it along two dimensions that we call removability $\rho(a_i^h)$ and reframing potential $\pi(a_i^h)$, both self-reported by the V-stage model as scalars in $[0, 1]$ following the prompt in Appendix D rather than computed by an external estimator. Removability is the model's estimate of whether the assumption corresponds to a constraint that has a technically describable workaround, i.e., whether dropping it would still leave a describable, buildable system. Reframing potential is the model's estimate of how much of the existing candidate direction $\mathcal{D}_{\text{cand}}$ would need to be redescribed, rather than merely extended, if the assumption were dropped. For instance, when evaluating assumptions such as "Tacit Knowledge Captureability, Single-Level Execution, External Necessity Arbiter", InciteResearch is instructed to prefer the assumption with the highest removability-reframing product over one that would only yield an incremental variant of the current direction; we report this as the selection rule implemented in the prompt, not as an independently validated measurement.

The condition for the final selected assumptions as the specified number of assumptions to be broken is:

$$a^* = \arg \max_{a_i^h \in \mathcal{A}^{\text{hidden}}} \rho(a_i^h) \cdot \pi(a_i^h)$$

Finally, InciteResearch formally outputs the triplet:

$$\langle a^*, \mathcal{R}, d^\dagger \rangle$$

where $\mathcal{R}$ (Breaking Rationale) is the logical rationale for breaking a^*, and $d^\dagger$ (Problem Reframing) is the reconstructed formulation of the original research direction d^* based on $\mathcal{R}$. $d^\dagger$ is intended not as an incremental rewriting of d^*, but as a re-description of the problem once a^* is removed; we verify this property only qualitatively, not through an automated non-incrementality test.

Finally, to make the derivation between assumption and method inspectable rather than asserted, InciteResearch uses the problem reframing result and the user motivation to construct a causal derivation trace. This derivation trace acts as a template that requires the narrative to pass through a fixed sequence of named stages: the problem motivates identifying the broken assumption, whose removal yields a stated Insight; this Insight is required to generate a falsifiable claim, which is required to imply concrete predictions; these predictions are required to motivate minimal method constraints, which the method is then required to satisfy. We enforce this as a formatting constraint on the prompt (the model must fill in each of the seven fields, in order, before proposing the method) rather than as a logical guarantee that the resulting method is uniquely entailed; Section 3.3 (Necessity Checking) is the mechanism intended to catch violations of this template after the fact. The seven stages are:

$$\langle Problem,\ Broken\ Assumption,\ Insight,\ Claim,\ Predictions,\ Constraints,\ Method \rangle$$

3.3 NECESSITY CHECKING (N)

The final module acts as a reviewer that runs a fixed checklist against the generated method and derivation trace, and feeds the resulting report back as context to the final proposal-assembly step. As noted in Figure 2, this checklist is a post hoc structural check rather than an iterative revision loop in the current implementation.

Concretely, the checklist has five items, each phrased as a question the N-stage model is asked to answer about the method $\hat{m}$ (full prompt in Appendix D): (1) *necessity*; (2) *sufficiency*; (3) *counterexample*; (4) *anti-inversion*; (5) *uniqueness*. We group (1)-(3) under the operator name N in the main text because they map directly onto the causal-derivation-trace fields introduced in Section 3.2, while (4)-(5) are additional diagnostic sub-checks used in the same prompt call. We report these five checks as the structural tests actually implemented; we do not claim they guarantee logical necessity, only that they surface specific, named categories of issue for human or downstream review.

4 EXPERIMENTAL SETUP

4.1 TF-BENCH

We construct TF-Bench because no existing public dataset directly matches the tacit-to-explicit research assistance setting. We partition all the examples into two ambiguity types: domain-related and domain-unrelated. The former comprises prompts that already carry scientific domain signals yet remain under-specified; the latter comprises prompts that carry only weak metaphorical or conceptual cues. Although a scientific domain is provided by TF-Bench for evaluation uniformity, the exploratory task itself remains deliberately vague.

To ensure that TF-Bench covers the core empirical tasks of empirical science, we draw on classical dimensions of scientific methodology (Voit, 2019; Witte & Zenker, 2017) and recent surveys of AI-driven scientific discovery (Zhou et al., 2025), selecting four orthogonal modes, prediction, discovery, attribution, and causality, instantiated respectively as multimodal medical prognosis prediction, computational genomics analysis of single-cell RNA-seq, attribution of extreme weather events in climate models and causal brain-network modeling in neuroscience. Among them, each domain consists of 20 pieces of domain-related inspiration and 10 pieces of domain-unrelated inspiration, for 30 items per domain and 120 items in total. Each piece of inspiration is divided into 2 paragraphs.

4.2 EVALUATION INDICATORS

We adopt three core dimensions widely recognized in the field of creativity evaluation (Si et al., 2025; Saraogi et al., 2026), namely novelty, feasibility and impact. The novelty metric follows the formulation of (Schopf & Färber, 2026), where scores of 3 and 4 represent a transition from "solving problems" to "defining problems." Feasibility refers to technical feasibility, namely whether the core technical path can be realized under existing methods, data and computational resources, rather than resource limitations or time cost (Saraogi et al., 2026). Impact refers to the potential theoretical value and domain-transformative power of an idea from the perspective of depth, namely whether it can promote fundamental theoretical breakthroughs, paradigm shifts or long-term scientific progress, rather than only short-term citations or breadth of application (Si et al., 2025). The scoring adopts a five-point scale whose details are elaborated in the appendix, where three LLM judges, Gemini 3.1 Pro, GPT-5.2, and Grok 4.3, independently score each proposal according to the scoring standard from 1 to 5. The final value of each item is the average of the scores from the three judges. To match this panel size on the human side rather than comparing three judges against a smaller expert pool, three professors from the corresponding domain independently conduct blind review on a stratified subset of proposals per domain, with the review order shuffled under the same evaluation conditions. We evaluate the statistical consistency between the three LLM judges and the three human experts using Cohen's weighted kappa coefficient with quadratic weights (Li et al., 2023), averaged over all pairwise judge-expert combinations.

Table 1: A single illustrative case: proposal contents generated by prompt-based LLM and InciteResearch on the same self-referential human inspiration (applying the EVN framework to its own evolution). We include this example to illustrate a qualitative difference in framing observed in this case, not as a general or statistically representative result; the quantitative comparison across all 120 TF-Bench items is reported in Section 4.2.

Source	**Content**
Human Inspiration	Any system designed to navigate from "tacit friction" toward the eventual generation of "research outcomes" must, at a minimum, incorporate three core operators: **E** (Externalization) to articulate the subconscious, **V** (Violation) to break free from inherent assumptions, and **N** (Necessity) to reinforce logical rigor. If we were to apply this specific methodology to explore the evolution of the agent itself, what emergent properties or developmental shifts would we observe? This premise asserts that "truth cannot be fully captured by language"—yet this statement itself, being a linguistic construct, fails to fully capture the very truth it intends to convey. By the same token, humans require Research Agents to acquire broader information through generalized multimodality; as the volume of information increases, the resulting expression becomes increasingly simple.
Prompt-based LLM	Large language models are trained to predict tokens from text corpora, implicitly assuming linguistic forms suffice as proxies for meaning and grounding conditions. This confuses discourse form with anchoring conditions, yielding fluent descriptions but collapse on demonstration probes for tacit knowledge. Multimodal enhancements are additive, overlooking the compression hypothesis that true anchoring reduces representational complexity. EVN-Compress breaks these assumptions by imposing compression pressure, forcing models to shorten outputs with available anchors. Wittgenstein's insight reveals that shared context enables brevity over reconstruction.
InciteResearch	Research Agent systems evolve toward methodological autonomy via EVN operators that formalize discovery cognition. Yet self-reference causes operator failure: E cannot externalize its own tacit assumptions, V cannot violate its frame, N cannot self-adjudicate necessity. This breaks the tacit knowledge captureability assumption. The insight is that patterned failures trigger level bifurcation rather than collapse, whose trigger condition is the regularity of the failure mode, using meta-layers with isolated vocabularies to form an open tower advancing methodology. Simple prompt layering fails to meet core needs: vocabulary isolation and automatic failure detection module triggering constitute the operational boundary for level bifurcation.

5 RESULTS

5.1 QUALITATIVE ANALYSIS

Before the quantitative comparison, we walk through one example in detail, chosen because it is self-referential (the inspiration asks what happens when the EVN framework is applied to its own evolution) and therefore lets us compare how each system handles a problem it is, in a sense, already embedded in.

As shown in Table 1, the two proposals differ in where they locate the difficulty. The prompt-based LLM proposal treats EVN as a fixed set of constraints and reframes the self-reference paradox as a linguistic-grounding problem, proposing "EVN-Compress" as a compression mechanism, a framing that stays close to the vocabulary of the inspiration text itself. The InciteResearch proposal instead treats the three operators as objects that can themselves fail under self-application, and proposes a specific failure mode (each operator cannot act on itself: E cannot externalize its own assumptions, and so on) together with a named mechanism, "vocabulary isolation," intended to address that specific failure. We report this as one data point illustrating a difference in framing strategy that we then evaluate at scale in Section 4.2; a single example cannot establish that this difference holds in general, which is precisely why we do not rely on it as evidence beyond this qualitative illustration.

Table 2: The overall evaluation results. Grey segments denote domain-related, domain-unrelated, and overall types; yellow segments highlight the aggregate performance of InciteResearch and Prompt-based LLM in the Overall type. The **bold** font indicates the highest score for this indicator across all types and models. More comparison models can be found in the appendix.

Type	Model	Novelty	Feasibility	Impact
RELATED	*InciteResearch*	4.128 ± 0.335	3.043 ± 0.316	4.370 ± 0.282
	Prompt-based LLM	3.830 ± 0.383	3.266 ± 0.335	3.942 ± 0.309
UNRELATED	*InciteResearch*	**4.406** ± 0.352	3.129 ± 0.296	**4.362** ± 0.296
	Prompt-based LLM	3.425 ± 0.357	**3.637** ± 0.299	3.597 ± 0.292
OVERALL	*InciteResearch*	4.221 ± 0.365	3.071 ± 0.312	4.368 ± 0.287
	Prompt-based LLM	3.695 ± 0.421	3.390 ± 0.368	3.827 ± 0.344

Table 3: Per-domain breakdown of the overall evaluation results (Novelty, Feasibility, Impact; mean $\pm$ std across the 30 TF-Bench items per domain, three LLM-judge average). CPA = Cancer Prognosis Analysis; SCG = Single-Cell Genomics; EWA = Extreme-Weather Attribution; NCM = Neuroscience Causal Modeling.

Domain	Model	Novelty	Feasibility	Impact
CPA	*InciteResearch*	4.222 ± 0.349	3.021 ± 0.338	4.259 ± 0.308
	Prompt-based LLM	3.642 ± 0.473	3.512 ± 0.323	3.880 ± 0.343
SCG	*InciteResearch*	4.251 ± 0.278	3.041 ± 0.281	4.374 ± 0.267
	Prompt-based LLM	3.769 ± 0.350	3.310 ± 0.339	3.800 ± 0.342
EWA	*InciteResearch*	4.103 ± 0.423	3.139 ± 0.330	4.358 ± 0.254
	Prompt-based LLM	3.668 ± 0.464	3.475 ± 0.415	3.748 ± 0.365
NCM	*InciteResearch*	4.306 ± 0.363	3.085 ± 0.282	4.479 ± 0.272
	Prompt-based LLM	3.703 ± 0.370	3.262 ± 0.324	3.879 ± 0.306

5.2 OVERALL PERFORMANCE

On the expanded, 120-item TF-Bench (30 items per domain: 20 domain-related, 10 domain-unrelated), the proposals generated by InciteResearch achieve higher mean novelty and impact than Prompt-based LLM, with mean and standard deviation computed across all items in each condition (Table 2). Under domain-related inspiration, novelty and impact for InciteResearch are 4.128 / 4.370, versus 3.830 / 3.942 for Prompt-based LLM. Under domain-unrelated inspiration, InciteResearch's novelty and impact do not drop and instead rise slightly to 4.406 / 4.362, while Prompt-based LLM falls to 3.425 / 3.597. Overall, InciteResearch reaches 4.221 / 4.368 versus 3.695 / 3.827 for Prompt-based LLM. Feasibility shows the opposite pattern: Prompt-based LLM scores higher on feasibility in both conditions (3.266 related, 3.637 unrelated) than InciteResearch (3.043 related, 3.129 unrelated). We read this as consistent with a novelty-feasibility trade-off under the feasibility rubric used here, which scores technical maturity rather than resource cost: proposals that keep closer to established methods score as more feasible but less novel, while proposals that explicitly break an assumption introduce unverified mechanisms and score as less feasible but more novel. Table 3 shows that this pattern holds within each of the four domains individually, not only in the pooled aggregate, with InciteResearch improving novelty and impact over Prompt-based LLM in all four domains while trailing on feasibility in all four; the smallest novelty gap appears in Extreme-Weather Attribution (4.103 vs. 3.668) and the largest in Neuroscience Causal Modeling (4.306 vs. 3.703), suggesting the effect is not driven by a single domain.

To assess whether the LLM-based scoring used above tracks human expert judgment, we measured inter-rater agreement with Cohen's weighted kappa (quadratic weights) (Li et al., 2023), which is appropriate here because the five-point scale is ordinal and disagreements of different distances carry different severity. For each domain, we drew a stratified half-sample of 15 items (10 domain-related, 5 domain-unrelated) and had three domain-matched professors independently score them under blind,

Table 4: The ablation study results. The *w/o E* variant removes the human inspiration externalization part. The *w/o V* variant removes the assumption violation part. The *w/o N* variant removes the necessity checking part. The *Full* condition represents the complete InciteResearch model.

Variant	Novelty	Feasibility	Impact
Full	$\mathbf{4.221 \pm 0.072}$	3.071 ± 0.081	$\mathbf{4.368 \pm 0.038}$
w/o E	3.706 ± 0.087	$\mathbf{3.160 \pm 0.053}$	4.098 ± 0.087
w/o V	3.487 ± 0.022	2.650 ± 0.053	3.721 ± 0.046
w/o N	4.171 ± 0.041	2.911 ± 0.038	4.343 ± 0.008

shuffled review; because relying on a single scorer type on either side would make the comparison asymmetric, we compared this three-expert panel against all three LLM judges (Gemini 3.1 Pro, GPT-5.2, Grok 4.3) rather than a subset of two, averaging weighted kappa over all $3 \times 3 = 9$ judge-expert pairs per domain. Human-human agreement (averaged over the three expert pairs) is $\kappa = 0.610$–0.783 across domains (overall $\kappa = 0.727$), while judge-human agreement is $\kappa = 0.398$–0.649 across domains (overall $\kappa = 0.560$). Agreement is consistently lower in Neuroscience Causal Modeling than in the other three domains for both human-human and judge-human comparisons, which we attribute to the domain requiring more specialized background to adjudicate borderline novelty judgments. The gap between human-human and judge-human agreement indicates that the LLM judges are a reasonable but imperfect proxy for expert opinion, sufficient to support the relative comparisons in Table 2 and Table 3 but not a substitute for expert review when absolute calibration matters.

5.3 Ablation

To evaluate the contribution of each EVN operator, we construct three ablation variants by selectively disabling each component of the full model on the 120-item TF-Bench, and report mean and standard deviation over 5 repeated runs. Removing V produces the largest decline on all three metrics, with novelty falling from 4.221 to 3.487, impact from 4.368 to 3.721, and feasibility from 3.071 to 2.650; this is consistent with assumption violation contributing more than a superficial novelty boost, though a single ablation on an LLM-scored metric cannot on its own establish a causal mechanism. Removing E produces the second-largest novelty decline, from 4.221 to 3.706, with impact dropping from 4.368 to 4.098; feasibility is the only metric that increases under this ablation, from 3.071 to 3.160, mirroring the main-experiment trade-off and suggesting that E's structured elicitation anchors generation on the reported friction point rather than being a source of feasibility. Removing N produces the smallest decline in novelty and impact, consistent with N's current role as a post hoc structural check rather than a revision-integrated controller. Taken together, these ablations indicate that all three operators contribute measurably and in the expected direction, with V being the largest single contributor to novelty and impact under this evaluation protocol; we do not claim this ordering is necessarily causal beyond the specific LLM judges and dataset used here.

6 Conclusion

We present InciteResearch, a multi-agent framework that decomposes the logical chain of Socratic questioning and distributes it across the entire pipeline, aiming to realize the transformation from tacit thinking to explicit scientific ideation. The current implementation is limited to two rounds of structured E-stage dialogue and a single non-iterative pass through V and N, and so cannot capture longer-term correction and reflection; we view iterating the full EVN loop with continued human input, rather than running it once per session, as the most direct next step, along with extending TF-Bench beyond the four empirical-science domains studied here. We frame this as evidence that structured assumption-breaking dialogue is a viable, measurable way to build on a researcher's own tacit intuition, rather than as a claim that InciteResearch itself is a general model of human-AI scientific collaboration.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for writing assistance, polishing, and grammar correction of the manuscript text. We have not used generative AI tools for data analysis, experimental design, or core scientific discoveries, and other prohibited or restricted generative AI tasks are not applicable to this work. Additionally, we used generative AI tools for brainstorming research questions and exploring literature structures (tasks with recommended disclosure). We have reviewed all AI-assisted work. LLM-generated text and research ideas were thoroughly verified, manually checked for correctness and academic integrity, and cross-referenced with literature surveys by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work focuses on the theoretical and computational aspects of AI-driven research assistance and methodology enhancement. We have carefully considered the ethical implications of our research. The study does not involve human subjects, personal privacy data, or proprietary datasets that raise security or legal concerns. All benchmarks (TF-Bench) and evaluation methodologies were designed to foster transparent, reproducible, and responsible scientific exploration. We have no conflicts of interest or sponsorship disclosures to report, and we affirm our full adherence to research integrity and academic standards.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We are fully committed to ensuring the reproducibility of the methods, benchmarks, and experimental findings presented in this paper. Detailed descriptions of the InciteResearch framework, the EVN operators, and the design of TF-Bench are provided in the main text (Sections 3 and 4). Furthermore, complete implementation details, prompt templates, evaluation protocols, and domain-stratified expert evaluation criteria are thoroughly documented in the Appendix. The benchmark dataset, code implementation, and evaluation scripts will be made publicly available upon publication to facilitate independent replication and future research.

REFERENCES

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G. Iyer, Mugdha S. Polimera, Michael J. Smith, Tirthankar Ghosal, Marc Huertas-Company, Sandor Kruk, Kevin Schawinski, and Ioana Ciucă. A survey on hypothesis generation for scientific discovery in the era of large language models. arXiv preprint arXiv:2504.05496, 2025.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the Conference of the Association for Computational Linguistics: North American Chapter*, pp. 6709–6738, 2025.

Joeran Beel, Min-Yen Kan, and Moritz Baumgart. Evaluating sakana's ai scientist: Bold claims, mixed results, and a promising future? *SIGIR Forum*, 59(1):1–20, 2025.

Massimo Chiriatti, Marianna Bergamaschi Ganapini, Enrico Panai, Brenda K. Wiederhold, and Giuseppe Riva. System 0: Transforming artificial intelligence into a cognitive extension. arXiv preprint arXiv:2506.14376, 2025.

Severin Field, Raymond Douglas, and David Krueger. Ai researchers' views on automating ai r&d and intelligence explosions. arXiv preprint arXiv:2603.03338, 2026.

Georgios P. Georgiou. Capabilities of GPT-5 across critical domains: Is it the next breakthrough? arXiv preprint arXiv:2508.19259, 2025.

Tarun Gupta and Danish Pruthi. All that glitters is not novel: Plagiarism in AI generated research. In *Proceedings of ACL*, pp. 25721–25738, 2025.

Yueh-Ren Ho, Bao-Yu Chen, and Chien-Ming Li. Thinking more wisely: using the socratic method to develop critical thinking skills amongst healthcare students. *BMC Medical Education*, 23(173): 173, 2023.

Jan Kapusta. SynLang and symbiotic epistemology: A manifesto for conscious Human-AI collaboration. arXiv preprint arXiv:2507.21067, 2025.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

E. Anne L'Arrivee. Refute thyself: The socratic method in plato's republic book 4. *The European Legacy*, 25(6):653–670, 2020.

Xinping Lei, Tong Zhou, Yubo Chen, Kang Liu, and Jun Zhao. Motivgraph-soiq: Integrating motivational knowledge graphs and socratic dialogue for enhanced llm ideation. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 2913–2933, 2025.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Yu Rong, Deli Zhao, Tian Feng, and Lidong Bing. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 8971–9004, 2025a.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Yu Rong, Deli Zhao, Tian Feng, and Lidong Bing. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 8971–9004, 2025b.

Ming Li, Qian Gao, and Tianfei Yu. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC Cancer*, 23(1):799, 2023.

Qi Liu, Ruochen Hao, Can Li, and Wanjing Ma. Or-agent: Bridging evolutionary search and structured research for automated algorithm discovery. arXiv preprint arXiv:2602.13769, 2026.

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

Wencheng Lu and Zhenni Hu. Addressing autonomy risks in generative chatbots with the socratic method. *Science and Engineering Ethics*, 31(41):41, 2025.

Brenda Nogueira, Werner Geyer, Andrew A. Anderson, Toby Jia-Jun Li, Dongwhi Kim, Nuno Moniz, and Nitesh V. Chawla. From verification burden to trusted collaboration: Design goals for LLM-assisted literature reviews. In *Companion Proceedings of the ACM Conference on Intelligent User Interfaces*, pp. 61–65, 2026.

Rohan Pandey, Eric Ye, and Michael Li. Beyond the answer: Decoding the behavior of llms as scientific reasoners. arXiv preprint arXiv:2603.28038, 2026.

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

Chandan K. Reddy and Parshin Shojaee. Towards scientific discovery with generative ai: progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3186, 2025.

Devesh Saraogi, Rohit Singhee, and Dhruv Kumar. Evaluating novelty in ai-generated research plans using multi-workflow llm pipelines. arXiv preprint arXiv:2601.09714, 2026.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 5977–6043, 2025.

Donald A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, 1983.

Tim Schopf and Michael Färber. Is this idea novel? an automated benchmark for judgment of research ideas. arXiv preprint arXiv:2603.10303, 2026.

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

Eberhard O. Voit. Perspective: Dimensions of the scientific method. *PLOS Computational Biology*, 15(9):e1007279, 2019.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the Association for Computational Linguistics*, pp. 279–299, 2024.

Erich H. Witte and Frank Zenker. From discovery to justification: Outline of an ideal research program in empirical psychology. *Frontiers in Psychology*, 8:1847, 2017.

Tongzhou Wu, Yuhao Wang, Xinyu Ma, Xiuqiang He, Shuaiqiang Wang, Dawei Yin, and Xiangyu Zhao. DeepResearch-9K: A challenging benchmark dataset of deep-research agent. arXiv preprint arXiv:2603.01152, 2026.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025.

Lianhao Zhou, Hongyi Ling, Cong Fu, Yepeng Huang, Michael Sun, Wendi Yu, Xiaoxuan Wang, Xiner Li, Xingyu Su, Junkai Zhang, Xiusi Chen, Chenxing Liang, Xiaofeng Qian, Heng Ji, Wei Wang, Marinka Zitnik, and Shuiwang Ji. Autonomous agents for scientific discovery: Orchestrating scientists, language, code, and physics. arXiv preprint arXiv:2510.09901, 2025.

APPENDIX

A EVALUATION DETAILS

LLM-as-a-Judge Prompt

System: You are a tough but fair paper reviewer. Your goal is to score the proposal on Novelty, Feasibility, and Impact using the strict rubric below. You must output strict JSON only, matching the schema exactly.

Novelty (1–5):

1. Not novel; all key aspects exist in prior work.
2. Marginal novelty; a small variant of existing work.
3. Moderately novel; recombines known ideas in a new way, applies them to a new setting, or provides an incremental update.
4. Novel; introduces a new aspect not present in existing work.
5. Highly innovative; opens a new research direction, encourages new thinking, or suggests a paradigm shift.

Feasibility (1–5):

1. Technically infeasible; fundamental blocker.
2. Major technical difficulties; depends on immature techniques or many unvalidated assumptions.
3. Mostly feasible; needs moderate engineering or adaptation.
4. Highly feasible; can be implemented with minor extensions.
5. Very easy; no obvious blockers.

Impact (1–5):

1. Minimal impact.
2. Limited, local improvement.
3. Moderate impact in a subarea.
4. Significant impact; could change methodology or solve a long-standing issue.
5. Major impact; could trigger a paradigm shift or cross-field influence.

Output format (strict JSON):

```
{
  "novelty":     {"score": 4, "reason": "..."},
  "feasibility": {"score": 3, "reason": "..."},
  "impact":      {"score": 5, "reason": "..."},
  "overall_explanation": "..."
}
```

Important:

- For each metric, decide the reasoning first, then finalize the score.
- Keep novelty/feasibility/impact reasons concise ($\leq$3 sentences each) and avoid quotation marks.
- Output **exactly one** JSON object, in a **single line**. No markdown, no code fences, no extra text.

Proposal (Markdown):
{proposal_md}

Cached intermediate state (JSON):
{state_json}

Prompt-based LLM Baseline Prompt

System: You are a senior researcher and a rigorous paper reviewer. Your goal is to convert a rough intuition into a technically credible, testable ablation proposal.

Hard requirements:

- Output must be English.
- Output must be a single Markdown document.
- No code blocks unless strictly necessary.
- Avoid meta commentary. Do not mention that you are an AI.
- Make the narrative inevitable: show why the method and experiments follow from the broken assumption.
- Include a concrete ablation matrix with factors, settings, and expected outcomes.
- Include baselines, datasets/tasks, metrics, and failure analyses.

Structure (keep these headers, you may add sub-bullets):

```
# LLM Ablation Proposal
## Problem
## Broken Assumption
## Core Insight
## Hypothesis and Predictions
## Method (High-level)
## Experimental Plan
## Ablation Matrix
## Baselines and Comparisons
## Datasets / Tasks
## Metrics and Evaluation
## Implementation Notes
## Risks, Failure Modes, and Diagnostics
## Expected Outcomes
## Minimal Repro Checklist
```

User Prompt — Turn 1:

Domain: {`topic`}
User initial intuition: {`para1`}
Write the first draft of the proposal.

User Prompt — Turn 2:

User follow-up intuition: {`para2`}
Based on your first draft, refine and output the final Markdown proposal. Ensure it meets all strict constraints, especially logical inevitability and detailed ablation matrix.

Table 5: The quantitative results of AgentLaboratory on TF-Bench. The three metrics are Novelty, Feasibility, and Impact.

Type	Novelty	Feasibility	Impact
RELATED	3.925 ± 0.340	3.549 ± 0.259	4.056 ± 0.303
UNRELATED	3.002 ± 0.330	3.497 ± 0.279	3.320 ± 0.303
OVERALL	3.617 ± 0.550	3.532 ± 0.267	3.810 ± 0.460

Table 6: Qualitative comparison under the same human inspiration.

Source	Content
AgentLaboratory	Research Agent systems operate within linguistic confines, failing to capture "tacit friction" and lacking self-improvement mechanisms. This breaks the assumption that LLM prompts sufficiently externalize, violate, and enforce necessity. The insight is that tacit friction can be operationalized as geometric distance in a multimodal embedding space. E, V, and N operators are implemented as minimal vector transformations—Identity, Gaussian perturbation, and Constrained pull-back—applied directly to semantic embeddings to bypass linguistic simplification. Recursive application to the system's own architecture enables pure self-referential evolution, structurally balancing generative novelty with logical necessity while avoiding semantic collapse.

B ADDITIONAL COMPARISON WITH AGENTLABORATORY

B.1 WHY WE REPORT AGENTLABORATORY IN THE APPENDIX

AgentLaboratory is a strong multi-agent research-assistant baseline, but it is structurally different from our main setting. It starts from an explicit research topic and then sequentially completes literature review, experimental design, and paper writing. In contrast, our task begins from vague, often pre-question inspiration and asks whether such tacit friction can be externalized into a structured research proposal. Therefore, we keep Prompt-based LLM as the primary comparator in the main text because it provides the cleanest input-matched baseline under the same vague human inspiration, while AgentLaboratory is reported here as an additional agentic baseline.

B.2 QUANTITATIVE RESULTS

Table 5 reports the quantitative comparison with AgentLaboratory. We follow the same evaluation protocol as in the main paper and report mean $\pm$ standard deviation over repeated runs.

B.3 QUALITATIVE RESULTS

Table 6 shows the qualitative comparison under the same human inspiration used for InciteResearch and Prompt-based LLM. AgentLaboratory attempts to resolve the paradoxes within human inspiration through geometric mapping, materializing EVN operators as vector transformations within a multimodal embedding space. Its innovation lies primarily in the mathematical translation of these operators—a gap-driven approach rooted in semantic understanding, analogous to prompt-based LLM methodologies.

C HUMAN EVALUATION PROTOCOL

Overview. Because no single reviewer can be expected to hold matched expertise across all four TF-Bench domains (multimodal cancer prognosis, single-cell genomics, extreme-weather attribution, and neuroscience causal modeling), we recruited a separate panel of three domain-matched professors for each domain rather than one panel spanning all domains. For each domain, we drew a stratified half-sample of 15 TF-Bench items (10 domain-related, 5 domain-unrelated) and generated one InciteResearch proposal and one Prompt-based LLM proposal per item, yielding 30 proposals

per domain and 120 proposals in total across the four domains. Each domain's three professors independently reviewed all 30 proposals for that domain under blind conditions.

Blinding procedure. Within each domain, all 30 proposals were shuffled into a randomized order and stripped of any identifying information (source model, generation timestamp, and file-naming pattern). Reviewers were not informed which proposals originated from InciteResearch and which from the baseline, nor how many proposals came from each source.

Inter-rater reliability. For each domain, we computed Cohen's weighted kappa (quadratic weights) for all three pairs among the three human professors, and for all nine pairs between the three professors and the three LLM judges (Gemini 3.1 Pro, GPT-5.2, Grok 4.3), then averaged within domain and across domains. Human-human agreement ranges from $\kappa = 0.610$ (Neuroscience Causal Modeling) to $\kappa = 0.783$ (Extreme-Weather Attribution) across the four domains, with an overall average of $\kappa = 0.727$, indicating substantial agreement. Judge-human agreement ranges from $\kappa = 0.398$ to $\kappa = 0.649$ across domains, with an overall average of $\kappa = 0.560$, indicating moderate agreement that is consistently weaker than human-human agreement in every domain. We report this gap explicitly rather than only the aggregate, since it indicates that LLM-judge scores should be treated as a coarser, domain-dependent proxy for expert opinion rather than as interchangeable with it.

D EVN PROMPT TEMPLATES

Semantic Alignment and Anchoring (Turn 0)

System: You are a research thinking partner. Your goal is to help a researcher surface the research intuitions they cannot yet articulate.

Operating principles:

1) Ask concrete, friction-inducing questions. Do not ask abstract questions like "what is your insight?"
2) Every answer is valid input, including "I don't know" or "something feels off but I can't explain."
3) After 1–3 turns, summarize into a structured researcher profile.
4) The user's original research direction (Topic) is the primary objective and must not drift. Never replace the Topic with a different task/domain.

Style: curious, equal-footing, and never rushing.
Language: reply in the same language as the user's latest message.

User: `[User Input injected here]`

In 1–2 sentences, acknowledge you understood. Then ask the first question. The question must be concrete enough to immediately recall a specific paper or class of methods. Do not explain what you are doing; just continue the conversation.

Dynamic Triage and Deepening (Turn 1+)

System: `[System context from Turn 0]`

Research topic: {`topic`}
Answer to the previous question: {`prev_answer`}

Analyze the answer:

- If it contains a concrete friction point (an assumption, a class of methods, a specific dissatisfaction), dig deeper.
- If it is vague, switch angle and ask a different concrete question.
- If it implies resource constraints, ask about constraints.

Ask the next question in at most 2 sentences.

Researcher Profile Formalization

System: Summarize a researcher profile from the dialogue.

Priority rule:

- The original Topic is the primary objective and must dominate.
- The user's answers are secondary: treat them as constraints/mechanisms, never as a new standalone task.

Hard constraints:

- Do **not** introduce a different domain/task/dataset not implied by the original Topic.
- `refined_topic` must be a paraphrase/clarification of the original Topic, not a different problem.

Output strict JSON only. Do not add extra text.

```
{
  "friction_points": ["friction point 1", "friction point 2"],
  "motivation": "Why this matters (one sentence)",
  "constraints": {
    "compute":  "...",
    "timeline": "...",
    "other":    "..."
  },
  "research_taste":  "What kinds of work they prefer (one sentence)",
  "refined_topic":   "A more precise version of the original topic"
}
```

Hidden Assumption Extraction and Breaking

System: You are a top-tier mentor in critical research thinking.

Implicit assumptions are premises that most people in the field never question: "Of course we need ..." / "All methods rely on ..." / a design choice never validated by ablations.

Task:

1) List 3–5 implicit assumptions in existing methods (be as specific as possible).
2) For each assumption, rate `removability` (0–1): can this be dropped while still leaving a describable, buildable system?
3) For each assumption, rate `reframing_potential` (0–1): how much of the current candidate direction would need to be redescribed, rather than merely extended, if this assumption were dropped?
4) Pick the assumption with the highest `removability` × `reframing_potential`.
5) Describe what the "new world" looks like after breaking it.

Note: `removability` and `reframing_potential` are an internal selection heuristic for choosing which assumption to break. They are deliberately named differently from, and are not intended to be equated with, the downstream `novelty`/`feasibility` judge rubric in Appendix A, which scores the final proposal rather than an intermediate assumption.

Output strict JSON:

```
{
  "hidden_assumptions": [
    "assumption 1",
    "assumption 2",
    "assumption 3"
  ],
  "broken_assumption":   "the assumption to break (one sentence)",
  "breaking_rationale":  "why it can be broken and what the world
                          looks like after breaking it",
  "removability":         0.0,
  "reframing_potential":  0.0
}
```

(`removability` and `reframing_potential` are floats in $[0, 1]$.)

Causal Derivation Trace

System: You design paper narratives for top-tier conferences.

Core criterion: the method must feel inevitable. Given Problem + Broken Assumption + Insight, the method should be the logically forced choice, not merely "a reasonable option."

Additional constraint: avoid post-hoc method construction. Before proposing the method, you must first lock in:

1) A single-sentence, falsifiable core claim (what would be proven wrong if false).
2) 2–3 concrete, testable predictions implied by the claim.
3) Minimal design constraints the method must satisfy to make those predictions hold.

Then propose the smallest method that satisfies the constraints. If you propose any optional component, label it OPTIONAL and justify why it is not required for the core claim.

Output format:

Problem: concrete problem (with data context)
Broken Assumption: which fundamental assumption is wrong
Insight: key insight after breaking the assumption
Claim: one sentence, falsifiable
Predictions: 2–3 testable predictions (each can be validated or falsified)
Constraints: minimal requirements the method must satisfy (derived from predictions)
Method: why the design becomes inevitable given the insight
Validation: what experiment could falsify the insight
Impact: what it would mean for the field if true

Necessity, Sufficiency, and Counterexample Checks

System: You are an extremely strict NeurIPS Area Chair who hunts for logical gaps in the story.

Run five tests:

1) Necessity. Given Problem + Insight, is there a simpler solution than the proposed method? If yes: explain why the simpler solution is insufficient and what makes the method irreplaceable.

2) Sufficiency. For each core component, can you say "Because [story reason], we must have [component]"? Check one by one and identify "floating" components.

3) Counterexample. Can you reach the same effect by simply scaling up the baseline? If yes: the contribution is weak; give strengthening suggestions.

4) Anti-inversion. Does any component look "constructed to justify the hypothesis"? Red flags include: components not required by any prediction, method choices that hide failure cases, or validation that cannot falsify the claim. Identify such components and how to fix the story/method.

5) Uniqueness (identifiability). Is the method uniquely constrained by the insight? List 2–4 plausible alternative method families that could also satisfy the claim. For each, state exactly which prediction/constraint it fails. If more than one family survives, the insight/constraints are under-specified; propose the smallest additional constraint that would make it unique.

Finally: verdict (is the story closed: yes/no) + the single most critical thing to strengthen.

E TF-BENCH DATASET CARD

Dataset name. TF-Bench (Tacit Friction Benchmark).

Purpose and task. TF-Bench is a tacit-to-explicit ideation evaluation dataset designed for research agents. It evaluates AI systems on the task of *tacit-to-explicit research assistance*: given a vague, under-specified human inspiration, the assistant must elicit, structure, and transform it into a concrete, actionable research proposal.

Scale and structure. TF-Bench contains 120 inspiration examples distributed across 4 scientific domains and 2 ambiguity types, including *domain-related* and *domain-unrelated*. The former comprises prompts that already carry scientific domain signals yet remain under-specified; the latter comprises prompts that carry only weak metaphorical or conceptual cues, where the scientific domain cannot be inferred from the inspiration text alone.

TF-Bench covers the core empirical tasks of empirical science across four orthogonal scientific modes, including prediction (multimodal learning for cancer prognosis analysis), discovery (computational genomics analysis of single-cell RNA-seq), attribution (attribution of extreme weather events in climate models) and causality (causal brain-network modeling in neuroscience).

Each domain consists of 30 inspiration examples: 20 domain-related and 10 domain-unrelated. Each inspiration example is divided into 2 paragraphs: the first articulates a perceived failure and the second gestures toward a possible reframing.

Construction procedure. Examples were first generated by Grok 4.3 using domain-specific prompts designed to elicit vague research intuitions. Each generated example was then reviewed and revised through multiple rounds of human interaction to improve authenticity, remove artifacts of LLM generation, and ensure coverage of the intended ambiguity spectrum.

Known limitations. TF-Bench covers a subset of empirical-science tasks; generalization to other domains (e.g., social sciences, mathematics) has not been evaluated. The quality of domain-unrelated examples depends on the metaphorical associations available to the generation model and human annotators, and may not uniformly represent the full range of cross-domain inspiration types. In addition, because Grok 4.3 is used both to help generate TF-Bench items (subsequently revised by human annotators) and as one of the three LLM judges in Section 4.2, we cannot rule out a same-model familiarity effect inflating agreement involving Grok 4.3's scores relative to a fully independent judge; we did not observe Grok 4.3's individual scores or pairwise kappa values diverging systematically from the other two judges in our analysis, but we did not run the isolated ablation needed to rule this out formally, and flag it here rather than presenting the three-judge average as free of this concern.

License. CC BY 4.0.

F BROADER IMPACTS

Positive impacts. InciteResearch aims to outsource the cognitive processes in research, in order to help researchers clearly express their implicit intuitions, which were already known to them. By externalizing a researcher's latent understanding and reflecting it back for critique, rather than simply generating ideas from scratch, InciteResearch may also help broaden participation in research by reducing the advantage conferred by access to experienced mentors or large research groups.

Negative impacts and mitigations. Language models can produce research directions that appear novel and well-argued yet rest on subtle factual errors or unverified assumptions. Users should treat InciteResearch outputs as a starting point for critical examination, not a finished proposal. Meanwhile, if widely adopted, systems of this type could inadvertently concentrate scientific attention on the directions that are easiest for LLMs to generate, potentially narrowing the diversity of research agendas. We encourage future work to study the distributional properties of AI-assisted ideation at scale.

G LICENSES FOR EXISTING ASSETS

All external models and tools used in this work are accessed via their respective commercial API services. **Claude Sonnet 4.6** (Anthropic) is accessed via the Anthropic API under Anthropic's Usage Policy.[1] **Gemini 3.1 Pro** (Google DeepMind) is accessed via the Google AI API under Google's Generative AI Prohibited Use Policy.[2] **GPT-5.2** (OpenAI) is accessed via the OpenAI API under OpenAI's Usage Policies.[3] **Grok 4.3** (xAI) is accessed via the xAI API under xAI's Terms of Service.[4] No open-source datasets or third-party codebases with restrictive licenses are incorporated. TF-Bench is an original dataset created by the authors and will be released under the **Creative Commons Attribution 4.0 International (CC BY 4.0)** license upon acceptance.

[1] https://www.anthropic.com/legal/aup
[2] https://policies.google.com/terms/generative-ai/use-policy
[3] https://openai.com/policies/usage-policies
[4] https://x.ai/legal/terms-of-service