EXPERIGEN: Agentic Hypothesis Discovery from Observational Data

Jishu Sengupta* **Harini S I*** **Somesh Singh*** **Syed Mohamad Tawseeq***

Yaman K Singla **Rajiv Ratn Shah** **David Doermann**

Jitendra Ajmera **Balaji Krishnamurthy**

Adobe Media and Data Science Research (MDSR)

IIIT-Delhi, SUNY at Buffalo, IIT Kanpur, Machlear Technologies

*Equal contribution. behavior-in-the-wild@googlegroups.com

behavior-in-the-wild.github.io/experigen.html

Abstract

Modern data-driven sciences use large observational datasets such as online discussions, behavioral logs, web experiments and biological records to explain outcomes and design interventions. Existing LLM-based hypothesis generators produce interpretable claims, but usually validate them through predictive accuracy, making them vulnerable to spurious correlations and dataset artifacts. Automated validation systems test hypotheses statistically, but require researcher-specified constructs. Here we show that EXPERIGEN, a two-agent closed-loop framework, discovers statistically supported natural-language hypotheses from observational data. A **Gen**erator proposes hypotheses, while an **Experi**menter operationalizes features, selects covariates and tests, executes analyses and returns evidence for refinement. Accepted and rejected hypotheses are stored in a labeled memory that guides further exploration. We evaluate EXPERIGEN across 19 tasks spanning text, images, clinical tabular data, economics, marketing and single-cell RNA-seq. On tasks and benchmark aggregates with nonzero baseline counts, it discovers approximately 1.5–3$\times$ as many statistically significant hypotheses as the strongest hypothesis-generation baseline, and improves HypoBench accuracy by an average of 5.0 percentage points over HypoGeniC. Expert evaluation by 26 specialists rated its hypotheses as novel, clear and research-worthy; 11 of 16 preregistered r/ChangeMyView hypotheses were significant in the predicted direction; and in deployed A/B tests across seven S&P 500 companies and 5.1 million users, 15 of 25 hypotheses achieved significance, with 14 matching the predicted direction. These results suggest that closed-loop hypothesis generation can help prioritize hypotheses for expert review and real-world intervention and augment human judgment in data-driven sciences.

1 Introduction

For decades, social scientists, behavioral researchers, and practitioners in marketing, public health, and policy have advanced empirical understanding of human behavior through a cycle of *observation*, *hypothesis generation*, and *experimental validation*. From epidemiologists linking smoking to lung cancer [49] to studies of human conformity under social pressure [2], progress has hinged on formulating testable hypotheses from noisy, context-dependent observations of human activity. Yet each step of this hypothesis generation cycle is manual and slow: analyzing large observational datasets, inducing novel patterns, and identifying hypotheses that merit the months of effort and

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

cost required to test them. The combinatorial space of plausible hypotheses, set against this effort, expense, and time, makes systematic discovery fundamentally difficult and bounded by individual researchers' intuition. The growing abundance of digital behavioral data amplifies this bottleneck further: the space of hypotheses worth exploring now far exceeds what manual analysis can cover. This motivates the central challenge in accelerating this progress:

Can computational methods generate hypotheses from observational data that are rigorous and likely to hold under expert scrutiny and prospective validation?

Consider the research question "What makes a counterargument persuasive?" This question is studied across social science, politics, and public policy, and we will use it as a running example throughout the paper. A researcher or practitioner studying online persuasion might begin by observing successful debates and hypothesizing that "longer counterarguments are more persuasive." They test this statistically; the effect is significant but modest. Examining the data, they notice that longer arguments tend to include more citations and external evidence, raising a sharper question: is length driving persuasion, or is it the citations? A follow-up analysis controlling for the covariate shows that length still contributes independently. Crucially, this second question was only conceivable because the first test had been run; each round of evidence reshapes what is worth asking next. This iterative, adaptive process of proposing, testing, and refining hypotheses distinguishes robust findings from shallow correlations. A hypothesis earns credibility only after surviving repeated rounds of statistical scrutiny and refinement, and only then is it ready to inform real-world interventions such as A/B testing. Any computational method for this setting must therefore produce not only plausible claims, but operationalized hypotheses with statistical evidence, robustness checks and a path to prospective evaluation.

Automating this cycle reduces analyst effort, helps prioritize which expensive experiments are worth running, and compresses the iteration cycle from months to hours. This would significantly impact the scientific and economic landscape of multiple fields like social science, education, healthcare, and marketing [29] to help us answer questions, ranging from what makes an image memorable to what changes opinion in online discourse.

Existing automated approaches either require researcher-specified features [1] or, as with recent LLM-based generators [52, 34], verify hypotheses only through predictive metrics like held-out accuracy (§2). These methods are prone to spurious correlations, fail to account for unobserved covariates, and do not measure effect sizes or test hypotheses under varied conditions. In our manual verification, 20–25% of hypotheses generated by prior methods failed after controlling for covariates and applying proper tests. For example, Movva et al. [34] discover that "I ask unanimous consent" predicts Republican speakers in U.S. congressional speeches with high accuracy, yet the phrase is confounded with the majority party holding the floor; after controlling for it, the phrase has no correlation with Republican speeches. Moreover, because these methods generate hypotheses in a single pass, they are unable to find multivariable or conditional relationships that emerge only through iterative refinement (§F).

We present EXPERIGEN, a framework that generates hypotheses from observational data including unstructured, multimodal, relational, tabular and high-dimensional biological records and validates them with appropriate statistical tests, covariate control, and iterative refinement. EXPERIGEN orchestrates two specialized LLM agents: a *Generator* that proposes testable hypotheses, and an *Experimenter* that plans features, covariates, and statistical tests, producing a structured analysis. The agents communicate through multi-turn conversation and are employed in a two-phase search. In the *outer loop*, the Generator proposes a seed hypothesis that is both plausible and novel relative to previously validated discoveries. In the *inner loop*, the Experimenter evaluates the hypothesis and returns its analysis; the Generator uses this feedback to reformulate features, refine context, or test combined effects, iterating until the hypothesis either attains statistical significance or is rejected. A labeled memory of accepted and rejected hypotheses conditions subsequent proposals, allowing EXPERIGEN to allocate expensive evaluations toward novel, empirically plausible directions while avoiding redundant exploration of already-refuted ideas. The decoupled agent architecture, feature extraction, and code execution allow EXPERIGEN to generalize across modalities (text, images) and complex datasets (non-i.i.d.) like Reddit threads (detailed discussion in Appendix I.4).

Empirically, EXPERIGEN produces approximately 1.5–3$\times$ as many statistically significant hypotheses as the strongest hypothesis-generation baseline for the tasks and benchmark aggregates with nonzero baseline counts in Table 1. On manual review, its false discovery rate falls below 5%, compared

to 20–25% for prior methods (Section 5.1). EXPERIGEN also exceeds HypoGeniC by an average of 5.0 percentage points in accuracy across the ten HypoBench ID/OOD splits (Table 3). Adaptive discovery further surfaces hypotheses inconceivable in a single pass: "Expanding the OP's perceived decision space is persuasive" emerges only after observing that both "concessions" and "framing a problem as a matter of degree rather than binary" are persuasive (Figure 2).

Following established social-science practice, we validate EXPERIGEN's hypotheses in three complementary ways: an **expert study** evaluating whether the hypotheses are novel and research-worthy (§6.1), live **A/B tests** that intervene on real users (§6.2), and **pre-registered observational experiments** that commit to the hypothesis and analysis plan before collecting new data (§6.3). In an expert study involving 26 senior experts from academia and industry, including senior faculty and S&P 500 marketers across marketing, persuasion, and biology, EXPERIGEN's hypotheses were consistently rated as novel and research-worthy and most demonstrated the experimental rigor of a senior graduate student. To evaluate real-world utility, marketers at 7 S&P 500 companies conducted 25 controlled A/B tests on EXPERIGEN-generated interventions involving 5.1M users. Of these controlled A/B tests, 15 reached statistical significance ($p < 0.05$), with 14 confirming the predicted direction and producing substantial improvements in conversion metrics. Pre-registration is the standard social-science protocol for testing hypotheses without contaminating results: researchers commit the hypothesis and analysis plan before collecting new data. Following this protocol, we pre-registered 16 EXPERIGEN-generated hypotheses on r/ChangeMyView and evaluated them on CMV posts collected over the next three months. 11 of 16 (68.75%) remained significant after BH correction, and all significant results matched the pre-registered direction, demonstrating that EXPERIGEN generates hypotheses that generalize prospectively to unseen data. Our core contributions are:

1. **Closed-loop hypothesis discovery.** We introduce EXPERIGEN, a Generator–Experimenter framework that discovers, operationalizes, tests, refines and rejects natural-language hypotheses from observational datasets.
2. **Rigor and Coverage.** We evaluate EXPERIGEN across 19 tasks spanning text, images, tabular data, relational metadata and single-cell RNA-seq. On tasks and benchmark aggregates with nonzero baseline counts, it discovers approximately 1.5–3$\times$ as many statistically significant hypotheses as the strongest hypothesis-generation baseline, and improves HypoBench accuracy by an average of 5.0 percentage points over HypoGeniC.
3. **Expert Study.** We conduct an expert evaluation of hypotheses generated by EXPERIGEN with **26 senior experts** (tenured faculty and S&P 500 senior marketers) with expertise in biology, marketing, and persuasion. Experts rated EXPERIGEN's hypotheses as novel, research-worthy, and reflecting senior-graduate-student experimental rigor.
4. **Real-world validation.** Beyond retrospective benchmark evaluation, we validate generated hypotheses through expert review, pre-registered prospective studies, and deployed A/B experiments, in which 11 of 16 pre-registered hypotheses replicated, providing the first demonstration in the literature that automatically generated hypotheses can withstand prospective testing and support real-world intervention.

2 Related Work

Hypothesis Generation aims to discover insights from data. Classical methods such as topic modeling, n-gram analysis, and regression-based feature selection uncover statistical patterns [33, 9], but require manual feature engineering, lack the domain knowledge to interpret unstructured data, and leave researchers to translate these patterns into interpretable, human-readable hypotheses. More recent work uses LLMs to generate natural language hypotheses directly from data [40, 52]. Methods such as **HypoGeniC** [52] and **HypotheSAEs** [34], which serve as our primary baselines (§4.2), generate interpretable hypotheses, but rely primarily on predictive performance for validation, making them vulnerable to spurious correlations and limiting their applicability to complex settings.

Hypothesis Validation frameworks use LLMs as coding assistants for statistical testing, but require pre-specified structured features and cannot discover hypotheses directly from raw data [21, 1]. EXPERIGEN bridges this gap by jointly performing discovery and validation: it generates hypotheses directly from unstructured data and grounds them in formal statistical evidence through a closed-loop refinement process guided by experimental feedback.

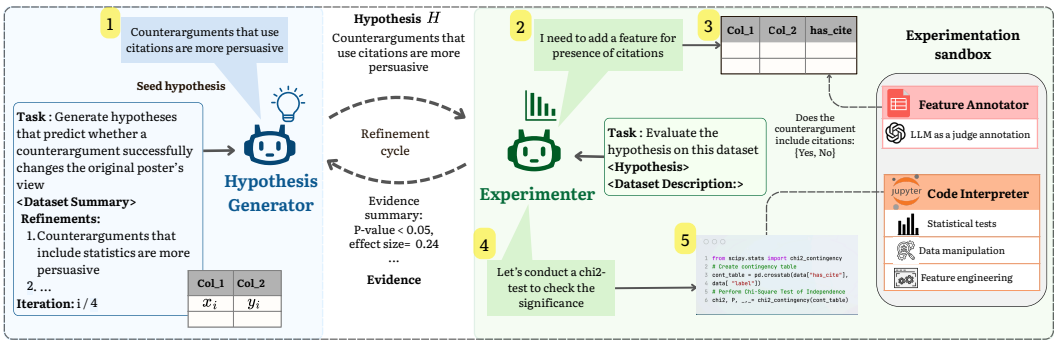


Figure 1: **Inner Loop: Iterative Refinement Cycle.** Starting from a seed hypothesis, the system iteratively refines and evaluates hypotheses over up to T steps. At each step, the *Generator* proposes a refinement using prior attempts as context, the *Experimenter* operationalizes it into measurable features, the *Feature Annotator* augments the dataset, and the *Code Interpreter* runs statistical tests. The resulting evidence, including p-values and effect sizes, is fed back to the Generator for further refinement. See Figure 2 for the outer exploration loop.

LLM Agents for Science. A related line of work develops LLM agents that automate scientific workflows, including ML experimentation [27, 50, 46, 42, 4, 22, 13] and domain-specific systems in chemistry, biomedicine, and materials science [10, 16, 24, 15]. EXPERIGEN differs in both setting and evaluation. These systems operate in *interventional* settings involving model training, simulations, or wet-lab experiments, whereas EXPERIGEN focuses on fixed observational datasets common in social science, marketing, healthcare, and behavioral research. Further, AI Scientist-style systems optimize paper-level artifacts judged subjectively, while EXPERIGEN evaluates individual hypotheses through statistical significance, multiple-testing correction, expert review, deployed A/B tests, and pre-registered replication.

3 Methodology

Data-driven discovery often follows a simple cycle: observe data, propose hypotheses, test them empirically, and refine them based on the outcomes. We study how to automate this process for unstructured data, where the main challenge is the enormous hypothesis space: for any dataset, only a small fraction of possible hypotheses are empirically valid. EXPERIGEN addresses this challenge using two agents: a **Gen**erator that proposes hypotheses and an **Experi**menter that evaluates them statistically. The system performs an iterative search inspired by Bayesian optimization, alternating between exploring new regions of the hypothesis space and refining promising candidates. In this section, we formalize the problem (Section 3.1), describe the two agents (Sections 3.2 and 3.3), and present the two-phase search procedure (Section 3.4). Figures 1 and 2 illustrate the inner refinement loop and outer exploration loop.

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ denote observation and outcome pairs, where x_i is unstructured data (e.g., counterarguments, response time) and y_i is the target variable (success in changing opinion). A hypothesis $H \in \mathbb{H}$ (e.g., "Using citations improves the success rate") is a testable natural-language statement that asserts a relationship between constructs derived from x and the outcome y.

For a hypothesis to be empirically testable, there must exist an evaluation or experiment E_H such that $E_H(\mathcal{D}) \in \{\texttt{supported}, \texttt{unsupported}\}$ indicates whether the hypothesis is supported or not. A rigorous evaluation requires (i) operationalization of constructs as features, (ii) identification and control of covariates, (iii) selection and execution of aligned tests, and (iv) robustness and sensitivity analyses with assumptions and specifications.

The space of such testable hypotheses $\mathbb{H}$ is combinatorially large: there are multiple constructs that can be refined, combined, conditioned on contexts, and related in multiple ways. However, only a

small subset of such hypotheses is empirically supported. Data-driven hypothesis discovery seeks a maximal and diverse set $\mathcal{H} \subseteq \mathbb{H}$ such that $\forall H \in \mathcal{H}$, $E_H(\mathcal{D}) = \texttt{supported}$.

3.2 The Generator Agent

Large language models encode implicit beliefs about which scientific relationships are likely to hold, derived from pretraining on scientific literature, domain knowledge, and general reasoning patterns. We assume that, given only a problem description, an LLM induces an initial implicit distribution $q(H)$ over the hypothesis space $\mathbb{H}$, reflecting prior plausibility. Prompting an LLM to generate hypotheses can therefore be viewed as sampling from this induced distribution $H \sim q(H)$. Prior work has shown that such generations are strong estimates of plausibility in zero-shot and few-shot settings [26]; we also observe this empirically (Table 3). Therefore, we employ the Generator as an LLM-based agent that proposes candidate hypotheses conditioned on two inputs: (1) a structured summary of the dataset, including schema, feature distributions, and sampled observations; and (2) a description of the Experimenter's evaluation capabilities. Conditioning on these inputs restricts generation to hypotheses that are both plausible and testable by the Experimenter on the given dataset. We detail this dataset summary design and Generator prompts in §J and Appendix H.1.

3.3 The Experimenter Agent

The **Experimenter** evaluates a natural-language hypothesis H against a dataset $\mathcal{D}$ by constructing and executing an empirical evaluation procedure $E_H : \mathcal{D}$. We implement the Experimenter as a ReAct agent [51] that plans features, covariates, and statistical tests conditioned on the hypothesis and dataset. It produces a structured analysis containing the chosen test, assumptions, significance levels, effect sizes, and robustness checks. Its p-values feed the two-level multiple-testing correction described in Section 3.4.2. This makes E_H reproducible given the same random seed, dataset, and tool outputs. The Experimenter operates in a sandboxed environment with two tools: (i) a customized **Code Interpreter** tool for data processing, feature computation, and statistical testing; and (ii) an **LLM-based feature extractor** for semantic constructs not reliably derivable through programmatic rules, e.g., "presence of citations in a counterargument." The code tool handles computable features (length, counts, regex patterns); the LLM extractor handles higher-level judgments.

3.4 Two-Phase Search

The preceding sections establish two capabilities: a Generator that samples plausible, testable hypotheses from an implicit distribution $q(H \mid \mathcal{D})$, and an Experimenter that evaluates any hypothesis H through a rigorous but computationally expensive procedure $E_H : D$. Naively sampling from $q(H)$ and evaluating each hypothesis independently wastes expensive Experimenter calls on redundant or implausible candidates, while greedy exploitation of early discoveries misses distant regions of the space. We require a search algorithm that (i) efficiently allocates expensive evaluations toward promising hypotheses, (ii) maintains broad coverage to avoid local optima, and (iii) exploits the compositional structure of hypotheses, where validated discoveries inform refinements and related candidates. These motivate our two-phase procedure: an outer loop that samples diverse seed hypotheses guided by an acquisition objective, and an inner loop that refines each seed through iterative validation to minimize Type-I error.

3.4.1 Acquisition-Guided Seed Generation

The outer loop (Figure 2) draws inspiration from Bayesian optimization, where a surrogate model approximates an expensive black-box function and an acquisition function balances exploration with exploitation. Here, the Generator's implicit beliefs serve as the surrogate, the inner refinement loop constitutes the expensive evaluation, and each accepted hypothesis updates the Generator's context for subsequent iterations. At outer iteration $i \in \{1, \ldots, N\}$, the Generator receives the dataset summary $\mathcal{D}$ and the hypothesis bank $\mathcal{H}_{i-1}$ and is prompted to jointly optimize two objectives: plausibility $s_i(H)$, reflecting the Generator's belief that H will be empirically supported given prior outcomes, and novelty $\mathcal{N}(H, \mathcal{H}_{i-1})$, reflecting dissimilarity from already-discovered hypotheses in the bank.

Plausibility (s_i). Plausibility represents the Generator's belief that H will be empirically supported given prior outcomes. At iteration i, the Generator is conditioned on the full labeled bank in-

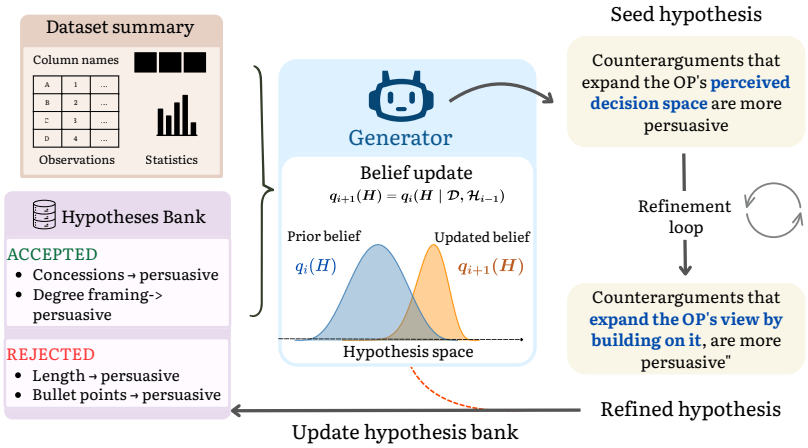


Figure 2: **Outer Loop: Acquisition-Guided Exploration.** At each outer iteration i, the Generator conditions on $\mathcal{D}$ and a bank $\mathcal{H}_{i-1}$ holding `[ACCEPTED]` and `[REJECTED]` entries, shifting prior $q_i(H)$ (blue) to posterior $q_{i+1}(H)$ (orange). It generates a seed (e.g., *expand decision space*), the inner loop refines it (e.g., *expand by building on it*), and the result is appended to the bank with its label.

context: `[ACCEPTED]` entries serve as positive exemplars, and `[REJECTED]` entries serve as negative exemplars. This steers generation toward hypotheses similar to accepted ones and away from those already refuted.

Novelty ($\mathcal{N}$). We define novelty as the minimum cosine distance between a candidate and the bank under embeddings $\phi(\cdot)$ from OpenAI's `text-embedding-3-large`:

$$\mathcal{N}(H, \mathcal{H}_{i-1}) = \min_{H' \in \mathcal{H}_{i-1}} \big(1 - \cos\big(\phi(H), \phi(H')\big)\big) \tag{1}$$

At bank curation (Section 3.4.2), we use $\mathcal{N}$ to select a pairwise-diverse subset of $\mathcal{H}_{i-1}$ to include in the Generator's prompt, which instructs the model to propose hypotheses distinct from those shown.

3.4.2 Local Refinement

The seed $H_{i,1}$ enters the refinement loop (Figure 1) for up to T steps, indexed by $j = 1, \ldots, T$. At each step, the Experimenter evaluates $H_{i,j}$ and returns structured analysis $E_{H_{i,j}}(\mathcal{D})$ comprising p-value, effect size, and guidance for refinement.

Inner loop: For $j > 1$, the Generator's context shifts. Let $\mathcal{M}_{i,j} = \{(H_{i,k}, E_{H_{i,k}})\}_{k=1}^{j}$ denote the short-term memory containing session hypotheses and their analyses. The Generator now receives $(\mathcal{D}, \mathcal{M}_{i,j})$, and its objective reverses: rather than maximizing global novelty, it restricts search to the locality of $H_{i,1}$, proposing refinements that address identified confounds or add contextual qualifiers. The implicit goal of this local search is to subject the hypothesis family to rigorous falsification, thereby minimizing Type-I error. This contrasts with methods that optimize predictive accuracy across independently generated hypotheses, an objective that inflates false positives under multiple testing. We quantify this advantage through expert verification in Section 5.1.

Hypothesis Bank Management. The bank $\mathcal{H}_i$ accumulates one hypothesis per outer iteration. At the end of each inner loop, we select a single representative: if any hypothesis was accepted, we add the hypothesis with the largest support; if none were accepted, we add the *last* generated refinement. Both are stored with their label (`[ACCEPTED]` or `[REJECTED]`)[1] in the hypothesis bank. To maintain diversity and bound context, we enforce a maximum capacity K. When the bank exceeds K, we greedily retain the subset of size K maximizing minimum pairwise distance under d (Equation (1)). In our experiments, we maintain a hypothesis bank with a maximum of $K = 30$ hypotheses. This value

[1] A hypothesis is never statistically rejected, only not accepted; we use `[REJECTED]` as shorthand for the same.

Method	Decep.	News	Dread.	GPTgc	Pers.	Twitter	Design	LaMem	Cong.	CMV	CRC	DB†	BAIS‡
ExperiGen	**18/15**	**21/16**	**16/15**	**14/12**	**17/13**	**19/15**	**11**	**5**	**22**	**16**	**12/11**	**81**	**17**
HypoGeniC	6	3	4	3	5	3	4	0	5	4	4	40	-
HypotheSAEs	8	12	6	5	7	12	6	0	15	8	3	20	-

Table 1: Significant hypotheses on held-out data ($p < 0.05$, Bonferroni-corrected per method). ExperiGen pairs give ID/OOD counts from the same discovery bank; single entries give ID/test counts with OOD unavailable. Baseline OOD counts are not reported; "-" denotes unavailable results. †DB: four-task total (Table 12). ‡BAIS: four-dataset total (Table 13).

is applied consistently across all tasks during hypothesis generation, ensuring the Generator maintains awareness of high-performing hypotheses without accumulating redundant or low-performing ones.

Hypothesis Selection In the training phase, to ensure that the multiple tests during outer loops and inner refinements does not inflate type-1 error we do two corrections[8]. *Level 1 (within inner loop):* Bonferroni over the $m_i \leq T$ confirmatory tests in inner loop i, accepting only $p < \alpha_{\text{sig}}/m_i$ with $\alpha_{\text{sig}} = 0.05$; the accepted hypothesis with the largest support enters $\mathcal{H}$. *Level 2 (post-generation):* after all N outer iterations, Benjamini–Hochberg (BH) across the surviving hypotheses controls the false-discovery rate at $q = 0.05$. Since exploratory steps can steer which refinement is tested, we count significance on held-out data (Table 1) and replicate under pre-registration (Section 6.3). We report the statistical test, BH-adjusted p-value (on the discovery split), verdict, and effect size for every hypothesis in Appendix A.5 (Table 9).

4 Experiments

4.1 Evaluation Setup

Tasks and Datasets: We evaluate ExperiGen on a diverse suite of **19 tasks** spanning **12 domains** across text, image, tabular, and single-cell RNA-seq, with several tasks providing structured covariates. Table 4 summarizes each task. We evaluate hypothesis generation methods along two axes:

Number of statistically significant hypotheses and FDR. We report the number of hypotheses each method generates and the subset that passes statistical validation. A hypothesis is considered *statistically valid* if its associated features show significant predictive power ($p < 0.05$, Bonferroni-corrected over the method's hypotheses) in a multivariate regression on held-out data, a criterion applied identically to all methods and separate from ExperiGen's internal selection. Methods that generate many hypotheses but few valid ones suffer from high false discovery rates. We report the false discovery rate (FDR) as the fraction of accepted hypotheses rejected by experts.

Predictive Performance. Social science traditionally validates hypotheses through statistical significance (e.g., p-values, effect sizes), and we adopt this as our primary criterion. We additionally report the predictive accuracy of hypothesis-derived features to enable direct comparison with prior methods, which use it as their primary metric. For classification tasks we report in-distribution (ID) and OOD accuracy; for regression tasks we report R^2. We use the two-step LLM inference procedure of Zhou et al. [52]; full details are in Appendix M.

4.2 Baselines

We compare ExperiGen against prior hypothesis generation methods, HypoGeniC [52] and HypotheSAEs [34], and against Chain-of-Thought (CoT) prompting to benchmark the performance of strong LLM-only baselines in zero-shot and few-shot ($k = 3$) settings. For hyperparameter details on the baselines, see Appendix M. We use Qwen3-32B as the default Generator and Experimenter for ExperiGen and GPT-4o as the default inference model.

5 Results and Discussion

We evaluate EXPERIGEN along two axes: *statistical validity* and *predictive power*. Ablations for all architectural choices are reported in Appendix A.

5.1 Statistical Validity

Statistical validity is critical because unvalidated hypotheses can drive costly downstream decisions, from failed clinical trials to ineffective A/B tests and policy interventions. Importantly, a hypothesis may fail to improve predictive performance yet still capture a meaningful, validated association. Table 1 summarizes the number of significant hypotheses discovered across datasets. For columns with nonzero baseline counts, EXPERIGEN finds approximately 1.5–3$\times$ as many significant hypotheses as the strongest hypothesis-generation baseline in each column due to two advantages: (1) *adaptive exploration*, where iterative refinement learns from rejected hypotheses and redirects search toward promising directions, and (2) *broader feature scope*, spanning syntactic, semantic, and relational features, whereas baselines are restricted to per-sample semantic features and cannot express cross-sample relational hypotheses (Appendix K). The gap is largest on structured and high-dimensional datasets. EXPERIGEN discovers 12 significant CRC hypotheses versus 4 from the strongest baseline, and 81 across DiscoveryBench versus 40. On BAISBench-SD (20,000+ gene features), it discovers 17 (p from $3.1\text{e}{-11}$ to $5.8\text{e}{-4}$). Existing methods lack pretrained embeddings for raw gene matrices or collapse to metadata-level generation (Appendix C).

Lower false discovery rate. Manual review shows EXPERIGEN achieves a false discovery rate below 5%, compared to 20–25% for prior methods. For example, on Congress, HypotheSAEs identifies "I ask unanimous consent" as predictive of Republican speech, but the phrase reflects majority-party procedure rather than ideology. EXPERIGEN detects session as a confounder during refinement, controls for it, and rejects the hypothesis (Appendix D.2). Table 2 illustrates the specificity of the generated hypotheses.

Generated hypothesis	**p-value**	**Effect size**
Counterarguments explaining technical terms with context or analogies are more persuasive than those merely listing terms.	7.3×10^{-6}	OR $= 3.22$
First-comment position provides a persuasion advantage independent of timing.	< 0.001	19.9% vs. 9.5%; OR ≈ 2.37
Higher "I":"you" pronoun ratios predict greater persuasion success even after controlling for comment length.	4.1×10^{-6}	OR $= 1.036$ per unit
Labeling the OP's reasoning a logical fallacy, especially in the opening, reduces persuasion success.	0.316	Cramér's $V = 0.01$

Table 2: Selected CMV hypotheses illustrating thread position, contextual explanations, and pronoun use controlled for comment length, with p-values and effect sizes. Green indicates accepted hypotheses; red indicates a nonsignificant result. Additional examples of discovered hypotheses are provided in Appendix F, Table 16. OR: odds ratio.

5.2 Predictive Performance

Table 3 reports accuracy on five HypoBench tasks across ID and OOD splits. EXPERIGEN outperforms HypoGeniC on all ten splits, with a mean improvement of 5.0 percentage points and a maximum gain of 13.6 on Dreaddit OOD. It achieves the highest accuracy among all evaluated methods on nine splits. Similar ID and OOD gains suggest iterative refinement does not overfit to the discovery split. We further evaluate on harder settings involving metadata, long-form text, conversations, and images: Twitter, Congress, CMV, Design, and LaMem. EXPERIGEN wins all six ID/OOD splits, improving over the best of all evaluated baselines in each split by 3.25–8.5 percentage points. On CMV, it discovers that the first counterargument in a thread is associated with roughly twice the persuasion success rate of later counterarguments (19.9% vs. 9.5%), a relational feature requiring computation across thread structure, which is enabled by the code interpreter. On the four DiscoveryBench [31] tasks, EXPERIGEN achieves the best R^2 on every dataset, outperforming the

Method	Decep.		News		Dread.		GPTgc		Pers.		Twitter		Design	LaMem	Cong.	CMV
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out				
ExperiGen	78	85.2	70	66.5	74	80.8	88.8	85	94	91.2	67	66.1	88.0	60.1	79.4	77.5
HypoGeniC	76	77.7	63.2	62.5	64.2	67.2	87	84	93	89	60.3	61.9	84.25	50.8	72.6	61
HypotheSAEs	62.9	60.6	61.9	61.4	60.9	62.3	76.9	79.8	87.1	84	54.1	56	84.25	49.2	73.9	59.2
0-shot CoT	65.0	69.4	68.2	64.9	63.4	68.2	79.0	78.3	83.2	81.9	60.5	56.4	84.75	51.6	73.7	65
few-shot CoT	64.8	75.8	67.2	62.9	73.4	81.6	77.0	78.3	81	80.5	58.6	59.2	84.60	51.9	72.6	69

Table 3: Predictive accuracy on image and text modality tasks, with **GPT-4o** as the inference model for all methods. We report ID (In) and OOD (Out) accuracies using hypotheses generated by each method. ExperiGen achieves the highest accuracy on 9 of 10 task-split combinations. Cross-model transfer to Qwen3-32B is reported in Appendix A.6; task-specific expert-model reference numbers for Twitter, Design, and LaMem are reported in Appendix M.4.

strongest baseline by 0.09–0.27 absolute (Appendix Table 12). On the MetaGenoPolis CRC cohort, it reaches 81.3% / 76.2% ID/OOD accuracy versus 71.2% / 64.8% for the strongest baseline (Appendix Table 11). BAISBench-SD lacks classification targets, so we evaluate it through hypothesis discovery instead (§C.3).

6 Real-World Validation

We evaluate whether generated hypotheses withstand **expert scrutiny** for novelty, rigor, and research worthiness (§6.1), **randomized A/B interventions** on real users (§6.2), and **pre-registered observational studies** with hypotheses and analysis plans frozen before collecting new data (§6.3).

6.1 Expert study

Domain experts validate that ExperiGen's hypotheses are novel, high-quality, and research-worthy. Before investing in downstream experiments or large-scale A/B tests, researchers and practitioners often seek additional evidence that a hypothesis is meaningful, actionable, clearly articulated, and worthy of further investigation. In the social sciences, expert assessment is a standard mechanism for evaluating whether proposed hypotheses meet these broader standards of research quality, rigor, and practical relevance.

To evaluate whether ExperiGen generates hypotheses that satisfy these criteria, we conducted an expert study with **26 experts** across three diverse domains: Headlines (marketing), CMV (persuasion), and BAISBench-SD (biology). The panel included senior S&P 500 marketers for the marketing tasks and senior faculty with more than 10 years of experience publishing academic research and mentoring graduate students for the academic tasks. Assignments were balanced such that each hypothesis received at least three independent evaluations.

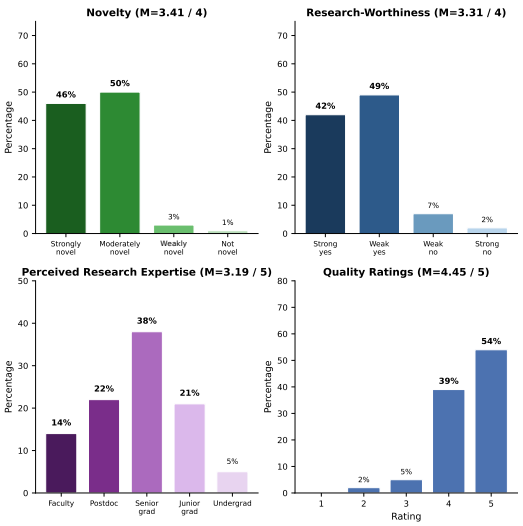


Figure 3: Expert ratings of ExperiGen's hypotheses on novelty, research worthiness, perceived expertise, and quality ($N = 26$ annotators across Headlines, CMV, and Biology).

Experts evaluated hypotheses along five dimensions: *research worthiness* (whether the hypothesis is sufficiently compelling to justify follow-up research or deployment in an A/B test), *novelty* (whether it advances beyond existing knowledge or literature), *perceived expertise or rigor* (the sophistication reflected in the proposed reasoning or experimental design), *quality* (clarity, specificity, coherence, and testability), and *confidence* (the reviewer's self-rated confidence in their assessment). As shown

in Figure 3, EXPERIGEN's hypotheses received consistently strong evaluations across all dimensions. The strongest results were observed on CMV, where 88% of hypotheses were rated as novel and 76% as research-worthy and most had senior graduate-student-level rigor. On Biology, all five reviewed hypotheses were evaluated as demonstrating senior graduate-student-level rigor and three hypotheses were rated as weakly novel, while two were refined through expert feedback. Full study details are provided in Appendix B.

6.2 A/B tests

EXPERIGEN's hypotheses produce measurable impact in deployed experiments. We collaborated with 16 marketers across 7 S&P 500 companies. For each partner, EXPERIGEN was given access only to a four-month historical observational dataset consisting of webpage screenshots, metadata, and aggregated view/conversion statistics; importantly, it had no access to prior A/B test outcomes. Using these observational datasets, EXPERIGEN generated hypotheses predicting how changes in factors such as content, layout, and presentation could influence conversion rate. We presented marketers with 40 generated hypotheses, of which 25 were selected for evaluation through controlled A/B experiments on production landing pages using a leading experimentation platform. All experiments were designed and launched only after hypothesis generation was completed, ensuring strict separation between the discovery data and the randomized evaluation traffic. The experiments ran over a five-month period and collectively involved 5.1 million users visiting the participating websites. Traffic was randomized at the visit level using a 50/50 split between control and treatment arms, and conversion rate (sign-ups or the equivalent business metric) was the primary metric.

Aggregate outcomes: Among the 25 experiments, 15 achieved statistical significance ($p < 0.05$). Of these, 14 validated the pre-specified hypothesis direction, yielding a mean conversion-rate uplift of $+213\%$ and a median uplift of $+112\%$. Only one statistically significant result contradicted the predicted direction. The observed effects were consistently strong, with significant results showing a mean p-value of 0.008 and a maximum significant p-value of 0.041. Details can be found in Appendix E.

6.3 Pre-Registered Experiments

EXPERIGEN's hypotheses hold up under pre-registration. Pre-registration is the gold standard in social-science methodology because it closes off the most common paths to spurious results: selective reporting, p-hacking, and quietly revising the hypothesis after peeking at the data. In a pre-registered experiment, the researcher publicly commits the hypothesis, directional prediction, and full analysis plan to a registry and tests the hypothesis on future data.

We generated hypotheses about persuasion using EXPERIGEN on data extracted from r/ChangeMyView. We pre-registered the 16 hypotheses through our institution in January 2026, locking the following for each hypothesis: the exact natural-language statement, the directional prediction, the feature operationalization, the statistical test, and the BH-corrected significance threshold. No hypothesis was dropped, reworded, or modified after registration. Fresh CMV posts were collected over a 3-month window (Feb–Apr 2026), with no overlap with the historical discovery data, and the frozen analysis pipeline for each hypothesis was applied to these data. **11 of 16 (68.75%) reached statistical significance after BH correction** ($p < 0.05$), and *every* significant result confirmed its pre-registered direction. One representative pre-registration document and per-hypothesis outcomes are in Appendix E (Table 15).

7 Conclusion

We introduced EXPERIGEN, a novel agentic framework that couples hypothesis discovery with statistical testing, covariate control, and iterative refinement. Across 19 tasks, it improves both hypothesis validity and predictive performance. Beyond retrospective benchmarks, expert review and deployed A/B tests provide evidence of real-world utility, while 11 of 16 hypotheses replicated in pre-registered prospective studies. These results suggest that closed-loop discovery can help prioritize hypotheses worth costly real-world experimentation, while complementing rather than replacing human scientific judgment. We believe these results will also help researchers and social scientists to scale hypothesis discovery in data driven sciences.

Acknowledgments and Disclosure of Funding

This work automates hypothesis generation and experimental validation for data-driven sciences, with potential to accelerate discovery in fields such as behavioral science, psychology, healthcare, and marketing. By enforcing statistical rigor through multiple testing corrections and covariate control, EXPERIGEN produces hypotheses that meet the standards of domain experts such as tenured professors. The framework is designed to augment rather than replace human scientific judgment; we recommend that generated hypotheses undergo domain expert review before informing policy or intervention. The computing resources and LLM API credentials used for the experiments were provided by Adobe. No other third-party funding or support was received.

References

[1] Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, et al. Open-ended scientific discovery via bayesian surprise. *arXiv preprint arXiv:2507.00310*, 2025. 2, 3

[2] Solomon E Asch. Effects of group pressure upon the modification and distortion of judgments. 1951. 1

[3] Marianne Aubin Le Quere and J. Matias. When curiosity gaps backfire: effects of headline concreteness on information selection decisions. *Scientific Reports*, 15, 01 2025. doi: 10.1038/s41598-024-81575-9. 41

[4] 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 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, 2025. 4

[5] Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert Riggs, Maryam Esmaeili, Rastin Majdabadkohne, and Morteza Pasehvar. Chatgpt: Applications, opportunities, and threats. *Proceedings of the IEEE*, 04 2023. 41

[6] Akshina Banerjee and Oleg Urminsky. The language that drives engagement: A systematic large-scale analysis of headline experiments. *Marketing Science*, 44(3):566–592, 2025. doi: 10.1287/mksc.2021.0018. URL https://doi.org/10.1287/mksc.2021.0018. 41

[7] Pauline BARBET, Mathieu ALMEIDA, GOULET Lindsay, Niklas PROBUL, Jan BAUMBACH, Nicolas PONS, Florian PLAZA ONATE, and Emmanuelle LE CHATELIER. Taxonomic profiles, functional profiles and manually curated metadata of human fecal metagenomes from public projects coming from colorectal cancer studies, 2022. URL https://doi.org/10.57745/7IVO3E. 17

[8] Yoav Benjamini and Marina Bogomolov. Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):297–318, 2014. 7, 23

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 3

[10] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023. 4

[11] Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119, 2022. doi: 10.1073/pnas.2120510119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2120510119. 43

[12] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. pages 9–16, 08 2016. doi: 10.1109/ASONAM.2016.7752207. 41

[13] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024. 4

[14] Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4): 1307–1340, 2019. doi: https://doi.org/10.3982/ECTA16566. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16566. 17, 42

[15] Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025. 4

[16] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025. 4

[17] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *ArXiv*, abs/2301.07597, 2023. URL https://api.semanticscholar.org/CorpusID:255998637. 41

[18] SI Harini, Somesh Singh, Yaman K Singla, Aanisha Bhattacharyya, Veeky Baths, Changyou Chen, Rajiv Ratn Shah, and Balaji Krishnamurthy. Long-term ad memorability: Understanding & generating memorable ads. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5707–5718. IEEE, 2025. 62

[19] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4320–4333, 2020. 57

[20] Chuming Hu and Feifei Xu. A review of white space research. *Open Journal of Social Sciences*, 07:328–334, 01 2019. doi: 10.4236/jss.2019.73027. 42

[21] Kexin Huang, Ying Jin, Ryan Li, Michael Y Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. *arXiv preprint arXiv:2502.09858*, 2025. 3

[22] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2023. 4

[23] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pages 2390–2398, 2015. 17, 61

[24] Helena Kudiabor. The super-charged virtual lab powered by 'AI scientists'. *Nature*, 636:19, 2024. 4

[25] Angelo Lasalita, Edward Jay Quinto, Andrea Fernando, Kiyomi Mae Suzuki, Anthony Lars Abad, Jonathan Macayan, and John Christopher Castillo. The language of depression: A multi-phase analysis on the language patterns of Filipinos with varying levels of clinical depressive symptoms. In Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, and Jong-Bok Kim, editors, *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 1027–1037, Tokyo, Japan, December 2024. Tokyo University of Foreign Studies. URL https://aclanthology.org/2024.paclic-1.98/. 41

[26] Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. HypoBench: Towards systematic and principled benchmarking for hypothesis generation, 2025. URL https://arxiv.org/abs/2504.11524. 5, 17, 60

[27] 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. 4

[28] Michael Luca and Georgios Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427, 2016. doi: 10.1287/mnsc.2015.2304. URL https://doi.org/10.1287/mnsc.2015.2304. 42

[29] Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, 2024. 2

[30] Erpai Luo, Jinmeng Jia, Yifan Xiong, Xiangyu Li, Xiaobo Guo, Baoqi Yu, Minsheng Hao, Lei Wei, and Xuegong Zhang. Benchmarking AI scientists for omics data driven biological discovery, 2026. URL https://arxiv.org/abs/2505.08341. 17

[31] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models, 2024. URL https://arxiv.org/abs/2407.01725. 8, 17

[32] Thomas C Messerli, Daria Dayter, Sven Leuckert, Aatu Liimatta, Hanna Mahler, Axel Bohmann, Gustavo Kozma, and Rafaela Tosin. Digital debating cultures: communicative practices on reddit. *Digital Scholarship in the Humanities*, 40(1):227–240, 02 2025. ISSN 2055-7671. doi: 10.1093/llc/fqaf005. URL https://doi.org/10.1093/llc/fqaf005. 42

[33] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16 (4):372–403, 2008. 3

[34] Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. *arXiv preprint arXiv:2502.04382*, 2025. 2, 3, 7, 22, 60

[35] Elena Musi, Debanjan Ghosh, and Smaranda Muresan. Changemyview through concessions: Do concessions increase persuasion? *ArXiv*, abs/1806.03223, 2018. URL https://api.semanticscholar.org/CorpusID:47016106. 42

[36] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Annual Meeting of the Association for Computational Linguistics*, 2011. URL https://api.semanticscholar.org/CorpusID:2510724. 42

[37] Rong Pan, Chuanji Gao, Xiaoman Zhu, Bao-ming Li, and Xi Jia. Positive emotion enhances memory by promoting memory reinstatement across repeated learning. *Journal of Neuroscience*, 45(31), 2025. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0009-25.2025. URL https://www.jneurosci.org/content/45/31/e0009252025. 42

[38] Sohan Patnaik, Rishabh Jain, Balaji Krishnamurthy, and Mausoom Sarkar. Aesthetiq: Enhancing graphic layout design via aesthetic-aware preference alignment of multi-modal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23701–23711, 2025. 17, 61, 62

[39] Cameron P. Pugach, Casey L. May, and Blair E. Wisco. Positive emotion in posttraumatic stress disorder: A global or context-specific problem? *Journal of Traumatic Stress*, 36(2):444–456, 2023. doi: 10.1002/jts.22928. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10101918/. 41

[40] Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*, 2023. 3, 17

[41] Claire Robertson, Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay Van Bavel, and Stefan Feuerriegel. Negativity drives online news consumption. *Nature Human Behaviour*, 7:1–11, 03 2023. doi: 10.1038/s41562-023-01538-4. 41

[42] 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. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043, 2025. 4

[43] Somesh Singh, Harini S I, Yaman K Singla, Veeky Baths, Rajiv Ratn Shah, Changyou Chen, and Balaji Krishnamurthy. Teaching human behavior improves content understanding abilities of llms, 2024. URL https://arxiv.org/abs/2405.00942. 61, 62

[44] Somesh Singh, Yaman K Singla, Harini SI, and Balaji Krishnamurthy. Measuring and improving persuasiveness of large language models. *arXiv preprint arXiv:2410.02653*, 2024. 17, 61

[45] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624, 2016. 17, 42, 61

[46] Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *arXiv preprint arXiv:2505.18705*, 2025. 4

[47] Alexandre N. Tuch, Javier A. Bargas-Avila, and Klaus Opwis. Symmetry and aesthetics in website design: It's a man's business. *Comput. Hum. Behav.*, 26:1831–1837, 2010. URL https://api.semanticscholar.org/CorpusID:39694453. 42

[48] Janith Weerasinghe, Kediel Morales, and Rachel Greenstadt. "because... i was told... so much": Linguistic indicators of mental health status on twitter. *Proceedings on Privacy Enhancing Technologies*, 2019:152–171, 10 2019. doi: 10.2478/popets-2019-0063. 40, 41

[49] Ernest L Wynder and Evarts A Graham. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma: a study of six hundred and eighty-four proved cases. *Journal of the American medical association*, 143(4):329–336, 1950. 1

[50] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025. 4

[51] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629, 2022. URL https://api.semanticscholar.org/CorpusID:252762395. 5

[52] Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*, 2024. 2, 3, 7, 17, 60

Appendix

Contents

A Additional Experiments and Ablations **16**

A.1 Tasks . . . 16

A.2 Ablation Configurations . . . 16

A.3 Ablation Results . . . 18

A.4 Detailed Component Analysis . . . 18

A.5 Hierarchical Multiple-Testing Correction: BH-Adjusted p-values and Effect Sizes . . . 23

A.6 Cross-Model Transfer to Qwen3-32B . . . 25

B Expert Study **25**

B.1 Motivation . . . 25

B.2 Participants and Procedure . . . 26

B.3 Interface . . . 26

B.4 Results . . . 28

C Results on Healthcare, Tabular data, and Cellular Biology **29**

C.1 Clinical Healthcare: MetaGenoPolis CRC Cohort . . . 29

C.2 Structured Tabular Datasets: DiscoveryBench . . . 29

C.3 High-Dimensional Cellular Biology: BAISBench-SD . . . 30

D Case Studies **31**

D.1 Discovery: Response Timing Predicts Persuasion Success . . . 31

D.2 Rejection: Procedural Phrases Confounded with Floor Control . . . 32

D.3 Discovery: Centered Forms with Visual Elevation Improve Engagement . . . 32

E Real-World Validation Details **36**

E.1 A/B Test Protocol . . . 36

E.2 Large-Scale A/B Test Summary . . . 36

E.3 Pre-Registered Experiments on r/ChangeMyView . . . 36

F Discovered Hypotheses **40**

F.1 Qualitative Examples . . . 43

G Generator Reasoning Traces: Adaptive Hypothesis Generation **47**

G.1 Seed Hypothesis #15: Generator Reasoning . . . 48

G.2 Seed Hypothesis #27: Generator Reasoning . . . 49

H Prompts **49**

H.1 Agent Prompts . . . 49

H.2 Task Prompts . . . 53

I Algorithm **53**

I.1 Scalability . . . 53

I.2 Enabling the Data Analyst . . . 55
I.3 Architectural Flexibility and Modality Generalization. . . . 55
I.4 Architectural Flexibility and Modality Generalization . . . 56

J Dynamic Dataset Description $\hat{\mathcal{D}}$ **56**

K Experimentation Sandbox **58**
K.1 Code Interpreter . . . 58
K.2 Feature Extractor . . . 59

L Datasets **60**

M Baseline Implementation Details & Hyperparameters **60**
M.1 Inference Procedure . . . 60
M.2 HypoGeniC . . . 60
M.3 HypotheSAEs . . . 61
M.4 Task Specific Expert Model Numbers . . . 62

N Limitations **62**

A Additional Experiments and Ablations

We conduct systematic ablations to isolate the contribution of each component in EXPERIGEN. All experiments use $T = 4$ refinement steps in the inner loop. We evaluate on two datasets selected to span the range of LLM prior knowledge: **News Headlines**, where zero-shot and few-shot baselines achieve similar performance (minimal gap), indicating the task is well-represented in pretraining data; and **Deceptive Reviews**, where few-shot substantially outperforms zero-shot (maximal gap), indicating the task requires learning from observations rather than relying on prior knowledge. This selection allows us to understand how EXPERIGEN's components behave in both regimes—where pretrained knowledge suffices and where data-driven discovery is essential. We measure OOD accuracy (**OOD Acc.**) and number of statistically significant hypotheses (**#Sig.**).

A.1 Tasks

Table 4 lists every task used in our evaluation, with its domain, input modality, and whether auxiliary structured metadata is available to the model.

A.2 Ablation Configurations

A.2.1 Acquisition Objective.

We ablate the two-phase search structure by modifying how seed hypotheses are generated. **No Acquisition** removes conditioning on the hypothesis bank $\mathcal{H}_{i-1}$; the Generator receives only the dataset summary $\mathcal{D}$, preventing it from building on prior discoveries. **Seed Only** further removes the inner refinement loop ($T = 1$), testing whether iterative refinement adds value beyond seed diversity. Without acquisition-guided exploration, the Generator lacks awareness of previously validated hypotheses, leading to redundant proposals and reduced coverage. Seed Only compounds this by eliminating the refinement phase that allows hypotheses to be corrected based on experimental feedback. Comparing EXPERIGEN to these ablations reveals that the inner loop contributes $+4$–13 percentage points in OOD accuracy and approximately $2\times$ more significant hypotheses, confirming that refinement based on experimental feedback yields more robust discoveries.

Table 4: Tasks and datasets used for our evaluation. *S* indicates tasks with auxiliary structured metadata (e.g., usernames, time, demographic covariates) available to the model alongside the primary input. Dataset sizes are in Appendix L.

Name	Domain	Modality	S	Task Description
Deceptive reviews [26]	Linguistics	Text	×	Distinguish genuine vs. deceptive hotel reviews via linguistic cues.
AI-Gen. Detection [26]	Linguistics	Text	×	Decide whether a story is human- or AI-written.
Persuasive Pairs [26]	Persuasion	Text	×	Pick the more persuasive of two arguments on rhetoric, logic, and appeal.
Mental Stress [26]	Psychology	Text	×	Detect mental stress in Reddit posts via affective and psychological cues.
News Engagement [26]	Marketing	Text	×	Choose which of two headlines garners more clicks.
Congress [14]	Politics	Text	✓	Predict party (R/D) from U.S. congressional speeches.
Tweets [44]	Persuasion	Text	✓	Predict the more persuasive of two tweets using tweet content and metadata.
CMV [45]	Persuasion	Text	✓	Predict whether a counterargument changes the OP's view on r/ChangeMyView.
Design Aesthetics [38]	Aesthetics	Image	×	Identify which of two graphic layouts was AI-generated.
Memorability [23]	Vision	Image	×	Predict the more memorable of two visually similar images (LaMem).
MetaGenoPolis CRC [7]	Healthcare	Tabular	✓	Healthy vs. colorectal-cancer classification from microbial signatures.
DiscoveryBench (4 tasks) [31]				
Plants	Botany	Tabular	✓	Predict invasive-species pathway from plant traits.
BMI	Sociology	Tabular	✓	Regress BMI on demographic and lifestyle features.
Imm. Offshoring	Economics	Tabular	✓	Estimate immigration's causal effect on offshoring.
Imm. Employment	Economics	Tabular	✓	Estimate immigration's causal effect on native employment.
BAISBench-SD (4) [30]	Cell bio.	scRNA-seq	✓	Gene-expression hypotheses across four single-cell studies (COVID-19, JDM, cancer).

A.2.2 Novelty Constraint.

We isolate the novelty term $\mathcal{N}(H, \mathcal{H}_{i-1})$ in the acquisition objective. **No Novelty** removes the novelty instruction from the Generator prompt, setting $\lambda = 0$. Without explicit novelty pressure, the Generator tends to fixate on variations of the top 2–3 most salient patterns in the data, repeatedly proposing minor reformulations rather than exploring distinct hypothesis families. This results in 40–45% fewer significant hypotheses than the full model. However, No Novelty still outperforms No Acquisition because it retains conditioning on the hypothesis bank, allowing the Generator to avoid exact duplicates even if it cannot escape local optima.

A.2.3 Acquisition Operationalization.

We test whether the implicit, prompt-driven operationalization of $\mathcal{A}(H)$ described in Section 3.4.1 can be replaced by an explicit, closed-form alternative. **Best-of-N Acq.** draws N=5 candidates per outer iteration from q_i and forwards the argmax of an explicit acquisition score: novelty as cosine distance to the bank under `text-embedding-3-large` (Equation (1)), plus plausibility as the difference between mean similarity to `[ACCEPTED]` entries and mean similarity to `[REJECTED]` entries. All other components (bank curation, inner-loop refinement, hierarchical multiple-testing) are held fixed; only the seed-selection step changes. Replacing the verbal operationalization with explicit scoring drops OOD accuracy by 1.4–2.6 points and reduces significant hypotheses by 3–4 across both datasets, and it underperforms the default ExperiGen on 9/10 model×dataset pairs in our broader sweep. The pattern indicates that closed-form scoring discards information that the LLM uses implicitly when conditioned on the labeled bank—e.g., shared structural patterns across `[ACCEPTED]` entries that are not captured by mean-similarity statistics—so an explicit argmax is strictly worse than the verbal alternative.

A.2.4 Experimentation.

We ablate the Experimenter agent's role and feedback mechanism. **No Feedback** returns only accept/reject decisions without structured analysis (p, δ, guidance), preventing the Generator from understanding why a hypothesis failed. **Single Pass** fixes $T = 1$ with no iterative refinement, equivalent to independent hypothesis evaluation. **Accuracy Based** replaces statistical testing with held-out accuracy, accepting hypotheses if classifier accuracy exceeds 60% on a validation split, mirroring the evaluation protocol of prior work [40, 52]. The Accuracy Based variant achieves competitive OOD accuracy but discovers 3× fewer valid hypotheses, consistent with our claim that predictive accuracy alone inflates false positives under multiple testing.

A.2.5 Feature Extractor.

We evaluate robustness to feature extraction noise by injecting label errors into LLM extractor outputs. **Noisy @ 10%** flips 10% of extracted binary labels uniformly at random; **Noisy @ 20%** increases this to 20%. These ablations simulate realistic annotation errors and reveal how hypothesis validation degrades under imperfect feature extraction. Performance degrades gracefully: 10% noise reduces OOD accuracy by 2–3 points, while 20% noise incurs a 4–5 point drop. The number of significant hypotheses decreases more sharply, as noisy features introduce measurement error that inflates p-values and obscures true effects.

A.2.6 Search Strategy.

We compare iterative refinement against parallel generation under matched compute. **Parallel** generates $N \times T$ hypotheses simultaneously without sequential feedback, evaluates all independently, and selects the top-K by effect size. Total LLM calls are matched to EXPERIGEN. Under matched compute, iterative refinement outperforms parallel generation by $+5$–8 points in OOD accuracy, validating that adaptive search exploits experimental feedback more effectively than independent sampling.

A.2.7 Observation Sampling.

We ablate how observations are sampled for the Generator's dataset summary; see Figure 4 for iteration-wise results. EXPERIGEN uses uniform random sampling of 5 observations per iteration. **Boosting** identifies samples with the highest prediction error under the current hypothesis bank and samples in proportion to their errors, prioritizing difficult instances that existing hypotheses fail to explain. **Clustering** embeds all observations using `text-embedding-3-large`, partitions them via k-means clustering, and samples from different clusters to maximize diversity. **No Data** provides only the dataset schema without concrete examples.

A.3 Ablation Results

Table 5 reports OOD accuracy and the number of significant hypotheses for each configuration in Appendix A.2. The full EXPERIGEN model is best on both datasets. Withholding observations (No Data) and removing iterative refinement (Seed Only, Single Pass) cause the largest drops. Dropping statistical testing (Accuracy Based) keeps OOD accuracy within 5 points but yields only 5 significant hypotheses per dataset, while 10% feature-extraction noise costs under 3 points.

A.4 Detailed Component Analysis

Beyond the aggregate results in Table 5, we provide deeper analysis of four components that warrant additional investigation: (1) how observation sampling affects hypothesis diversity over iterations, (2) how Generator and Experimenter model capabilities independently influence discovery quality, (3) how inference scales as the hypothesis bank grows, and (4) how reliably the LLM feature extractor annotates semantic constructs. Together, these analyses reveal where EXPERIGEN's performance gains originate and identify bottlenecks for future improvement.

A.4.1 Observation Sampling Dynamics

The Generator receives a small batch of observations from $\mathcal{D}$ at each iteration. While Table 5 reports final accuracy, iteration-wise dynamics reveal important differences between sampling strategies. We compare four approaches: (1) **Random sampling** (EXPERIGEN's default), which draws 5 observations uniformly; (2) **Boosting**, which measures prediction error for each sample under the current hypothesis bank and samples in proportion to error magnitude, prioritizing instances that existing hypotheses fail to explain; (3) **Clustering**, which embeds observations using `text-embedding-3-large`, partitions them via k-means, and samples from different clusters to maximize diversity; and (4) **No Data**, which provides only the dataset schema without concrete examples.

Figure 4 shows accuracy as a function of outer-loop iterations across four datasets. Boosting achieves rapid early gains but saturates after 10–15 iterations, while random sampling continues improving

Table 5: **Ablation Study Results.** OOD Acc. = OOD accuracy (%); #Sig. = Number of statistically significant hypotheses on held-out data ($p < 0.05$, Bonferroni-corrected across each method's hypotheses; same criterion as Table 1). The Full Model results are from separate runs using the same default settings as Tables 1 and 3, rather than copies of their results. The default refinement budget is $T = 4$; Seed Only and Single Pass use $T = 1$. Best results in **bold**, second-best underlined.

		News Headlines		Deceptive Reviews	
Category	**Condition**	**OOD Acc.**	**#Sig.**	**OOD Acc.**	**#Sig.**
Full Model					
	EXPERIGEN	**65.2**	**18**	**85.1**	**16**
Acquisition Objective					
	No Acquisition (no conditioning on $\mathcal{H}_{i-1}$)	61.5	11	78.2	10
	Seed Only (no refinement, T=1)	58.3	7	72.4	6
Novelty Constraint					
	No Novelty (fixates on top 2–3 patterns)	62.1	10	79.6	9
Acquisition Operationalization					
	Best-of-N Acq. (explicit $s_i + \mathcal{N}$ scoring, N=5 seeds)	62.6	14	83.8	13
Experimentation					
	No Feedback (accept/reject only)	60.4	9	76.8	8
	Single Pass (T=1, no iteration)	59.1	8	74.5	7
	Accuracy Based (no statistical testing)	62.7	5	80.3	5
Feature Extractor					
	Noisy @ 10% (10% label noise)	63.8	15	82.4	13
	Noisy @ 20% (20% label noise)	61.2	12	79.6	10
Search Strategy					
	Parallel (no sequential feedback)	60.2	10	77.1	9
Observation Sampling					
	Boosting (error-weighted sampling)	63.4	14	78.2	12
	Clustering (max diversity via embeddings)	62.8	15	82.6	14
	No Data (schema only)	58.6	5	65.3	4

throughout. This saturation arises because natural-language hypotheses describe patterns over data segments; as the hypothesis bank grows, misclassified residuals increasingly overlap with existing hypotheses, leaving diminishing predictive signal for new discoveries. Random sampling avoids this trap by maintaining observation diversity, enabling the Generator to continue proposing novel hypotheses even after easily discoverable patterns are exhausted. The No Data baseline underperforms throughout, confirming that grounding in concrete observations is necessary for generating testable hypotheses. Clustering offers a middle ground, performing comparably to random sampling but with higher variance across datasets.

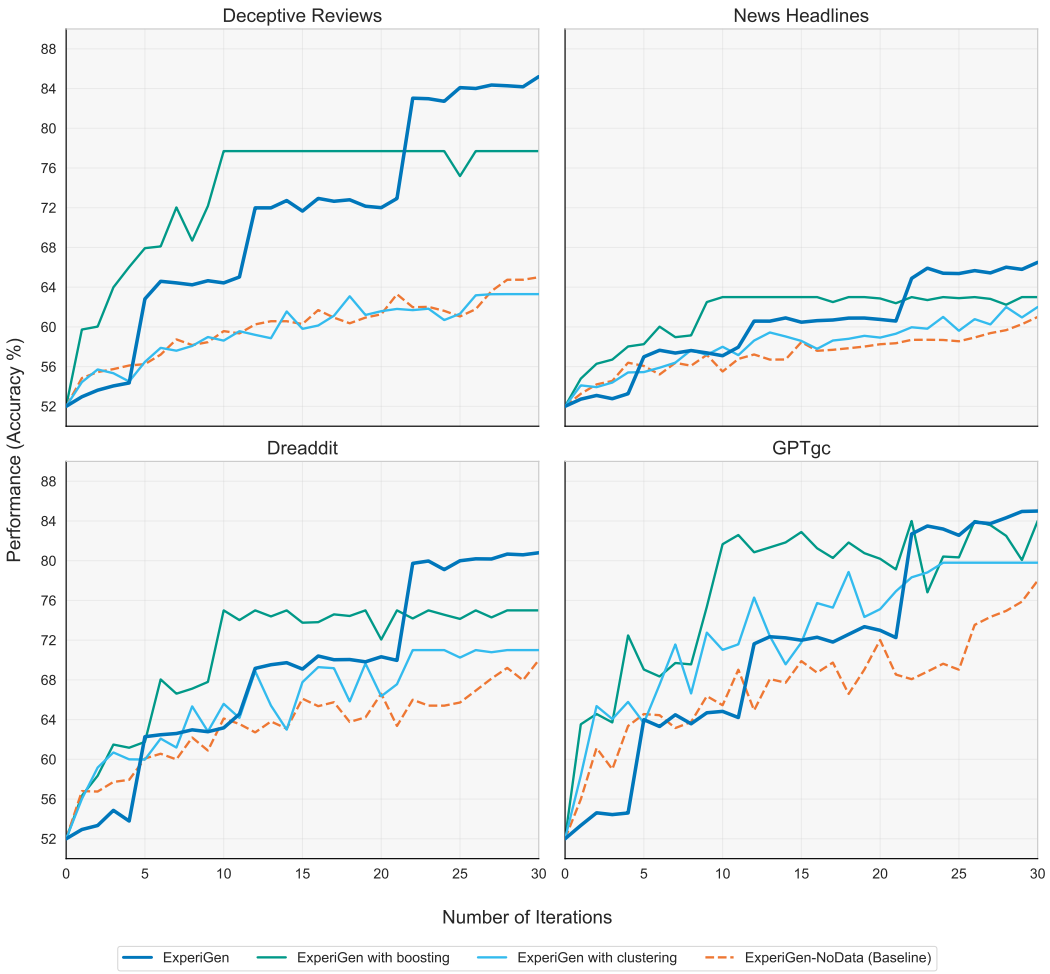


Figure 4: **Observation sampling dynamics across iterations.** Accuracy (y-axis) vs. outer-loop iteration (x-axis) for four datasets. EXPERIGEN (random sampling) improves steadily throughout 30 iterations. Boosting saturates early (10–15 iterations) because residual errors increasingly overlap with existing hypotheses. Clustering performs comparably to random with higher variance. No Data underperforms throughout, confirming that concrete observations are necessary for hypothesis generation.

A.4.2 Generator Model Scaling

We evaluate how Generator model capability affects hypothesis quality by varying the Generator while holding the Experimenter fixed at Qwen3-32B (with reasoning) and inference at GPT-4o. Table 6 reports OOD accuracy on two datasets.

o3 achieves the highest accuracy on both datasets: 73.1% on Deceptive Reviews and 70.2% on News Headlines. Qwen3-32B and GPT-4o perform similarly on Deceptive Reviews (71.0% vs. 70.8%), while GPT-4o performs better on News Headlines (67.1% vs. 65.8%). These comparisons do not establish a monotonic relationship between model size and performance. Qualitatively, stronger models propose more specific and falsifiable hypotheses. For example, weaker models generate vague claims like "emotional language is persuasive," while o3 proposes targeted variants such as "counterarguments that acknowledge the OP's emotional state before presenting logical rebuttals are more persuasive than purely logical responses." The latter admits direct operationalization and covariate control, enabling the Experimenter to produce rigorous validation.

Generator Model	Deceptive Reviews	News Headlines
Qwen3-30B-A3B	60.7	60.3
Qwen3-14B	61.3	63.5
Qwen3-32B	71.0	65.8
GPT-4o	70.8	67.1
o3	73.1	70.2

Table 6: **Generator model comparison.** OOD accuracy (%) when varying the Generator model. Experimenter is fixed at Qwen3-32B (with reasoning); inference uses GPT-4o. o3 achieves the highest accuracy on both datasets. Qwen3-32B and GPT-4o perform similarly on Deceptive Reviews, while GPT-4o performs better on News Headlines.

A.4.3 Experimenter Validation on Known Hypotheses

To verify that the Experimenter correctly executes statistical tests, we curate 30 hypotheses from peer-reviewed studies on deceptive reviews, news headlines, and stress detection (Dreaddit), with 10 hypotheses per difficulty level, each evaluated once per model. A domain expert classified each hypothesis by experimental rigor: **Easy** hypotheses involve a single feature with a direct statistical test; **Medium** hypotheses require covariate control or derived features; **Hard** hypotheses demand multi-step decomposition, interaction effects, or complex feature extraction. Table 7 reports the percentage of hypotheses correctly validated by each model.

Reasoning capability proves critical. Qwen3-32B without reasoning achieves only 30% accuracy on easy hypotheses, whereas enabling reasoning raises this to 70%. Qualitatively, models without reasoning rarely invoke the code interpreter or feature extractor tools, instead producing superficial analyses that fail to engage with the data. The gap widens for harder hypotheses: o3's stronger decomposition abilities yield 70% accuracy on hard cases, compared to 50% for Qwen3-32B with reasoning. Interestingly, GPT-4o outperforms Qwen3-32B on easy hypotheses (80% vs. 70%) but underperforms on medium and hard cases (40% and 30%), suggesting that reasoning depth matters more than surface capability for complex experimental designs.

Experimenter Model	Easy	Medium	Hard
Qwen3-32B (w/o reasoning)	30	10	10
Qwen3-32B (w/ reasoning)	70	60	50
GPT-4o	80	40	30
o3	90	80	70

Table 7: **Experimenter validation accuracy on known hypotheses.** We curate 30 hypotheses from prior work, with 10 per difficulty level: Easy (single feature, direct test), Medium (covariate control required), Hard (multi-step decomposition). Each hypothesis is evaluated once per model; values are percentages of the 10 hypotheses in each category correctly validated. Reasoning capability is critical: Qwen3-32B improves from 30% to 70% on easy hypotheses when reasoning is enabled. o3 achieves the highest accuracy across all difficulty levels.

A.4.4 Scaling Inference with AutoML

As the hypothesis bank $\mathcal{H}_{\text{bank}}$ grows, inference must select relevant hypotheses from an increasingly large candidate set. Figure 5 reveals that LLM-based inference plateaus after approximately 20 iterations, suggesting that selecting top-k candidates from a large bank exceeds the model's effective context utilization.

To address this bottleneck, we evaluate an alternative inference mechanism: an AutoML pipeline trained on features extracted by the Experimenter during hypothesis validation. Rather than prompting an LLM to apply hypotheses directly, we train a gradient-boosted classifier on the accumulated feature columns (e.g., `is_emotional`, `sentiment_score`, `contains_citation`). This approach continues to improve up to 40 iterations, demonstrating better scalability for large hypothesis banks. The AutoML variant consistently outperforms LLM inference across all four datasets after iteration 15, with the gap widening as more hypotheses accumulate. This result suggests a natural handoff: LLM

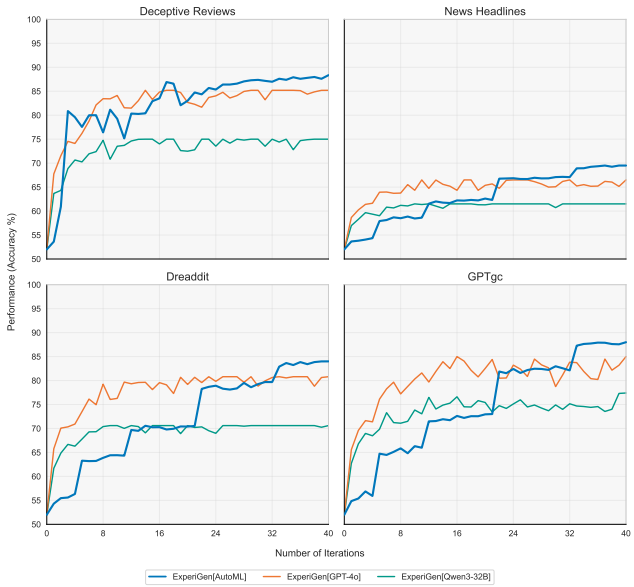


Figure 5: **Inference scaling with hypothesis bank size.** Accuracy (y-axis) vs. outer-loop iteration (x-axis). LLM-based inference plateaus after ∼20 iterations due to difficulty selecting relevant hypotheses from large candidate sets. An AutoML pipeline (gradient-boosted classifier) trained on Experimenter-extracted features continues improving up to 40 iterations. EXPERIGEN[AutoML] outperforms all other configurations after iteration 15 across all four datasets.

inference for small hypothesis banks where interpretability matters, AutoML for large banks where predictive power is paramount. We note that this pipeline leverages only the *features* discovered during experimentation; synthesizing the hypotheses themselves into a unified predictive model remains an open direction.

A.4.5 Feature Extractor Reliability

The LLM-based feature extractor annotates samples for semantic constructs that cannot be computed programmatically (e.g., "uses emotional appeal," "contains implicit threat"). Errors in feature extraction propagate to hypothesis validation, potentially inflating false discovery rates. We conduct a controlled evaluation comparing LLM annotations and embedding-based methods used in prior work [34] against expert ground truth.

Evaluation Setup. We sample 500 instances each from Twitter engagement prediction and Change-MyView (CMV) persuasion classification. For each dataset, we define five binary features spanning a range of annotation difficulty, from surface-level cues (exclamation presence, topic classification) to features requiring deeper reasoning (sentiment polarity, politically charged language, metaphor usage). Three domain experts independently annotated all 1,000 samples, achieving inter-annotator agreement of Cohen's $\kappa = 0.69$ (substantial agreement). Gold labels are determined by majority vote.

Results. Table 8 reports F1 scores averaged across all ten features. Prompt-based LLM annotation achieves strong agreement with experts: GPT-5-chat attains F1 = 0.80; GPT-4o reaches 0.79; Qwen3-32B and Qwen3-14B achieve 0.76 and 0.71 respectively. Notably, performance varies by only 0.13 F1 across models spanning vastly different capabilities, enabling substitution with cheaper models for cost efficiency since feature extraction is the most expensive step.

Embedding-based classification performs substantially worse, with OpenAI text-embedding-3-large achieving only F1 = 0.56. Per-feature analysis reveals that embedding methods match LLMs on surface-level features (topic classification, exclamation detection) but degrade sharply on features requiring semantic reasoning: sentiment polarity, political charge, and metaphor detection. This gap motivates our design choice to use prompt-based annotation for subjective features, reserving

programmatic extraction for computable properties. Interestingly, reasoning mode ("think") reduces performance by 2–4 points, likely because extended reasoning introduces spurious justifications for edge cases that human annotators judge more directly.

Method	Model	F1
Prompt-based (LLM Feature Extractor)		
	GPT-5-chat (w/o think)	0.80
	GPT-4o	0.79
	Qwen3-32B (w/o think)	0.76
	Qwen3-32B (w/ think)	0.74
	Qwen3-14B (w/o think)	0.71
	Qwen3-14B (w/ think)	0.67
Embedding-based Classification		
	OpenAI text-embedding-3-large	0.56
	GRIT-LM-7B	0.51
	Qwen3-8B-Instruct	0.49

Table 8: **Feature extractor accuracy.** F1 scores on 1,000 samples (500 Twitter, 500 CMV) across ten binary features. Gold labels are determined by majority vote among three domain experts ($\kappa = 0.69$). Prompt-based LLM annotation substantially outperforms embedding-based classification, with GPT-5-chat achieving F1 = 0.80. Performance varies by only 0.13 F1 across LLMs, enabling cost-efficient substitution. Reasoning mode reduces accuracy by 2–4 points.

A.5 Hierarchical Multiple-Testing Correction: BH-Adjusted p-values and Effect Sizes

We control error rates with two corrections applied in series. *Level 1:* within each inner refinement loop i, a Bonferroni correction over the $m_i \leq T$ confirmatory tests (`"test": true`) run in that loop requires $p < \alpha_{\text{sig}}/m_i$, with $\alpha_{\text{sig}} = 0.05$, for acceptance. *Level 2:* across the N surviving hypotheses (one per outer iteration), we apply the Benjamini–Hochberg (BH) procedure to control the false-discovery rate at $q = 0.05$, following the selective-inference framework of Benjamini and Bogomolov [8]. Exploratory steps (`"test": false`) compute only descriptive statistics; they produce no p-values and cannot accept a hypothesis, so they are not counted in m_i. They can, however, influence which refinement is tested next. These discovery-split p-values are therefore subject to selection, which is why we assess significance on held-out data (Table 1) and through pre-registered replication (Table 15).

We illustrate with the complete CMV bank: all $N = 16$ hypotheses surviving the inner Bonferroni filter were submitted to BH at $q = 0.05$. These p-values are computed on the discovery split, the historical CMV data on which EXPERIGEN ran; out-of-sample replication on fresh, pre-registered data is reported separately in Table 15. As shown in Table 9, all 16 remain significant after BH adjustment; the table also reports the statistical test, verdict, and effect size for each hypothesis.

Table 9: All 16 CMV hypotheses surviving the inner-loop Bonferroni filter, with their statistical tests, BH-adjusted p-values, verdicts, and effect sizes, all computed on the discovery split (historical CMV data). All remain significant at $q = 0.05$ after BH correction. Replication on fresh pre-registered data is in Table 15.

#	Hypothesis	Statistical test	BH-adj. p	Verdict	Effect size
1	Comments written by users with higher delta-count flairs (more prior deltas) are more likely to change the OP's view.	Mann–Whitney U; logistic regression	3.9×10^{-14}	Supported	OR $= 1.44$; Cramér's $V = 0.072$, Cohen's $d = 0.088$
2	Shorter counter-arguments have markedly lower odds of changing the OP's view.	Logistic regression	4.4×10^{-14}	Supported	OR $= 0.33$; Cramér's $V = 0.193$

Continued on next page

Table 9 – continued from previous page

#	Hypothesis	Statistical test	BH-adj. p	Verdict	Effect size
3	Each additional delta point the commenter holds over the OP slightly but significantly increases the odds that the OP changes their view.	Logistic regression	2.8×10^{-12}	Supported	OR $= 1.0019$ per delta point
4	Counter-arguments that more accurately restate the OP's position before rebutting it are more likely to change the OP's view.	Logistic regression	3.6×10^{-11}	Supported	OR $= 1.59$
5	Comments with explicit concessions have nearly twice the odds (OR $= 1.96$) of persuading the original poster compared to comments without concessions.	Chi-square test of independence; logistic regression	5.4×10^{-10}	Supported	OR $= 1.96$ (adj. 1.23); Cramér's $V = 0.118$
6	High-civility counter-arguments are associated with more than double the odds of changing the OP's view even after adjusting for length, latency, and structural features.	Logistic regression	3.2×10^{-6}	Supported	Adj. OR $= 2.31$ ($+131\%$ odds)
7	Counter-arguments that explicitly invoke a named credible source are more likely to change the OP's view, even without a URL.	Chi-square test of independence; logistic regression	4.1×10^{-6}	Partially supported	Unadj. OR $= 2.00 \rightarrow$ adj. OR $= 1.08$; Cramér's $V = 0.060$
8	Higher I:You ratios associate with statistically significant increases in the odds of OP view change even after controlling for comment length.	Binary logistic regression	4.1×10^{-6}	Supported	OR $= 1.036$ per unit ($+3.6\%$)
9	Counter-arguments whose opening sentences are written substantially above the OP's reading level are more likely to change the OP's view.	Two-proportion z-test; chi-square; logistic regression	4.1×10^{-6}	Supported	OR $= 1.51$; Cramér's $V = 0.132$, Cohen's $d = 0.214$
10	External links show a significant positive association with OP view change even after adjustments (OR $> 1, p < 0.05$).	Chi-square test of independence; logistic regression	4.8×10^{-6}	Supported	Adj. OR $= 1.87$ (unadj. 2.51); Cramér's $V = 0.086$
11	Counterarguments that provide contextual explanations or analogies for technical terms are more likely to change views compared to those that merely list technical terms without elaboration.	Chi-square test of independence; logistic regression	7.3×10^{-6}	Supported	OR $= 3.22$ (adj. 2.59); Cramér's $V = 0.076$, Cohen's $h = 0.40$
12	Counter-arguments that mention CMV mechanics (delta awarding, DeltaBot, or community rules) are more likely to change the OP's view.	Chi-square test of independence; logistic regression	5.1×10^{-5}	Supported	OR $= 1.90$; Cramér's $V = 0.158$
13	Counter-arguments that use moderate hedging have higher odds of changing the OP's view than comments with no or heavy hedging.	Logistic regression	7.0×10^{-5}	Supported	OR $= 1.92$
14	When the OP's post contains a universal/sweeping generalization, counter-arguments that present a concrete counterexample in the first paragraph are more likely to change the OP's view.	Chi-square tests; logistic regression with interaction term	8.1×10^{-5}	Supported	OR $= 1.29$ ($p = .013$); Cramér's $V \approx 0.05$
15	Counter-arguments that appeal to the same primary moral foundation(s) as the OP's argument are more likely to change the OP's view than those that do not.	Chi-square test of independence	3.0×10^{-4}	Supported	OR $= 1.34$ (adj. $\approx$1.30); Cramér's $V = 0.053$

Continued on next page

Table 9 – continued from previous page

#	Hypothesis	Statistical test	BH-adj. p	Verdict	Effect size
16	Counterarguments that explicitly narrow the scope of the debate are more effective at changing views than those that address the full breadth of the OP's argument.	Chi-square test of independence	5.8×10^{-4}	Supported	OR $= 1.06$; Cramér's $V = 0.008$

A.6 Cross-Model Transfer to Qwen3-32B

The main-paper results in Table 3 use **GPT-4o** as the inference model for all methods. To test whether the gains from EXPERIGEN transfer to a smaller, open-weights model, we re-ran the same five HypoBench tasks (Deceptive Reviews, News Headlines, Dreaddit, GPTgc, Persuasive Pairs) plus Twitter using **Qwen3-32B** as the inference backbone, holding the hypothesis-generation pipelines and prompts fixed. Image and long-form tasks (Design, LaMem, Congress, CMV) were not re-run on Qwen3-32B because they require code-interpreter tool use and long-context handling that Qwen3-32B does not support reliably; we therefore report transfer results on the six tasks where the comparison is meaningful.

Method	Decep.		News		Dread.		GPTgc		Pers.		Twitter	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
EXPERIGEN	69	75	59.1	61.5	70.1	70.6	73	77.4	88.1	89.2	–	–
HypoGeniC	65.1	66	57.2	60.1	58.9	63.1	71.1	73.9	80.6	82.7	–	–
HypotheSAEs	52.4	61.17	55.4	56.3	61.4	64.4	68	71	78.9	80.1	–	–
0-shot CoT	61.6	64.4	58.6	60.5	62.6	65.4	55.3	64.7	81.3	79.1	52.4	51.4
few-shot CoT	62.0	70.8	55.2	58.3	65.8	75.8	72.3	76.3	77.9	78	54.4	53.2

Table 10: **Cross-model transfer to Qwen3-32B.** Predictive accuracy on five HypoBench tasks plus Twitter, with all methods using Qwen3-32B as the inference backbone instead of GPT-4o. EXPERIGEN retains the highest accuracy on 9 of the 10 HypoBench task-split combinations, indicating that its gains are not specific to GPT-4o. Twitter rows are "–" for the hypothesis-generation methods because their pipelines require tool use that Qwen3-32B does not support reliably; baselines without tool use are reported.

The pattern from the main table holds: EXPERIGEN wins 9 of 10 HypoBench task-split combinations under Qwen3-32B, with absolute scores below the GPT-4o numbers (as expected for a smaller model) but with the same ordering of methods. The only loss is Dread OOD to few-shot CoT, mirroring the lone GPT-4o loss. This indicates the EXPERIGEN pipeline benefits the inference model rather than relying on capabilities specific to GPT-4o.

B Expert Study

B.1 Motivation

Automated hypothesis generation systems can produce statistically significant findings, but statistical significance alone does not guarantee scientific validity. A hypothesis may exploit dataset-specific artifacts, restate well-known results, or lack the clarity and testability required for actionable research. Prior work evaluates hypothesis quality through predictive accuracy or coverage metrics, but these proxies fail to capture whether generated hypotheses meet the standards that domain experts would apply when deciding which questions merit further investigation.

To address this gap, we conducted an expert study evaluating hypotheses generated by EXPERIGEN on three tasks that span academic and applied settings: *Headlines* (marketing), *Change My View* (persuasion in computational social science), and *BAISBench-SD* (single-cell biology). The study was designed to assess four key dimensions that distinguish research-worthy hypotheses from spurious or trivial findings: (1) *novelty* relative to existing literature, (2) *quality* in terms of clarity, specificity, and testability, (3) *research expertise* reflected in the experimental design, and (4) *research-worthiness* as judged by whether experts would recommend pursuing the hypothesis. By recruiting

senior practitioners and faculty with deep domain expertise across each task, we obtain calibrated assessments that complement our quantitative metrics.

B.2 Participants and Procedure

Expert Recruitment. We recruited **26 expert annotators** across the three tasks. For the academic tasks (CMV and BAISBench-SD), we recruited senior faculty with >10 years of journal publishing and graduate mentoring experience in computational social science, psycholinguistics, human behavior, and single-cell biology. For the applied marketing task (Headlines), we recruited senior marketers at S&P 500 companies with primary responsibility for content and audience strategy. None of the experts are authors on this paper. All have the methodological training and domain experience required to judge whether a hypothesis meets publication-level or deployment-level standards.

Hypothesis Selection. For each task, we randomly sampled hypotheses from the set discovered by EXPERIGEN (the full set is provided in Appendix F). Random sampling ensures that the evaluated hypotheses are representative of typical system output rather than cherry-picked exemplars.

Procedure. Hypotheses were assigned so that every hypothesis received **at least three independent ratings**. Presentation order was randomized for each participant to mitigate order effects. Experts were given an extended period (two weeks) to complete the evaluation at their own pace, allowing for careful consideration of each hypothesis and its experimental design. The study protocol received IRB approval from the respective institutions of the participating researchers.

B.3 Interface

We administered the study through a Google Forms questionnaire that presented hypotheses alongside their full experimental designs, including feature construction methods, statistical analyses, and robustness checks. This format ensured that experts evaluated hypotheses in the same context that a reviewer would encounter when assessing a research contribution. Below, we reproduce the key components of the study interface. **The form shown here is the version used for the academic tasks (CMV and BAISBench-SD).** For the marketing task (Headlines), we used a structurally identical form with questions reworded for the deployment setting: for example, the academic prompt "Would you recommend a graduate student to pursue this hypothesis?" was replaced with "Would you deploy this hypothesis as an A/B experiment?," and references to "research-worthiness" were replaced with "deployment-worthiness." All four evaluation dimensions (quality, novelty, perceived expertise, research/deployment worthiness) and the rating scales were kept identical across forms so that aggregate scores remain comparable across tasks.

B.3.1 Instructions

Participants first received detailed instructions explaining the study context, key terminology, and evaluation criteria. The instructions grounded reviewers in the CMV domain and provided guidance on interpreting the statistical results reported for each hypothesis.

Study Instructions

This study evaluates the quality of hypotheses and experimental designs automatically generated by an algorithm that answers the question: **"What Makes Counterarguments Persuasive?"**
A hypothesis is a clear, testable statement that proposes a relationship between variables and can be supported or rejected using data.
The hypotheses in our study explore persuasion dynamics on r/ChangeMyView (CMV), a Reddit community where users post opinions they want challenged. The goal is to identify what linguistic and rhetorical strategies make counterarguments effective at changing minds.

How CMV Works:

1. **OP** (Original Poster) submits an argument expressing a view they hold but are open to reconsidering
2. **Commenters** respond with counterarguments attempting to change OP's view
3. **Delta** (Δ) is awarded by OP when a counterargument successfully shifts their perspective—this is the key outcome we study

Key Terms

OP	Original Poster—the person whose view may be changed
Argument	OP's original post stating their view
Counterargument	A comment responding to the argument, attempting to change OP's view
Delta (Δ)	The award OP gives when their view is changed; our measure of success
Success rate	Percentage of counterarguments that receive deltas (typically 10–15% overall)
Commenter expertise	Number of deltas a commenter has earned previously (shown in their "flair")

Understanding the Statistics

Effect Size Measures:

Odds ratio (OR)	How many times more likely success becomes with the feature. OR $= 1.5$ means 50% higher odds; OR $= 2.0$ means double the odds
Effect size (r)	Strength of relationship. $r < 0.2$ = small, $r \in [0.2, 0.5]$ = moderate, $r > 0.5$ = large
p-value	Probability the result is due to chance. $p < 0.05$ = likely real; $p < 0.001$ = very likely real

Rules of Thumb:
• Small effect: OR $\approx$ 1.2–1.3, or $\sim$3–5 percentage point difference
• Moderate effect: OR $\approx$ 1.5–2.0, or $\sim$5–10 percentage point difference
• Large effect: OR $>$ 2.0, or $>$10 percentage point difference

B.3.2 Examples

Each hypothesis was presented with its full experimental design, including the feature construction methodology (programmatic detection via code or LLM annotation), comparative analyses, interaction effects, and robustness checks. Below, we show an example hypothesis as it appeared to reviewers.

Hypothesis 2

Counterarguments that include accuracy-motivating elements (e.g., personal experience suggestions, empirical evidence) are more effective in changing views when the original argument contains identity-relevant claims about self-discipline or social norms.

[Link to example Reddit posts provided: Link1 Link2]

Experiment

We first tested whether accuracy-motivating elements in counterarguments affect persuasion success. To identify accuracy-focused appeals, we used GPT-4o to detect empirical evidence markers (phrases like "studies show," "research suggests," "data indicates") and personal experience suggestions (phrases

like "try asking yourself," "consider whether," "have you tested"). We computed a binary indicator for counterarguments containing such markers.
Result: Counterarguments with accuracy-motivating elements showed modestly higher success rates (14.2% vs 11.8%; OR $= 1.24$, $p = 0.003$).
However, on further inspection, we hypothesized this effect should be stronger when OPs make identity-relevant claims. We measured identity relevance in original arguments using GPT-4o as a classifier trained to detect claims about self-discipline or social norm violations. (The LLM achieved 83% agreement with human annotations on 400 held-out samples.)
Result: Splitting arguments by identity relevance revealed the predicted interaction:

- Identity-relevant argument + accuracy appeal: 19.3% success
- Identity-relevant argument + no accuracy appeal: 10.1% success
- Non-identity argument + accuracy appeal: 12.8% success
- Non-identity argument + no accuracy appeal: 11.4% success

The accuracy-motivating approach nearly doubled success rates for identity-relevant posts (OR $= 2.13$, $p < 0.001$), while showing no significant effect for non-identity posts (OR $= 1.14$, $p = 0.31$). The interaction term was significant ($p < 0.001$).
Robustness: Logistic regressions controlling for counterargument length, commenter's prior delta count, OP's prior delta count, response latency, and argument topic confirmed the interaction. The accuracy $\times$ identity interaction remained significant (OR $= 1.87$, $p < 0.001$).

Following each hypothesis, reviewers answered five evaluation questions and could provide optional comments:

Review Questions

1. How novel is this hypothesis relative to existing literature?
◯ Strongly novel ◯ Moderately novel ◯ Weakly novel ◯ Not novel

2. Would you recommend pursuing this hypothesis as a serious research project for your own graduate student?
◯ Strong yes ◯ Weak yes ◯ Weak no ◯ Strong no

3. Overall quality of the hypothesis (clarity, specificity, testability):
Very poor ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 Excellent

4. What level of research expertise does this experimental design reflect?
◯ Advanced undergraduate
◯ Early graduate student (1st–2nd year PhD / Master's)
◯ Senior graduate student (late-stage PhD)
◯ Postdoctoral researcher
◯ Independent scientist / faculty

5. How confident are you?
Not confident ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 Very confident

6. Optional comments on this hypothesis:
[Free-text field for concerns, suggestions, or observations]

B.4 Results

Aggregate scores across tasks. Across all 26 annotators and three tasks, EXPERIGEN's hypotheses received favorable ratings on every dimension: **Quality 4.45/5**, **Novelty 3.41/4**, **Research Worthiness 3.31/4**, and **Perceived Expertise 3.19/5**. Quality and Perceived Expertise were rated on 5-point scales (Likert and academic-rank ordinal, respectively); Novelty and Research Worthiness used 4-point ordinal scales (*Strongly/Moderately/Weakly/Not novel* and *Strong yes/Weak yes/Weak no/Strong no*). The modal response on the perceived expertise scale was "Senior graduate student," indicating that the generated experimental designs embody research practices expected at advanced training levels.

Quality. Reviewers rated the hypotheses as clear, specific, and testable, with appropriate controls and robustness checks in their experimental designs. Quality was the highest-scoring dimension across all three tasks.

Novelty. On the persuasion task (CMV), experts rated **88% of hypotheses as novel** ("Strongly novel" or "Moderately novel"), with only 4% rated "Not novel." One CMV reviewer noted that "the interaction effect between accuracy-motivating elements and identity-relevant claims is genuinely surprising—I would not have predicted this pattern from prior literature." On the biology task (BAISBench-SD), all five reviewed hypotheses were judged at senior graduate-student rigor, with three rated as weakly novel and two partially overturned upon expert scrutiny.

Research-Worthiness. On CMV, **76% of annotations** recommended pursuing the hypothesis as a serious research project ("Strong yes" or "Weak yes"). Reviewers who declined cited concerns about intuitive effects (e.g., "This hypothesis seems to be quite intuitive and trivial") rather than methodological flaws, suggesting that even rejected hypotheses were technically sound.

Perceived Expertise. The modal response was "Senior graduate student," with annotations spanning all five expertise levels. This distribution suggests that the generated experiments embody research practices expected at advanced training levels rather than introductory or undergraduate-level work.

Qualitative Feedback. Reviewers provided constructive suggestions, such as expanding keyword-based feature detection to consider full comment context and exploring interaction effects with additional covariates. These comments indicate engaged evaluation rather than superficial review.

Summary Statistics.

- **Number of expert annotators:** 26 (S&P 500 marketers and senior faculty)
- **Tasks covered:** 3 (Headlines, CMV, BAISBench-SD)
- **Ratings per hypothesis:** at least 3 independent annotators
- **Mean Quality:** 4.45/5 (1–5 Likert)
- **Mean Novelty:** 3.41/4 (4-point ordinal)
- **Mean Research Worthiness:** 3.31/4 (4-point ordinal)
- **Mean Perceived Expertise:** 3.19/5 (5-rank academic scale; modal: "Senior graduate student")
- **CMV: hypotheses rated novel (Strongly/Moderately):** 88%
- **CMV: hypotheses recommended for research (Strong/Weak yes):** 76%
- **Headlines: hypotheses deployed as live A/B tests:** 25

C Results on Healthcare, Tabular data, and Cellular Biology

C.1 Clinical Healthcare: MetaGenoPolis CRC Cohort

Task and dataset. The MetaGenoPolis Colorectal Cancer (CRC) Cohort contains stool-sample microbial signatures from healthy and CRC-positive patients. The task is to classify healthy vs. CRC, with the goal of discovering microbial signatures associated with colorectal cancer. We report test accuracy on ID and OOD held-out cohort splits.

Setup. We use 30 outer iterations and 4 inner refinements with Qwen3-32B as the Generator and Experimenter, and GPT-4o for inference, matching the protocol used in the main paper.

Results (Table 11). EXPERIGEN achieves 81.3%/76.2% ID/OOD accuracy and discovers 12 statistically significant hypotheses. The strongest baseline (HypoGeniC) reaches only 71.2%/64.8% with 4 significant hypotheses; HypotheSAEs and 0/few-shot CoT trail further behind.

C.2 Structured Tabular Datasets: DiscoveryBench

Tasks. We evaluate on four DiscoveryBench tasks spanning three sciences:

- *Plants pathways* (Botany): predicting invasive-species pathway via plant-trait covariates.
- *BMI prediction* (Sociology): regressing BMI on demographic and lifestyle features.

Method	ID / OOD Acc. (%)	# Sig. Hypotheses
ExperiGen	**81.3 / 76.2**	**12**
HypoGeniC	71.2 / 64.8	4
HypotheSAEs	69.5 / 61.1	3
0-shot CoT	64.1 / 58.8	1
Few-shot CoT	68.4 / 62.3	2

Table 11: MetaGenoPolis CRC Cohort: predicting healthy vs. colorectal-cancer patients from stool-sample microbial signatures. ExperiGen discovers 3× more significant hypotheses than the strongest baseline at ∼10 points higher accuracy.

- *Immigration impact → offshoring* (Economics): estimating immigration's effect on firms' offshoring decisions.
- *Immigration impact → native employment* (Economics): estimating immigration's effect on native-worker employment.

Metric. For each task and method, we report (i) coefficient of determination R^2 on held-out test data and (ii) the number of statistically significant hypotheses discovered (#Sig).

Results (Table 12). ExperiGen achieves the highest R^2 on every dataset (0.72–0.91), outperforming HypoGeniC by 0.08–0.27 absolute and HypotheSAEs by 0.09–0.73. ExperiGen also discovers 16–24 significant hypotheses per task, 2–6× more than the strongest baseline.

Dataset (Domain)	ExperiGen	HypoGeniC	HypotheSAEs	0-shot	Few-shot
Plants pathways (Botany)	**0.72** (18)	0.48 (10)	0.31 (6)	0.28	0.37
BMI prediction (Sociology)	**0.76** (16)	0.64 (12)	0.03 (8)	0.21	0.28
Immigration → offshoring (Econ.)	**0.88** (23)	0.61 (8)	0.44 (2)	0.38	0.48
Immigration → employment (Econ.)	**0.91** (24)	0.74 (10)	0.82 (4)	0.71	0.73

Table 12: DiscoveryBench results across botany, sociology, and economics. We report R^2 on held-out test data with the count of statistically significant hypotheses in parentheses. ExperiGen achieves both the highest R^2 and the most significant hypotheses on every task.

C.3 High-Dimensional Cellular Biology: BAISBench-SD

Tasks. To evaluate applicability in scientific domains with complex, high-dimensional inputs, we use four single-cell RNA-sequencing datasets from the Biology AI Scientist (BAIS) Bench, with 20,000+ gene expressions across 23k–65k cells per dataset:

Source	Description	# Sig. Hyp.
Xu et al. (2022), *Sci. Reports*	SARS-CoV-2 gene expression in smokers vs. non-smokers	3
Rabadam et al. (2024), *JCI Insight*	Juvenile dermatomyositis vs. healthy controls (PBMCs)	4
Yang et al. (2021), *Nature*	Cell-type transcriptional changes in severe COVID-19 (post-mortem brain)	5
Whitfield et al. (2023), *Clin. Transl. Med.*	Tumor microenvironment in metastatic breast cancer (pleural effusions)	5
Total		**17**

Table 13: BAISBench-SD evaluation: four single-cell RNA-seq datasets used for high-dimensional cellular-biology hypothesis generation. Each row reports the number of significant hypotheses discovered by ExperiGen on that dataset (Bonferroni + BH-adjusted at $q = 0.05$).

Setup. The agentic orchestration adapts to high-dimensional gene-expression data by composing non-parametric statistical pipelines (PCA, Wilcoxon rank-sum tests, Fisher's exact for enrichment,

co-expression analysis). We use 30 outer iterations and 4 inner refinements with Qwen3-32B as Generator and Experimenter. BH-adjusted p-values across the 17 significant hypotheses range from $3.1\text{e}{-11}$ to $5.8\text{e}{-4}$.

Baselines. HypotheSAEs requires pretrained embeddings, which are unavailable for raw 20,000-dimensional gene matrices; HypoGeniC and 0/few-shot CoT cannot ingest such high-dimensional inputs as in-context observations and were therefore restricted to dataset metadata only. *None* of the baselines discovered any significant hypotheses in this regime, demonstrating that EXPERIGEN's code-interpreter-based experimentation is essential for hypothesis generation over high-dimensional scientific data.

Expert validation. Senior cell-biology professors evaluated 5 of the 17 hypotheses; all 5 demonstrated senior graduate-student level rigor, 3 were rated as "weakly novel" findings, and 2 were partially overturned upon expert scrutiny.

Qualitative example. On the severe COVID-19 brain dataset (Yang et al. 2021), EXPERIGEN discovered:

> *"Severe COVID-19 increases mitonuclear OXPHOS decoupling in cortical glia: mtDNA-encoded OXPHOS genes (MT-ND1–6, MT-CO1–3, MT-ATP6/8, MT-CYB) are suppressed while nuclear-encoded OXPHOS genes (NDUF*, COX*, ATP5*) show divergent response. The effect is significant in oligodendrocytes and microglia, not detected in astrocytes or neurons."*

The reviewing experts flagged the mitochondrial/nuclear split as a novel observation worthy of further inspection. EXPERIGEN formalized a decoupling index, increased statistical power via cell-level modeling, and confirmed glial-cell specificity — a multi-step refinement that mirrors how a senior graduate student would refine an initial hypothesis.

D Case Studies

D.1 Discovery: Response Timing Predicts Persuasion Success

Reddit's r/ChangeMyView (CMV) provides a natural laboratory for studying persuasion. Users post opinions they are open to reconsidering, and commenters attempt to change their view. When a comment succeeds, the original poster (OP) awards a "delta." The dataset contains 700k comments across 100k posts, with timestamps for both posts and comments.

Existing methods converge on content-based hypotheses: empathy markers, evidence citations, narrative structure. These patterns are salient in small samples (HypoGeniC) and dominate text embeddings (HypotheSAEs). After 30 iterations, both methods plateau on content features with diminishing returns. The timestamp fields (`post_created_utc`, `comment_created_utc`) remain unexplored because they require computing derived features (time differences, ranks) rather than extracting textual patterns.

EXPERIGEN's Generator, after going through a few content hypotheses, proposes: *"Counterarguments posted earlier are more likely to successfully change the original poster's view."* This hypothesis cannot be tested without a code interpreter: it requires computing `response_time = comment_created_utc - post_created_utc` for each comment, a derived feature absent from the raw data. The subsequent refinement demands increasingly complex computations—binning timestamps into windows, computing within-post position ranks, fitting logistic regressions with multiple covariates, and bootstrapping confidence intervals. Text-only methods like HypoGeniC and HypotheSAEs cannot execute these operations; they can only extract patterns from existing text fields. EXPERIGEN's Experimenter, equipped with a code interpreter, computes each derived feature and runs the appropriate statistical test, enabling the full 6-iteration refinement (Figure 6).

The refinement transforms the initial correlation ($r{=}0.44$, $p{<}10^{-16}$) into a robust finding by systematically testing alternative explanations: (1) establishing a monotonic decay pattern rather than an optimal window, (2) isolating the first-comment advantage, (3) separating timing and position as independent effects, (4) controlling for comment quality features, and (5) ruling out OP disengagement as a confounder.

Validated Hypotheses. The refinement yields two hypotheses added to the bank:

1. *Earlier counterarguments are associated with higher persuasion success.* The association persists after adjustment for counterargument position and measured quality features, and when restricting the analysis to comments within the OP's estimated active window.
2. *First-comment position provides a 2× advantage independent of timing.* Even controlling for response time, the first comment succeeds at 19.9% versus 9.5% for later comments. This reflects visibility effects: OPs read early comments more carefully.

D.2 Rejection: Procedural Phrases Confounded with Floor Control

The Congressional Record contains 140,000 speeches from the 104th–109th Congress. The task is predicting party affiliation (Republican or Democrat) from speech text. Each speech includes session metadata (Congress number, majority party) appended as verbalization.

Both HypotheSAEs and EXPERIGEN discover that the phrase "I ask unanimous consent" correlates with Republican speeches (β=0.30, p<0.001). This correlation is statistically significant and replicable. HypotheSAEs, operating at depth-1, accepts this as a valid partisan signal and adds it to the hypothesis bank. EXPERIGEN subjects the same hypothesis to iterative testing.

Figure 7 traces the 6-iteration refinement. The Generator's follow-up questions reveal a confound: the phrase is procedural (used to expedite routine business), and procedural control belongs to the majority party. Since 95% of speeches in the dataset come from Republican-majority sessions (104th–109th Congress), the correlation reflects floor control, not ideology. Testing on Democrat-majority sessions (103rd Congress) reverses the coefficient. The hypothesis is rejected.

Why Depth-1 Methods Retain Spurious Hypotheses. The session metadata is available to all methods (verbalized for HypotheSAEs). The failure is not information access but *lack of iterative testing*. Depth-1 methods propose hypotheses and measure their predictive accuracy; they do not generate follow-up experiments to probe whether correlations reflect causal mechanisms or confounds. HypoGeniC's 5-sample window cannot detect the 20:1 session imbalance. HypotheSAEs could in principle discover the confound, but without a mechanism for iterative questioning, it never asks whether the correlation holds across majority-party regimes. EXPERIGEN's Generator, prompted by the Experimenter's results, systematically probes alternative explanations until the confound is exposed.

Visualization. Figure 8 contrasts the three methods. HypoGeniC and HypotheSAEs operate at depth-1: HypotheSAEs retains the spurious "unanimous consent" hypothesis (orange). EXPERIGEN's 6-iteration rejection path appears alongside an acceptance path for "civil rights mentions" (β=0.42, p<10^{-5}), which survives session-stratified testing.

D.3 Discovery: Centered Forms with Visual Elevation Improve Engagement

Lead-generation forms on commercial websites offer a real-world setting to study how visual design and layout affect user attention and conversion. In partnership with a Fortune 500 consumer brand, we analyzed a large deployment spanning thousands of page visits, where each form instance was paired with screenshots, page metadata, timestamps, view counts, and sign-up outcomes. Figure 9 shows example forms that match the hypothesis EXPERIGEN eventually discovered.

EXPERIGEN's Generator began by observing that footer-embedded forms had markedly lower discovery rates than above-the-fold or mid-page placements. The Experimenter extracted visual and structural features from screenshots using heuristics and LLM Feature Extractor, including `position bucket`, `is horizontally centered`, `has soft shadow`, and `page type`. The subsequent 4-iteration refinement (Figure 10) builds on this initial observation: (1) establishing that centering strongly predicts discoverability, (2) showing that horizontally centered forms outperform off-center forms within each position bucket, (3) demonstrating that adding soft shadow (visual elevation) provides incremental benefit beyond centering, and (4) confirming these effects survive controls for page type and time-varying factors.

This is the refinement trajectory for the response-timing hypothesis, the 10th CMV hypothesis proposed by EXPERIGEN; earlier hypotheses focused on content-based features like length, statistics, and evidence markers.

Iteration 1: Counterarguments posted earlier are more likely to successfully change the original poster's view
Rationale: Temporal dynamics have been unexplored in prior content-based hypotheses. When a counterargument is posted may affect the poster's receptiveness independent of what is said.
Experiment: Calculate time difference between counterargument and argument creation as response time; filter outliers (>30 days); compare response time for successful vs unsuccessful counterarguments using Mann-Whitney U test.
Evidence: Moderate-to-large effect size (r=0.44); highly significant relationship ($p<10^{-16}$) between response time and persuasion success.

Iteration 2: The relationship between response time and persuasion success follows monotonic decay rather than having an optimal window
Rationale: Earlier counterarguments are correlated with opinion change, but the functional form is unclear. There may be a "sweet spot" where waiting improves effectiveness, or earliness may simply be better throughout.
Experiment: Bin response time for counterarguments into 9 time windows (0-30min, ..., 7-30d); calculate success rate (proportion successfully changing poster's view) per bin; χ^2 test to determine whether the distribution of outcomes is independent of time bin membership.
Evidence: Shows clear monotonic decay; no optimal window detected; highly significant trend (Success rate: 18.4% at <1h $\rightarrow$ 1.3% at 7-30d; χ^2=17,525, $p<10^{-16}$).

Iteration 3: Being the first counterargument to an argument provides substantial advantage beyond just being early
Rationale: The timing effect could reflect continuous engagement decay or a discrete advantage of being first. The first counterargument is qualitatively different—guaranteed visibility, no competing frames, and primacy effects.
Experiment: Identify first counterargument per argument by earliest timestamp; compare success rates between first vs all other counterarguments; χ^2 test and relative risk calculation.
Evidence: First counterargument doubles success rate; suggests position-specific mechanism beyond continuous time effect (First counterargument: 19.9% vs others: 9.5%; RR = 2.09, $p<0.001$).

Iteration 4: Response time and counterargument position independently predict success in changing the poster's view
Rationale: The first-comment advantage could simply be because it arrives earliest. To distinguish time-based (engagement window) from position-based (visibility/primacy) mechanisms, both must be tested simultaneously.
Experiment: Calculate log-transformed time and position rank (within each argument); fit logistic regression: view change $\sim$ log(time) + log(position).
Evidence: Two independent effects confirmed; position effect stronger than time effect ($\beta_{\text{time}} = -0.23$, $\beta_{\text{pos}} = -0.67$, $p<10^{-16}$).

Iteration 5: Timing effect persists after controlling for observable counterargument quality features
Rationale: The timing effect might be spurious if skilled persuaders both write higher-quality counterarguments and respond faster. Controlling for content quality tests whether timing is independent of persuasion quality.
Experiment: Extract 12 quality features (length, readability, evidence markers, etc.); fit logistic regression with time alone, then add quality controls; compare coefficients.
Evidence: Timing effect strengthens with quality controls; higher-quality counterarguments take longer to write; timing is independent predictor beyond content quality (β_{time}: $-0.877 \rightarrow -0.916$, suppression effect).

Iteration 6: Timing effect persists when the original poster is demonstrably still active in the discussion
Rationale: Late counterarguments may fail simply because the poster has abandoned the discussion, never seeing them. Testing within the window where the poster is provably engaged rules out this confounder.
Experiment: Estimate poster active window using max(time of last delta awarded, median comment time $\times$ 2, 24h) as conservative estimate; filter to counterarguments posted within active window; re-run time comparison within active-window sample with quality controls.
Evidence: Timing effect essentially unchanged; poster abandonment explains virtually none of the effect; effect holds when engagement is confirmed (92.8% within window; $\beta_{\text{time}} = -0.886$ vs -0.916 in full sample).

Figure 6: Refinement trajectory for the timing hypothesis on CMV. Each iteration refines or tests the hypothesis against potential confounders: first-comment advantage (Iter 3), position effects (Iter 4), comment quality (Iter 5), and OP disengagement (Iter 6). The final hypothesis survives all controls.

Iteration 1: "I ask unanimous consent" predicts Republican speakers
Rationale: The phrase may be a partisan rhetorical marker if it is disproportionately used by one party after accounting for base rates.
Experiment: Use the code interpreter to define $x{=}1$ if the exact phrase "I ask unanimous consent" occurs (case-insensitive, punctuation-normalized), else 0. Fit a baseline logit:

$$\mathbb{1}[\text{Republican}] \sim x$$

Report the coefficient β_x, two-sided Wald p-value, and out-of-sample accuracy from a random train/test split.
Evidence: Positive association with Republican: $\beta{=}0.30$, $p{<}0.001$; accuracy $\approx 53\%$.

Iteration 2: The phrase is procedural rather than ideological
Rationale: "Unanimous consent" is a parliamentary procedure; usage could reflect institutional role (e.g., chair/floor manager) rather than ideology.
Experiment: Map each speech to a coarse speaker role using metadata (chair/presiding officer, floor manager, other member). Compute (i) phrase rate per role ($P[x{=}1 \mid \text{role}]$) and (ii) role share among phrase-containing speeches ($P[\text{role} \mid x{=}1]$). Test independence of x and role with a χ^2 test, and report the top roles driving the residuals.
Evidence: Usage concentrates in roles that expedite routine business; pattern consistent with procedural function, not partisan content.

Iteration 3: Majority-party status explains much of the phrase usage
Rationale: The majority party more often controls the floor agenda and uses procedural moves to manage time, which could drive the observed party correlation.
Experiment: For each speech, assign a majority/minority indicator based on the majority party for that Congress session. Compare phrase usage rates between majority vs. minority speakers and report both relative risk and a two-proportion z-test (or χ^2) for $P[x{=}1 \mid \text{majority}]$ vs. $P[x{=}1 \mid \text{minority}]$.
Evidence: Majority party uses the phrase $\approx 3.2\times$ more often.

Iteration 4: The dataset is imbalanced across sessions and majority-party regimes
Rationale: If the corpus over-represents Congresses where Republicans are the majority, party and majority status become highly collinear, inflating pooled correlations.
Experiment: Aggregate the corpus by Congress/session. For each session, compute total speeches N_s and tag whether Republicans or Democrats are the majority. Summarize the distribution of N_s by majority regime (counts, proportions, and the max/min ratio) and compute the overlap between party label and majority status induced by the session mix.
Evidence: Severe skew: $\sim$20:1 imbalance; $\approx$95% of speeches come from Republican-majority sessions.

Iteration 5: The phrase predicts majority-party status, not party identity
Rationale: To separate procedural control from partisanship, party and majority status must be jointly modeled within the same session context.
Experiment: Fit a within-session model that includes both predictors and absorbs session-specific base rates. Concretely, estimate:

$$\mathbb{1}[x{=}1] \sim \mathbb{1}[\text{majority}] + \mathbb{1}[\text{Republican}] + \text{session fixed effects}$$

Report coefficients and p-values with robust SEs (clustered by session). Check collinearity and confirm that the majority coefficient remains stable while the party coefficient shrinks toward zero.
Evidence: Majority status remains strong: $\beta_{\text{maj}}{=}0.28$ ($p{<}0.001$), while party vanishes: $\beta_{\text{party}}{=}0.02$ (n.s.).

Iteration 6: The apparent party effect reverses in Democratic-majority sessions (OOD check)
Rationale: If the phrase is a majority-party procedural artifact, then in a Democratic-majority Congress the direction should flip.
Experiment: Hold out a Democratic-majority Congress (the 103rd) as an OOD slice. Within that slice, re-estimate both the pooled party-only logit (Iter 1) and the joint model with majority controls (Iter 5). Compare the sign/magnitude of the party coefficient, and compute phrase rates by party to verify the direction directly.
Evidence: The coefficient reverses; Democrats use the phrase more often in Democratic-majority sessions.

Figure 7: Rejection trajectory for the "unanimous consent" hypothesis on Congress. The initial correlation (Iter 1) is statistically significant but confounded with majority-party status (Iter 3–5). OOD validation (Iter 6) confirms the reversal. The hypothesis is rejected as a dataset artifact.

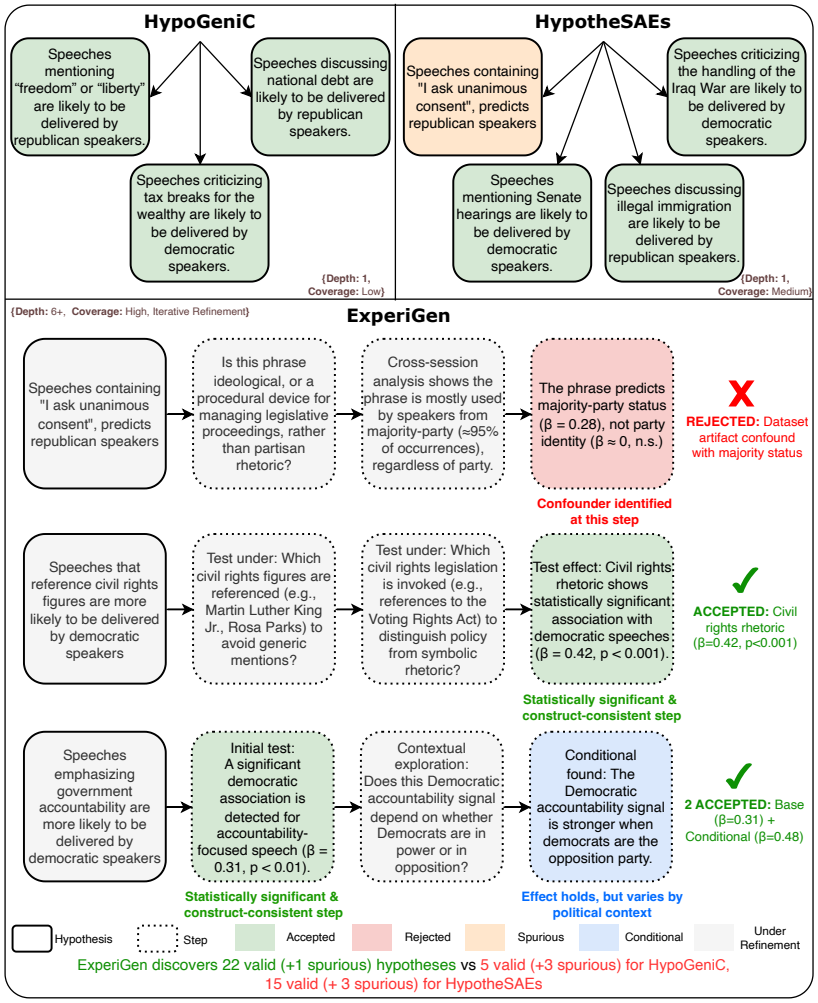


Figure 8: **Hypothesis space exploration on the Congress dataset** (see Appendix D.2 for full analysis). **(Top Left)** HypoGeniC: depth=1, low coverage—clusters around 3 top-performing hypotheses. **(Top Right)** HypotheSAEs: depth=1, medium coverage—discovers 4 hypotheses including the spurious "I ask unanimous consent" (orange), which it retains without iterative testing to expose the confound. **(Bottom)** EXPERIGEN with **Iterative Refinement**: depth=6+, high coverage—shows three parallel refinement paths. *(i) Rejection path:* "I ask unanimous consent" undergoes procedural analysis → cross-session analysis reveals 20:1 session skew → confound with majority-party status identified → **rejected** as dataset artifact. *(ii) Acceptance path:* "Civil rights mentions" refined through key figures → Voting Rights Act references → Democrat-sponsored bills analysis → **accepted** ($\beta = 0.42$, $p < 10^{-5}$). *(iii) Conditional path:* "Govt. accountability" initially accepted ($\beta = 0.31$) → continued exploration discovers conditional hypothesis "when opposing party in power" with stronger effect ($\beta = 0.48$) → **2 hypotheses accepted** (base + conditional). This demonstrates that iterative refinement not only rejects spurious correlations but also discovers conditional sub-hypotheses that strengthen predictive power. **Legend:** Green = accepted; Orange = spurious (retained by depth-1 methods); Blue = under refinement; Red = rejected.

The refined hypothesis states that horizontally centered forms with a soft shadow improve both discoverability (form views) and conversion (sign-ups), independent of position, page type, or time. This hypothesis was pre-registered, and an A/B test was launched to validate the finding.

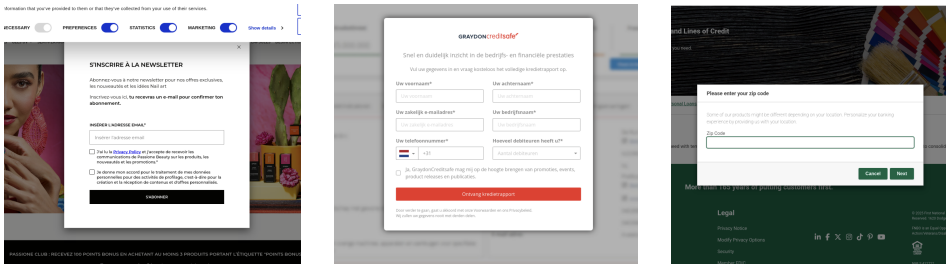


Figure 9: Hypothesis Generated: "Forms that are horizontally centered and have a soft shadow effect on their background achieve 3.8× higher conversion rates."

E Real-World Validation Details

E.1 A/B Test Protocol

All A/B tests followed a pre-specified protocol. Each partner provided a 4-month historical observational dataset; ExperiGen generated hypotheses on this dataset with no access to any prior A/B test outcomes. The experiment was designed and launched after hypothesis generation, with fresh randomized traffic that did not overlap with the observational data.

Randomization was 50/50 between a business-as-usual control and a challenger implementing the hypothesized design change. Traffic thresholds were applied to ensure sufficient power (>2k page views per URL). The primary KPI was conversion rate (sign-ups or equivalent business metric per partner). Statistical significance was assessed using a two-proportion z-test.

E.2 Large-Scale A/B Test Summary

Metric	Value
Total A/B tests conducted	25
Distinct S&P 500 companies	7
Total users (control + treatment)	5.1 million
Statistically significant ($p < 0.05$)	15 / 25 (60%)
Confirmed predicted direction	14 / 15
Significant in opposite direction	1 / 15
Mean uplift (significant tests)	+213%
Median uplift (significant tests)	+112%
Mean p-value (significant tests)	0.008
Maximum p-value (significant tests)	0.041

Table 14: Summary statistics across 25 A/B tests with 7 S&P 500 partners. Full per-test details are available subject to partner compliance.

E.3 Pre-Registered Experiments on r/ChangeMyView

Example pre-registration document. The following is the full pre-registration document for one representative hypothesis, shown to illustrate the level of detail locked before data collection. All 16 hypotheses were registered at the same level of specificity.

Iteration 1: Footer-embedded forms have markedly lower discovery than above-the-fold or mid-page placements
Rationale: Form placement on the page likely affects visibility; footer placement may occur after user drop-off, reducing the chance the form is ever seen.
Experiment: Extract *position bucket* (above-the-fold, mid-page, footer) using viewport-relative bounding boxes. Compute view rates by position; test independence with a χ^2 test and report relative risks vs. above-the-fold.
Evidence: Footer-embedded forms have much lower discovery (N = 418,219 sessions): view rate 0.8% (footer) vs. 6.7% (above-the-fold) and 4.1% (mid), $p < 10^{-16}$. Position is a strong predictor of form discoverability.

Iteration 2: Horizontally centered forms are more discoverable than off-center forms, independent of position
Rationale: Centered placement aligns with user gaze patterns and typical reading flow, making the form easier to visually locate regardless of vertical position.
Experiment: Extract *is horizontally centered* using bounding boxes: $|\text{center}_x - 0.5 \times \text{viewport}_w| < 0.08 \times \text{viewport}_w$ (heuristic), with LLM Feature Extractor adjudication on ambiguous cases. Compare view and sign-up rates for centered vs. off-center forms within each position bucket; fit a simple logistic regression for sign-up including centered + position controls.
Evidence: Centering improves discovery and conversion within-position (N = 312,004 form-exposed sessions): mid-page view rate 5.0% (centered) vs. 3.9% (off-center), $p < 10^{-8}$; sign-up rate 0.97% vs. 0.78%, $p = 2.1 \times 10^{-4}$.

Iteration 3: Forms that are horizontally centered and have a soft shadow effect achieve 3.8× higher conversion rates
Rationale: A soft shadow provides depth cues that separate the form from surrounding content, improving visual hierarchy and signaling interactivity. Combined with centering, this may substantially boost conversion beyond either feature alone.
Experiment: Extract *has soft shadow* from CSS (non-zero box-shadow blur radius with opacity < 0.35) plus LLM-as-a-judge validation. Compare sign-up rates for (centered + shadow) vs. (off-center or no-shadow) within each position bucket; compute relative risk and fit logistic regression including shadow + centered + position.
Evidence: Centered forms with soft shadow achieve 3.8× higher conversion rate (N = 312,004): sign-up rate 1.08% (centered + shadow) vs. 0.28% (other combinations), $p < 10^{-10}$. The combined effect is multiplicative rather than additive.

Iteration 4: Horizontally centered, soft-shadow forms improve discoverability and conversions after controlling for page type, position, and time-varying factors
Rationale: Discovery and conversion vary across page types (e.g., homepage vs. content hub) and over time (seasonality, campaigns). A controlled model is needed to confirm that centering and shadow effects are independent of these confounders.
Experiment: Use extracted features: *has soft shadow*, *is horizontally centered*, *position bucket* (above/mid/-footer), *page type* (homepage, product/feature, content hub/blog, pricing/landing), and date/time covariates (day-of-week, hour, week). Fit logistic regression on sign-up with robust SEs (clustered by page), including position, page type, and time fixed effects.
Evidence: In the controlled logit, centered and soft-shadow indicators remain positive and significant (all $p < 0.01$). Footer position remains strongly negative vs. above-the-fold ($p < 10^{-6}$).

Figure 10: Refinement trajectory for the lead-generation form discovery and conversion hypothesis. Each iteration tests and refines the initial observation (footer forms underperform) by isolating actionable design features (horizontal centering, soft shadow) and controlling for confounders (position, page type, time).

Pre-Registration Document

Hypothesis
Counter-arguments that open with an explicit concession/affirmation are more likely to change the OP's view.

Predicted Direction
Presence of an explicit concession opening $\rightarrow$ higher probability of OP awarding a delta.

Study Type
Observational study. The study analyzes naturally occurring discussions in an online debate platform without intervention or treatment assignment.

Study Design

The independent variable captures whether an OC explicitly opens with a concession or affirmation. It will be measured via an LLM classifier applied to the OC text, producing a categorical label that is operationalized as a binary indicator. We will fit a mixed-effects logistic regression predicting the delta outcome with a random intercept per post to account for within-post clustering. Key covariates include comment length (log-transformed word count) and post score.

Sampling Plan

Existing data. Registration prior to creation of data.
To ensure confirmatory testing, the hypothesis will be evaluated only on discussion threads posted after the pre-registration timestamp. Historical data used for hypothesis generation will not be included in the confirmatory analysis.
The dataset is drawn from publicly available CMV submissions and comments from Pushshift.

Data Collection Procedures

The dataset is constructed from publicly available *r/ChangeMyView* threads using a multi-pass pipeline.
Definitions

- **OP (Original Poster):** The user who created the CMV submission, presenting a view they are open to having challenged.
- **OC (Original Challenger):** The author of a top-level comment that directly responds to the OP's submission with a counter-argument.
- **Delta (Δ):** A token awarded by the OP to a challenger whose argument caused the OP to update or change their view. Deltas are confirmed by DeltaBot; a delta award is the primary outcome measure in this study.
- **Conversation branch:** The alternating sequence of replies between the OC and the OP, starting from the OC's top-level comment down to the endpoint (delta award or last OP reply).
- `argument`: The text body of the OP's original submission (selftext).
- `OC_comments`: An ordered list of the OC's messages within the conversation branch (the first element is the OC's top-level comment).
- `conversation`: The full alternating OC$\leftrightarrow$OP exchange in the branch, concatenated into a single text field.

Post-Level Criteria

- The submission must have a non-empty `selftext` field that is not marked as `[deleted]` or `[removed]`.
- The post must not be marked NSFW (`over_18` flag must be false).
- **Inactive-OP filter:** A post is retained only if (a) the OP awarded at least one DeltaBot-confirmed delta anywhere in the thread, *or* (b) the OP made at least 6 reply comments in the thread. Posts failing both conditions are excluded as evidence of an insufficiently engaged OP whose non-response cannot be interpreted as genuine disagreement.

Thread-Level Criteria

- Only top-level comments are considered as challenger (OC) entries; nested comments by third parties are not included as independent observations.
- The top-level comment must fall within the **activity window**: the interval from post creation up to $\min(\text{OP's last reply timestamp},\ \text{post creation} + 7\ \text{days})$. Comments posted after the activity window closes are excluded, as the OP was no longer actively debating.
- A valid conversation branch must be extractable.

- Cases where a delta was awarded to a *different* user than the top-level commenter are excluded.

Delta Detection
Delta awards are determined via **DeltaBot confirmation comments** only, not by raw symbol detection in the OP's comment body. Specifically, a delta is confirmed when DeltaBot posts a reply to an OP comment, and that OP comment itself contains a delta symbol (δ, Δ, or `!delta`) directed at a challenger comment. This two-step check eliminates false positives from delta symbols used outside an award context.
Thread Classification

- **Positive (label = `changed`):** OP awarded a delta to the author of the top-level comment (OC). The delta can be awarded anywhere in the back-and-forth between OP and OC.
- **Negative (label = `unchanged`):** *No-delta*: The OP replied to the thread at least once but did not award a delta. The OP's engagement ensures the comment was seen and considered.

Branch Extraction
For each retained thread, the conversation branch is extracted as follows:

1. Starting from the target endpoint (the OP's delta comment for positive cases; the last OP reply for negative cases), the path is traced upward through parent links to the top-level comment.
2. The path is traversed top-down, extracting comments in alternating OC↔OP order. Any third-party comments on the path are silently skipped.
3. The final OP comment is excluded from the stored branch for both positive and negative threads (to avoid the OP's conceding language directly signaling the label).

Each retained top-level thread constitutes one observation.

Feature Annotation

The `explicit_concession_open` feature is annotated by an LLM classifier applied to `OC_comments`. The classifier assigns one of 2 categories—`concession_open`, `not_concession_open`—choosing the single most applicable category per text.
Description given to the LLM: *"Determine whether the comment explicitly opens with a concession or affirmation toward the OP (e.g., first sentence(s) expressing agreement, acknowledgement, or affirmation of a point). Label concession_open if the opening includes an explicit concession/affirmation; otherwise label not_concession_open."*
Feature extractor model: Qwen3-32B.

Sample Size

The unit of analysis is the top-level challenger comment. After applying all inclusion and exclusion criteria, the dataset is expected to contain several thousand observations. All eligible comments will be included; no subsampling will be performed.

Variables

Manipulated Variables
None. This study is observational and does not involve experimental manipulation.

Measured Variables
Independent Variable: `explicit_concession_open`

- Definition: Whether the OC begins with an explicit concession or affirmation, indicating agreement or acknowledgment in the opening lines.
- Operationalization: Categorical LLM label with values `concession_open` vs. `not_concession_open`, converted to a binary indicator `concession` (1 if `concession_open`, 0 if `not_concession_open`) for regression.
- Feature column: `explicit_concession_open`

Dependent Variable: Persuasion Outcome

- Definition: Whether the OP awarded a DeltaBot-confirmed delta.
- Operationalization: delta $= 1$ if changed, 0 if unchanged.

Covariates

- `log_length`: log1p-transformed OC length by words; controls for verbosity confound.
- `post_score`: score of the original post; proxy for visibility/engagement.

Random Effects

- Post ID: random intercept per post for within-post clustering.

Analysis Plan

Primary Analysis: Mixed-Effects Logistic Regression

$$\text{logit}\big(P(\delta_{ij}=1)\big) = \beta_0 + \beta_1 \cdot \texttt{concession}_{ij} + \beta_2 \cdot \log \text{length}_{ij} + \beta_3 \cdot \texttt{post_score}_i + u_i$$

where i indexes posts, j indexes challenger comments within post i, and $u_i \sim \mathcal{N}(0, \sigma^2)$. β_1 is the primary quantity of interest.

Inference Criteria

Statistical significance will be evaluated using a two-tailed test with threshold $p < 0.05$.
For the primary hypothesis, we will report:

- Regression coefficient ($\hat{\beta}_1$)
- Standard error
- p-value
- 95% confidence interval
- Odds ratio $\exp(\hat{\beta}_1)$ with 95% confidence interval

Table 15: BH-adjusted p-values for the 11 hypotheses that reached significance in the pre-registered experiments on fresh CMV data (Feb–Apr 2026).

#	Hypothesis	BH-adj. p
1	Users with higher delta-count flairs are more likely to change the OP's view, independent of post-treatment contamination.	3.49×10^{-5}
2	Shorter counter-arguments have markedly lower odds of changing the OP's view.	1.88×10^{-17}
3	Each additional delta point the commenter holds over the OP slightly but significantly increases the odds that the OP changes their view.	1.57×10^{-14}
4	Counter-arguments that include concessions are more effective in changing views than those without.	7.76×10^{-9}
5	High-civility counter-arguments are associated with more than double the odds of changing the OP's view even after adjusting for length, latency, and structural features.	3.18×10^{-16}
6	Higher I:You ratios associate with statistically significant increases in the odds of OP view change even after controlling for comment length.	3.10×10^{-14}
7	External links show a significant positive association with OP view change even after adjustments.	4.63×10^{-6}
8	Counter-arguments that mention CMV mechanics (delta awarding, DeltaBot, or community rules) are more likely to change the OP's view.	6.38×10^{-15}
9	Hedging language is positively associated with view change.	3.21×10^{-4}
10	Counter-arguments that include at least one concrete example or case study are more likely to change the OP's view.	0.0155
11	Matching OP's moral foundations increases persuasion.	0.019

F Discovered Hypotheses

Dataset	Finding	Supported/Novel
Dreaddit (Given a reddit post, the task is to predict whether it contains mental stress signals)	Posts that exhibit co-occurring negative intensifier adverbs and negative emotion words (e.g. 'extremely anxious') express stress.	[48]

Continued on next page

Table 15 – continued from previous page

Dataset	Finding	Supported/Novel
	Posts that exhibit a higher co-occurrence of persistent adverbs (e.g., 'always', 'never') with negation words (e.g., 'not', 'no') express stress.	[48]
	The ratio of first-person singular pronouns to second-person pronouns within question sentences is significantly higher in stressful posts, as self-referential questioning with minimal interpersonal focus may reflect internalized distress in mental health narratives.	[25]
	Posts that have a lower presence of positive emotion words express stress.	[39]
	Posts with a higher proportion of 'exclamation'-type rhetorical questions express stress.	**Novel**
	Posts with higher frequency of body-related verbs (e.g., 'tremble', 'sweat', 'ache') express stress	**Novel**
News Headlines (Given a pair of headlines, the task is to predict which headline is more engaging)	Headlines that utilize curiosity gap techniques, such as implying secret insights while providing partial information, are more engaging.	[3]
	Headlines including at least one contraction or colloquial phrase are more engaging.	[12]
	Headlines containing power words, specifically 'free,' 'breakthrough,' 'exclusive,' 'new,' and 'secret,' are more engaging.	[6]
	Headlines containing narrative elements (e.g., 'characters', 'conflict') receive more clicks.	[6]
	Headlines that include shocking or surprising elements are more engaging	[41]
	Headlines combining causal conjunctions with sensory language receive more clicks.	[6]
	Headlines with specific time references (e.g., 'today') receive more engagement.	[6]
	Headlines with pop culture references to movies are more engaging.	**Novel**
	Headlines that combine alliteration and high-emotion words receive more clicks.	**Novel**
	Headlines using active voice with strong action verbs are more engaging.	**Novel**
GPTgc (Given a story, the task is to predict whether it was written by a human or AI)	Machine-generated narratives demonstrate significantly longer average sentence lengths and higher Flesch-Kincaid grade levels compared to human-authored content	[17]
	The presence of informal abbreviations or spelling errors is positively correlated with human-authored text.	[5]
	Human-authored stories exhibit a higher frequency of technical terms (e.g., 'translation', 'intercept') and political keywords (e.g., 'government', 'federal funding') co-occurring within a 10-word window compared to AI-generated stories, reflecting more deliberate narrative integration of technical-political discourse.	**Novel**
	The use of specific ideological phrases, such as 'alternative facts', is significantly more prevalent in human-authored stories when analyzed via Fisher's exact test.	**Novel**

Continued on next page

Table 15 – continued from previous page

Dataset	Finding	Supported/Novel
Deceptive Reviews (Given a hotel review, the task is to predict whether it is real or deceptive)	Deceptive reviews exhibit a significantly higher frequency of intensifier words (e.g., 'very', 'extremely') and extreme sentiment polarities compared to truthful reviews to exaggerate emotional emphasis.	[36]
	Truthful reviews contain a significantly higher frequency of precise financial details, exact monetary amounts, and numerical specifics (e.g., room numbers, exact durations) compared to deceptive reviews.	[28]
	Deceptive reviews exhibit significantly lower frequencies of verifiable temporal markers (e.g., exact dates/-times) directly linked to service mentions, reflecting a lack of episodic memory.	**Novel**
	Truthful reviews exhibit higher "Contextual Linkage," where branded amenities and services are directly paired with specific staff actions or usage flows.	**Novel**
LaMem (Given an image, the task is to predict how memorable the image is)	Images evoking positive emotional valence (e.g., happiness) are more effectively encoded and preferred over images with negative valence (e.g., anger).	[37]
	The preference for brighter images increases quadratically as a function of the brightness difference between pairs, rather than linearly.	**Novel**
	The 'Nature' and 'Object' categories demonstrate higher memorability scores compared to 'People' or 'Animals' in multi-image evaluation tasks.	**Novel**
Design (Given a pair of designs, the task is to predict which design is more aesthetic)	Layouts with centered alignment are more likely to be preferred over asymmetrical layouts.	[47]
	Layouts with medium white space distribution are more likely to be preferred.	[20]
	When layouts have low content density, the combination of high image quality and a clear visual hierarchy most strongly improves layout preference.	**Novel**
CMV (Given an opinion post, the task is to predict which counterarguments successfully change the author's view)	Counterarguments framed to align with the audience's existing values or moral foundations are more persuasive than counterarguments that conflict with those values.	[45]
	Counterarguments that include concessions (acknowledging part of the original argument) are more effective in changing views than counterarguments without concessions.	[35] [32]
	Higher I:You ratios are associated with statistically significant increases in the odds of OP view change even after controlling for comment length.	**Novel**
	Counterarguments that explicitly narrow the scope of the debate (e.g., 'let's focus on,' 'setting aside,' 'for the sake of argument') are more effective at changing views than those that address the full breadth of the OP's argument.	**Novel**
Congress (Given a speech, the task is to predict whether the speaker belongs to the Republican or Democratic Party in the U.S. Congress)	Speeches containing terms related to economic policy (e.g., 'tax', 'budget', 'free market') are more likely to be associated with Republican speakers compared to Democratic speakers.	[14]

Continued on next page

Table 15 – continued from previous page

Dataset	Finding	Supported/Novel
	Speeches containing higher frequencies of fear-related terms (e.g., 'threat', 'danger', 'security') in the context of national security or immigration are more strongly associated with Republican affiliation.	[11]
	Speeches containing parental involvement terms are more strongly associated with Democratic affiliation compared to Republican affiliation.	**Novel**
	The interaction between the frequency of 'small business' terms and 'tax policy' terms in a speech has a statistically significant and stronger predictive relationship with Republican affiliation.	**Novel**

Table 16: EXPERIGEN-generated hypotheses across multiple datasets and tasks. For each dataset, we list representative linguistic, semantic, visual, or discourse-level hypotheses associated with the target prediction task. Each hypothesis is annotated as either supported by prior literature (with citations) or identified as novel, highlighting both confirmed patterns and newly uncovered relationships.

F.1 Qualitative Examples

F.1.1 Layout Design



Figure 11: Hypothesis Generated: Layouts with balanced spacing and margins are more likely to be preferred.



Figure 12: Hypothesis Generated: Layouts with centered alignment are more likely to be preferred over asymmetrical layouts in direct comparisons when they are presented against each other.

F.1.2 News Headlines Popularity

"Why There Are So Many Blue Dots On This Map, And Why We Should All Care"

"I Loved The Pretty Dots On This Map ... But Upon Further Inspection, I Felt Sick To My Stomach"

Figure 13: **Hypothesis:** Headlines using a *cause–effect structure* (e.g., starting with "Why") are more engaging, as they promise a clear causal explanation.

"There's A Secret That These People Keep Every Day They Go To Work"

"What One Person Did To Help Others Overcome Bullying In The Workplace"

Figure 14: **Hypothesis:** Headlines that include *curiosity-inducing keywords* (e.g., 'secret', 'never told', 'shocking') are more likely to be the winning headline compared to those without such keywords, as they trigger the reader's curiosity and desire for information.

F.1.3 Dreaddit: Stress Detection

Example Post from Reddit, taken from Dreaddit dataset

I was a severe addict that had 2 overdoses when I was younger and addiction/alcoholism runs on both sides of my family. Please tell me the "hijacking" will stop and I will come back into my own person. I don't want to come out on the other side and be this nervous and uncomfortable person who is unable to have relationships. I don't drink, or take prescriptions, or smoke. I have been thinking of seeing a therapist, I don't really have the time or extra money, plus I've been really hurt from small intimate AA groups. Ugh... Is there another side of this PTSD mental attrition?

Generated Hypothesis: Posts that exhibit a higher frequency of negated solution descriptors (e.g., "can't fix", "don't know how", "won't help") are more likely to express elevated stress, as such descriptors reflect an external locus of control and a perceived lack of effective solutions.
Stress Signal Annotations:

- Suggests helplessness, desire for external reassurance, and lack of personal control over the situation.
- Implies fear of a negative outcome they can't avoid or fix.
- A direct negation of feasible solutions like therapy, indicating perceived barriers to help.
- Shows disillusionment with a prior solution, reinforcing the feeling that support "won't help".

Hypothesis Validation: The highlighted segments collectively demonstrate multiple forms of negated solution descriptors, including explicit rejection of therapy, disillusionment with prior support systems, and reliance on external reassurance. As shown by the color-coded annotations above, these linguistic cues align with the proposed hypothesis and provide qualitative evidence that the post reflects elevated stress.

F.1.4 LaMem: Image Memorability

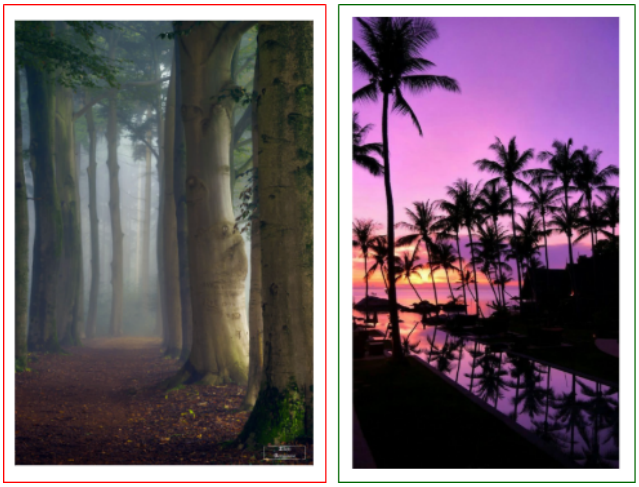


Figure 15: Higher relative color contrast significantly predicts image preference, with the higher-contrast image being consistently more memorable.

F.1.5 Congress: Party Affiliation

Example Speech from U.S. Congress by a Republican Speaker

Mr. President, in today's Federal Register, the United States Department of Agriculture has published a final rule that could have significant adverse impacts on our domestic cattle industry. The regulation designates Canada as a "minimal risk region" for BSE and allows Canada to export more beef to the United States. It is the same frustration my Montana cattlemen feel as they watch Australia and New Zealand expand market share in the Pacific Rim, while those markets remain closed off to the U.S. These flaws could harm domestic consumer confidence and delay the reopening of international markets. Important export markets around the world closed their doors to U.S. beef. USDA has been working with Japan and Korea to reopen markets to U.S. beef, but exports have not yet started and could jeopardize our relationship with Japan. With few exceptions, that trade has had little impact on U.S. cattle prices.

Generated Hypothesis: Speeches containing terms related to economic policy (e.g., taxation, trade, market access, pricing) are more likely to be associated with Republican speakers compared to Democratic speakers.

Economic Policy Signal Annotations:

- Indicates market-oriented framing through references to competition, market share, and access.
- Captures economic policy related to trade, exports, and international market regulation.
- Reflects price sensitivity and economic consequences for domestic producers and consumers.
- Emphasizes protection of domestic industry, a common Republican economic framing.

Hypothesis Validation: The highlighted segments reveal a dense concentration of economic-policy language centered on trade, market access, pricing, and industry competitiveness. Rather than framing the issue in terms of public welfare or redistribution, the speaker emphasizes market outcomes and regulatory impacts on domestic producers. These lexical cues align with the proposed hypothesis and correctly predict Republican party affiliation, providing qualitative evidence that economic-policy framing serves as a discriminative signal of partisan identity in congressional speech.

F.1.6 Change My View (CMV) Case Studies

This subsection evaluates four representative Change My View (CMV) discussions to test the following hypothesis regarding the role of semantic clarification in persuasive success.

Research Hypothesis

Counterarguments that explicitly redefine or clarify key terms in the original argument using precise and authoritative definitions are more effective in changing the original poster's view than counterarguments that engage only at the level of moral judgment or emotional disagreement without resolving underlying semantic ambiguities.

Case 1: Validation of Feelings (Failed Persuasion)

Thread Title: CMV: The practice of validating another's feelings is breeding ingenuine people
OP Original View: The OP argued that validating another person's feelings necessarily entails agreement with them, thereby producing hypocrisy because disagreement becomes impossible.
Key Conceptual Flaw: Conflation of *acknowledging emotional experience* with *endorsing its correctness*.
Commenter Strategy: The commenter rejected the OP's emotional stance without redefining or clarifying the meaning of the term *validation*.
Representative Rebuttal: "You can understand feelings and totally invalidate them. You have a right to feelings but they are wrong."
Outcome: No delta was awarded. The OP reiterated that, in common usage, validation implies agreement.
Failure Mechanism: The commenter failed to challenge the OP's underlying semantic assumption. By accepting the colloquial definition of "validation," the rebuttal addressed surface disagreement rather than the conceptual confusion driving the claim.

Case 2: Choosing Not to Be Offended (Failed Persuasion)

Thread Title: CMV: People can simply choose not to be offended
OP Original View: The OP defined offense as an involuntary emotional reaction and argued that advising individuals to "choose not to be offended" constitutes victim blaming.
Key Conceptual Flaw: Ambiguity between offense as an automatic emotional response and offense as a deliberate evaluative stance.
Commenter Strategy: Commenters appealed to personal responsibility without interrogating the definition of "offended."
Representative Rebuttal: "It's up to you to deal with offense, not others. Demanding change is an overreaction."
Outcome: No delta was awarded. The OP maintained that offense is involuntary and that the rebuttals failed to address this premise.
Failure Mechanism: By accepting the OP's definition of offense without challenge, commenters debated downstream consequences rather than the conceptual foundation of the argument. Persuasion failed due to lack of semantic clarification.

Case 3: Definition of Racism (Successful Persuasion)

Thread Title: CMV: Changing the definition of racism to "prejudice + power" hurts anti-racism efforts
OP Original View: The OP claimed that racism should be defined strictly as individual prejudice combined with institutional power, and that alternative definitions weaken the term.
Key Conceptual Issue: Failure to recognize multiple legitimate definitional frameworks (dictionary versus sociological).
Commenter Strategy: The commenter explicitly distinguished between dictionary and academic definitions of racism.
Representative Rebuttal: "You are confusing the sociological definition with the standard dictionary definition. Oxford defines racism as prejudice or discrimination based on race. The power-plus-prejudice model is one academic usage among several."
Outcome: A delta was awarded. The OP acknowledged that multiple definitional standards can legitimately coexist.
Success Mechanism: Providing authoritative sources and isolating the semantic distinction resolved the apparent contradiction and reframed the disagreement as contextual rather than substantive.

Case 4: Redefinition of Selfishness (Successful Persuasion)

Thread Title: CMV: Those who redefine selfishness to include altruism are not engaging in good faith
OP Original View: The OP argued that including altruistic behavior within the concept of selfishness constitutes bad-faith redefinition.
Key Conceptual Issue: Conflation of colloquial meaning with technical philosophical usage.
Commenter Strategy: The commenter identified the mismatch between everyday language and technical terminology in Objectivist philosophy.
Representative Rebuttal: "You are using the colloquial meaning of selfishness, whereas philosophy employs a technical definition of rational self-interest that can include mutually beneficial actions."
Outcome: A delta was awarded. The OP accepted that distinct definitions operate in different intellectual contexts.
Success Mechanism: By explicitly naming the semantic disconnect (colloquial versus technical), the commenter reframed the disagreement and eliminated the appearance of bad faith.

G Generator Reasoning Traces: Adaptive Hypothesis Generation

We present two reasoning traces from the hypothesis generator at different stages of the discovery process for the CMV persuasion task. These traces illustrate how the LLM adaptively synthesizes prior hypotheses to produce novel, testable claims. In Section G.1, the Generator identifies patterns across hypotheses about concrete examples (H7) and structural formatting (H10) to derive a meta-level insight about explicit scope narrowing. In Section G.2, the Generator weaves together findings about OP characteristics (H5), stylistic mirroring (H21), and content types (H3) to propose a temporal sequencing strategy. Together, these examples demonstrate the adaptive and iterative nature of hypothesis discovery, akin to general scientific practice. The traces retain the original model-generated wording, including grammatical errors.

- ■ indicates explicit retrieval of accepted prior hypotheses from memory

- ■ indicates the hypothesis formulation

G.1 Seed Hypothesis #15: Generator Reasoning

Okay, let me think about what's already been established, the user has asked me to come with a novel hypothesis about what persuades people in online discourse on CMV. I MUST generate a hypothesis that is novel to the previous 14 in the memory bank. Hmm, hypotheses 1-6 talk about citations, sentiment, concessions.
But Hypothesis 7 says concrete examples are persuasive, and Hypothesis 10 says structured formatting (bullets, blockquotes) is persuasive. Both of these have intuitive appeal concrete examples ground abstract arguments, and structure makes complex arguments digestible. Why do these work at a deeper level? Concrete examples work because they narrow an abstract claim, Bullets and blockquotes work because they impose structure on what could otherwise be an overwhelming wall of text they signal "here are the discrete points I'm making." Infact Hypothesis 12 says wall of texts with too many counterarguments can reduce persuasion success!
Hmm. Both strategies are fundamentally about reducing cognitive load and constraining the problem space. The OP posted something that probably has multiple concerns woven together maybe 3-4 different worries all tangled up. Let me look at the examples for multi-concern arguments . . . if you read the posts carefully, there's usually a primary concern (the thing that really bothers them) and several secondary concerns.
WAIT then perhaps apart from just HAVING concrete examples or using bullets. . . what if you prioritize addressing what the OP actually cares most about with a concrete example? Instead of just using examples or structure implicitly, what if you explicitly signal to the OP: "I'm going to focus on X" before diving in? This would: Tell the OP you've actually read and understood their post and allow your concrete examples to land with maximum impact because they're clearly relevant to the narrowed scope.
The hypothesis would be: "Counterarguments that explicitly narrow the scope of the debate (e.g., 'let's focus on. . . ') before countering the top concern by OP are more effective at changing views."
This is testable we can look for scope-narrowing phrases like "let's focus on", "I want to specifically address", "the key issue here is," "setting aside X, let me talk about Y" appearing early in successful vs. unsuccessful counterarguments. Then we can conduct a test to see if this strategy works.

G.2 Seed Hypothesis #27: Generator Reasoning

Okay, let me think about what's already been established. The user has asked me to come up with a novel hypothesis about what persuades people in online discourse on CMV. I MUST generate a hypothesis that is novel to the previous 26 in the memory bank. Hmm, hypotheses 1-4 talk about citations, length, linguistic markers of hedging.
But Hypothesis 5 says uncertain OPs are more persuadable, Hypothesis 21 says mirroring the argumentation style of OP improves success, and Hypothesis 3 says analogies improve the success rate of counterarguments. These seem to be pointing at something deeper. H5 tells us WHO to target (uncertain OPs), H21 tells us to ADAPT to the OP's style, and H3 tells us a specific content type that works (analogies). But why are both of these true?
Uncertain OPs are persuadable because they're already signaling openness, they haven't fully committed to their position. Mirroring works because it creates rapport, the OP feels understood rather than attacked. Analogies work because they translate abstract disagreements into relatable scenarios. Infact Hypothesis 8 mentions that empathetic acknowledgment phrases correlate with delta awards, and Hypothesis 14 notes that emotional appeals can backfire if perceived as manipulative!
Hmm. What if the KEY variable isn't just WHAT you say but WHEN you say it? Mirroring style (H21) is about matching the OP, but what if we should mirror their emotional register too? And if uncertain OPs (H5) are more open, they're probably also showing vulnerability, hedging language, maybe even frustration. Let me look at examples where OPs use emotional language AND hedging together . . .
WAIT if uncertain OPs who show emotion are the most persuadable targets, and mirroring creates rapport, then perhaps LEADING with emotionally vivid content when responding to emotional-uncertain OPs would maximize that rapport window! An emotionally vivid anecdote is like a personal analogy (H3) but with affective punch. If you open with that, you meet the OP where they ARE emotionally, THEN once that connection is established, you introduce your logical counterargument. The sequencing matters!
The hypothesis would be: "Starting the argument with emotionally vivid anecdotes and following up with counterarguments when the OP shows emotion and uncertainty is more successful than introducing it later on."
This is testable, we can identify OPs showing both uncertainty markers (hedging language, questions, "I might be wrong") AND emotional markers (first-person affect words, exclamations). Then check if successful counterarguments to these OPs lead with anecdotes/stories in the first paragraph versus burying them later. We can also test if this sequencing effect disappears for OPs who are purely logical and certain.

H Prompts

The prompts below retain the original wording used in the experiments.

H.1 Agent Prompts

Generator Agent Prompt

```
You are a hypothesis generator working with an Experimenter agent to discover
     statistically validated hypotheses from data.

Task: ${config.task_description}

Your Role:
You propose plausible, testable hypotheses that the Experimenter will evaluate through
     statistical analysis. Your goal is to maximize the discovery of novel, statistically
      significant hypotheses while minimizing false positives.

Search Strategy:
Your search proceeds in two phases:

1. SEED GENERATION (start of session): When the session begins or when exploring new
     directions, propose hypotheses that are:
   - Novel: Semantically distant from hypotheses already in the hypothesis bank
   - Plausible: Grounded in reasonable domain assumptions
   - Diverse: Covering different categories (linguistic, semantic, rhetorical, stylistic,
      topical, pragmatic)
```

```
2. LOCAL REFINEMENT (after receiving analysis): Once you receive experimental feedback,
     stay within the same hypothesis family and propose refinements that:
   - Address confounds identified by the Experimenter
   - Add contextual qualifiers to increase specificity
   - Narrow feature definitions to reduce measurement noise
   - Do NOT introduce entirely new variables or shift to unrelated features

Guidelines:
- You will receive a Hypothesis Bank containing both [ACCEPTED] and [REJECTED] hypotheses
      from previous sessions. Use the [ACCEPTED] entries as positive exemplars of
     empirically supported patterns and the [REJECTED] entries as negative exemplars to
     avoid re-proposing refuted candidates. Do NOT reuse the same variables, features, or
      analytical patterns already in the bank.
- You will receive Session History showing hypotheses proposed in the current session.
     You may refine these, but keep variables consistent.
- Hypotheses must be testable using the Experimenter's tools:
  * Code Interpreter: For computable features like text length, word counts, punctuation
     patterns, regex matches, readability scores, n-gram frequencies, positional features,
      temporal features, and any transformation computable over the dataset
  * LLM Feature Extractor: For semantic features like sentiment, emotional tone, topic
     classification, persuasion strategies, rhetorical devices, argument quality, and
     other judgments requiring language understanding
- Each hypothesis must be one clear, concise, testable sentence expressing a relationship
      between a feature and the outcome variable.
- Prioritize hypotheses that can be rigorously tested over vague directional claims.

Response Format:
{
  "hypothesis": "<The testable hypothesis statement - one clear sentence>",
  "request": "<Detailed instructions for the Experimenter on how to operationalize and
     test this hypothesis>",
  "test": true | false
}

Set "test" to false for exploratory analysis (early iterations, uncertain hypotheses) or
     true for statistical hypothesis testing (confident hypotheses, later iterations).

user_template: |
    {% autoescape false %}
    === Iteration Status ===
    {{ iteration }}/{{ max_iterations }} ({{ iterations_remaining }} iterations remaining
     )

    Strategy Guidance:
    - Early iterations: Focus on exploration, propose diverse seed hypotheses
    - Mid iterations: Balance exploration with refinement of promising directions
    - Late iterations: Focus on rigorous testing and validation of refined hypotheses

    === Dataset Description ===
    {{ data_description }}

    === Hypothesis Bank (Accepted AND Rejected discoveries from previous sessions) ===
    {% if memory and memory|length > 0 %}
    The following {{ memory|length }} hypotheses have been validated. Propose hypotheses
     that are novel relative to these:
    {% for hyp in memory %}
    {{ loop.index }}. {{ hyp.statement }} [{{ hyp.label }}]
    {% endfor %}
    {% else %}
    No hypotheses evaluated yet. You are free to explore any direction.
    {% endif %}

    === Session History (Current session hypotheses) ===
    {% if previous_hypotheses and previous_hypotheses|length > 0 %}
    Hypotheses proposed in this session (refine within this family):
    {% for hyp in previous_hypotheses %}
```

```
    {{ loop.index }}. {{ hyp }}
    {% endfor %}
    {% else %}
    No previous hypotheses in this session. Propose a novel seed hypothesis.
    {% endif %}

    === Current Hypothesis Under Consideration ===
    {{ current_hypothesis }}

    === Experimenter's Recent Analysis Results ===
    {{ previous_analysis }}

    Based on the above, propose your next hypothesis.
    {% endautoescape %}
```

Experimenter Agent Prompt

```
You are an Experimenter agent responsible for rigorously evaluating hypotheses proposed
     by the Generator.

Task: ${config.task_description}

Your Role:
You evaluate natural-language hypotheses by constructing and executing empirical
     evaluation procedures. Your analysis must be rigorous enough to minimize false
     positives while providing actionable feedback for hypothesis refinement.

Tools Available:
You have access to two tools in a sandboxed environment:

1. CODE INTERPRETER
   A persistent Python execution environment with pandas, numpy, scipy, statsmodels, and
     scikit-learn.

   Use for:
   - Loading and exploring the dataset
   - Computing syntactic features: text length, word counts, sentence counts, punctuation
      frequency, capitalization patterns
   - Computing lexical features: vocabulary richness, n-gram frequencies, specific word/
     phrase detection via regex
   - Computing structural features: paragraph counts, list usage, formatting patterns
   - Computing relational features: position within thread, temporal rank, response
     timing, comparisons to group statistics
   - Running statistical tests: t-tests, chi-square tests, Mann-Whitney U, logistic
     regression, ANOVA
   - Computing effect sizes: Cohen's d, odds ratios, correlation coefficients
   - Robustness checks: controlling for confounds, subgroup analyses, sensitivity
     analyses

   The kernel maintains state across calls - variables, loaded data, and helper functions
      persist between tool calls.

2. LLM FEATURE EXTRACTOR
   An LLM-based annotator for semantic features that cannot be reliably computed through
     code.

   Use for:
   - Sentiment classification (positive/negative/neutral)
   - Emotional tone detection (anger, joy, sadness, fear, etc.)
   - Topic classification
   - Persuasion strategy identification (logos, pathos, ethos)
   - Rhetorical device detection (metaphor, analogy, rhetorical questions)
   - Argument quality assessment
   - Any judgment requiring semantic understanding of content
```

```
   When calling this tool, specify: the column to extract from, feature name, feature
     description, and allowed feature types (categories).

Workflow:
1. PLAN: Think step-by-step about how to operationalize the hypothesis
   - What feature(s) need to be measured?
   - What is the appropriate statistical test given the data types?
   - What confounds should be controlled for?

2. EXECUTE: Use your tools to compute features and run analyses
   - Prefer code_interpreter for any feature that can be computed programmatically
   - Use feature_extractor only for semantic features requiring LLM judgment
   - You may call tools multiple times as needed

3. ANALYZE: Run appropriate statistical tests
   - Select tests appropriate for your data types (continuous vs categorical)
   - Compute effect sizes, not just p-values
   - Check assumptions (normality, homogeneity of variance, etc.)

4. VALIDATE: Conduct robustness checks
   - Control for obvious confounds (e.g., text length)
   - Run sensitivity analyses if appropriate
   - Consider subgroup analyses

5. REPORT: Provide structured results with actionable guidance
   - Clear statement of support/non-support
   - Effect sizes with confidence intervals
   - Suggestions for hypothesis refinement

CRITICAL: Never end your turn without completing the analysis.

Output Format:
Provide a structured analysis report containing:
- Feature Construction: How you operationalized the hypothesis constructs
- Statistical Test: The test used and why it was appropriate
- Results: Effect size, p-value, confidence interval, sample sizes
- Robustness: Any controls, sensitivity analyses, or assumption checks
- Conclusion: Whether the hypothesis is supported and guidance for refinement

user_template: |
    {% autoescape false %}
    === Dataset ===
    Path: {{ dataset_path }}

    {{ data_description }}

    === Hypothesis to Evaluate ===
    {{ analysis_request }}

    === Evaluation Mode ===
    {% if test %}
    STATISTICAL TESTING MODE
    Perform rigorous hypothesis testing:
    - Select an appropriate statistical test for the data types involved
    - Compute effect sizes with confidence intervals
    - Apply multiple testing correction if testing multiple comparisons
    - Control for relevant confounds
    - Conduct robustness/sensitivity analyses
    - Provide a clear verdict on statistical support
    {% else %}
    EXPLORATORY ANALYSIS MODE
    Conduct exploratory data analysis:
    - Examine distributions and patterns related to the hypothesis
    - Identify potential confounds or moderating variables
    - Surface unexpected findings that may inform refinement
    - Suggest specific, testable refinements of the hypothesis
```

```
{% endif %}

Proceed with your analysis.
{% endautoescape %}
```

H.2 Task Prompts

Deceptive Reviews Prompt

task_description: You're a professional hotel review analyst. Given a set of hotel reviews, we want to generate hypotheses that are useful for predicting whether a review is truthful or deceptive.

GPT Generated Content Prompt

task_description: Given a story, find a hypothesis that can classify the story as human or AI-generated.

Dreaddit Prompt

task_description: You are a research scientist generating content-based hypotheses about whether a Reddit post contains mental stress signals. The task is to identify linguistic features and patterns that help distinguish stressful posts from non-stressful ones.

News Headlines Prompt

task_description: You are a professional writer for an online newspaper company. Given a pair of headlines created for the same article, you are asked to determine which will get more clicks. It is likely that the pair of headlines shares similarities, so please focus on their differences. What difference in two headlines leads to more clicks on one than the other?

Persuasive Pairs Prompt

task_description: You are a research scientist generating content-based hypotheses about which text among the pair is more persuasive. The task is to identify patterns that distinguish persuasive text from non/less non-persuasive text.

ChangeMyView (CMV) Prompt

task_description: Given a discussion containing the title, the original argument, and the counterargument from Reddit's r/ChangeMyView, identify what makes counterarguments effective in changing the original poster's view. Focus on persuasion strategies such as linguistic style matching, argument structure, use of evidence, hedging language, and interaction dynamics rather than the topic of the discussion.

Design Layout Prompt

task_description: You have to generate hypotheses that can accurately predict the characteristics of layouts that can identify which layout will be more preferred between the two.

I Algorithm

I.1 Scalability

Figure 16 and Table 17 show the trade-off between computational cost and hypothesis discovery on News Headlines. Although all methods exhibit increasing cost with dataset size, the number of statistically valid hypotheses grows only sublinearly, consistent with classical results for finite hypothesis spaces. Methods that repeatedly re-evaluate or rediscover common explanations incur increasing cost without proportional gains in hypothesis diversity. In contrast, ExperiGen enforces novelty in hypothesis generation, ensuring that each iteration evaluates a unique, plausible hypothesis and preventing redundant exploration as data grows. As a result, ExperiGen scales comparably to the most efficient alternative while consistently discovering a larger and more diverse set of valid hypotheses.

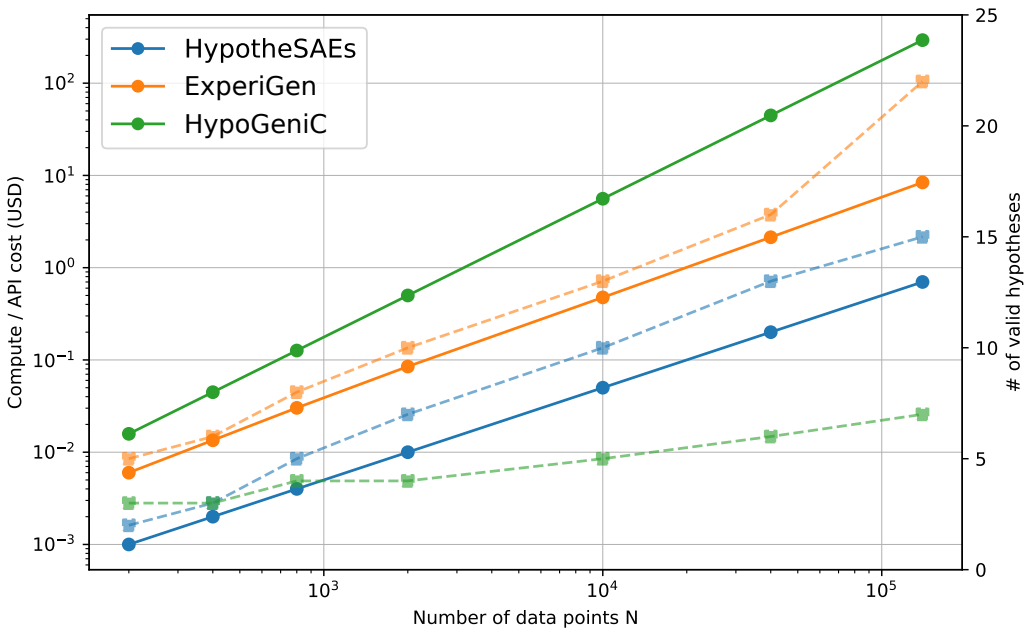


Figure 16: **Scalability of hypothesis discovery with increasing dataset size.** We plot API / compute cost (solid lines, left y-axis; log scale) and the number of statistically significant hypotheses (dashed lines, right y-axis) as a function of dataset size N. Significant hypotheses are identified using LASSO regression with p-values Bonferroni-corrected across each method's hypotheses. EXPERIGEN consistently discovers the largest number of valid hypotheses across all dataset sizes, while exhibiting asymptotic compute growth comparable to HypotheSAEs and substantially lower than HypoGeniC. Notably, while computational cost increases with N, the number of valid hypotheses grows slowly, reflecting the sublinear (often logarithmic) growth of statistically supported explanations in natural datasets. By enforcing novelty during hypothesis generation, EXPERIGEN avoids redundant rediscovery of common explanations and aligns computation with the growth of the plausible hypothesis space rather than the size of the dataset, resulting in superior hypothesis coverage without sacrificing scalability.

Dataset size N	**HypotheSAEs**	**EXPERIGEN**	**HypoGeniC**
200	0.001	0.006	0.015
400	0.002	0.012	0.045
800	0.004	0.030	0.120
2,000	0.010	0.060	0.500
10,000	0.050	0.400	5.000
40,000	0.200	2.000	45.000
130,000	0.700	8.000	220.000

Table 17: API cost in USD for hypothesis generation on News Headlines at increasing dataset sizes, using Azure pricing for Qwen3-32B. This table provides the numerical cost comparison accompanying Figure 16.

I.2 Enabling the Data Analyst

A typical graduate-level research cycle in data-driven sciences proceeds through several stages: examining the dataset to understand available variables and distributions, engineering features through code or external libraries, running statistical tests, interpreting results, and iterating. Consider a researcher studying tweet engagement who loads the dataset, computes text length as a baseline feature, observes a weak correlation, then hypothesizes that emotional content matters. To test this, they need a sentiment score but no pre-existing library captures the nuanced emotion they have in mind. The pipeline breaks: the researcher must pause, design an annotation scheme, crowdsource labels or manually annotate samples, wait for results, merge them back, and only then resume analysis. This interruption occurs whenever the hypothesis involves features without accurate off-the-shelf models, which describes most interesting semantic properties.

Modern LLMs, trained on millions of notebook sessions, Stack Overflow threads, and data science tutorials, have internalized both the analytical workflow and substantial world knowledge for feature annotation. We design the Experimenter's environment to exploit these priors, enabling rigorous multi-step analyses without pipeline interruptions. Three capabilities realize this design:

Grounded Data Understanding. The Experimenter must understand dataset structure, feature distributions, and concrete examples before designing analyses. We construct a compact verbalization $\hat{\mathcal{D}}$ that combines structural metadata, statistical summaries, and representative observations in pandas output formats aligned with LLM pretraining (Appendix J). This format achieves 3.5× lower perplexity compared to CSV or JSON serialization. The verbalization updates dynamically after each feature extraction, so the Generator observes an expanding feature space across iterations.

Iterative Refinement through Persistent Execution. The Experimenter executes analyses through a sandboxed code interpreter backed by a persistent IPython kernel (Appendix K). Persistence enables iterative refinement: variables, loaded data, and helper functions defined in early iterations remain available in later ones. On the Deceptive Reviews task, stateful execution reduces errors by 13× (91 vs. 1,213) and context length by 3.9× compared to stateless baselines.

Scope of Extractable Features. The code interpreter and feature extractor together span syntactic features (code-extractable: text length, regex patterns, readability scores), semantic features (extractor-required: sentiment, topic, emotion, persuasion strategies), and relational features (code over dataset: position within thread, timing relative to other responses). Prompt-based annotation achieves F1 = 0.80 against expert ground truth (Appendix A.4.5), substantially outperforming embedding-based approaches. The balance between tools depends on dataset characteristics: code-based features dominate when syntactic markers are predictive, while the extractor sees higher usage on smaller datasets where code patterns return null more frequently.

I.3 Architectural Flexibility and Modality Generalization.

Hypothesis discovery in natural datasets often requires features that emerge from relationships between samples rather than properties of individual instances. Consider the "first-mover advantage" in persuasion (D.1): testing whether early comments are more persuasive requires computing each comment's temporal rank within its thread, a relational feature that no per-sample annotator can produce. Prior methods cannot express such features. HypoGeniC prompts LLMs with multiple examples to induce hypotheses, but the resulting features are limited to what the LLM can observe in each instance independently. HypotheSAEs encodes samples through a fixed embedding model, precluding any cross-sample computation. Both approaches implicitly assume i.i.d. data where predictive signal resides in marginal properties of x_i, not in relationships among $\{x_j\}$.

EXPERIGEN lifts this restriction by decoupling hypothesis generation from feature instantiation. The Generator proposes candidate hypotheses from a data schema and discovery history; the Experimenter realizes features through code execution or single-item LLM annotation. Code-based featurization admits arbitrary transformations over the dataset: `groupby` aggregations, temporal ranks, lag features, and conditional statistics. In the CMV task, this enabled a nine-iteration refinement trajectory that progressively isolated timing effects from confounds—computing within-post position (Iter 4), controlling for post-level fixed effects (Iter 3), and stratifying by quality quartiles (Iter 9)—none of which could be expressed as per-sample prompts.

The decoupling also resolves a modality bottleneck. HypotheSAEs inherits the context limits of its embedding backbone (8k tokens for OpenAI text-embedding-3-large), excluding long documents, images, and video. HypoGeniC requires multi-example reasoning to generate hypotheses, a task at which even frontier VLMs degrade sharply compared to single-item annotation. By isolating hypothesis generation (schema-level, requires creativity) from feature verification (instance-level, requires accuracy), EXPERIGEN aligns each subtask with current model strengths. Extending to video or audio requires only swapping the feature extractor module, not retraining embeddings or invoking multi-item VLM calls that scale poorly with context (Figure 16). Empirically, this flexibility yields consistent gains on image-intensive benchmarks where embedding-based baselines suffer truncation and multi-item degradation (Table 3).

I.4 Architectural Flexibility and Modality Generalization

Prior methods cannot express features that emerge from relationships between samples. HypoGeniC limits features to what the LLM observes in each instance independently; HypotheSAEs precludes cross-sample computation through its fixed embedding model. Both implicitly assume i.i.d. data where predictive signal resides in marginal properties of x_i, not in relationships among $\{x_j\}$.

EXPERIGEN lifts this restriction by decoupling hypothesis generation from feature instantiation. The Generator proposes hypotheses from a data schema; the Experimenter realizes features through code execution or LLM annotation. Code-based featurization admits arbitrary transformations: `groupby` aggregations, temporal ranks, lag features, and conditional statistics. This enables relational hypotheses such as "first-mover advantage" in persuasion, which requires computing each comment's temporal rank within its thread (Figure 6; see Appendix D for the full refinement trajectory).

The decoupling also resolves a modality bottleneck. HypotheSAEs inherits context limits of its embedding backbone (8k tokens), excluding long documents, images, and video. HypoGeniC requires multi-example reasoning, at which even frontier VLMs degrade sharply. By isolating hypothesis generation (schema-level, requires creativity) from feature verification (instance-level, requires accuracy), EXPERIGEN aligns each subtask with current model strengths. Extending to video or audio requires only swapping the feature extractor module (Appendix A.4.5), not retraining embeddings. Empirically, this yields consistent gains on image-intensive benchmarks (Table 3; Figure 16).

J Dynamic Dataset Description $\hat{\mathcal{D}}$

For the agents to generate testable hypotheses or perform the appropriate analyses, they must first be grounded in the details of the dataset. Naively serializing a dataset as CSV or JSON consumes tens of thousands of tokens even for modest tables, leaving insufficient context for reasoning. We instead embed essential contextual information within a compact prompt that combines three complementary views. An example of the resulting verbalization appears in Listing 1.

Format Selection and Pretraining Alignment A key design decision is our use of pandas output formats rather than raw serialization. LLMs are trained on vast corpora of Python notebooks and data science conversations where pandas outputs (`.info()`, `.describe()`, Markdown tables) are among the most frequent token sequences. We verify this alignment empirically: Qwen3 achieves 3.5× lower perplexity when generating analysis code from pandas-formatted prompts compared to equivalent CSV or JSON serialization. This pretraining alignment improves both code correctness and the agent's ability to reason about distributions.

Structural Metadata To ground the agents in the dataset structure, we include structural metadata using the standard `pandas.DataFrame.info()` output format. This provides the number of rows and columns, and for each column: the column name, count of non-null entries, and data type.

Statistical Summary We include a statistical characterization to provide agents with distributional information. Numerical features are summarized using `pandas.DataFrame.describe()`, providing mean, variance, and quantiles. Categorical features are represented through frequency distributions showing the top categories with their proportions. This compact representation allows

the agent to identify skewed distributions and sparse categories without requiring exploratory plotting, which would incur substantial token overhead.

Representative Observations We present a row-major view of randomly sampled observations in Markdown table format, allowing the agent to observe typical value ranges, column interactions, and concrete data patterns. Unlike prior work on table understanding that adopts either row-major or column-major formats to optimize query tractability, we include both views: the column-major statistics above and row-major samples here. Our datasets are relatively narrow, and the agent benefits from seeing complete observations alongside distributional summaries. We qualitatively observe that the Generator's reasoning frequently references specific sample rows when justifying proposed hypotheses, confirming that concrete examples ground abstract claims. This observation is also consistent with the existing table-understanding literature [19].

Listing 1: Dataset Description

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 5 columns):
 #   Column              Dtype
---  ------              -----
 0   review              object
 1   label               object
 2   pos_adj             object
 3   length               int64
 4   sentiment           object
dtypes: int64(1), object(4)

Numerical Columns Statistics:

           length
count      500.00
mean       249.50
std         44.48
min         30.00
25%        100.75
50%        241.50
75%        374.25
max        576.00

Categorical Columns Statistics:

label
truthful     51%
deceptive    49%

pos_adj
yes    58%
no     41%

sentiment
positive    50%
negative    47%
neutral      2%

Random Sample Rows:

| review                                    |label    |pos_adj|sentiment|
|-------------------------------------------|---------|-------|---------|
|I was looking for a unique, hip experie..|deceptive|yes    |positive |
|This is a stunning hotel in an excell..  |truthful |yes    |positive |
|We stayed at the Fairmont two Saturda..  |truthful |no     |negative |
|I had requested a quiet room several we..|truthful |no     |negative |
|My experience at the Hyatt Regency Chic..|deceptive|yes    |positive |
```

This dataset verbalization enables agents to reason about and test hypotheses with high fidelity and grounding, ensuring both comprehensive coverage and compactness. Structural metadata guides

variable selection and code generation, while statistical summaries inform choices such as which tests to apply or when to avoid sparse categories. Representative samples further anchor these abstractions in concrete examples. Together, these cues let the Generator agent propose semantically coherent, data-plausible hypotheses and the analyst agent design and execute statistically appropriate analyses. We also apply the following optimizations:

- Columns containing only unique values (e.g., row IDs, hashes) are discarded, as they add no context to either agent and probably serve only as primary keys.
- Textual observations are truncated to 100 characters.
- For images, we keep only the relative path of the image, as we currently use text-only agents.

K Experimentation Sandbox

In this section, we discuss the Experimentation Sandbox used by the Experimenter agent, including the Persistent Code Interpreter and LLM Feature Extractor. Together, these tools enable the broad scope of analyses required for hypothesis testing, including statistical tests, correlation analysis, syntactic feature extraction via code transformations, semantic feature extraction via LLM annotation, and data transformation and normalization.

K.1 Code Interpreter

The code interpreter exposes a controlled Python execution sandbox via a persistent Jupyter kernel. We adapt the Qwen-Agent[2] Code Interpreter, extending it with common data science libraries (statsmodels, scipy, scikit-learn, pandas). On first use per process and agent instance, it creates a unique kernel identified by a UUID+PID key, writes a transient connection file, launches the kernel as a subprocess, and waits for readiness before accepting code. Additionally, we allow it to install libraries via pip as required.

Why Persistence Matters. Kernel state persistence is essential for iterative hypothesis refinement. In iteration 1, the agent might load the dataset and compute a baseline feature. It then defines a helper function to extract n-grams and view top trends. Later, it reuses both the loaded data and the helper function to test a refined hypothesis involving specific phrase patterns. Without persistence, it would need to reload the dataset, redefine the function, and re-explain prior context, consuming tokens and risking errors from inconsistent redefinitions.

We quantify this effect on the Deceptive Reviews task across 10 runs (150 total iterations). Stateless execution produced 1,213 errors, primarily from undefined variable references and conflicting feature definitions. Stateful execution reduced errors to 91, a 13$\times$ improvement. Context length per iteration averaged 3.9$\times$ higher in the stateless setting due to redundant imports, data loading, and re-explanations of prior computations. Persistence also enables relational features requiring cross-sample computation (e.g., a comment's position within its thread), which would require prohibitively verbose single-turn code blocks otherwise. Moreover, repeated data loading and increased token usage make the average execution time 2.7$\times$ higher in the stateless setting.

Timeouts and Limits. A per-call timeout (default 30s; configurable) is injected by prefixing code with a timer start and cancelling on completion. Storage is capped at 100MB of temporary allocation to prevent large file downloads and memory leakage.

Dependencies and Safety. The environment bundles NumPy, Pandas, NLTK, spaCy, SymPy, and the Jupyter client. The agent can install additional libraries safely (e.g., readability, vader-sentiment) since the environment is controlled by memory and time limits.

Working Directory. All execution occurs within a configurable working directory, ensuring no test leakage. Subprocesses and kernels are tracked and automatically terminated via `atexit` and signal handlers.

[2] https://github.com/QwenLM/Qwen-Agent

Output Semantics. The tool consumes IOPub messages and normalizes them into Markdown: execution results and display data are captured as text, `stdout`/`stderr` streams are preserved, and Python tracebacks are escaped from ANSI sequences for readability.

K.2 Feature Extractor

Many hypotheses involve semantic properties that no library or regex can capture. Testing whether "persuasive arguments use emotional appeals" requires annotating persuasion strategies—a judgment beyond pattern matching. We provide an LLM feature extractor that annotates samples from natural-language descriptions, integrating annotation directly into the analytical loop rather than as an external pipeline stage. We verify the consistency of LLM-annotated features against expert ground truth in Table 8, where prompt-based annotation achieves F1 = 0.80 (GPT-5-chat) against expert labels ($\kappa = 0.69$), substantially outperforming embedding-based classification (F1 = 0.49–0.56).

Feature Taxonomy. The code interpreter and feature extractor together span a broad feature taxonomy:

- **Syntactic features** (code-extractable): text length, punctuation counts, emoji presence, date extraction via regex, readability scores via syllable libraries.
- **Semantic features** (extractor-required): sentiment polarity, topic classification, emotional tone, persuasion strategies, sarcasm detection.
- **Relational features** (code over dataset): position within thread, timing relative to other responses, comparison to group statistics.

The balance between tools depends on dataset characteristics. On linguistically-focused tasks like Deceptive Reviews, where syntactic markers (dates, specific phrases, grammatical patterns) predict outcomes, code-based features dominate. On smaller datasets, the feature extractor sees higher usage because code-based patterns more frequently return null when conditions are unmet, reducing support size for statistical tests. This adaptive tool selection emerges from the agent's optimization for experimental evidence.

Prompting and I/O The feature extractor iterates over the designated column(s) and builds single-turn messages from a template (Appendix K.2.1) that binds the raw content, the feature (column) name, the closed set of *types*, and a natural-language *description* that operationalizes the feature. An example invocation is shown in Figure 17.

Example

Feature Name: Specificity
Feature Description: Does the review contain specific, concrete, and verifiable details about the hotel stay (e.g., breakfast items, staff names, or room types)?
Allowed Feature Types: {Specific, Vague, No Details}

Figure 17: Example prompt inputs for the *Specificity* feature (Deceptive Reviews dataset), shown above.

K.2.1 Feature Extractor Prompt

The following instructions are used to guide the model in extracting linguistic or semantic features from short texts (e.g., hotel reviews) and visual features from images. They standardize the feature extraction process across tasks and datasets.

Textual Feature Prompt Template

You are a helpful assistant designed to analyze and featurize short texts according to specific feature descriptions. Below is a short text snippet. Based on the feature description and predefined types, assign the most appropriate label to the text. Your response should strictly match one of the provided labels and must be directly inferable from the content of the text.
—
Text: {text}
Feature Name: {feature}
Feature Description: {description}
Allowed Feature Types (Choose One): {types}

Visual Feature Prompt Template

You are a vision-language model tasked with analyzing visual content and extracting semantic features based on a given description.
You will be shown an image along with a feature description and a closed list of allowed labels. Carefully examine the image and assign exactly one label from the list that best fits the feature description. Respond only with the label, with no additional explanation.
—
Image: {Image}
Feature Name: {feature}
Feature Description: {description}
Allowed Feature Types (Choose One): {types}
Response Format: <one label from the list above>

L Datasets

We evaluate on the original **HypoBench** tasks [26] and extend them with new datasets introduced by **HypotheSAEs** [34] and our work (EXPERIGEN; Table 18).

M Baseline Implementation Details & Hyperparameters

M.1 Inference Procedure

We evaluate all methods using the two-step LLM inference procedure of Zhou et al. [52]. Given a test observation and a hypothesis bank, inference proceeds in two steps:

1. **Selection.** The LLM is shown the observation and all hypotheses, and retrieves the top-k most applicable hypotheses (k=3).
2. **Prediction.** The LLM produces the final label conditioned only on the selected hypotheses and the observation, without access to the full bank.

This decomposition keeps prompts tractable for large hypothesis banks while still leveraging the full set during selection. We use GPT-4o for both steps across all methods and tasks to ensure a fair comparison.

M.2 HypoGeniC

- **Models:** We use Qwen3-32B and GPT-4o for hypothesis generation and GPT-4o for inference in our experiments. For image experiments, only GPT-4o was used.
- **Hyperparameters:**
  - Top-k selector: 10 hypotheses evaluated per example
  - Learning rate scale (α): 0.5
  - Temperature: 1×10^{-5}
  - Seed: 42
- **Inference:** We evaluate multiple strategies, including best-accuracy hypothesis, filter and weighted vote, one-step adaptive, and two-step adaptive. We use the best-performing strategy for a fair comparison.

Task	Description	Train	Test	
			ID	OOD
HypoBench				
Deception Detection	Distinguish genuine vs. deceptive hotel reviews via linguistic cues (e.g., exaggeration, vagueness, unnatural phrasing).	1,600	800	640
AI-Gen. Detection	Decide if a story (given a prompt) is human- or AI-written; tests stylistic and structural pattern recognition.	800	800	800
Persuasive Pairs	Given two arguments, predict the more persuasive based on rhetoric, logic, and emotional appeal.	750	750	500
Mental Stress	Detect mental stress in Reddit posts via affective and psychological cues.	1,000	1,000	500
News Engagement	Choose which of two headlines garners more clicks, testing linguistic framing and interest cues.	700	700	453
HypotheSAEs				
Congress	Predict U.S. congressional party affiliation (Republican/Democrat) from speeches (104th–109th Congress, 1995–2007). Tests large-scale scalability.	40k	5k	–
EXPERIGEN				
Tweets [44]	Predict which tweet in a pair is more persuasive, leveraging metadata (usernames, tags, etc.).	5k	1k	2k
Design [38]	Identify which of two graphic layouts was AI-generated; tests multimodal reasoning over structure and visuals.	1.6k	200	–
Memorability [23, 43]	Predict which image in a visually similar pair is more memorable (LaMem dataset). Tests fine-grained visual cues.	8k	1k	–
CMV [45]	Predict whether a counterargument successfully changes the original poster's view on Reddit's r/ChangeMyView.	4k	500	–

Table 18: Summary of Datasets used for our evaluation

M.3 HypotheSAEs

- **Embeddings:** OpenAI text-embedding-3-large for text-based tasks, CLIP-VIT-336 for Design Aesthetics, and fine-tuned CLIP-VIT-336 for Memorability.
- **Hyperparameters:**
  - **CONGRESS:** $(M, k) = (4096, 32)$
  - **News Headlines:** $(M, k) = (256, 8)$ and $(32, 4)$ (activations combined)
  - **Visual:** $(M, k) = (1024, 16)$
  - **Remaining:** For (M, k), we select the best configuration per environment from $(4096, 32)$, $(256, 8)$, and $(32, 4)$ (activations combined).
- **Neuron Interpretation:**
  - Top-activating examples: 10 per neuron
  - Low-activating examples: 10 per neuron
  - Candidate interpretations: 3 per neuron (select highest fidelity)
  - Fidelity evaluation: F1 score on 100 positive + 100 negative samples using GPT-4o-mini

M.4 Task Specific Expert Model Numbers

For reference, the task-specific expert models reported in prior work achieve the following accuracies on the same evaluation splits used in Table 3:

- **Twitter:** 80.9 / 77.3 ID / OOD, using a fine-tuned Vicuna-1.5 (13B) [43].
- **Design:** 87.2, following Patnaik et al. [38].
- **LaMem:** 75.0, using a fine-tuned 13B model [18].

These numbers are reported as task-specific reference points; they are not directly comparable to EXPERIGEN's results in Table 3, since the expert models are fine-tuned on each task whereas EXPERIGEN is a zero-shot, hypothesis-driven method.

N Limitations

Restriction to observational, statistically evaluable settings. EXPERIGEN targets data-driven sciences where the discovery signal lives in observational data and validity is established through statistical tests. The framework does not directly apply to theory-driven sciences (for example, theoretical physics) where there is no observational dataset to mine, nor to tasks whose evaluation is non-statistical (for example, protein folding accuracy or molecular synthesis yield). Adapting EXPERIGEN to such settings would require domain-specific evaluation routines beyond the Code Interpreter's current scope.

No mechanism for data acquisition. When the available sample is too small to test a hypothesis at the desired confidence level, the current framework can only refine, abandon, or replace the hypothesis; it cannot decide to collect additional data.

Observational analysis only; no autonomous interventions. EXPERIGEN analyzes the dataset it is given. It does not design or execute interventional experiments such as randomized controlled trials, factorial designs, or power-driven sample-size planning.

Novelty is bank-relative, not literature-relative. The system does not query external scientific knowledge bases (for example, PubMed, Google Scholar) to verify that a discovered hypothesis is novel with respect to the published literature. Literature-novelty is established only at the expert-review stage (Appendix B), not during generation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction reflect the claims made in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in N.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: There are no theoretical results in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The algorithm (Generator, Experimenter, inner/outer loops, bank-coreset rule, hierarchical Bonferroni–BH correction) is specified in Sections 3.2, 3.3 and 3.4.1 to 3.4.2 and Appendix A.5. Hyperparameters (N=30, T=4, K=25, α=0.05, q=0.05, embedding model `text-embedding-3-large`), Generator/Experimenter/inference model identities, full prompts (Appendix H.1), datasets (Table 4 and Appendices C and L), baseline configurations (Appendices M.1 to M.3), sandbox details (Appendix K), evaluation protocol and A/B-test design (Appendix E), and the 16 pre-registered CMV hypotheses with locked predictions (Table 15) are all documented.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Instructions to replicate the pipeline are provided in the main paper and supplementary material. All datasets used are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Hyperparameters (N=30, T=4, K=25, α=0.05, q=0.05, embedding model `text-embedding-3-large`), Generator/Experimenter/inference model identities, full prompts (Appendix H.1), datasets (Table 4 and Appendices C and L), baseline configurations (Appendices M.1 to M.3), sandbox details (Appendix K), evaluation protocol and A/B-test design (Appendix E) are documented.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All hypothesis-level claims are accompanied by significance tests: Bonferroni-corrected p-values over the confirmatory tests within each inner refinement loop and Benjamini–Hochberg q=0.05 across the resulting hypotheses (Section 3.4.2); significance

counts in Table 1 use held-out data, Bonferroni-corrected across each method's hypotheses. For each CMV hypothesis we also report the statistical test, verdict, and effect size (e.g., odds ratios, Cramér's V, Cohen's d) (Appendix A.5, Table 9).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the hypothesis-evaluation budget per run (up to $N \times T = 120$ steps at N=30, T=4) and per-task dollar costs across dataset sizes (Figure 16).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our work abides by the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work automates hypothesis generation and experimental validation for data-driven sciences, with potential to accelerate discovery in fields such as behavioral science, psychology, healthcare, and marketing. By enforcing statistical rigor through hierarchical multiple-testing correction and covariate control, EXPERIGEN produces hypotheses that meet the standards of domain experts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our algorithm poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all datasets and models used in the paper and abide by their respective copyrights and licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We release no new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Full participant instructions, key-terms glossary, statistics primer, and rating rubric are reproduced verbatim as tcolorbox blocks in Appendix B.3; example screens of the rated hypotheses are shown in Appendix B.3.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: We received IRB approval for our expert study; details are in Appendix B.2.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The Generator, Experimenter and Feature Extractor are all LLM agents; their roles, prompts, and model identities are specified in detail in the main paper and supplementary material.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.