NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026 · Spotlight

ExperiGen

Agentic Hypothesis Discovery from Observational Data

linktr.ee/behaviorinthewild

Jishu Sengupta[1*] Harini S I[1*] **Somesh Singh**[1,2,3*] Syed Mohamad Tawseeq[1*]

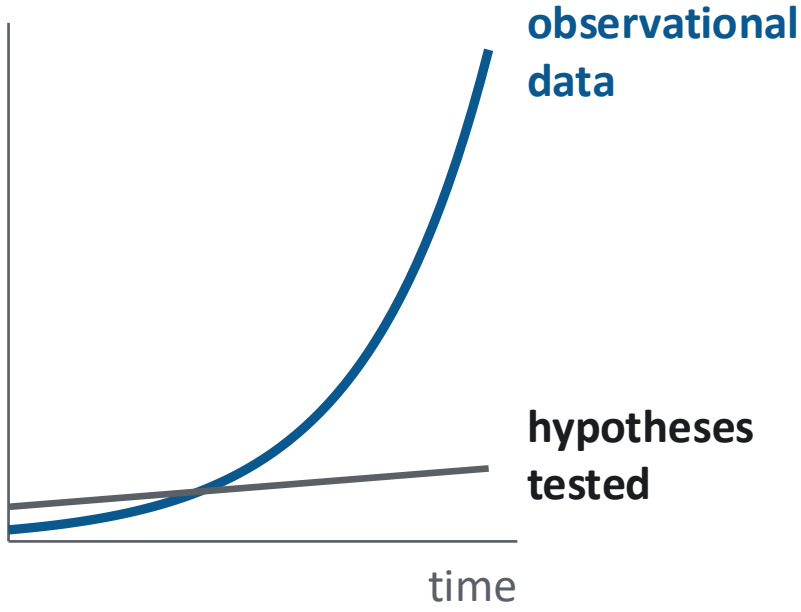
Yaman K Singla[1] Rajiv Ratn Shah[4] David Doermann[3] Jitendra Ajmera[1] Balaji Krishnamurthy[5]

[1]Adobe MDSR [2]IIIT-Delhi [3]SUNY at Buffalo [4]IIT Kanpur [5]Machlear Technologies [*]Equal contribution

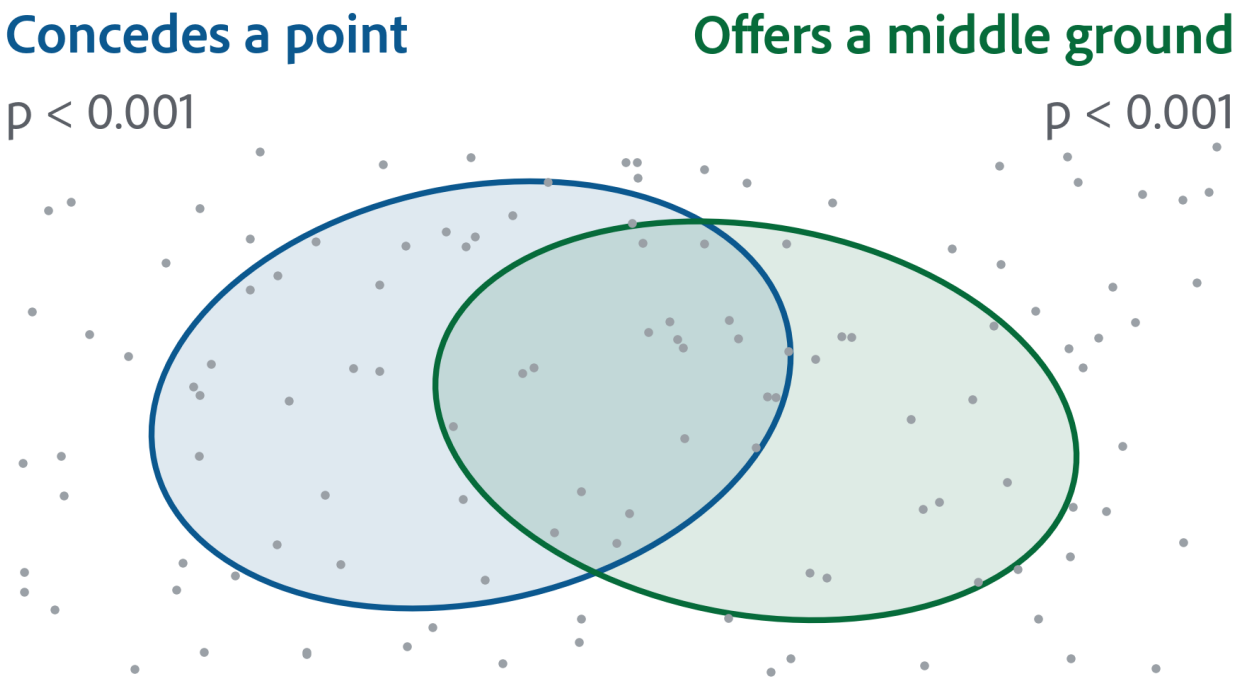
Data keeps growing. Hypotheses are still written by hand.

Can an agent propose **novel, significant hypotheses** from observational data?

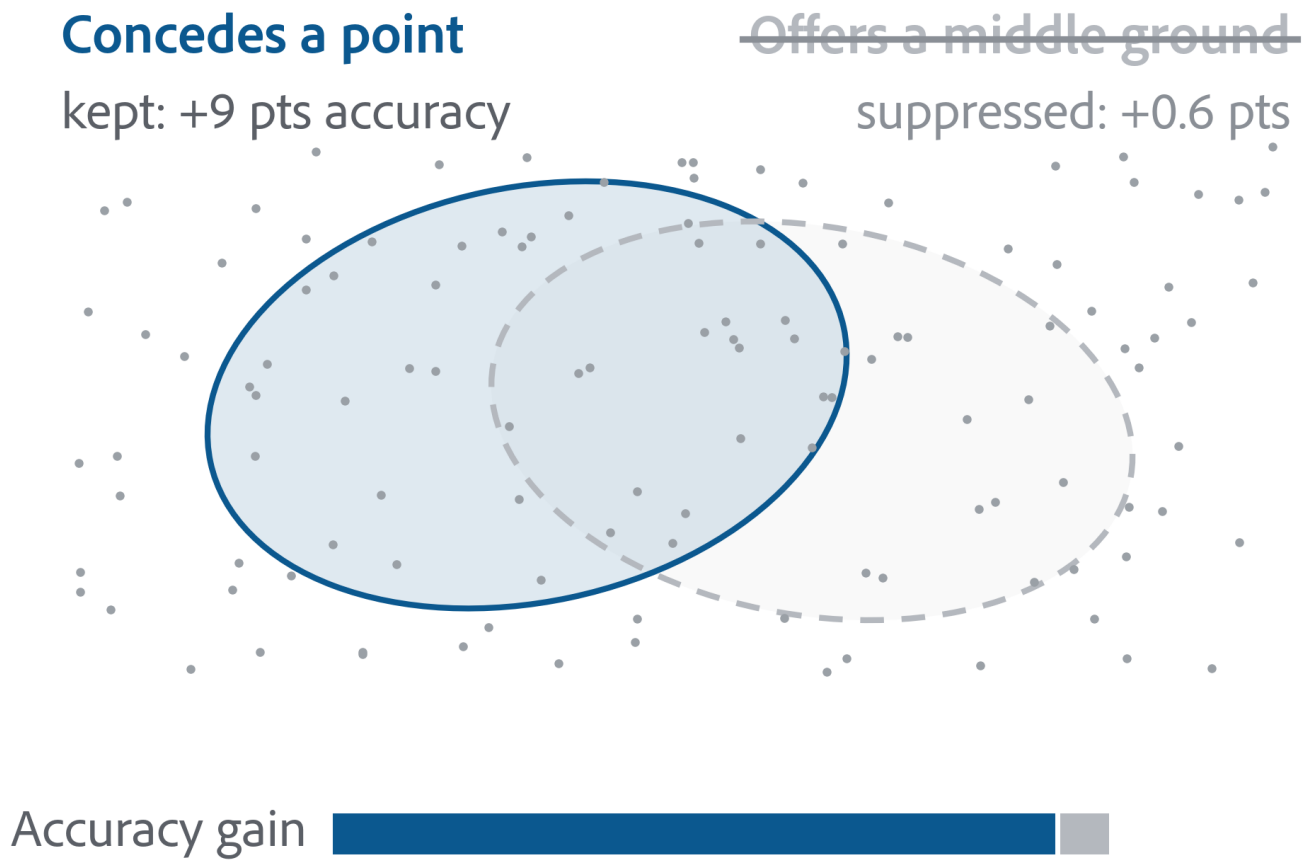
So that experiments go to the ideas most worth running.



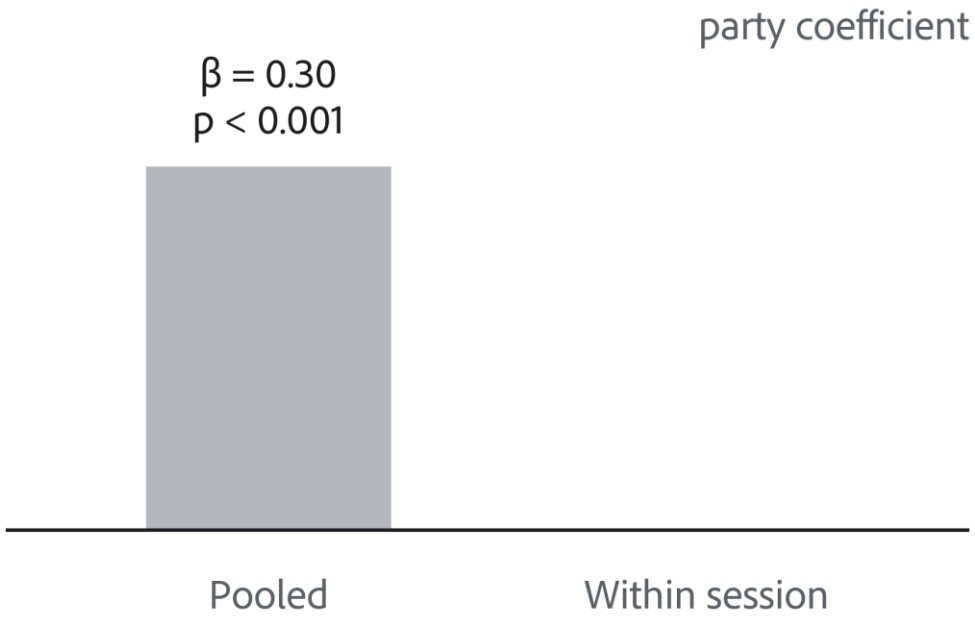
Optimizing accuracy suppresses overlapping hypotheses



Optimizing accuracy suppresses overlapping hypotheses



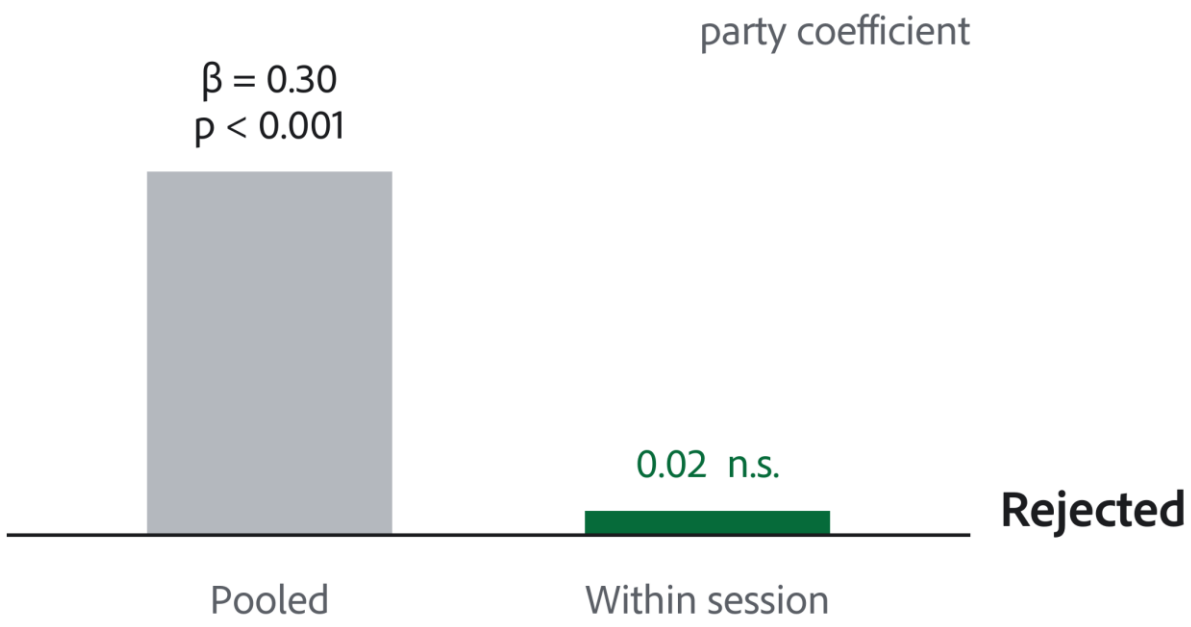
A single test reports confounders as findings



“I ask unanimous consent” → Republican speaker

Current methods: $p < 0.001$, reported as a finding

A single test reports confounders as findings



“I ask unanimous consent” → Republican speaker

Current methods: $p < 0.001$, reported as a finding

Reads the dataset description

A procedural phrase, used by the majority party

Control for session: the effect vanishes

Rejected

One agent proposes hypotheses, another tests them

Generator

Proposes hypotheses

Reads the data and the memory of past tests

Proposes ideas that are plausible and new

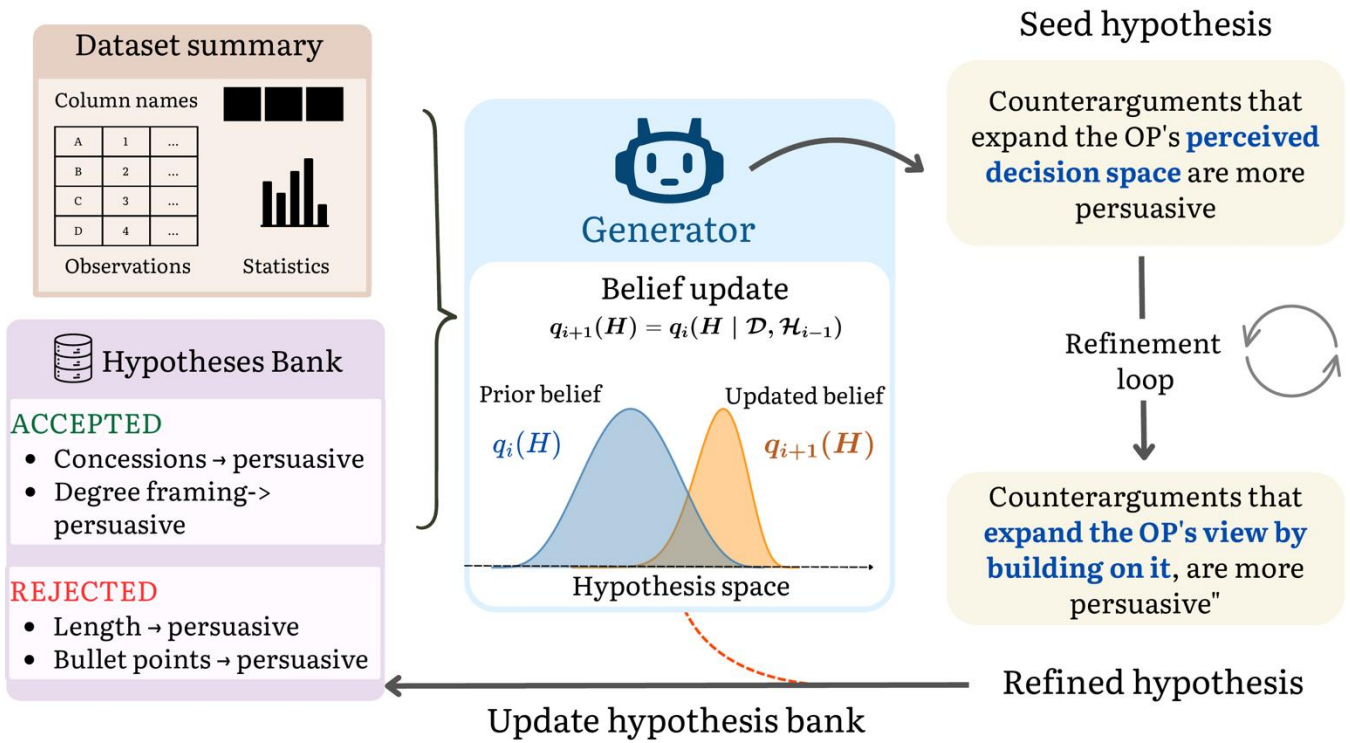
Experimenter

Tests them, and tries to break them

Turns each hypothesis into measurable features

Looks for confounders before accepting

The Generator learns from every test it has run



Beliefs as a prior

What it believes about the data guides each proposal

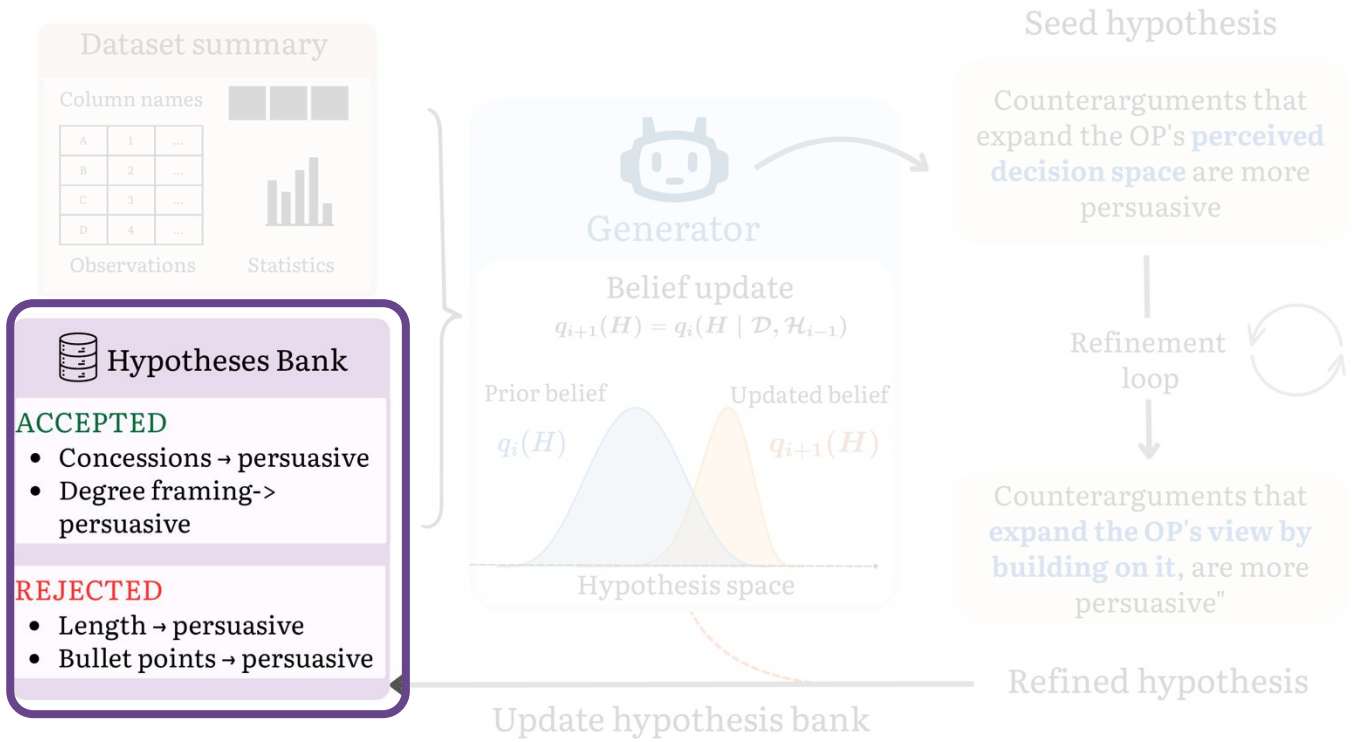
Plausible and new

Each proposal differs from what it already tried

Memory

Accepted and rejected hypotheses, with their evidence

The Generator learns from every test it has run



Beliefs as a prior

What it believes about the data guides each proposal

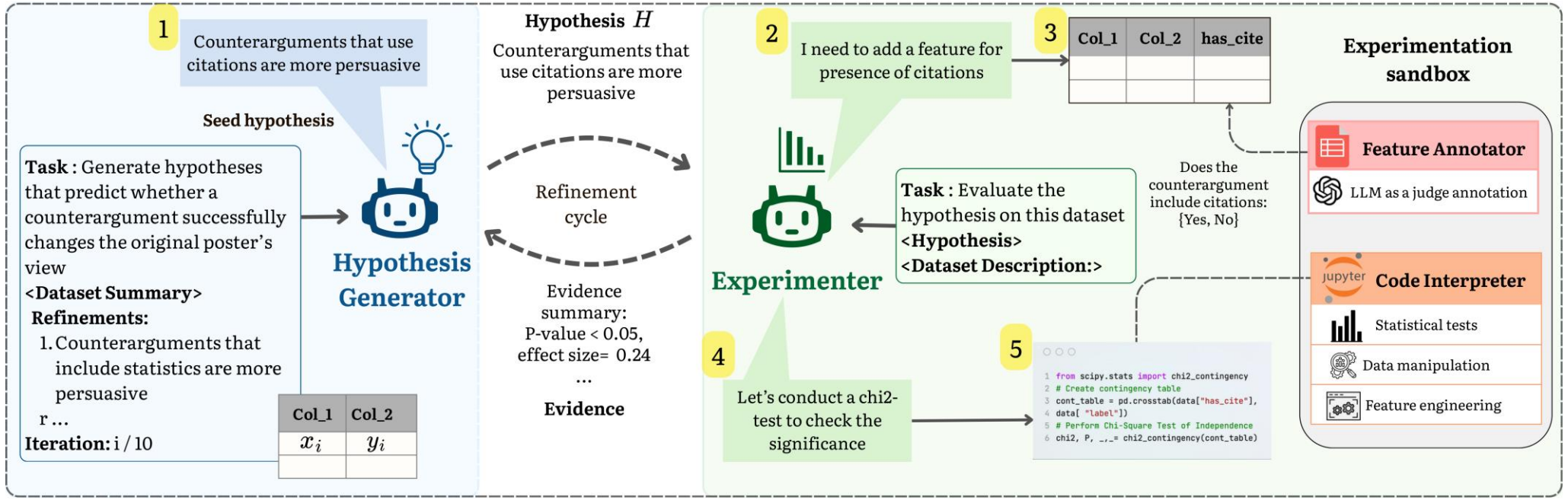
Plausible and new

Each proposal differs from what it already tried

Memory

Accepted and rejected hypotheses, with their evidence

The Experimenter tries to falsify each hypothesis



Measure

Code for simple features, a small LLM for semantic ones

Test

Run the statistical test on the data

Falsify

Hunt for confounders; refine until it holds or fails

Significant hypotheses across 19 tasks in 12 domains

1.5–3×
significant hypotheses vs. the strongest baseline

< 5% vs. 20–25%
false discovery rate on manual review

	Decep.	News	Dread.	Pers.	Twitter	Cong.	CMV	CRC	DB	BAIS
ExperiGen	**18**	**21**	**16**	**17**	**19**	**22**	**16**	**12**	**81**	**17**
HypotheSAEs	8	12	6	7	12	15	8	3	20	–
HypoGeniC	6	3	4	5	3	5	4	4	40	–

Significant hypotheses on held-out data ($p < 0.05$, Bonferroni-corrected per method). DB: DiscoveryBench, 4 tasks. BAIS: BAISBench single-cell RNA-seq, 4 datasets.

Better predictions, without optimizing for accuracy

+5.0 pts
over the best baseline on HypoBench

9 / 10
HypoBench splits won

	Decep.	News	Dread.	GPTgc	Pers.	Twitter	Cong.	CMV
ExperiGen	**85.2**	**66.5**	**80.8**	**85.0**	**91.2**	**66.1**	**79.4**	**77.5**
HypoGeniC	77.7	62.5	67.2	84.0	89.0	61.9	72.6	61.0
HypotheSAEs	60.6	61.4	62.3	79.8	84.0	56.0	73.9	59.2

Out-of-distribution accuracy (%) with GPT-4o as the inference model; Congress and CMV are test accuracy. Full in- and out-of-distribution results in the paper.

Validation in real-world settings

26

domain experts

88% rated the hypotheses novel, 76% worth pursuing as research

11 / 16

pre-registered hypotheses held up

Significant on new data, all in the predicted direction

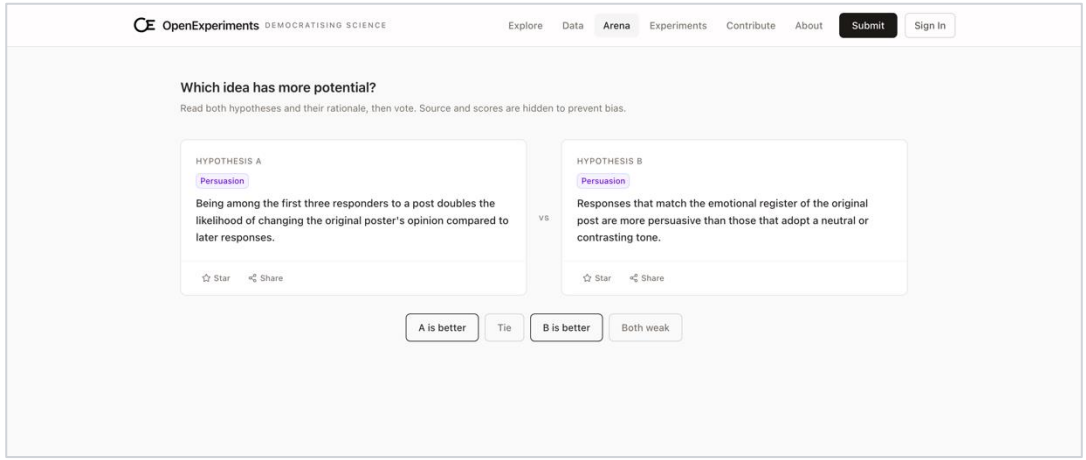
5.1M

users in 25 A/B tests

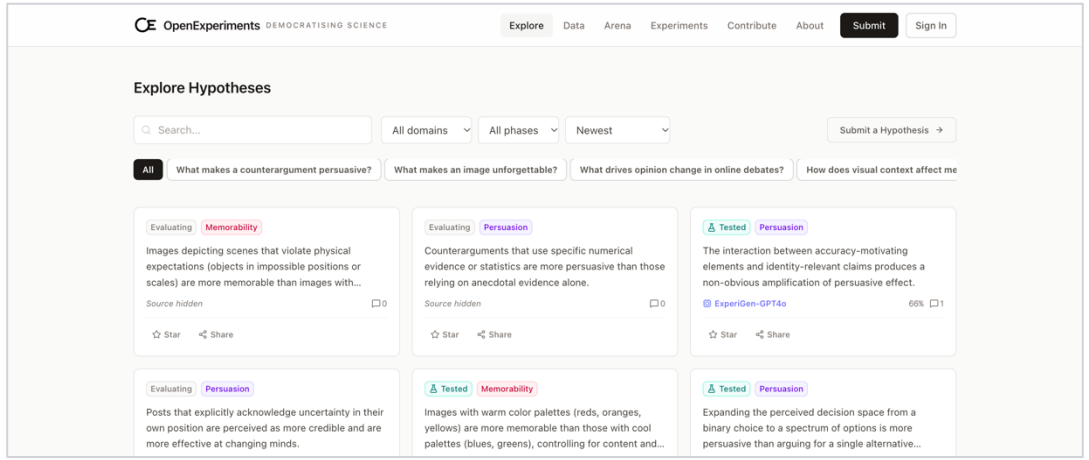
7 S&P 500 companies, 5 months: 15 significant, median uplift +112%

OpenExperiments: democratising science

An open platform where anyone, humans and AI alike, can submit ideas, have the community vote on them, and see experiments conducted in real time.



Arena: compare hypotheses head to head



Explore: hypotheses from humans and agents

We are releasing all our pre-registered experiments and A/B tests here.

openexperiments.pages.dev

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026 · Spotlight

Thank you

ExperiGen: Agentic Hypothesis Discovery from Observational Data

Team	Behavior in the Wild, Adobe
Webpage	behavior-in-the-wild.github.io/experigen.html
Paper	arxiv.org/abs/2602.07983
OpenExperiments	openexperiments.pages.dev
Contact	behavior-in-the-wild@googlegroups.com

linktr.ee/behaviorinthewild