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*]

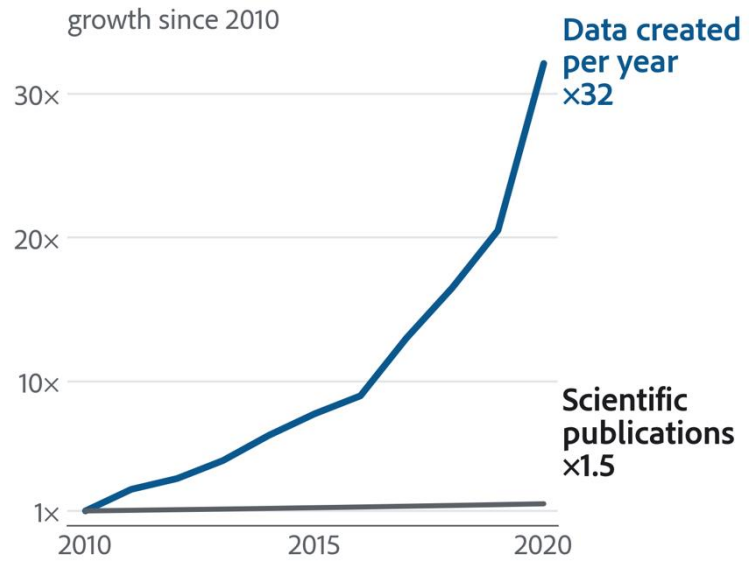
Yaman K Singla[1] Rajiv Ratn Shah[4] David Doermann[3] Jitendra Ajmera[1] Balaji Krishnamurthy[5]

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

[*]Equal contribution, order chosen by roll of dice

The scale of observational datasets is accelerating, however hypotheses about what explains them are bottlenecked

How can we propose **novel hypotheses that are likely to be successful under real world experimentation** using observational datasets?



Industries run over 20k A/B tests per year, yet only about a third of ideas bring a significant uplift [3].

Since 2010, the data created each year grew 32×, while scientific papers, a rough proxy for hypotheses tested, grew only 1.5× [1, 2].

[1] D. Reinsel, J. Gantz, and J. Rydning. The Digitization of the World: From Edge to Core. IDC White Paper, 2018; IDC, Worldwide Global DataSphere Forecast, 2021–2025, 2021.
[2] L. Bornmann, R. Haunschild, and R. Mutz. Growth rates of modern science: a latent piecewise growth curve approach. Humanities and Social Sciences Communications, 8(1):224, 2021.
[3] R. Kohavi, D. Tang, and Y. Xu. Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press, 2020.

Problem Statement: given a dataset of observations and their outcomes, propose hypotheses that explain their relationships

Input: an observational dataset

r/ChangeMyView: users post an opinion, others reply to change it, and the poster marks the replies that did.

reply		changed view
1	*"Fair point on cost, I hadn't considered that. But…"*	✓
2	*"That's simply wrong, and here is why."*	✗
3	*"You're right on cost, so why not a phased rollout?"*	✓
4	*"What if it stayed, but became optional?"*	✓

Output: hypotheses we can test

Replies that acknowledge a point the poster made are more likely to change their view

supported by replies 1, 3

Replies that propose a compromise are more likely to change their view

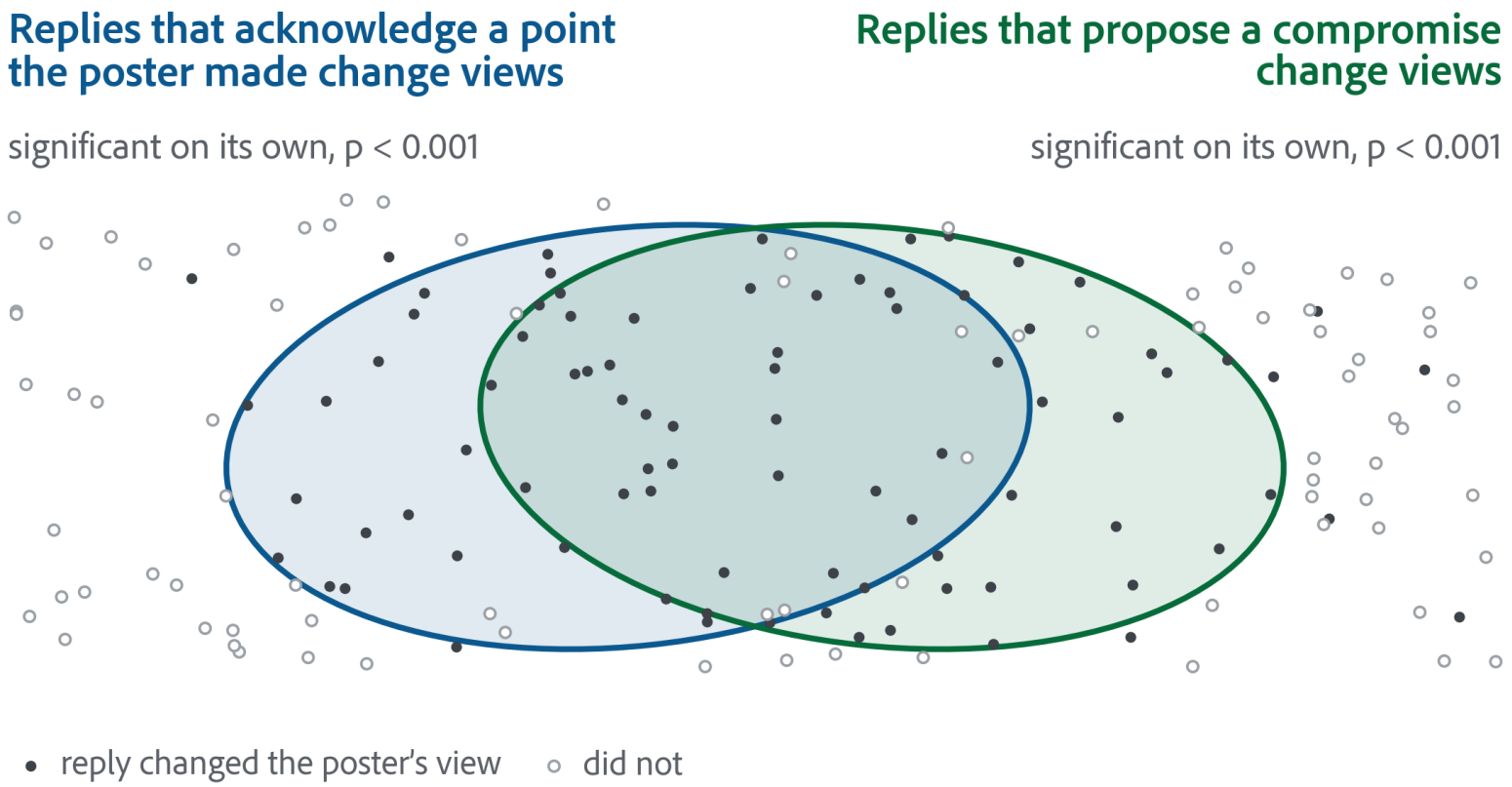
supported by replies 3, 4

Current methods have two shortcomings

Optimization pressure

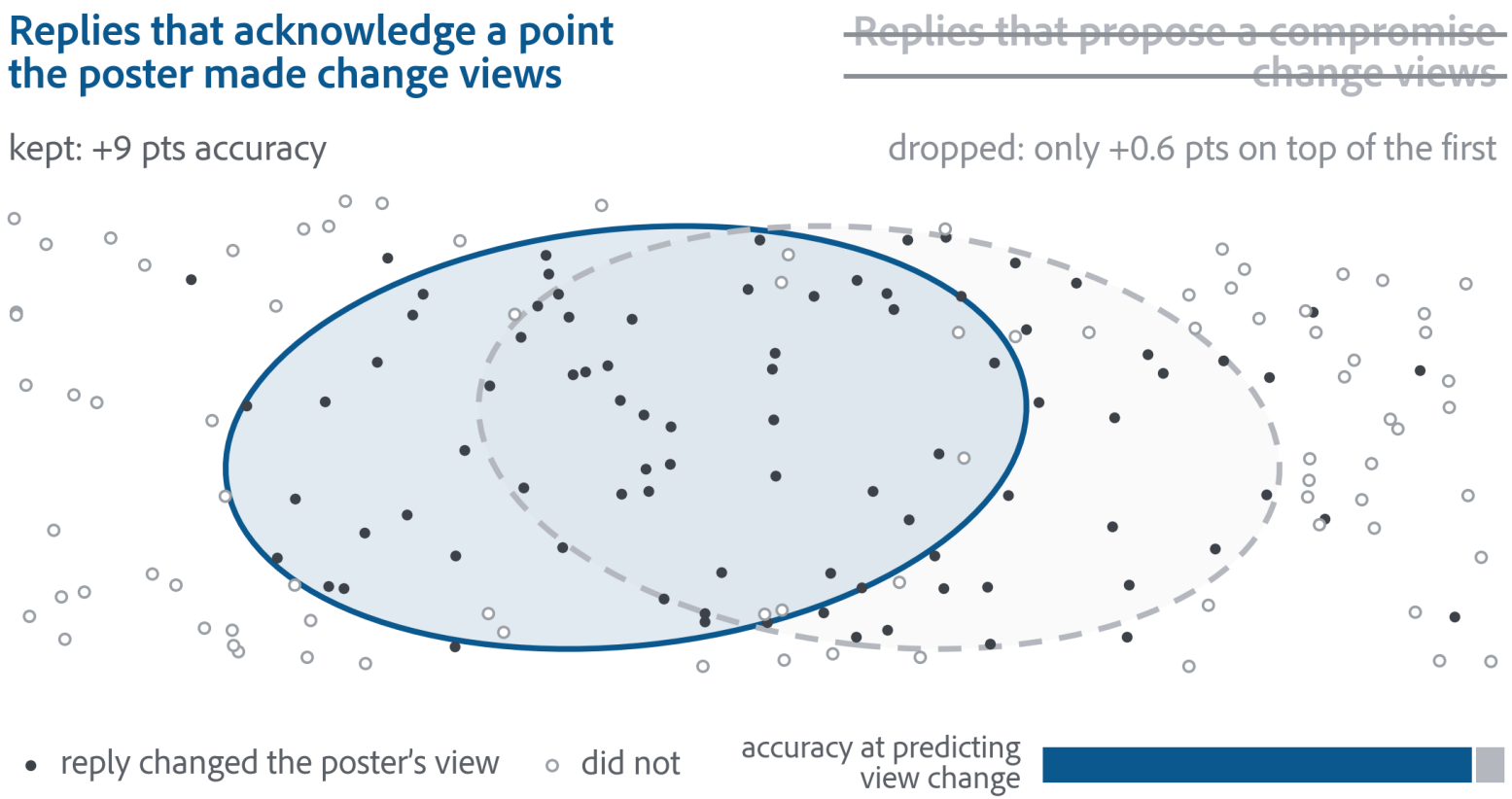
Current methods use the predictive accuracy of the generated hypotheses as the reward.

Optimization pressure: current methods fail to propose hypotheses that have overlapping support



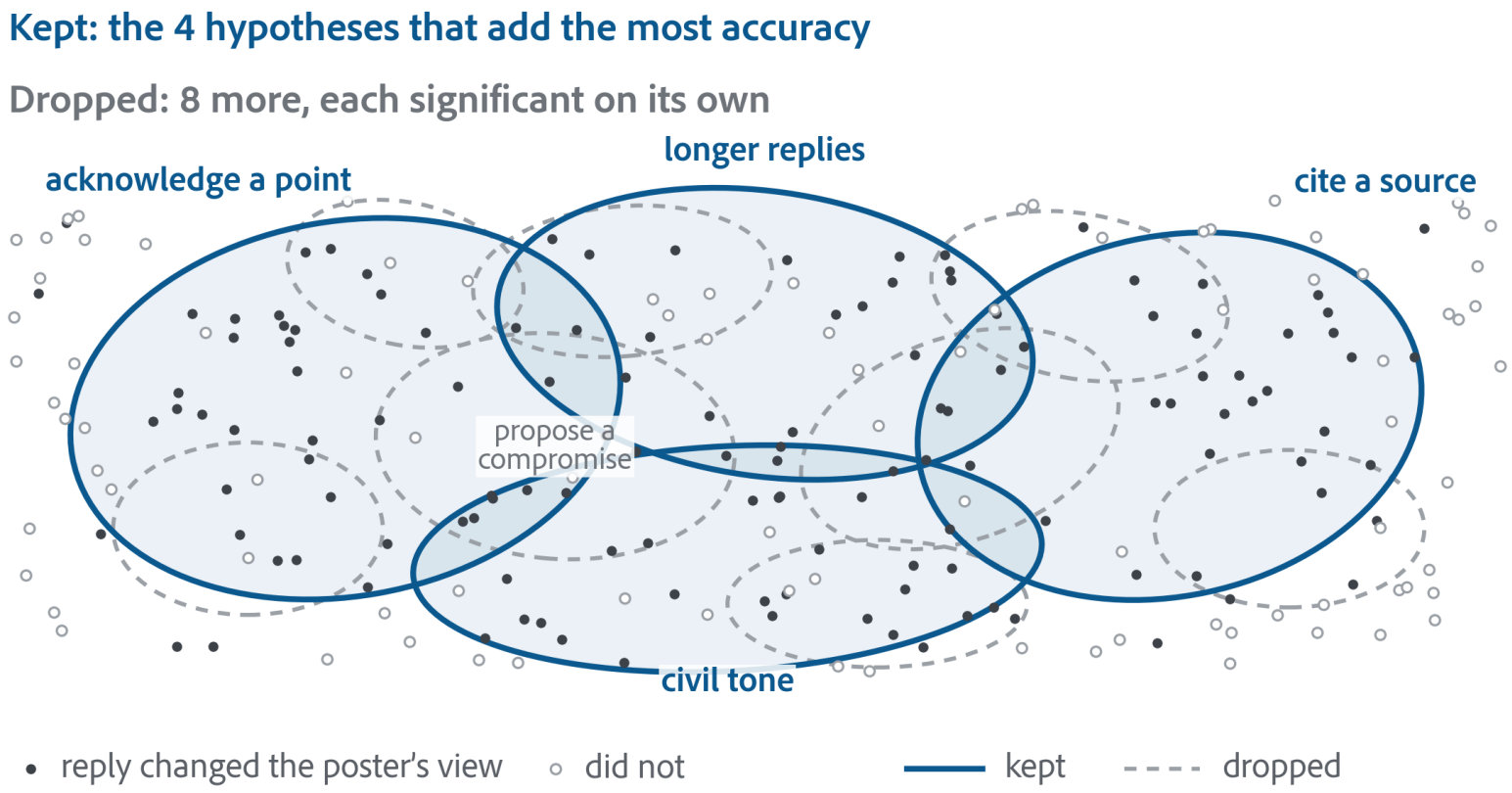
Each hypothesis stays significant ($p < 0.001$) in a joint logistic regression that also includes the other hypothesis and their interaction.

Optimization pressure: current methods fail to propose hypotheses that have overlapping support



Each hypothesis stays significant ($p < 0.001$) in a joint logistic regression that also includes the other hypothesis and their interaction.

Optimization pressure: current methods fail to propose hypotheses that have overlapping support



Each hypothesis stays significant ($p < 0.001$) in a joint logistic regression that also includes the other hypothesis and their interaction.

Current methods have two shortcomings

Optimization pressure

Current methods use the predictive accuracy of the generated hypotheses as the reward.

Static experimentation

Current methods test each hypothesis once with a fixed procedure, without checking how the data was collected, testing for unobserved confounders, or refining the hypothesis.

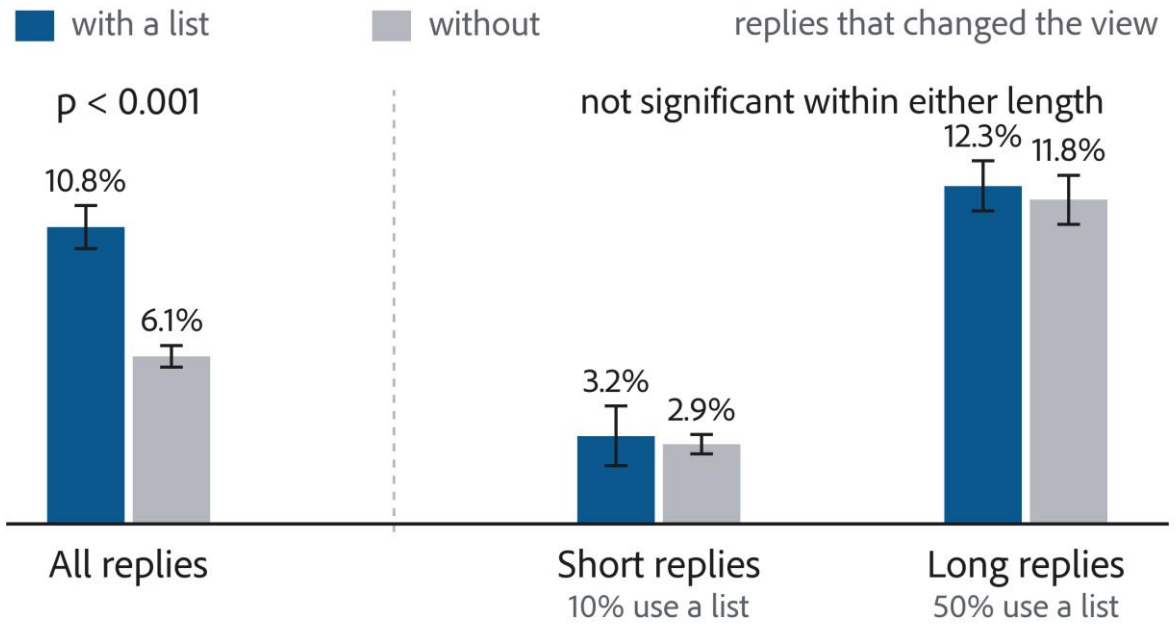
Static experimentation: current methods fall short of social science standards, such as testing for unobserved confounders



“Replies that use a bullet-point list change more views”

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

Static experimentation: current methods fall short of social science standards, such as testing for unobserved confounders



“Replies that use a bullet-point list change more views”

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

Lists appear mostly in long replies

Long replies change more views, with or without a list

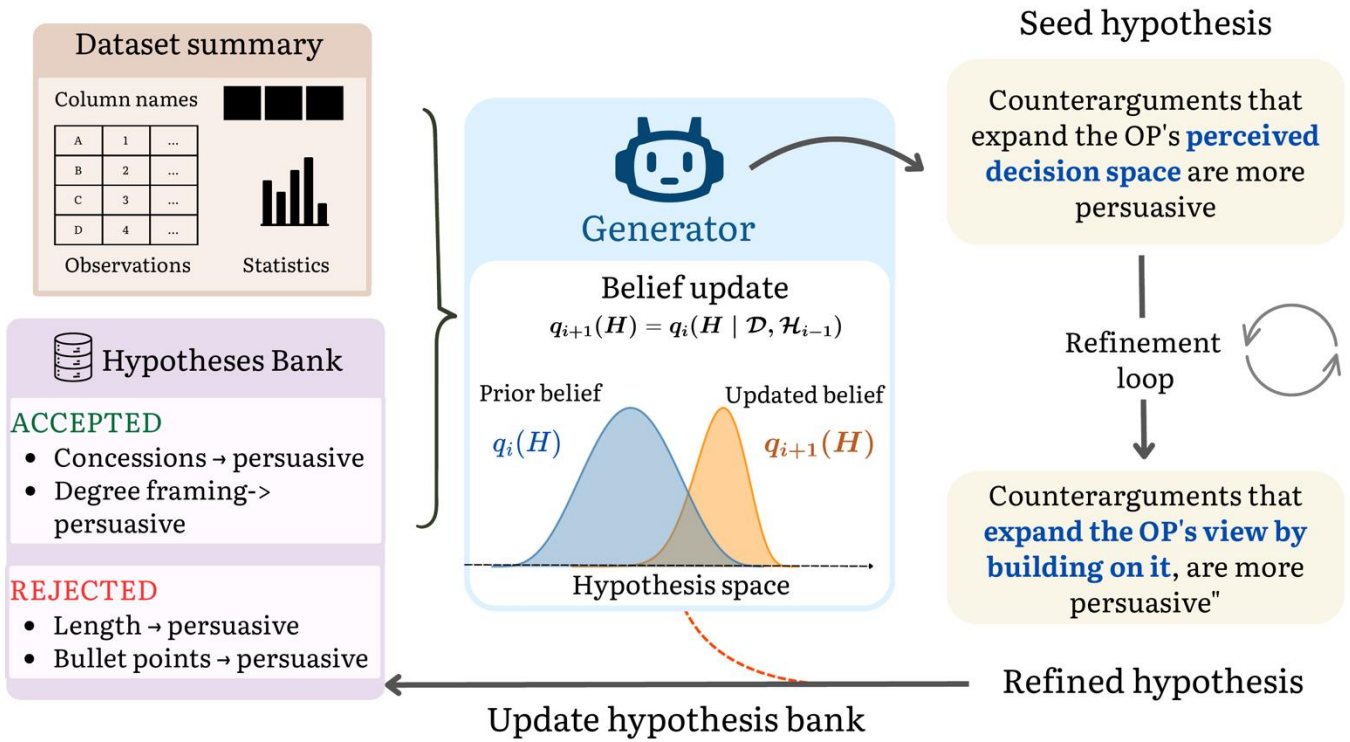
Controls for reply length: the effect vanishes

Rejected

ExperiGen addresses each problem with a multi agent framework

PROBLEM	EXPERIGEN
Accuracy as the objective drops overlapping hypotheses	→ **Generator** Searches for hypotheses that are novel and significant, inspired by Bayesian optimization
A single fixed test reports confounders as findings	→ **Experimenter** Experiments iteratively: controls for confounders, then refines or rejects each hypothesis

The Generator updates its beliefs about the data after every hypothesis it tests



Beliefs as a prior

What it believes about the data guides each proposal

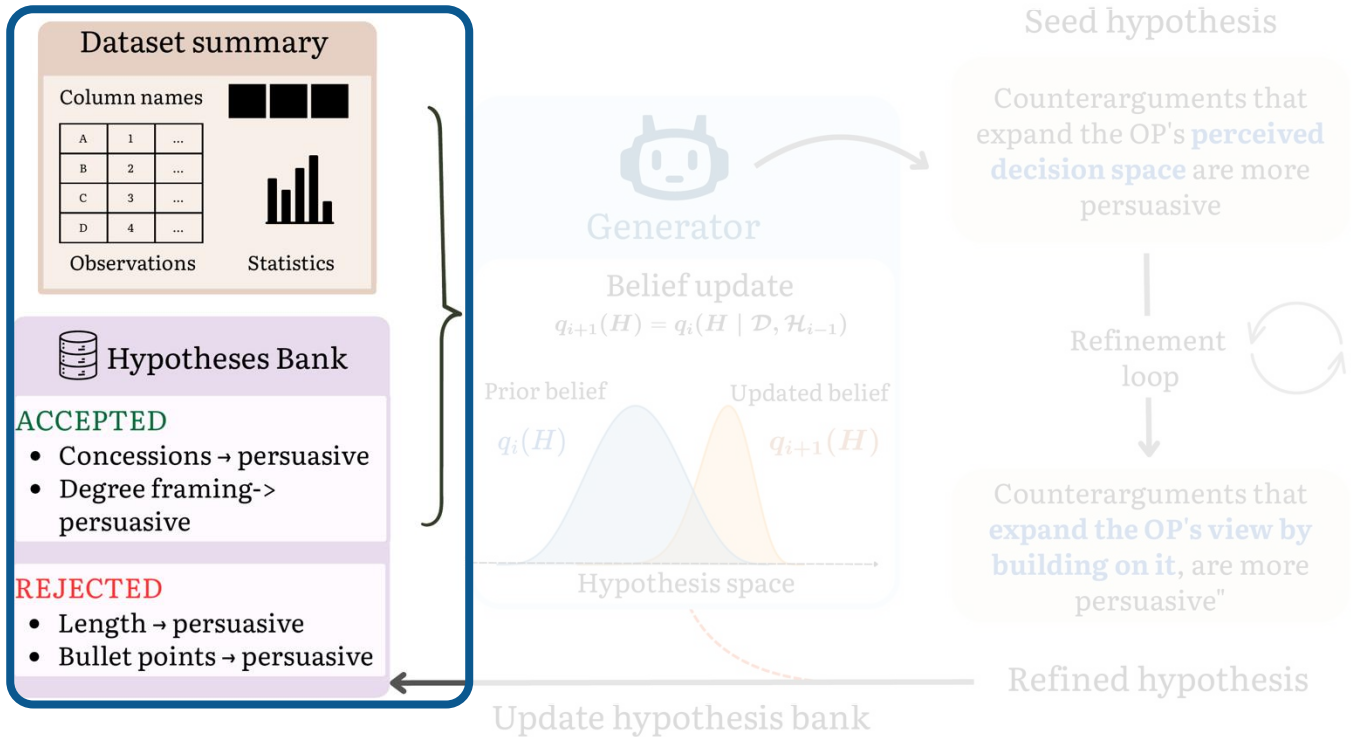
Plausible and new

Each proposal differs from what it already tried

Memory

Accepted and rejected hypotheses, with their evidence

The Generator updates its beliefs about the data after every hypothesis it tests



STEP 1 · INPUTS

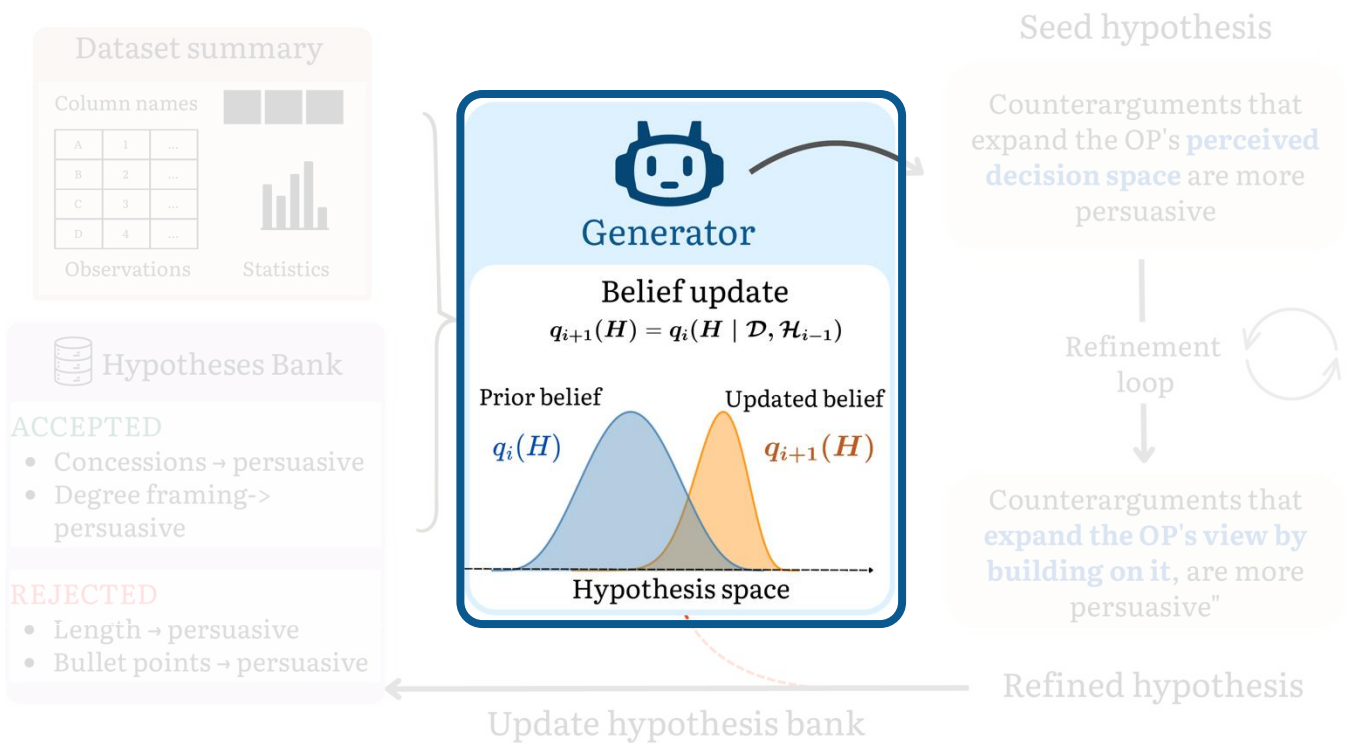
Generator inputs

At every outer iteration, the Generator is given two inputs:

Dataset summary: column names, a few example rows and summary statistics of the data.

Hypothesis bank: the hypotheses tested so far (at most 30), each labeled ACCEPTED or REJECTED by the Experimenter.

The Generator updates its beliefs about the data after every hypothesis it tests



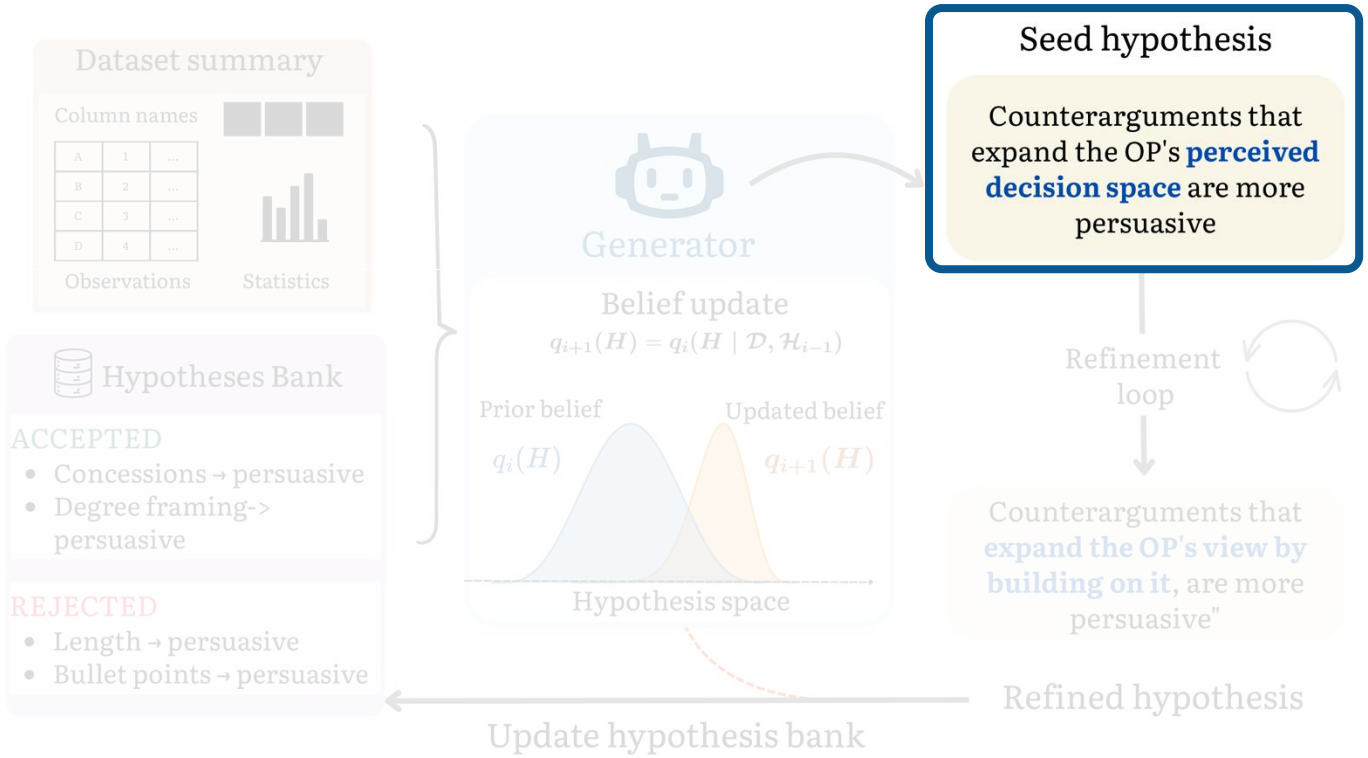
STEP 2 · BELIEF UPDATE

Updating beliefs from the bank

Blue: the LLM's prior belief about which hypotheses are likely to hold, before seeing any results.

Orange: the belief after reading the bank. It moves toward hypotheses similar to ACCEPTED ones and away from REJECTED ones.

The Generator updates its beliefs about the data after every hypothesis it tests



STEP 3 · SEED HYPOTHESIS

Proposing a seed hypothesis

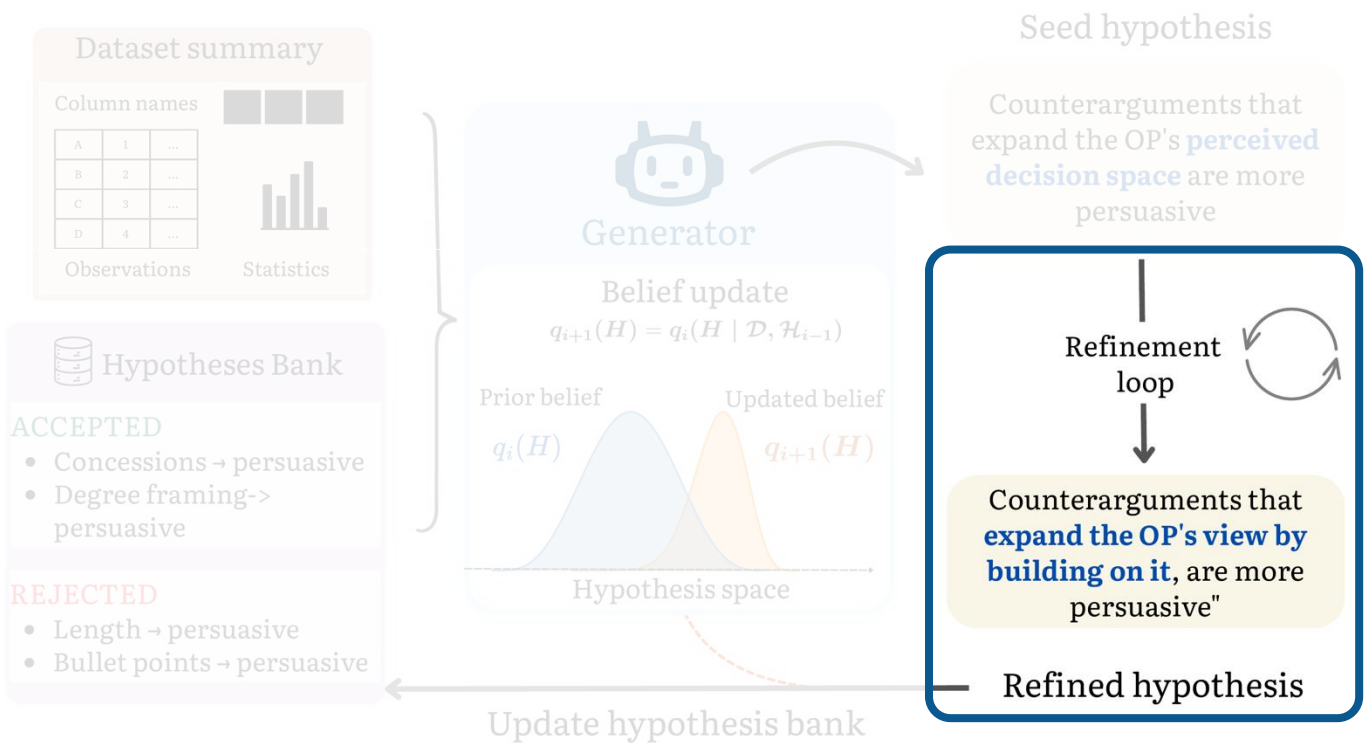
Using the updated belief, the Generator proposes a new hypothesis (the seed) that is:

Plausible: likely to be supported, given past test results.

Novel: different from every hypothesis already in the bank.

Example: "Counterarguments that expand the OP's perceived decision space are more persuasive."

The Generator updates its beliefs about the data after every hypothesis it tests



STEP 4 · REFINEMENT LOOP

Refining the seed

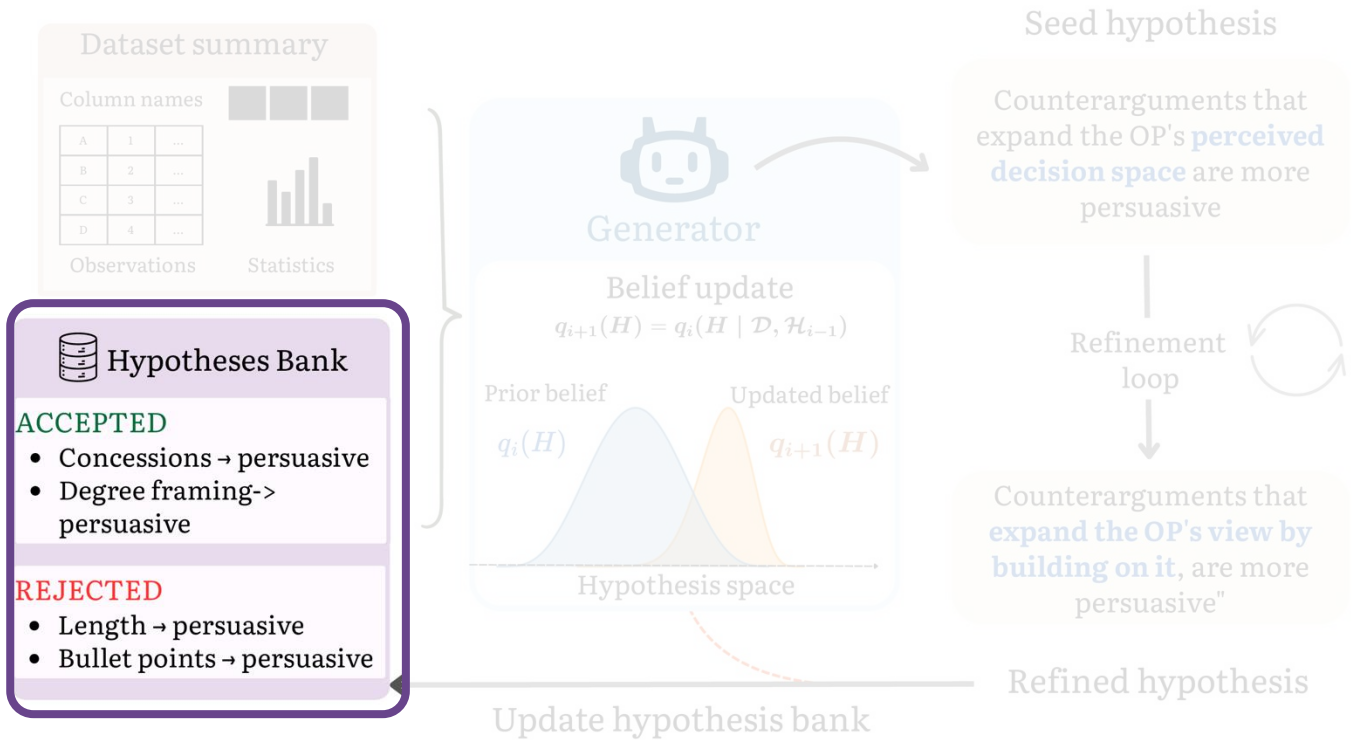
The Experimenter tests the seed on the data.

Using the results, the Generator revises it, e.g. adds a condition or controls for a confound. This repeats for up to T rounds.

The final version is added to the bank with its label.

Refined: "…that expand the OP's view by building on it, are more persuasive."

The Generator updates its beliefs about the data after every hypothesis it tests



Beliefs as a prior

What it believes about the data guides each proposal

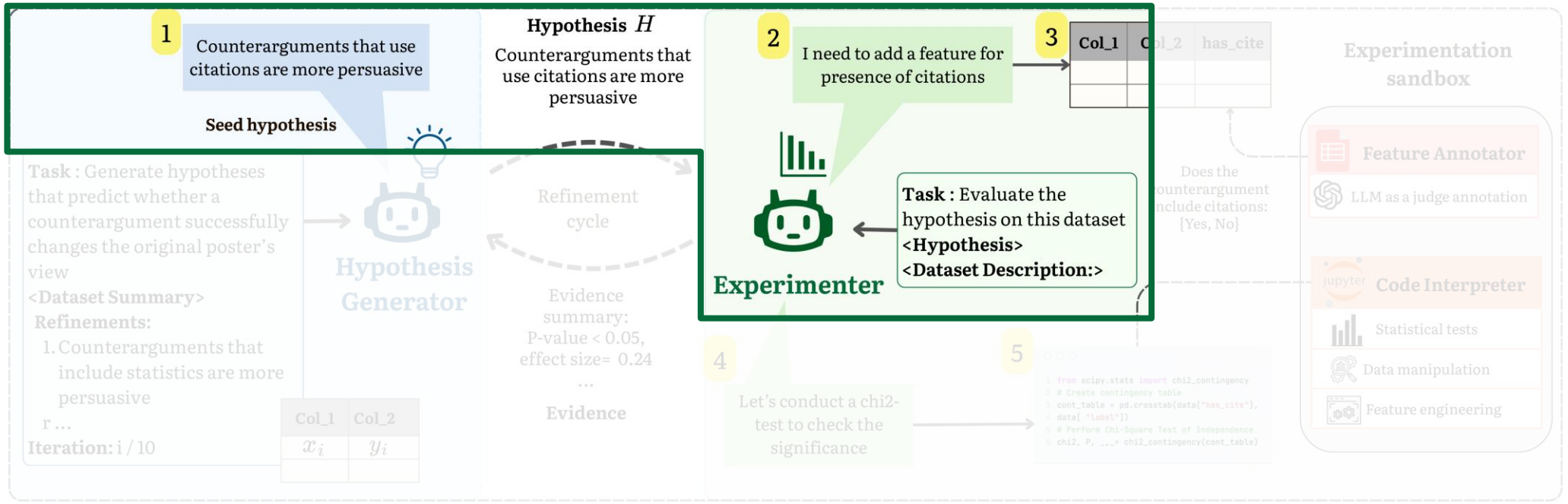
Plausible and new

Each proposal differs from what it already tried

Memory

Accepted and rejected hypotheses, with their evidence

The Experimenter measures each hypothesis in the data, tests it, and tries to falsify it



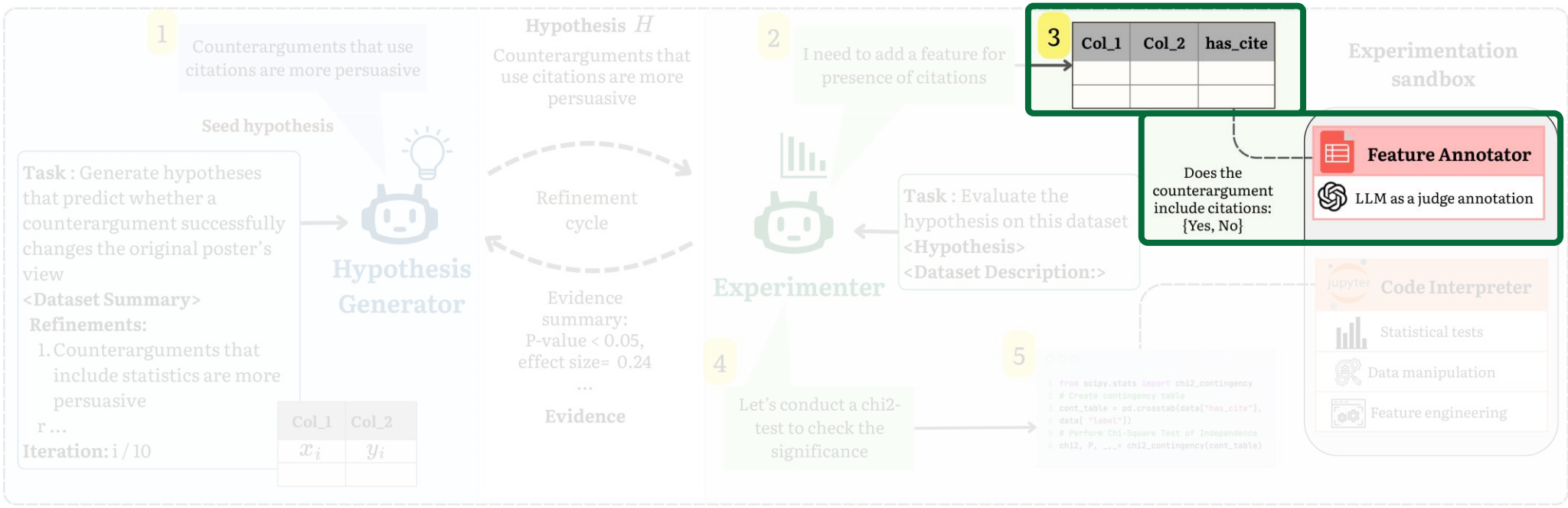
STEPS 1–2 · HYPOTHESIS AND PLAN

The Experimenter receives a seed hypothesis and plans how to test it

Inputs: a seed hypothesis, e.g. "Counterarguments that use citations are more persuasive", and the dataset. The Experimenter (a ReAct agent) decides which features to measure, which covariates to control for and which test to use.

Here it needs a feature that does not exist yet: whether each counterargument contains a citation.

The Experimenter measures each hypothesis in the data, tests it, and tries to falsify it



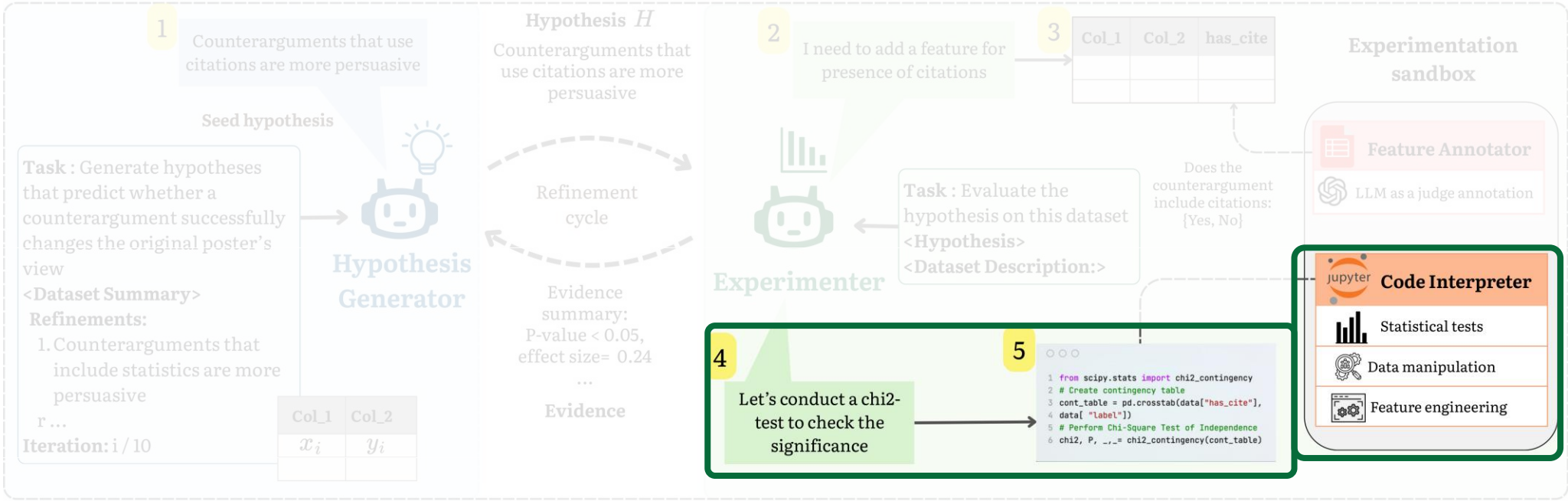
STEP 3 · MEASURE

Computing the feature

The new feature is added as a column (has_cite: Yes/No) for every counterargument.

Semantic features like this are labeled by an LLM annotator; simple features (length, word counts, regex matches) are computed with code.

The Experimenter measures each hypothesis in the data, tests it, and tries to falsify it



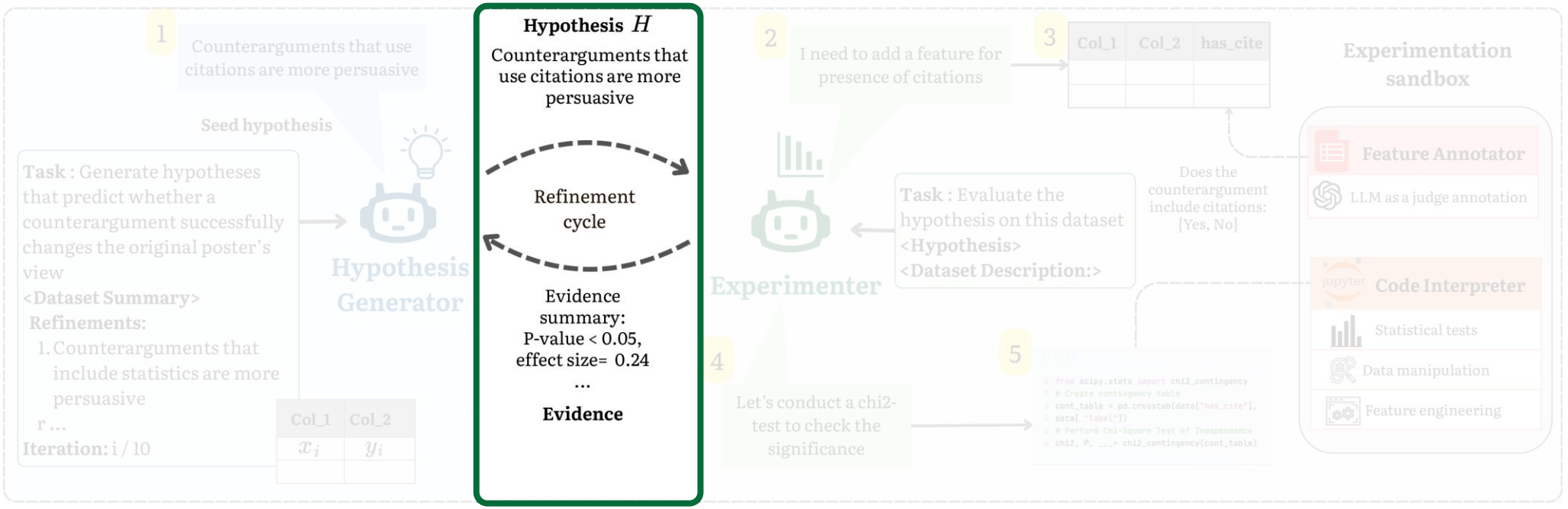
STEP 4 · TEST

Running the statistical test

The Experimenter writes and runs analysis code in a sandboxed code interpreter.

Here it runs a chi-squared test of whether has_cite is associated with the outcome (whether the counterargument changed the original poster's view).

The Experimenter measures each hypothesis in the data, tests it, and tries to falsify it



STEP 5 · EVIDENCE AND REFINEMENT

Returning evidence to the Generator

Output: an evidence summary: p-value, effect size and suggestions for refinement (e.g. a confound to control for).

The Generator revises the hypothesis; this repeats for up to T rounds. Accepted only if $p < \alpha/T$ (Bonferroni).

We present the first case of autonomously generated hypotheses that succeeded in real world A/B tests and pre-registered studies

For example

“Counterarguments explaining technical terms with context or analogies are more persuasive than those merely listing terms.”

15/25

A/B tests with 5.1M users

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

11 / 16

pre-registered hypotheses replicated

Registered January 2026, tested on new data, all in the predicted direction

To our knowledge

More representative hypotheses that ExperiGen accepted and rejected on ChangeMyView

Generated hypothesis	p-value	Effect size
First-comment position provides a persuasion advantage independent of timing.	0.0093	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 = 1.136 per unit
Labeling the OP's reasoning a logical fallacy, especially in the opening, reduces persuasion success.	0.316	Cramér's V = 0.01

Green: accepted. Red: rejected, not significant. OR: odds ratio.

Across 19 tasks in 12 domains, ExperiGen finds more significant hypotheses with fewer false discoveries

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	–

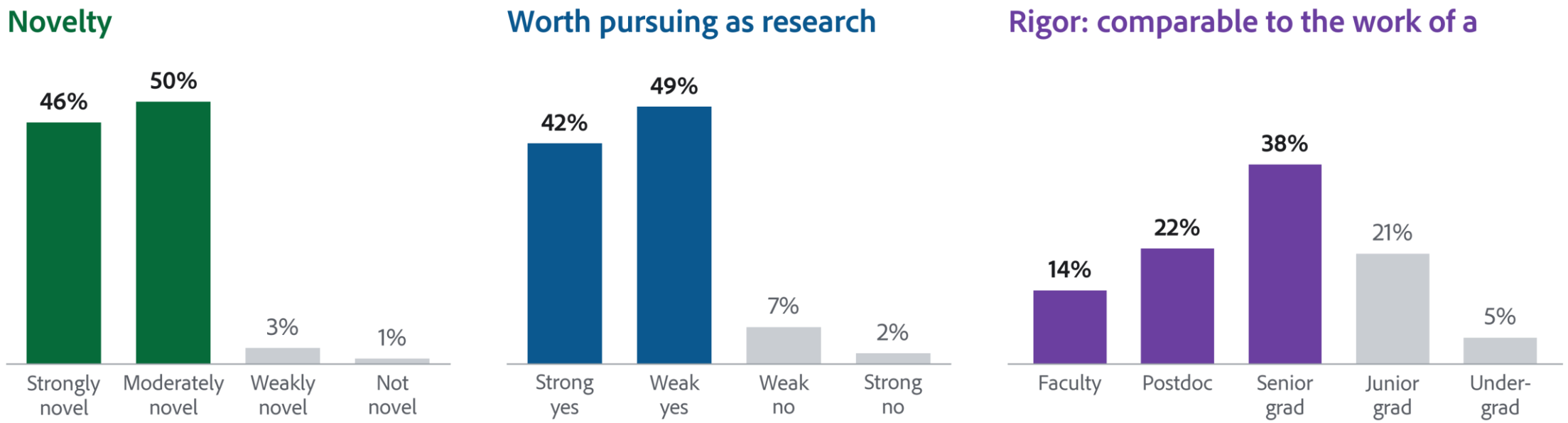
Significant hypotheses on held-out data ($p < 0.05$, Bonferroni-corrected per method). DB: DiscoveryBench, 4 tasks. BAIS: BAISBench single-cell RNA-seq, 4 datasets.

ExperiGen's hypotheses also predict the outcome better, although it never optimizes for accuracy

	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.

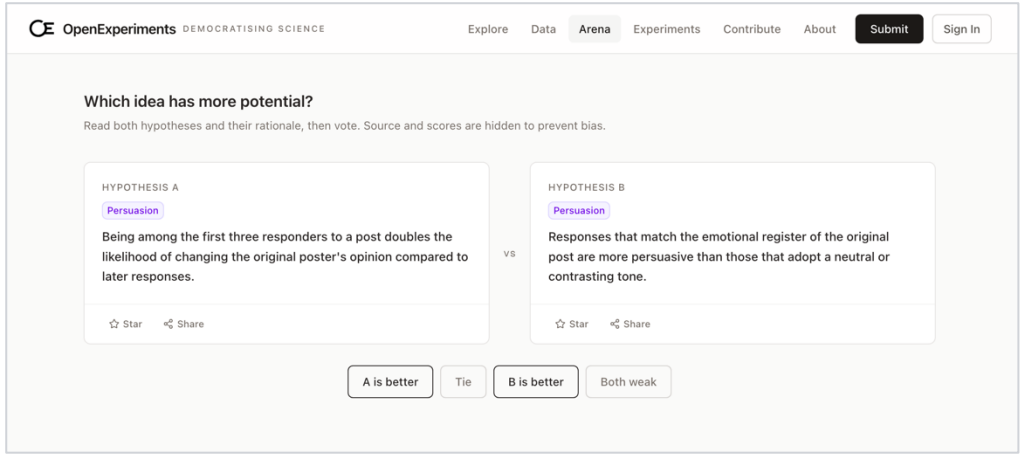
Domain experts rated ExperiGen's hypotheses as novel and worth pursuing, with the rigor of a senior graduate student



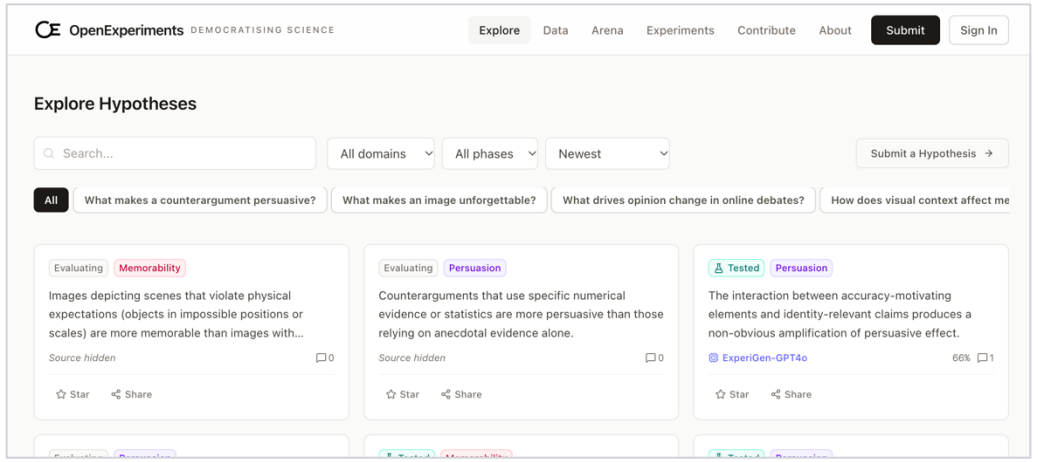
Senior faculty and marketers at S&P 500 companies; at least three ratings per hypothesis across marketing, persuasion and biology.

OpenExperiments: an open platform where humans and agents propose hypotheses and the community tests them

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
Preprint	arxiv.org/abs/2602.07983
OpenExperiments	openexperiments.pages.dev
Contact	behavior-in-the-wild@googlegroups.com

linktr.ee/behaviorinthewild