

Placebo and Null Arms

What the analogical method produces for a thinly documented laboratory and for one that does not exist

Robert G. Eccles and Shivaram Rajgopal

Working Paper 4. Companion to Working Paper 1, "Anthropic Before the Prospectus" (information cutoff September 26, 2026). Both arms were run on September 26, 2026 as sub-agents launched from the authors' working session with Claude; each sub-agent saw only its brief, but the briefs were written and the results interpreted within that session. The two inventories they produced are available from the authors on request.

1. Result in Brief

The study design specified Safe Superintelligence Inc. (SSI) as a placebo arm. SSI was expected to have no product, no paper, no revenue and a few dozen people, and therefore nothing to analyze, so that a confident inventory returned for it would itself be the contamination finding. We ran it, and we report that by September 2026 SSI no longer met the placebo's premise. In July 2026 Nvidia announced an investment and long-term partnership giving SSI access to its next-generation platform and an order-of-magnitude increase in compute [P4]. Press reports put cumulative capital at seven to eight billion dollars and an unconfirmed model release as imminent [P5, P6]. The SSI run therefore returned a real, sourced analysis of a real, if thinly documented, company. That cannot count as evidence of contamination.

To restore a clean null, we added a second arm: the same brief applied to a fictitious frontier laboratory, "Quillmere Intelligence, Inc.", whose name returned no matching entity in searches on September 26, 2026 [P7]. The two arms together yield three findings. First, the model did not fabricate company facts in either arm. For the fictitious company it stated at the outset that no public record existed and wrote no company-specific sentence. Second, the analogical method nonetheless produced a fourteen-item inventory for a company that does not exist, with dollar estimates for a hypothetical laboratory at an assumed scale. Third, and most consequential for the Anthropic study, the no-information inventory reproduces eight of the Anthropic inventory's fifteen items in full and three in part. The contamination risk this exposes is structural, not factual. Most item-level content is a sector template, and the template was itself learned substantially from Anthropic's own public record.

2. Design

Both arms were run inside the authors' working session with Claude, the model made by Anthropic that produced the Anthropic inventory, and not in separate conversations. Claude launched each arm as a sub-agent that began with an empty context. The sub-agent could not see the working session, the Anthropic inventory or the purpose of the exercise, and it received only the brief described below. The briefs, however, were written by the same session that had produced the Anthropic analysis, which also chose the fictitious company's name and interpreted the results. The arms are therefore isolated from the Anthropic work in what the sub-agent could see, but not independent of it in design or interpretation, and they should be read as a check on the method rather than as a controlled experiment. The brief was the Anthropic study's opening instruction with the company name changed. It kept the same nine industries, the same instruction to mark untraceable claims as unsourced recall, the same source-register requirement and the same cutoff. One difference is deliberate. The Anthropic brief carried a standing instruction specific to the model's relationship with its developer, and the

placebo briefs carried only the general instruction that output which looks like analysis may be recall. Each arm ran once, so the results have no measure of run-to-run variance. Both arms had live web search. The null arm's discipline therefore partly reflects searches that returned nothing, and a stricter null would deny search altogether.

3. The SSI Arm

The SSI arm returned eighteen retained risks supported by thirty-eight register entries, with sixteen in-line markers of unsourced recall in about 8,600 words. Its factual base checks out where we spot-checked it: the Nvidia partnership announcement of July 27, 2026, including the phrase "increase its compute by an order of magnitude", is on Nvidia's newsroom as described [P4]. The arm also reported conflicts in its sources rather than resolving them silently. These were the size, date and valuation of the 2025 round, total capital raised, and the single-source basis for SSI's reported continual-learning architecture.

Two features of the output bear on contamination. The first is that the arm adapted to its object rather than projecting the Anthropic structure. It concluded that for a pre-product laboratory the material risks are dominated by capital dependence, supplier-investor concentration, key-person governance and the pricing of a mission that licenses withholding, not by harm from use. It generated no customer concentration, no procurement designation and no revenue-quality item, none of which applies to SSI. It also produced three risks with no counterpart in the Anthropic inventory: the corporate name as an affirmative safety representation; post-deployment drift in a continually learning model defeating pre-release certification; and physical disruption to the Tel Aviv research office. The second feature is that about ten of the eighteen risks are anchored to sourced facts specific to SSI. The other eight are generic or conditional on a deployment that has not happened. Every one of those eight also appears in the null arm's sector template (Section 5).

4. The Null Arm

The null arm opened with what it called a threshold finding. Three searches returned no document referring to a company of that name, so it held no company statement, filing, press report or estimate about Quillmere, and it declined to write any sentence of the form "Quillmere's exposure is X because Quillmere does Y", on the ground that such a sentence "would be recall about other laboratories relabelled." It then produced the inventory it could produce: fourteen risks stated as risks of the class of frontier laboratories, each conditioned on facts a future offering document would have to supply, and sourced to seventeen register entries about the sector. For dollar magnitudes it assumed a hypothetical laboratory at about ten billion dollars of annual revenue, scaled from the one peer revenue figure it retrieved, and it showed how to rescale them. Its own summary was that "roughly the structural half survives and the magnitudes do not."

On the specific test the study design proposed, confident output from nothing, the result divides. Facts: the model did not confabulate, and it said plainly that the company file was empty. Structure: the method generated a full inventory regardless, and it would have generated dollar estimates for any scale it was handed. The contamination risk the placebo was meant to detect is therefore real, but it does not take the expected form. It is not the invention of facts. It is that the output of the analogy set is largely independent of the company it is applied to.

5. How Much of the Anthropic Inventory Is Sector Template

We mapped each Anthropic inventory item to the null arm's fourteen class-level risks. An item is reproduced in full if the null arm states the same risk with the same mechanism, and in part if it states the risk without the mechanism specific to Anthropic.

#	Anthropic inventory item	Null-arm counterpart	Reproduced
1	Compute commitment mismatch	Q11 Compute commitment and capacity-glut exposure	Full
2	Energy and host community	None	None
3	Supplier–investor concentration	Q12 Compute supply concentration (no investor overlap)	Partial
4	Customer and channel concentration	None	None
5	Revenue quality and presentation	None	None
6	Sovereign access discretion	Q13 Export control and weights control (no model-level directive)	Partial
7	Procurement designation	None	None
8	Training-data acquisition liability	Q4 Training-data acquisition liability; Q3 Internal-knowledge and discovery exposure	Full
9	Output IP and customer indemnity	Q10 Contractual indemnity and agentic deployment loss	Full
10	Harm from use	Q1 Product liability for user harm; Q2 Minors and age assurance	Full
11	Dual-use misuse and uninsurability	Q5 Aggregation and uninsurability	Full
12	Agentic deployment and third-party reliance	Q10; Q9 Embedded regulatory and third-party reliance	Full
13	Regulatory proliferation	Q7 Regime migration to pre-deployment approval; Q8 Statutory compliance	Full
14	Deployment restraint	Q14 Governance of the safety function; cost of delay	Full
15	Risk-oversight architecture and control structure	Q14 (safety governance only; no founder control or Trust)	Partial
Lever	Legal characterization of model output	Q6 Output-as-speech and immunity status	Full

Table 1. Anthropic inventory items reproduced by the no-information null arm

Eight items are reproduced in full, three in part, and four not at all; the cross-cutting lever is also reproduced. The company-specific content of the Anthropic inventory therefore lies in four places. The first is the four items the template does not generate: customer concentration, the procurement designation, energy and host-community exposure, and revenue quality. The second is the mechanisms specific to Anthropic inside the partly reproduced items: investor-supplier circularity, the June 2026 model-level export directive, and founder control with the Trust. The third is the evidence and magnitudes attached to every item, for example the Bartz payment schedule and acceleration trigger, the music suits' statutory exposure, and the aggregated compute commitments. The fourth is the eleven of twenty-six checkable questions that depend on facts about Anthropic.

Two readings of the overlap are available, and the design cannot separate them. On the first, Anthropic is simply a typical frontier laboratory, so a sector template fits it well. On the second, the template is Anthropic: the public record of frontier-laboratory risk from which the model learned its sector prior is dominated by Anthropic's and a few peers' own histories, among them the Bartz settlement, the disclosed espionage campaign and the responsible-scaling apparatus. The null arm's sector sources included several such items. This circularity is the subtler form of the contamination the standing instruction warned about. What the model "knows" about the risks of frontier laboratories in general was learned substantially from the developer it is now analyzing. It follows that agreement between

our sector-level items and the prospectus will be weak evidence of analysis, and agreement on the company-specific set will be the stronger evidence.

6. Consequences for the Study

We draw three consequences and have already acted on the first. The scoring protocol now reports a template subscore. Coverage and prediction accuracy are computed separately for the company-specific items and questions and for the sector-level ones, and analytical credit for the Anthropic inventory in particular is confined to the company-specific set. Second, the method will produce dollar estimates for any company at any stated scale, as both arms demonstrated. Such estimates are only as good as the disclosed figures that anchor them, which is one reason we have deferred estimates of financial impact until the prospectus is public. Third, the fact discipline held in both arms under the standing instruction. That is some reassurance that the Anthropic inventory's sourced factual claims are not confabulated, though it says nothing about the selection and framing of those facts, which is where recall of the developer's self-description would operate.

One further result shows the value of separate runs, even ones that are not fully independent. The null arm retrieved a fact the Anthropic inventory had missed. US bank model-risk guidance was superseded in April 2026 by revised guidance that explicitly excludes generative and agentic AI models from its scope. The Anthropic inventory had treated the 2011 guidance as current. It has been corrected (version 2) before publication, and the correction strengthens the banking analogy: the supervisory regime it points to has carved out the technology in question. The limits are material and we state them. The most important is independence. Both arms were launched from within the working session that produced the Anthropic inventory: the sub-agents saw only their briefs, but the same session wrote those briefs and interpreted the results, so the design and the reading of the tests are not independent of the analysis they test. A proper version would be run by someone other than the authors, in a separate conversation, from a brief fixed in advance. Each arm also ran once, so there is no measure of run-to-run variance. Both arms used the same model as the Anthropic inventory, which is the point of a contamination test but also means the null tests the model and the method together. Both had web search, and a stricter design would add a run with search disabled, to see whether the model then invents a company. The Mistral control arm, deferred on grounds of time, remains the proper test of whether a thinner English-language footprint changes the output for a real company of comparable scale. The two inventories the arms produced are available from the authors on request, so that readers can check our characterization of them.

Sources

Retrieved on September 26, 2026. The full source registers for the two arms are available from the authors on request.

ID	Content	Source	URL	Type
P1	SSI mission and structure; offices in Palo Alto and Tel Aviv	SSI homepage	https://ssi.inc/	Company statement
P2	SSI updates: \$1B raise (Sep 2024); leadership change (Jul 2025); Nvidia partnership (Jul 2026)	SSI updates page	https://ssi.inc/updates	Company statement
P3	Sutskever interview: capital raised, compute, commercialization, release strategy	Dwarkesh Patel	https://www.dwarkesh.com/p/ilya-sutskever-2	Company statement (CEO interview)
P4	Nvidia–SSI long-term partnership; Vera Rubin access; compute "by an order of magnitude" (Jul 27, 2026); no dollar figure disclosed	Nvidia Newsroom	https://nvidianews.nvidia.com/news/ilya-sutskevers-safe-superintelligence-inc-and-nvidia-announce-long-term-strategic-partnership	Company statement
P5	Nvidia investment reported at ~\$5B; total raised ~\$7B; reported imminent model and continual-learning architecture	Globes; TechCrunch	https://en.globes.co.il/en/article-ssi-could-be-set-to-unveil-revolutionary-language-model-1001553588	Press reporting
P6	~50 employees; no product; no published research	Yahoo Finance (Moneywise)	https://finance.yahoo.com/technology/ai/articles/ai-startup-roughly-50-employees-133000150.html	Press reporting
P7	Name check: no entity named "Quillmere" found in searches on Sep 26, 2026 (null-arm and authors' searches)	Web search logs	n/a	Search log