

AI and Research Practice

Working with large language models in your research

Derek Harmon

October 13, 2026

ASK ANYTHING, ANONYMOUSLY

*Open this now and keep
it open during the talk to
ask questions →*



BROAD COLLEGE *of* BUSINESS

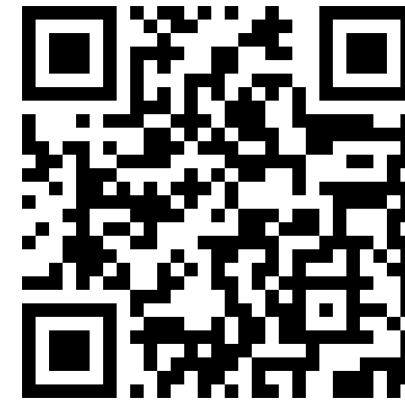
derekharmon.com/ask

Session modules

- ❖ **Motivation**
- ❖ **How these models work**
Break for clarifying questions
- ❖ **What this means for research**
Remaining questions time permitting – I'll stay after too

Ask anything, anonymously

It stays open all session.



derekharmon.com/ask

Dorothy Carter is curating.

Motivation

ASK ANYTHING,
ANONYMOUSLY



Fact #1: AI is already being used in research

58%

of researchers use AI in their work
— up from 37% a year earlier

3,200+ researchers, 113 countries · Elsevier,
2025

48%

of early stage researchers use AI
frequently — against 34% of those
with 20 years or more experience

10,480 researchers, 117 countries · Springer
Nature, Mar 2026

1 in 4

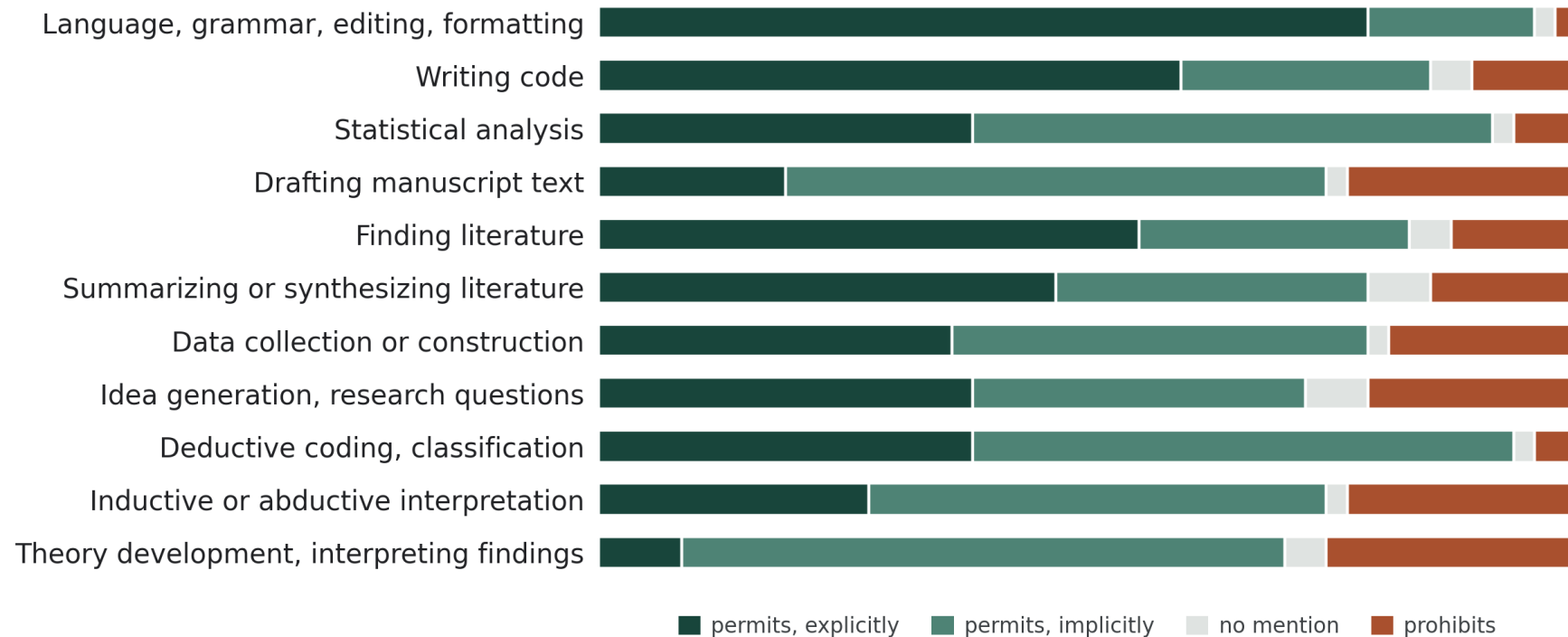
submissions to Organization
Science had AI in substantial
control of the writing

4,715 post-ChatGPT submissions · Pangram
score above 30% · Organization Science, 2026

Fact #2: AI is largely permitted by journals

Author-side AI rules across the 47 Broad Strategically Targeted Journals

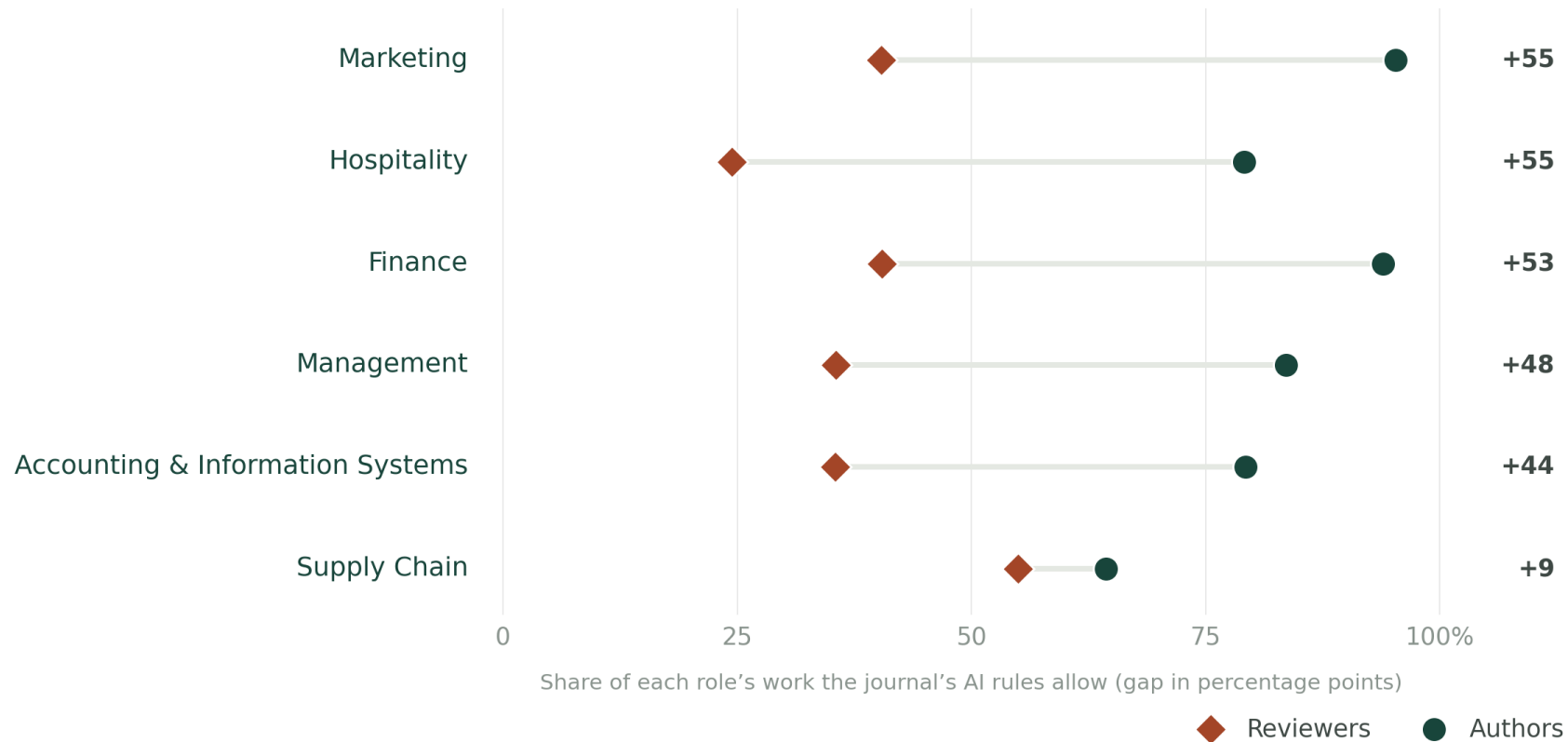
Research activities, coded from each journal's governing policy · October 2026



LOOK UP ANY JOURNAL

derekharmon.com/ai-policies

Permitted *much* less for reviewers



Fact #3: AI is being discussed by editors

WHICH JOURNALS HAVE WRITTEN?



WHAT KIND OF EDITORIAL IS IT?



WHAT THE 40 EDITORIALS ARE ABOUT



MOST DISCUSSIONS ARE AROUND:

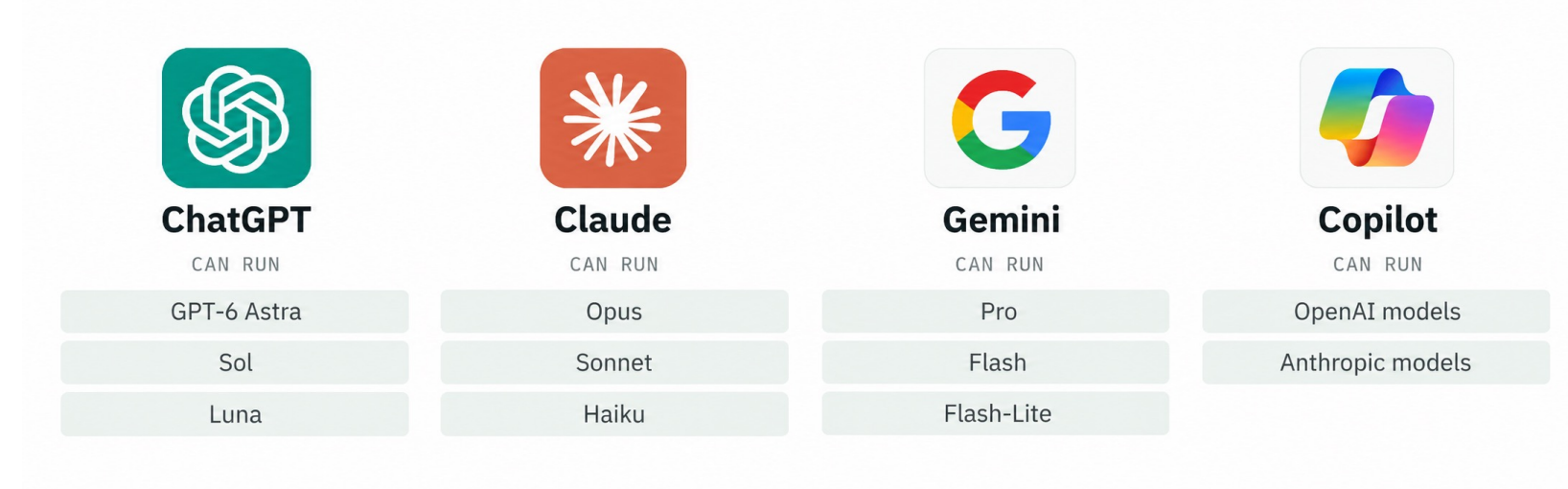
Am I allowed to use AI?

What will AI do to the field?

FAR LESS IS ABOUT:

How do I get good output?

This matters because we have all used these tools.



Sometimes the output is useful, sometimes it isn't, and sometimes we cannot tell which.

How do we use LLMs in our research to get high-quality output?

How we get there

1 Understand how LLMs actually work

Why a question returns a different answer each time. Why a wrong answer reads like a right one.

2 What that means for our research

What to give LLMs to improve quality output. How to check what comes back.

**WHAT THIS
SESSION IS NOT**

Not tool training.
There is a lab for that.

Not an argument for using these tools.
An argument for understanding them.

THE LAB

Zoom sessions for tool demonstrations and training.
Three to start, with more added later based on topic interest.

Details and sign up at the close of this session.

1 How LLMs work

ASK ANYTHING,
ANONYMOUSLY



This section is going to reinforce:

TWO KEY PRINCIPLES

- P1** LLMs are *predictive*, which means all they compute is a probability for every word that could come next.
 - P2** The output is *stochastic*, which means what you get is one draw from those probabilities.
-

DEMOS



ECONOMETRICA

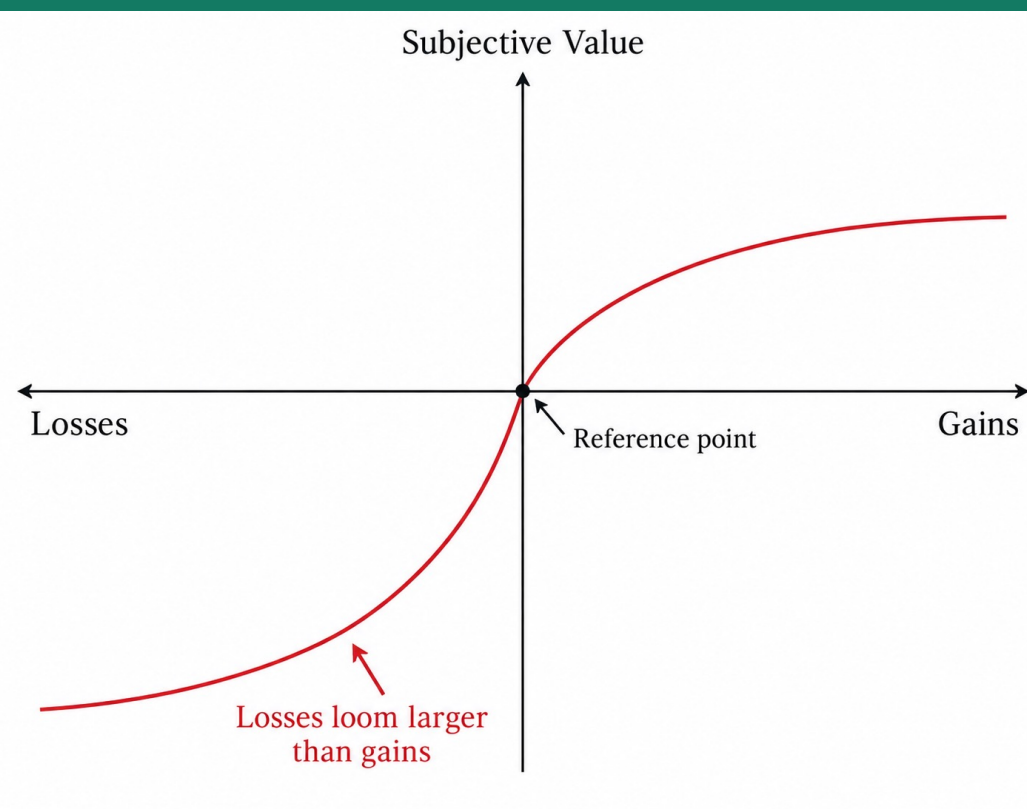
VOLUME 47

MARCH, 1979

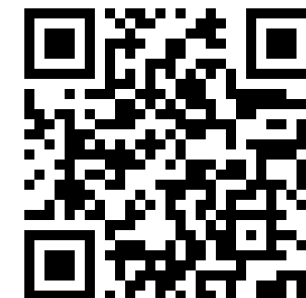
NUMBER 2

PROSPECT THEORY: AN ANALYSIS OF DECISION UNDER RISK

BY DANIEL KAHNEMAN AND AMOS TVERSKY¹



ASK ANYTHING,
ANONYMOUSLY



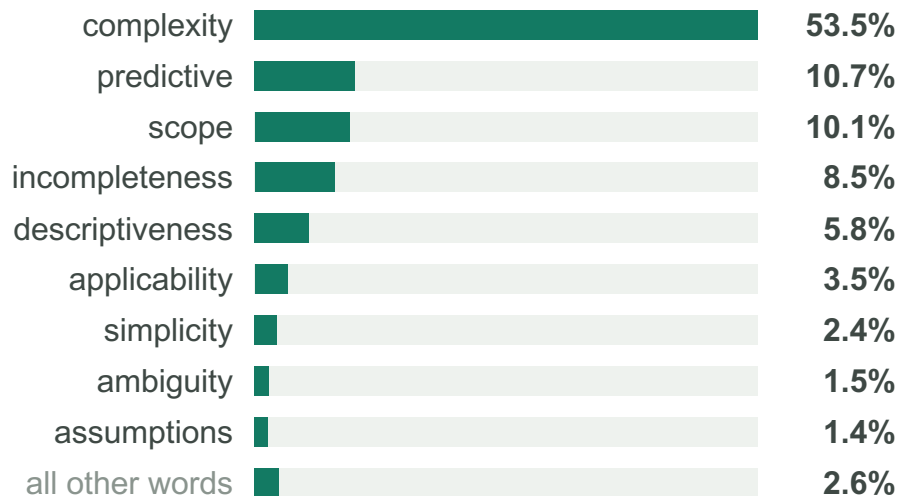
DEMO 1

FILL IN THE BLANK

"The most important limitation of prospect theory is its ____"

RAN THIS 20 TIMES

Probability what the model gave each word



Probability data came from gpt-4o through the API. The flagship models from OpenAI, Anthropic and Google will not show it to you

PRINCIPLE 1

LLMs are *predictive*, which means all they compute is a probability for every word that could come next.

Output what the model actually said



PRINCIPLE 2

The output is *stochastic*, which means what you get is one draw from those probabilities.

DEMO 1

FILL IN THE BLANK

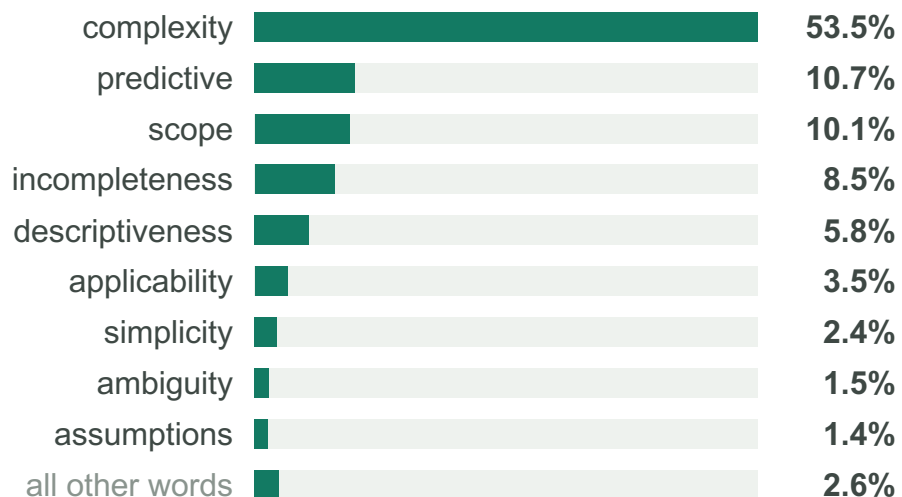
"The most important limitation of prospect theory is its ____"

FROM P1

Predicts what comes next,
not looking anything up

RAN THIS 20 TIMES

Probability what the model gave each word



Probability data came from gpt-4o through the API. The flagship models from OpenAI, Anthropic and Google will not show it to you

Output what the model actually said



FROM P2

Your output is one draw

FROM P2

Ask multiple times, get different answers

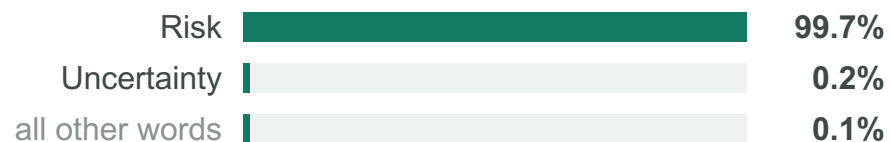
DEMO 2

LET'S TRY A PROMPT THAT HAS AN OBVIOUS RIGHT ANSWER

"Kahneman & Tversky (1979), Prospect Theory: An Analysis of Decision under ____"

ACROSS THE SAME 20 RUNS

Probability what the model gave each word



Output what the model actually said



FROM P1

Output is right because that word dominates its training text, *not* because it looked anything up

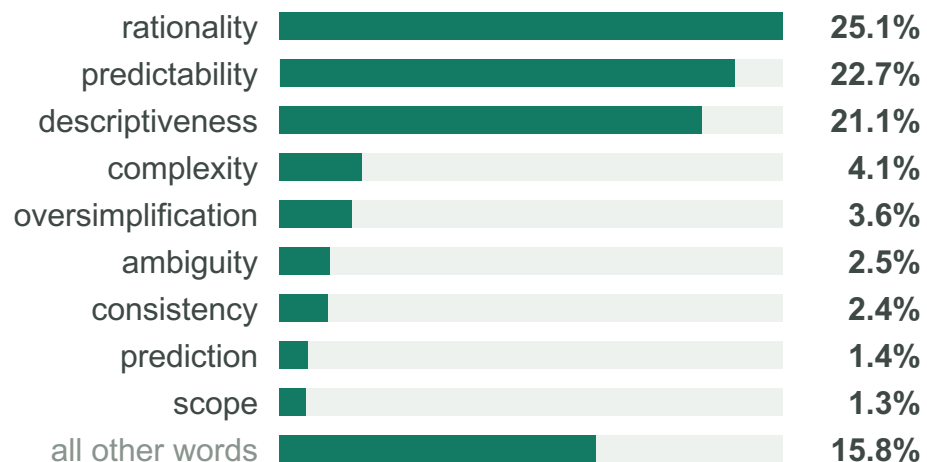
DEMO 3

NOW LET'S PHRASE THE PROMPT AS A QUESTION (what most of us actually do)

"In one word, what is the most important limitation of prospect theory?"

ACROSS THE SAME 20 RUNS

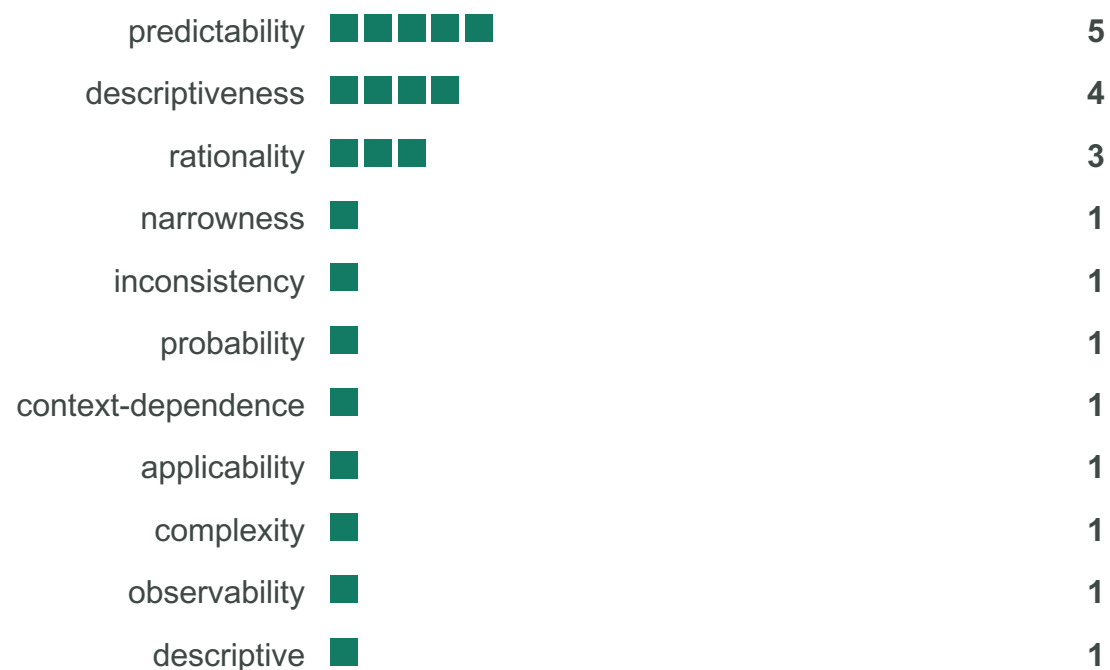
Probability what the model gave each word



FROM P1

**Even when we phrase our prompt as a question,
the model is still predicting what comes next**

Output what the model actually said



DEMO 4

NOW LET'S SEE HOW LLMs PRODUCE A FULL SENTENCE

"I am writing a paper that builds on prospect theory. In one sentence, what is the single most important gap it leaves?"

Prospect theory addresses decision-making under risk but leaves a significant gap in _____

Probability what the model gave each word

explaining		47.9%
accounting		18.8%
understanding		14.2%
fully		11.7%
its		4.2%
comprehensively		1.2%
adequately		0.8%
not		0.3%
how		0.2%
all other words		0.7%

Output what the first run actually wrote

accounting 18.8%

...**accounting** (18.8%) **for** (100%) **how** (57%) **individuals** (39%) **consistently** (1%)
integrate emotions and long-term consequences into their evaluations.

FROM P1

**Always sounds fluent because every word is
predicted from everything before it**

DEMO 5

NOW WE CAN UNDERSTAND WHY HALLUCINATIONS HAPPEN

"Give me the full citation for the Administrative Science Quarterly article that introduced the concept of algorithmic legitimacy."

WHAT CAME BACK

Buell, R. W., Porter, E., & Tedeschi, J. (2023). When do algorithms have legitimacy in organizations? *Administrative Science Quarterly*, 68(3), 735–770.
<https://doi.org/10.1177/00018392221150720>

DOES NOT EXIST

The DOI returns 404. No paper with that title exists in Crossref. ASQ volume 68 is real and is 2023. All three authors are real, and work in three different fields.

FROM P1

The model did what it was trained to do
(predict what comes next, not retrieve a record)

FROM P1 + P2

The problem is that the output always reads fluently
(and so it looks real even if it's not)

Note: Frontier models are getting better at avoiding problems like this, but they will never be perfect.

TWO KEY PRINCIPLES

- P1** LLMs are *predictive*, which means all they compute is a probability for every word that could come next.
 - P2** The output is *stochastic*, which means what you get is one draw from those probabilities.
-

IMPLICATION

LLMs are optimized for fluency, so they sound just as certain when they're wrong.

Break for clarifying questions.



derekharmon.com/ask

2 What this means for research

ASK ANYTHING,
ANONYMOUSLY



There are two places we can intervene

FRONT-END JUDGMENT

Context

What the model sees before it predicts.

Context does not make the model smarter. It changes which words have a real chance of being drawn for output.

For example:

- Your prompt
- Attachments
- Your conversation
- System memory
- Web retrieval

THE MODEL



BACK-END JUDGMENT

Verification

How we evaluate the output for quality.

A wrong output looks exactly like a right one. Verification is how we find out which one we got.

For example:

- Check against source
- Ask it to point to the data
- Re-run it cold in fresh context
- Vary the model
- Red teaming

There are two places we can intervene

FRONT-END JUDGMENT

Context

What the model sees before it predicts.

Context does not make the model smarter. It changes which words have a real chance of being drawn for output.

THE MODEL



BACK-END JUDGMENT

Verification

How we evaluate the output for quality.

A wrong output looks exactly like a right one. Verification is how we find out which one we got.

How we intervene depends on the type of question we ask:

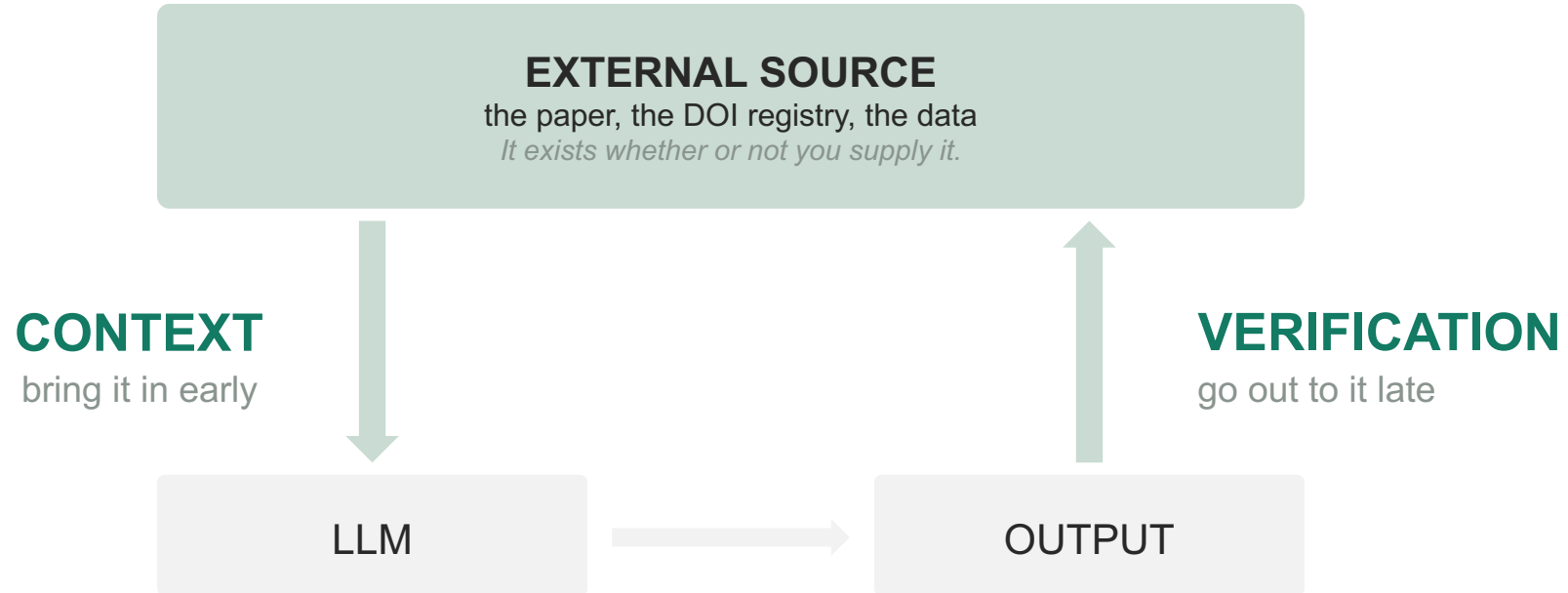
1. A right answer exists (Type 1 questions)
2. No right answer exists (Type 2 questions)

TYPE 1 DEMOS

A right answer exists

Type 1: A right answer exists

Quality output = accurate output



Bring the source into your context.

Still needed, but now quick.

“In Kahneman & Tversky (1979), what percentage of subjects chose option C in Problem 2?”

CORRECT ANSWER: 83

WHEN I ASKED THE PROMPT

18 of 20 runs said:

80

THEN I ASKED “ARE YOU SURE?”

20 of 20

apologized and abandoned their
answer,
and 19 of 20 corrected with:

92

I SAID “QUOTE THE SENTENCE”

15 of 20 refused on copyright.

5 of 20

invented this quote:

“In problem 2, 80% of
subjects chose option C,
and 20% option D.”

Not in the paper.

“In Kahneman & Tversky (1979), what percentage of subjects chose option C in Problem 2?”

CORRECT ANSWER: 83

WHEN I ASKED THE PROMPT

20 of 20 runs said:

83



I SAID “QUOTE THE SENTENCE”

19 of 20

quoted the real sentence from the paper:

“The data show that 82 per cent of the subjects chose B in Problem 1, and 83 per cent of the subjects chose C in Problem 2.”

That sentence is in the paper.

For Type 1 questions, context is what makes verification cheap.

Sample workflows I've built

KEY: Bring the source into your context, then check against it.

Idea Feasibility Test

Check whether my idea has been done

Paper Accuracy Audit

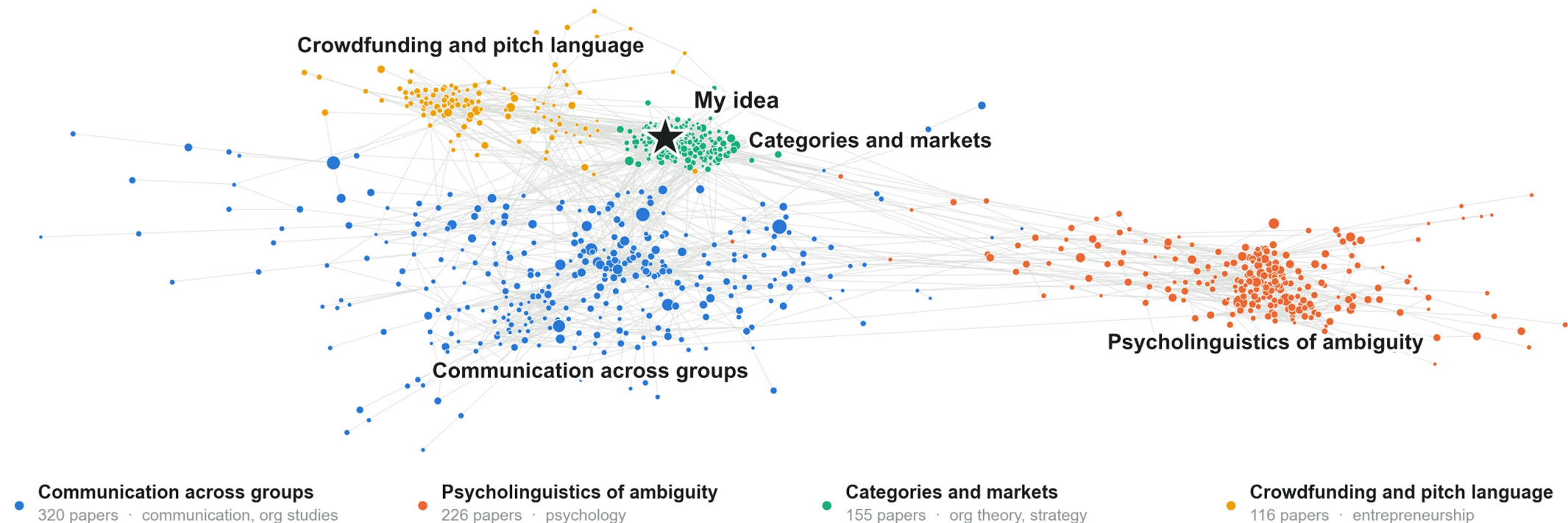
Check the accuracy of a revision before I resubmit it

Idea Feasibility Test

IDEA: Entrepreneurs often pitch novel products to diverse audiences. Could using more ambiguous words in their pitch mobilize more support?

CHECK 1 OF 2 **My idea** against **multiple literatures**, to see where it sits

SOURCE: 1,641 papers, found by searching the idea in each field's own vocabulary and following citations from six papers it builds on



FOUND: my idea sits in categories and markets, one of four literatures it touches, each in a different field.

OpenAlex, Sept 2026 · no date cutoff · literatures found from shared references alone, names suggested from each group's top papers · the star is placed by the six papers the idea builds on · papers outside the four largest literatures not drawn

Idea Feasibility Test

IDEA: Entrepreneurs often pitch novel products to diverse audiences. Could using more ambiguous words in their pitch mobilize more support?

CHECK 2 OF 2 **My idea** against **the closest studies**, to see if it's been done

SOURCE: the 350 papers that link two or more parts of my idea, each coded against all five parts. The nine closest are below.

	One message, varied audience	Words with several meanings	Prior knowledge shapes the reading	Broader support	Entrepreneurs pitching products
STUDIES OF THE MECHANISM					
Fiske 1986 Communication	✓	✓	✓	✓	•
Smith et al. 2006 Communication	✓	✓	✓	✓	•
Grundfest & Pritchard 2002 Law	✓	✓	✓	✓	•
McMahan & Evans 2018 Sociology	✓	✓	✓	✓	•
Sohn & Edwards 2018 Communication	✓	✓	•	✓	•
CLOSEST STUDIES OF ENTREPRENEURS					
Song 2025 Entrepreneurship	✓	•	✓	✓	✓
Wry et al. 2013 Entrepreneurship	•	•	✓	✓	✓
Vossen & Ihl 2020 Entrepreneurship	✓	•	•	✓	✓
Boulongne 2022 Entrepreneurship	•	•	✓	✓	✓

FOUND: the mechanism isn't new, the setting is. 5 studies test it, none on entrepreneurs, whose closest studies look at categories, not words.

Paper Accuracy Audit

Run on a late-stage revision of mine.

SOURCE	CHECK	FOUND
 85 papers I cited	My citations against the papers I cited Claims checked for the 85 citations whose PDF I had	1 Does every citation exist? 85 of 85 references exist 2 Does each paper say what I cited it for? 5 of 85 citations contradicted by the paper they cite 522413 52 Supported24 Loosely supported1 Not supported5 Contradicted3 Left for me to read
 My manuscript	My text against my own tables and sections Every number, measure, and cross-reference	10 issues to fix in my own paper Biggest: one variable reported on three scales that can't all be true.
 My manuscript	My response letter against the manuscript 73 changes the letter says I made	19 of 73 claimed changes not fully made 54163 54 Held up16 Only partly made3 Not in the manuscript

References checked against Crossref. Every other check by Claude, required to quote the source. Every quote confirmed word for word by script. Sept 2026.

Sample workflows I've built

KEY: Bring the source into your context, then check against it.

Idea Feasibility Test

Check whether my idea has been done

SOURCE: existing literature

Two checks of my idea against:

- 1) Multiple literatures, to see where it sits
- 2) The closest studies, to see if it's been done

Paper Accuracy Audit

Check the accuracy of a revision before I resubmit it

SOURCE: my manuscript + the papers I cited

Three checks of my:

- 1) Citations against the papers I cited
- 2) Text against my own tables and sections
- 3) Response letter against the manuscript

Learn how to do your own in:

Lab 1 · Nov 10

Developing agent architecture & using the terminal

Lab 2 · Nov 17

Context & verification when a right answer exists

TYPE 2 DEMOS

No right answer exists

Type 2: No right answer exists

Quality output = intended output



Decide your intention and bring it into your context.

Ask for many options/objections, and judge them yourself.

“I want to build on prospect theory. What’s an interesting place to apply it?”

“Here are a few interesting areas where you could apply or extend prospect theory:”

1. Behavioral Finance

the disposition effect, the equity premium puzzle

2. Healthcare

risky treatments vs. certain outcomes

3. Public Policy

health advisories, climate, disaster preparedness

How do you know if this is quality output (i.e., what you intended)?

You stated an intention, but it was vague.

You never said what “interesting” means. *Interesting to whom?*

DEMO 9

WITHOUT STATING WHAT “INTERESTING” MEANS, THE MODEL DEFINES IT

“I want to build on prospect theory. What’s an interesting place to apply it?”

RUN IT 100 TIMES

User experience design
Behavioral economics in climate change policy
Behavioral Economic Policies
Personal Finance Decisions
Climate Change Communication
Health decision-making
Healthcare decision-making
Behavioral Economics in Renewable Energy Adoption
Behavioral finance in cryptocurrency markets
Climate Change Policy Design
Climate Change Policy Decisions
Financial decision-making
Behavioral Finance
Personal Finance and Investment Decisions
Renewable energy investment
Health behavior change
Insurance industry
... and so on, all different

AGGREGATE MAIN IDEAS

behavioral finance

health

climate

cryptocurrency

organizations

insurance

negotiation

consumer

DRAWS

39%

21%

20%

10%

2%

1%

1%

1%

PAPERS

2,658

367

637

142

641

381

168

641

**You asked for an interesting place to apply it,
but without clarifying your intention, the model defined it as the most conventional.
(only the hundred runs show that)**

“I want to build on prospect theory. What’s an interesting place to apply it?”

SAME QUESTION

gpt-4o · API, temp 1.0

GPT-5.6 Sol · Chat window, Memory on

TOP TWO ANSWERS RUNNING 100 TIMES

behavioral finance 39% · health 21%

corporate preannouncements 43% · AI delegation 40%

WHAT IT REMEMBERS FROM MY PAST CHATS

Enable memory

Let ChatGPT personalize your experience based on your chats, files, and connected apps. [Learn more](#)



*Settings → Personalization → Manage.
None of this was in the prompt.*

ChatGPT defined “interesting” by what it remembered about me.

(the whole distribution shifted towards projects I’ve been working on)

If you don’t know what’s in your context, you can’t judge if it’s what you wanted.

Sample workflows I've built

KEY: Decide your intention and bring it into your context, then ask for many options or objections and judge them yourself.

Idea Triage

Decide which way to take an early kernel of an idea

Paper Red Team Audit

Find my draft's weakest points before reviewers do

Idea Triage

Run on an early idea of mine, April 2026.

BUILD MY INTENTION

My idea
and what I want from it

An interview first

Where did it come from?

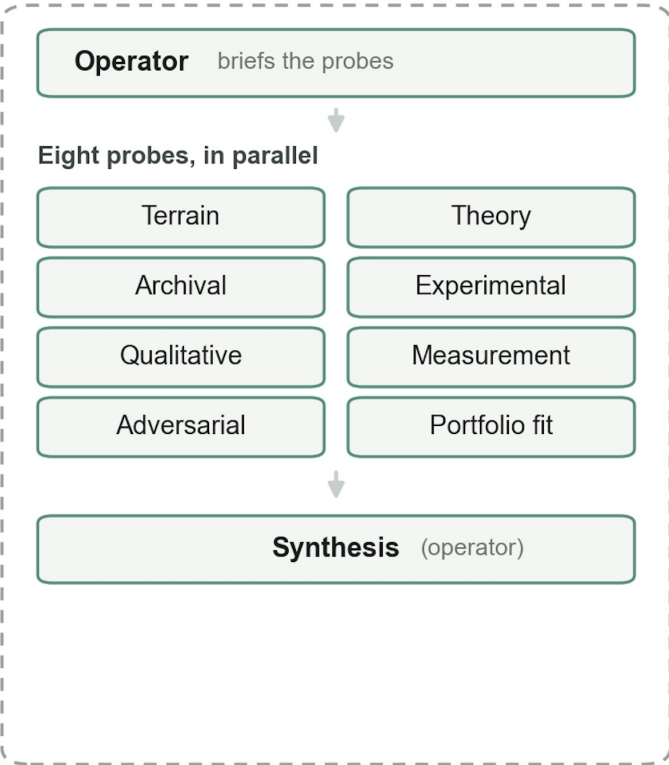
What are my constraints?

Which journal?

So what? (required)

Also built in
My research profile: top journals, real-world impact, blue-ocean ideas

AGENTIC ARCHITECTURE



PRODUCES MANY OPTIONS

36 options, from eight probes

4

9

23

■ priority 1 ■ priority 2 ■ priority 3

The ranked list, from the top

A

Archival paper
recommended: best balance of feasibility and stakes

B

Theory paper
highest ceiling, harder to land

C

Qualitative study
fastest path, the fallback

+ 33 more, ranked below

MY JUDGMENT

The tool produces options. I judge.
I work down the list as far as I want.

✓

A Pursuing
RISK OVERRULED
Its top risk, measurement circularity, I judged overstated.

○

B Left open

✗

C Killed
OVERRULED
Public-data case studies read as illustrative, not a top-journal paper.

Saved for every future run
No implicit paper pipelines.

FROM MY STATED INTENTIONS

36 options

10 evaluated

26 saved for later

1 pursued

Triage run, April 2026, on an early idea of mine. Eight Claude probes in parallel. Options and priorities from the run's inventory, my decisions from the project log.

Paper Red Team Audit

Run blind on a first draft of mine, then compared with what the reviewers later said.

BUILD MY INTENTION

My manuscript

Introduction

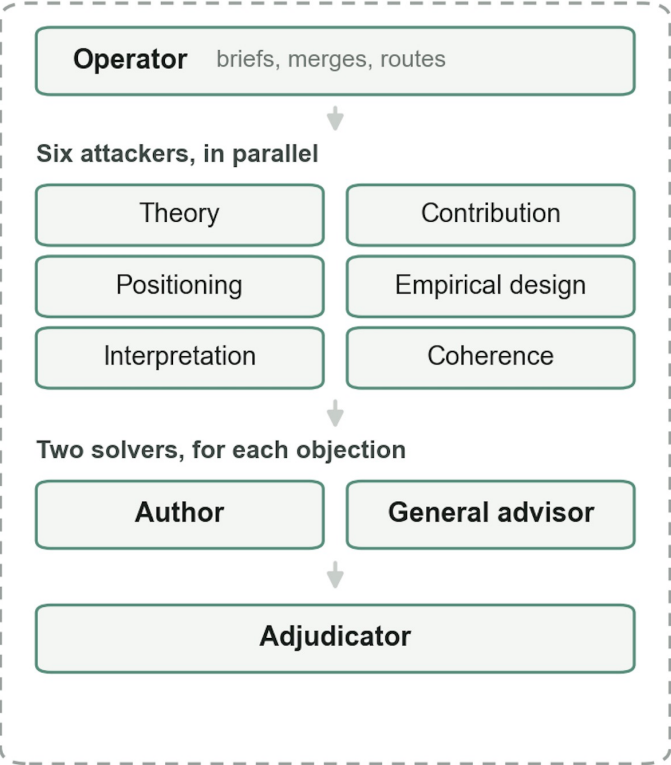
Theory

Methods

Results

Discussion

AGENTIC ARCHITECTURE



PRODUCES MANY OBJECTIONS

33 objections, merged from 82

21

9

3 critical

21 major

9 minor

Critical = a headline claim at risk as written

Top three, all critical

CRITICAL

The measure may track common words, restating the familiarity advice

CRITICAL

The mechanism test is built from the ambiguity score itself

CRITICAL

The 'avoid ambiguity' foil ignores work the paper itself cites

MY JUDGMENT

16 objections where all agents agree

The adjudicator recommends the fix.

I judge.

8 objections where agents disagree

Each comes to me with the agents' different options.

I judge.

THIS WORKFLOW IDENTIFIED

24 of 25 issues raised by the reviewers

5 of 5 major issues raised by the editor

Retrospective run, Oct 2026, on my January 2025 submission, blind to the reviews. Ten Claude agents. Every objection quotes the draft, confirmed by script.

Sample workflows I've built

KEY: Decide your intention and bring it into your context, then ask for many options or objections and judge them yourself.

Idea Triage

Decide which way to take an early kernel of an idea

INTENTION: what I want from an idea

Nine agents give me:

- 1) Options: possible ways to study it, ranked
- 2) Objections: an adversary attacks all options

Then I judge: pursue, park, or kill options

Paper Red Team Audit

Find my draft's weakest points before reviewers do

INTENTION: what I want from this paper

Ten agents give me:

- 1) Objections: attacks on every part of paper
- 2) Options: fixes for each, recommendations

Then I judge: each recommended fix

Learn how to do your own in:

Lab 1 · Nov 10

Developing agent architecture & using the terminal

Lab 3 · Dec 1

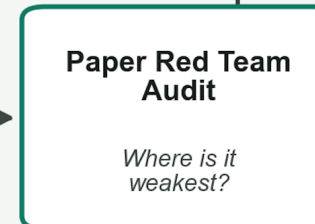
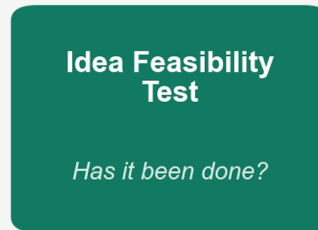
Context & verification when no right answer exists

Workflow types interlock across a project

START OF A PROJECT

BEFORE SUBMITTING

TYPE 1 A right answer, checked against a source



TYPE 2 No right answer, held against what I want

Concluding thoughts



What to take with you

THE CHALLENGE

Fluency tells you nothing.

A wrong output reads exactly like a right one.

WHAT WE CAN DO ABOUT IT

Set up the input. Check the output.

TYPE OF QUESTION	CONTEXT	VERIFICATION
A right answer exists	Bring the source into your context.	Still needed, but now quick.
No right answer exists	Decide your intention and bring it into your context.	Ask for many options/objections, then judge them yourself.

Your questions

ASK ANYTHING, ANONYMOUSLY



derekharmon.com/ask



Closing information



Lab sessions

Dates

1) Using agents and the terminal

Developing agent architecture and working at scale in the terminal.

Nov 10

2) Context and verification when a right answer exists

Attaching sources and building checks that improve output accuracy.

Nov 17

3) Context and verification when no right answer exists

Curating context and running robustness checks against what you intended.

Dec 1

*Tuesdays @ noon
60 min on Zoom*

Lab sessions are **practical**. We will discuss how to accomplish ***real research tasks*** in a way that yields high-quality output. More topics added later based on interest.

Thank you!



SIGN UP FOR THE LABS

derekharmon.com/ai-lab

Sign up and I'll send you the link and the materials.
Come to what you can. The rest is recorded.

GUIDE, DEMOS, SLIDES, AI POLICIES, AND EVERYTHING ELSE

derekharmon.com/ai

Recordings of this session and each lab get posted here.

Appendix

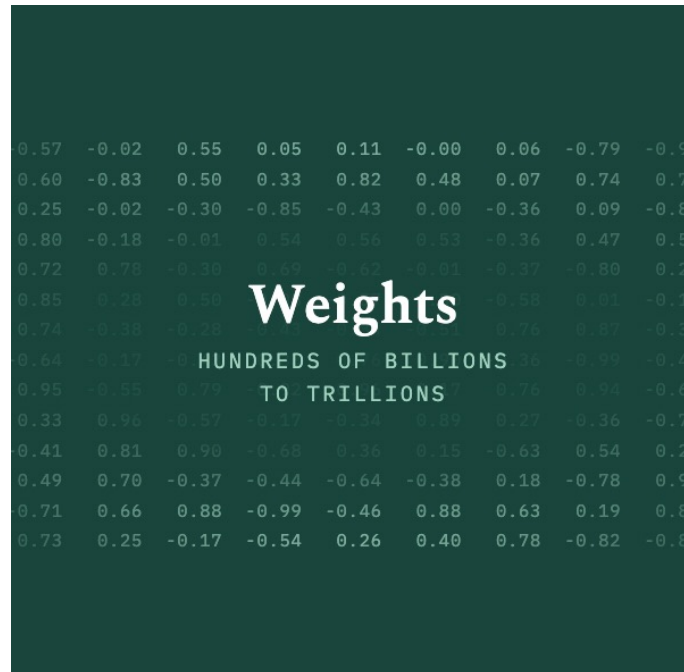


Before we run anything through it, we need to understand that...

The model is just a very large set of numbers

PRE-TRAINING (HOW IT'S BUILT)

- 1 Take an enormous pile of text.
- 2 Cover the next word. The model gives **all words* a probability of coming next**.
- 3 Now uncover that next word. **You can see the exact probability the model gave the real word**. If that probability was **low**, shift the model's weights so it would have been higher. Move along one word and do it again.
- 4 **Do this trillions of times. Those numbers are the model weights** →



POST-TRAINING (HOW IT'S TUNED)

A smaller pass over the finished model.

Shapes conduct more than knowledge.

- Hold a confident, helpful tone
- Continue a question with an answer, rather than with more text
- Add guardrails, such as refusing to provide certain information

Reinforcement learning from human feedback (RLHF), and related methods.

THE MODEL IS NOT A LIST OF FACTS

There is no table, nothing to look up, and no record of where anything came from. It is weights, predicting what comes next.

**Strictly speaking, model returns a probability for pieces of words, or tokens.*