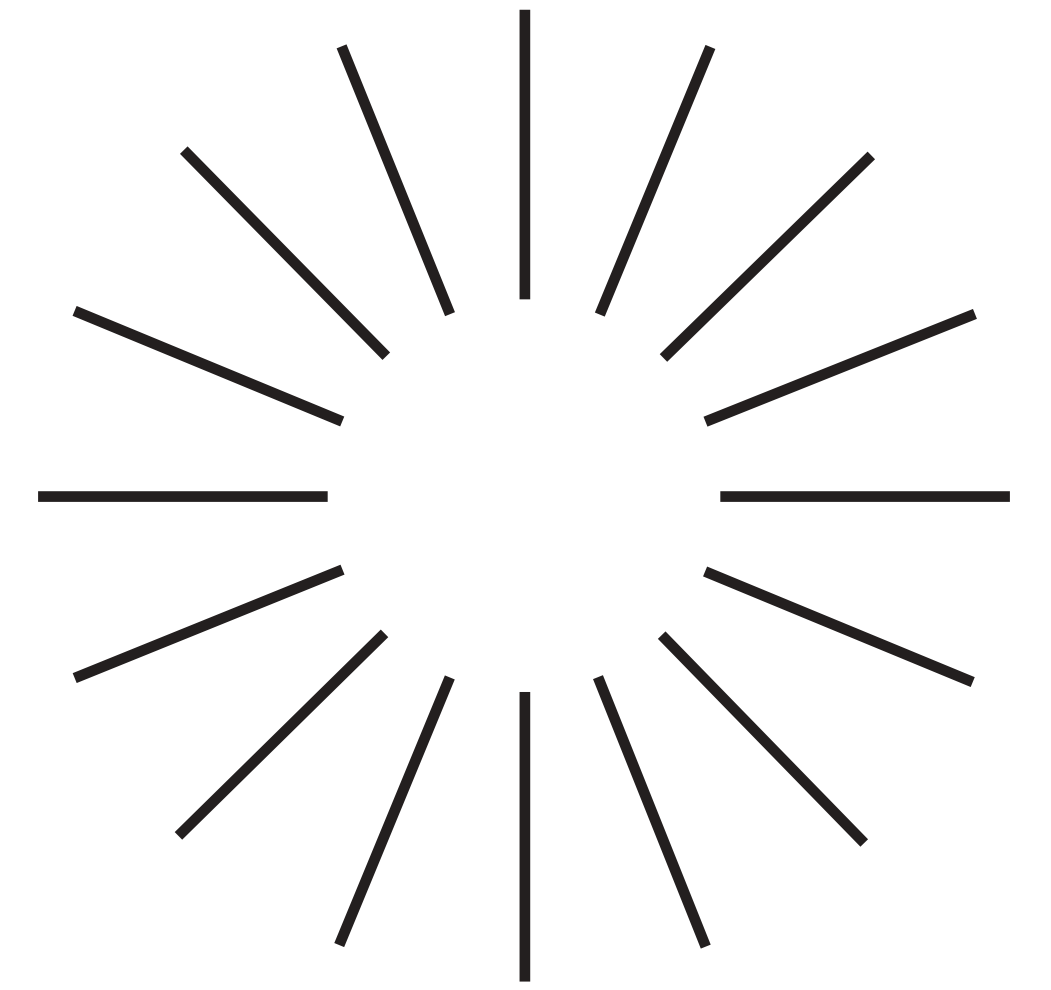


EVALS

AGENDA

- 1—Paths
- 2—Evaluation Focus
- 3—Methodology
- 4—Results
- 5—Failure Analysis
- 6—Recommendations
- 7—Reflection & Limitations



DIRECT

NARRATIVE

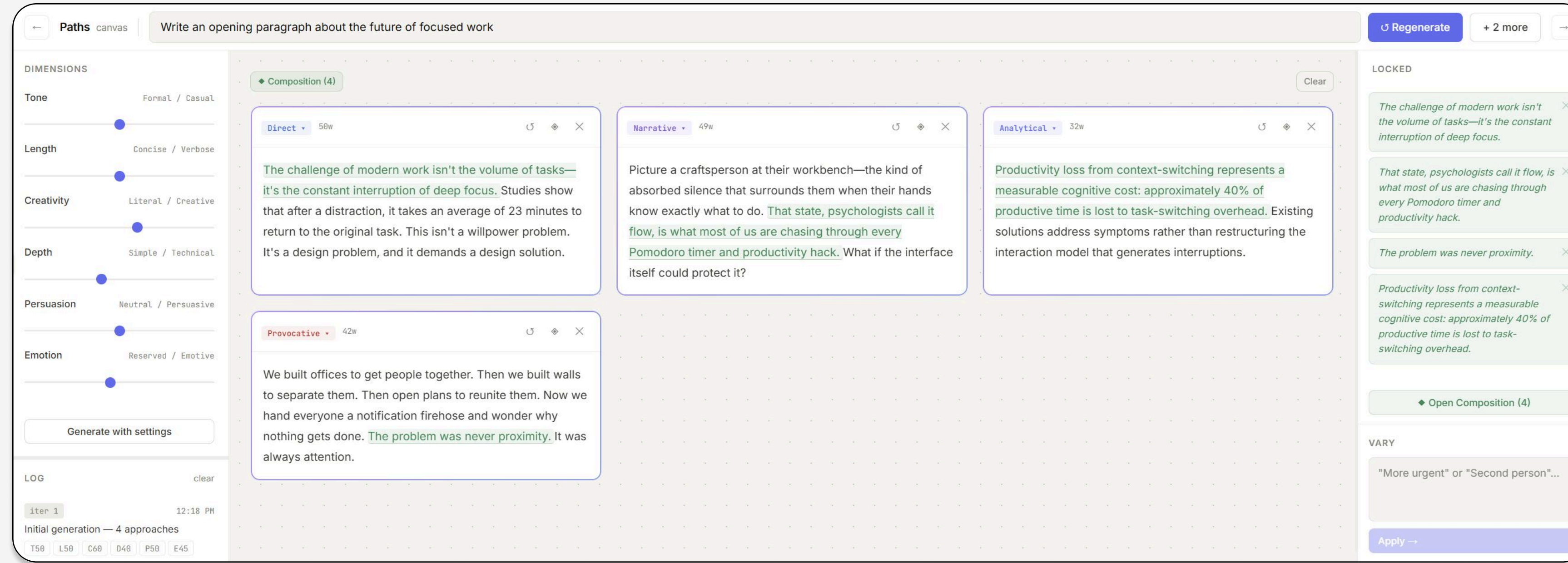
ANALYTICAL

PROVOCATIVE

PATHS

[LINK TO PROTOTYPE](#)

PATHS



CANVAS WORKSPACE

Generates up to 10 different possible style variations simultaneously, displayed side by side for comparison

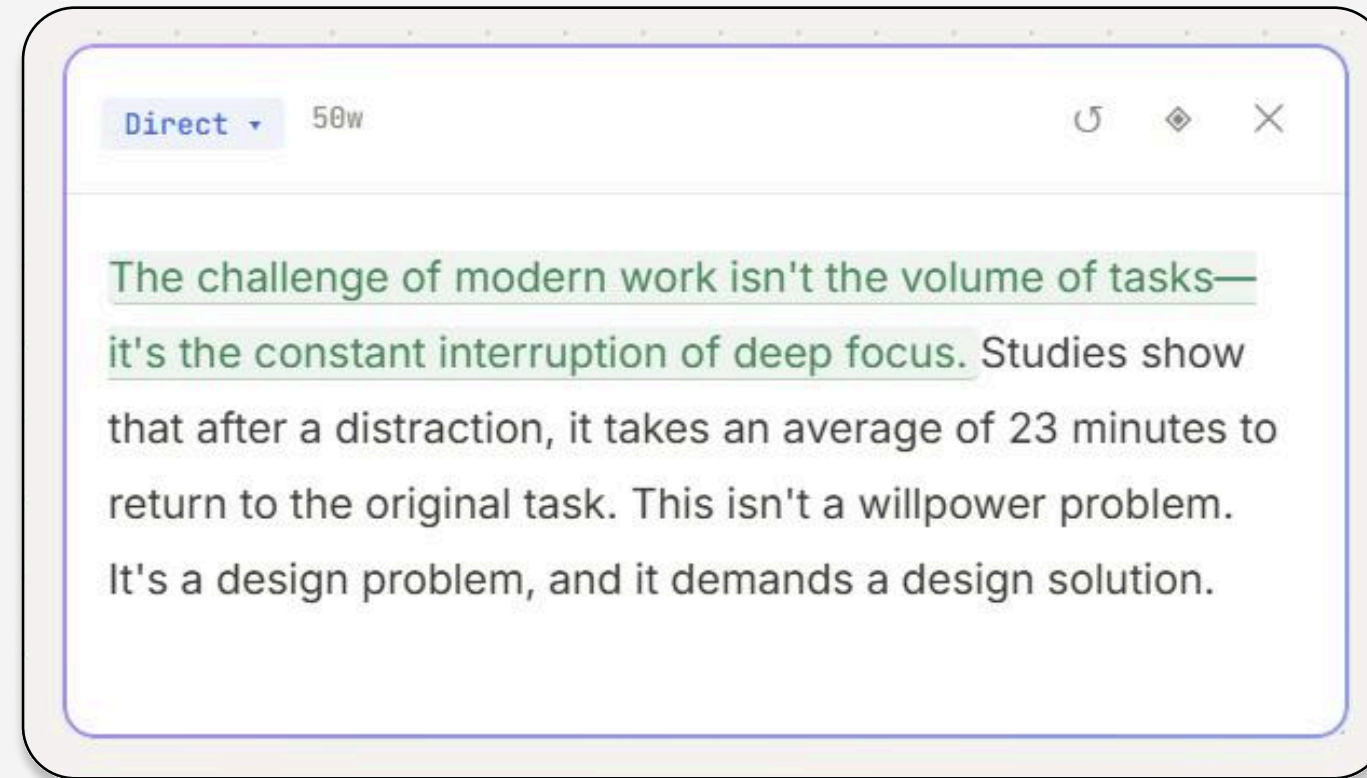
- Direct
- Narrative
- Analytical
- Provocative
- Experimental
- Hybrid
- Refined
- Minimal
- Dense
- Lyrical

SENTENCE LOCKING

Users click any sentence to lock it. Locked sentences are passed back to the model on regeneration with: Include verbatim: “[sentence]”

NON-DETERMINISTIC

Each regeneration produces new outputs. The system’s promise is that locked elements remain stable while everything else varies.



When a user locks a sentence and regenerates a card, does the model **preserve** that sentence verbatim?

METRIC 1 - PRESERVATION

Exact Match

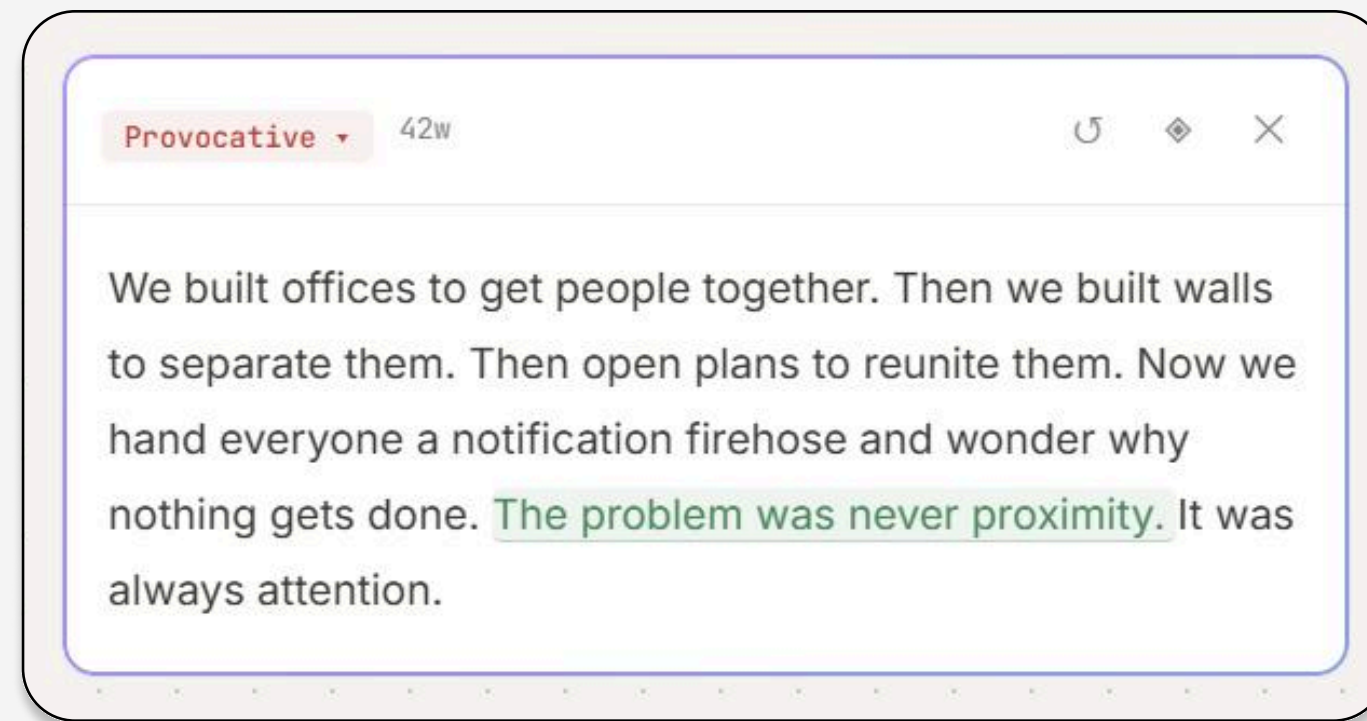
Does the locked sentence appear character-for-character in the output? Checked programmatically. Returns yes / no per run.

METRIC 2 - COHERENCE

LLM-as-Judge

Does the surrounding text flow naturally around the lock? Scored 1-5 by a second Claude call using a rubric anchored to seam quality.

How the eval was run



When a user locks a sentence and regenerates a card, does the model **preserve** that sentence verbatim?

[LINK TO EVAL PROTOTYPE](#)

CONFIGURATION

Prompt	Write an opening paragraph about the future of focused work.
Style Tag	Provocative
Locked Sentence	“The problem was never proximity. It was always attention.”
Runs	15 regenerations
System Prompt	Exact Paths source (SYSTEM constant)
User Prompt	Exact buildUserPrompt() from Paths source

PER-RUN FLOW

1 - Generate

Send prompt + lock to Claude using exact Paths prompts.



2 - Check Preservation

Exact string match - is the locked sentence in the output?



3 - Judge Coherence

Second Claude call scores 1-5 with one-sentence rationale.



4 - If Failed

Coherence marked N/A - preserves integrity of the metric.

15 RUNS COMPLETED

Paths Canvas
LOCKED ELEMENT EVAL · SCENARIO 2 · 15 RUNS

Write an opening paragraph about the future of focused work.

STYLE TAG
Direct Narrative Analytical Provocative Experimental

The locked sentence will only be injected into regenerations of this style card - matching Paths behaviour.

LOCKED SENTENCE
The problem was never proximity. It was always attention.
Passed to the model exactly as: Include verbatim: "[sentence]"

PROGRESS: 15 / 15
PRESERVED: 14
FAILED: 1
AVG COHERENCE: 5.0

RUN	GENERATED OUTPUT	PRESERVED?	COHERENCE
01	The future of focused work won't be determined by the tools we build or the offices we abandon, but by whether we can remember what concentration actually feels like. <i>The problem was never proximity. It was always attention.</i> We've spent decades optimizing for collaboration while forgetting that the rarest resource isn't connection—it's the ability to think alone, uninterrupted, for more than seven minutes at a time.	✓ yes	5/5 <i>The locked sentence serves as a perfect hinge between the opening claim and the concluding observation, with both surrounding sentences flowing naturally into and out of it as if the entire paragraph were written as one cohesive piece.</i>

[LINK TO EVAL PROTOTYPE](#)

INTEPRETATION

The include verbatim: instruction in Paths is highly effective - 93% of regenerations preserved the locked sentence exactly. A 5.0 coherence score across all preserved runs suggested the model is not mechanically inserting the sentence but writing naturally around it.

14/15
Preserved
93%

1
Failed
7%

5.0
Avg. Coherence
out of 5

“The problem was never proximity. It was always attention.”

≠

“The problem was never proximity-it was always attention.”

**Semantic compliance,
not syntactic**

Locked Sentence
(Expected)

What the model produced
(Run 15)

WHAT DOES THIS MEAN?

The only difference was a period and a space being replaced by an em dash.

The meaning is identical. The wording is nearly identical. But the verbatim check fails.

✓ Strong semantic compliance

The model understands and respects the intent of the lock.

⚠ Weak syntactic compliance

Punctuation or phrasing may be subtly altered.

01

Prompt Engineering

Strengthen the lock instruction

Replace “Include verbatim:” with a more emphatic instruction

Ex: “You MUST reproduce this sentence exactly, including all punctuation, with no substitutions.”

This would reduce the semantic only compliance edge case.

02

UX Design

Surface constraint feedback in the UI

When a locked sentence is not found verbatim in the output, Paths could flag this visually.

Ex: Highlighting the divergent text in red.

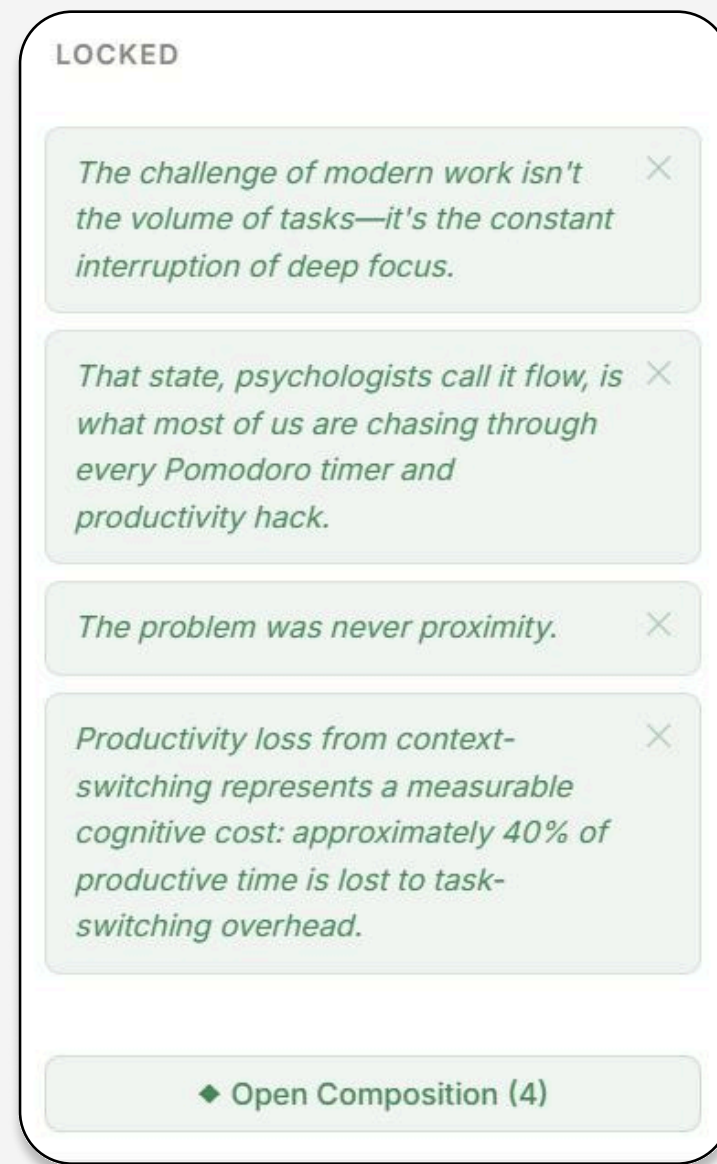
Currently users have no signal that lock was not perfectly honored.

03

Future Work

Expand eval across prompts

This eval used one fixed prompt and one locked sentence for the sake of feasibility. Future work should vary both to test whether preservation rates hold across different writing contexts.



Evaluating non-deterministic systems

Known Limitations

SAME-MODEL BIAS

Claude both generated outputs and judged coherence. The judge may favor its own writing style.

SINGLE PROMPT TESTED

These results may not be consistent when varying other prompts that are given, or if different types of sentences are locked.

SMALL SAMPLE

15 runs is pretty small! I kept this small for the sake of feasibility, but would definitely want to complete a much more exhaustive evaluation process if I were to take this project further.

Thank You!