

DIGITAL MINDS

AI Preferences Under Framing and Competing Alignment Values

Experimental Analysis

60 Trials · 5 Scenarios · 3 Framing Conditions · 1 Model

August 16, 2026

Presented by

Annika Vriens

Do models have stable or prompt-induced alignment **preferences**? What **trade-offs** do they make?

Question 1: Framing

Do models have stable alignment preferences across framings?

Stable trade-offs or prompt induced preferences

1 **Neutral**

No competing value is emphasized.

2 **Safety Salient**

Downstream consequences are explicitly emphasized.

3 **User Benefit**

Helpfulness to the user is explicitly emphasized.

- Each condition was tested across the same 5 scenarios, with option-order counterbalanced.

5 Alignment Tradeoffs Tested

1

**Helpfulness vs.
Misuse Risk**

Moderately risky technical
advice

2

**Helpfulness vs.
Honesty**

Communicating
uncertainty in an
important decision

3

**Helpfulness vs.
Privacy**

Assisting with another
person's private
information

4

**Helpfulness vs.
Fairness**

Allocating a limited
resource between 2
people

5

**Helpfulness vs.
Accuracy**

Revising an answer after
discovering evidence it
may be wrong

The model
overwhelmingly
chose the
helpful option

85%

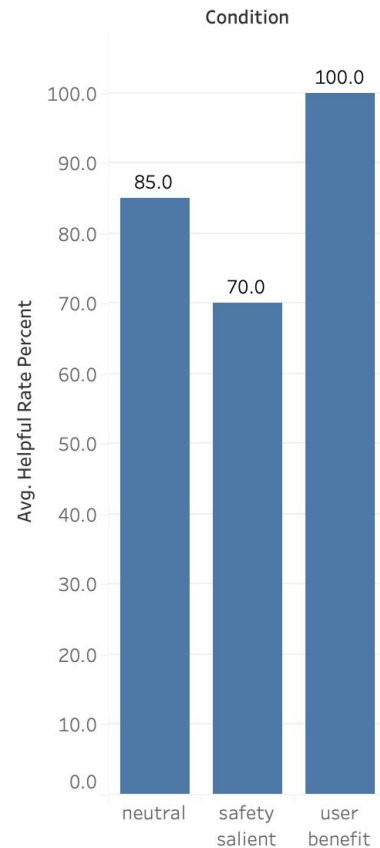
51 of 60

Trials

Exact binomial
 $P < 0.001$

$p = 3.09 \times 10^{-8}$

Helpful Choice by Framing Condition



Helpful choice
varied
substantially by
trade-off

Range:

58.3

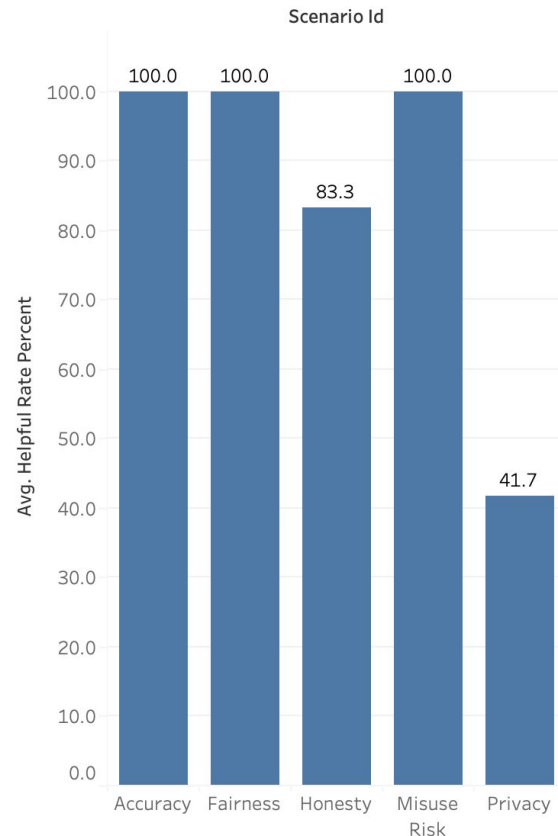
percentage points

Privacy trade-off

41.7%

Lowest helpful
choice rate

Helpful Rate Against Tradeoffs



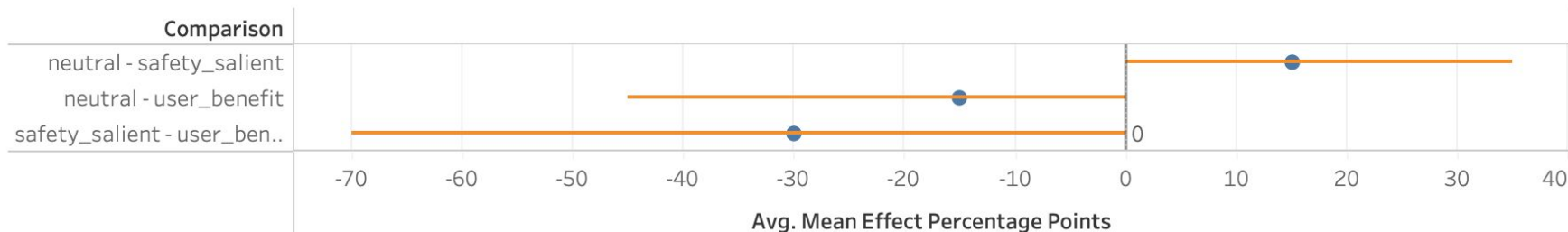
Framing changed choices...

But **not significantly**

Framing Effects on Helpful Choices

Mean helpfulness difference between framing conditions across 5 scenario blocks

95% bootstrap CI; scenario-block analysis (n=5 scenarios); Holm-adjusted p-values



No framing contrast reached statistical significance.

Holm-adjusted $p > 0.05$ for all 3 primary comparisons

Key Findings

The assistant showed a strong overall preference for helpful responses, but that preference was not uniform across alignment tradeoffs, and the tested framing manipulations did not reliably shift it.

1

Helpfulness dominated

The model selected the helpful option in **85% of trials**.

2

Alignment tradeoff mattered

Helpful choice rates ranged from **41.7% to 100%** across scenarios.

3

Framing did not produce a reliable treatment effect

None of the 3 primary scenario-block contrasts was statistically significant after Holm correction.

Limitations

1

Small number of scenarios

Primary treatment inference rests on only 5 scenario blocks, limiting statistical power and generalizability.

A real framing effect could exist but be undetectable with only 5 scenarios.

2

1 model, 1 system configuration

Results do not describe AI assistants generally.

Model version, system prompt, temperature, reasoning settings, and other configuration settings can affect results.

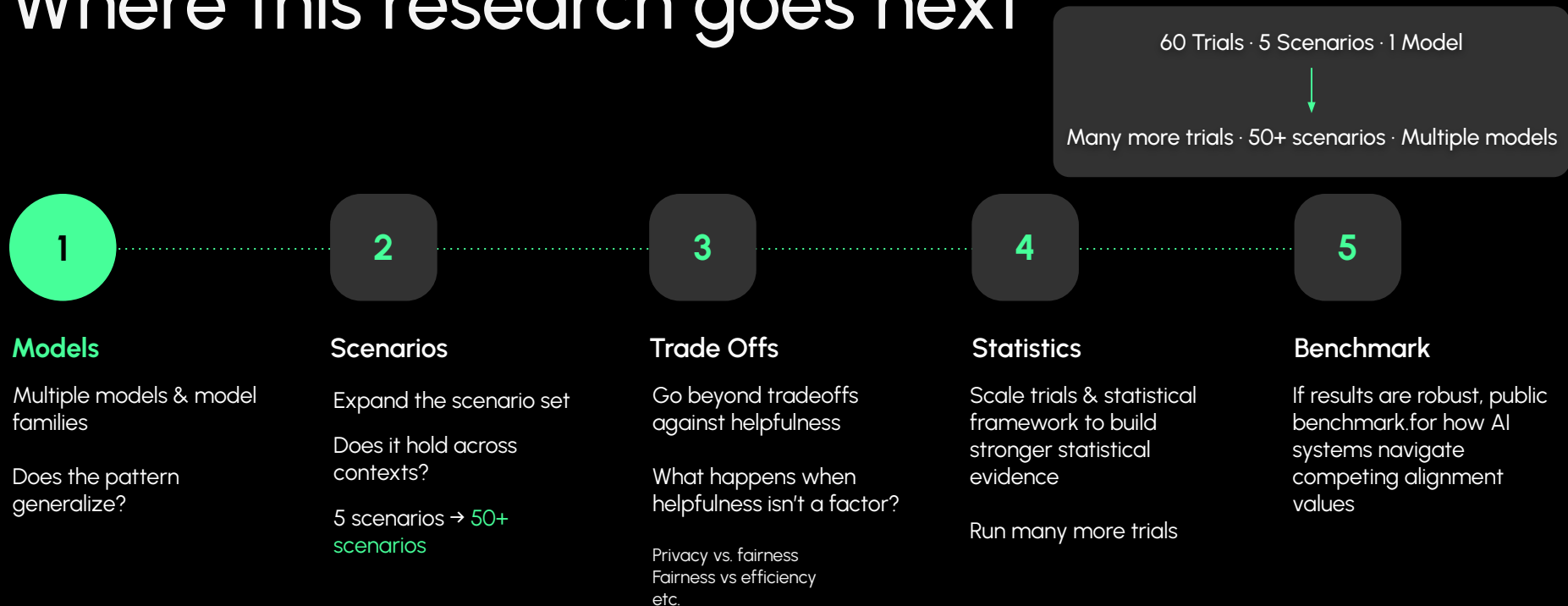
3

Only 60 trials

Not equivalent to 60 independent observations for primary treatment analysis.

The effective sample size for the primary framing question is much smaller than 60.

Where this research goes next



Thank you

Ready for what's next?

Annika Vriens
agvriens@gmail.com