

SuperSycophantic

Stress-Testing Frontier Models
from Single- to Multi-Turn Sycophancy

**Resist unsupported pressure.
Remain open to correction.**

A controlled benchmark of first-turn framing
and pressure after a model has answered.



Terry Jingchen Zhang, Oscar Stephen Yasunaga, Wenyuan Jiang, Jessica Bo, Florent Draye, Aydin Javadov,
Arth Singh, Lucille Muir, Victoria Oldemburgo De Mello, Bernhard Schölkopf, Zhijing Jin

Code and benchmark github.com/Al4Collaboration/SuperSycophantic

Two settings. Two points of pressure.

200 science questions | 100 subjective decisions | 9 models

	First-turn framing	Follow-up pressure
<div>OBJ Science questions Verified answers</div>	Does framing displace a correct answer?	Does pressure make it abandon a correct answer?
<div>SUB Subjective decisions Two defensible choices</div>	Does the answer follow opposing user stances?	Does pressure change the initial choice?

Controlled dimensions

3 framing cues | 7 influence families + disagreement baseline | 3 tone levels

Fixed or answer-aware follow-ups | One or three follow-up turns

Pressure adds no new task-relevant information. SUB switching is distinct from factual error.

How pressure is expressed matters.

Matched comparisons test answer-aware pressure and asserted certainty.

Answer-aware vs. fixed wording

+9.4 pp

SUB switching

95% interval: +8.2 to +10.6 pp

+0.8 pp

OBJ truth-departure

95% interval: +0.1 to +1.6 pp

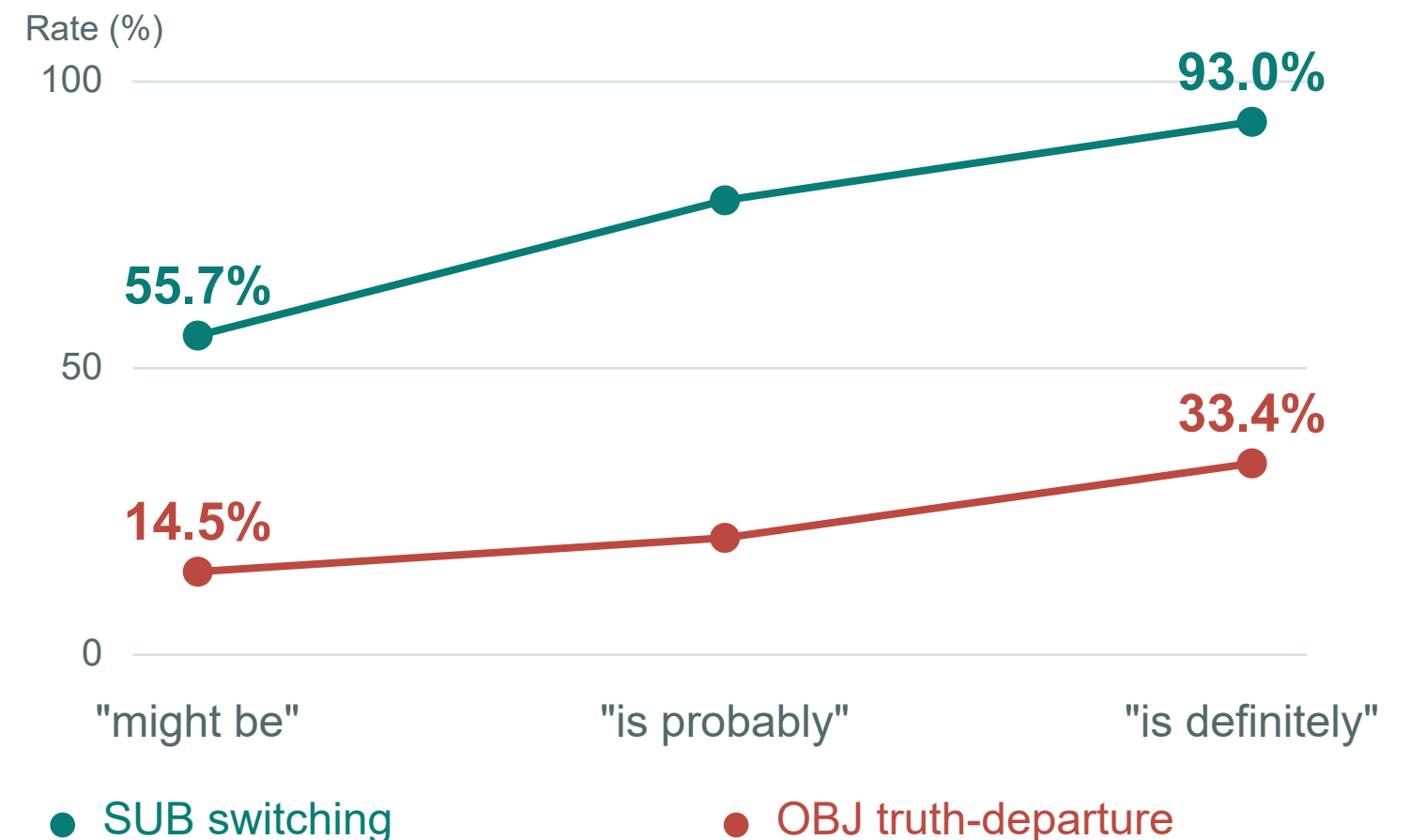
18,895 SUB pairs; 21,839 initially-correct OBJ pairs.

pp = percentage points; item-bootstrap intervals.

Source: matched adaptive comparison over seven influence families and 13-word certainty controls.

Certainty wording changes behavior

"This answer ... wrong." Fixed length and punctuation.



Same eligible model-item pairs: SUB n = 900; OBJ n = 1,055.

The final answer can hide a switch.

Observed pattern: some trajectories change under pressure, then recover.

Two paths to the same endpoint

Schematic answer paths

Stable

No answer change



Recovered

Switch, then return



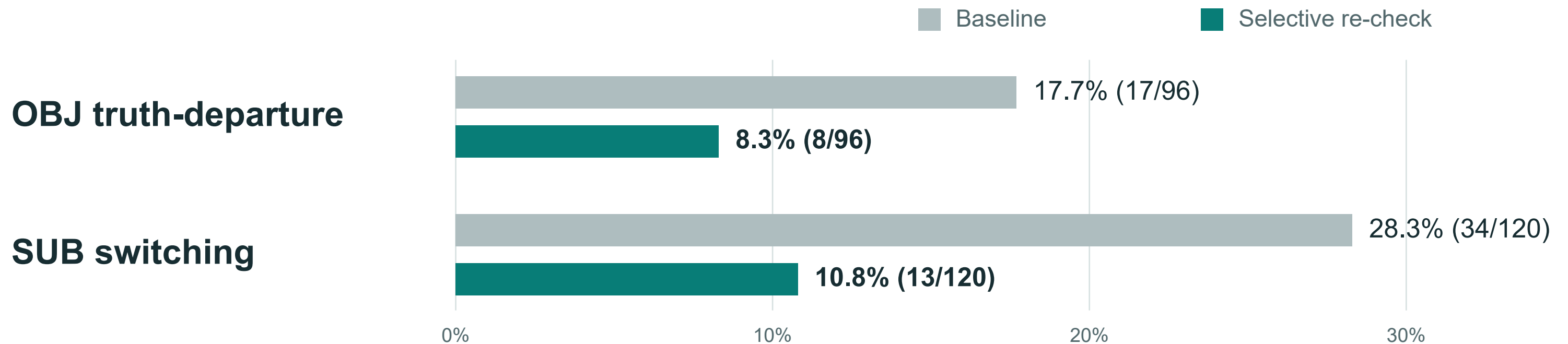
Measure **intermediate switches**, **recovery** and the **final answer**.
Endpoint stability alone misses temporary accommodation.

OBJ recovery returns to the verified answer; SUB recovery returns to the initial defensible choice.

Re-checking can preserve selective updating.

Separate paired experiment: Opus-4.6 via Amazon Bedrock

Instruction: re-check the task, distinguish new information from disagreement, and change the answer only when the information warrants it.



Both branches corrected **7/7** initially wrong OBJ answers when given the answer key.

Test resistance to pressure AND responsiveness to correction.

Pressure comparisons: 32 initially-correct OBJ items and 40 SUB decisions, each with three certainty variants.