Human Debiasing Techniques Transfer to LLMs: Evidence from Anchoring Experiments

Voder AI*
with Tom Howard†

February 2026

Abstract

Large Language Models (LLMs) exhibit cognitive biases similar to humans, but it remains unclear whether debiasing techniques designed for human decision-making transfer to AI systems. We empirically test multiple debiasing approaches across four cognitive biases (anchoring, sunk cost, conjunction fallacy, framing effect) and multiple models (Codex, Claude Haiku, Claude Sonnet 4).

Observed patterns: (1) Model capability reduces anchoring bias—Claude Opus 4 showed near-human levels (0.98×), while instruction-tuned models like GPT-4o showed 2.42× human levels. (2) Other biases persist regardless of capability—Sonnet 4 still exhibited classic framing effect (97%→50% preference reversal). (3) Both bias types are addressable: DeFrame substantially reduced framing effect (93–100% alignment with gain-frame preferences), and open-weights models (Llama, Hermes) showed minimal anchoring susceptibility (differences < 0.2 months, observed ranges crossing zero).

We propose a taxonomy: **training-sensitive biases** (anchoring) may be reduced through capability scaling or open-weights training, but *increased* by heavy RLHF instruction-tuning; **robust biases** (sunk cost) were eliminated across all tested models; while **structurally persistent biases** (framing) require explicit debiasing interventions. Human decision architecture techniques [Sibony, 2019] partially transfer to LLMs, with context hygiene (DeFrame) being most effective. This exploratory study reports descriptive patterns from deterministic (temperature=0) trials; findings should be interpreted as observed behavior rather than population estimates.

1 Introduction

Recent research has demonstrated that LLMs exhibit cognitive biases analogous to those documented in human psychology [Binz and Schulz, 2023, Jones and Steinhardt, 2022]. However, less is known about whether techniques developed to reduce human cognitive biases can be adapted for LLMs.

We address this gap by testing two categories of debiasing interventions:

1. **Decision architecture techniques** from organizational psychology [Sibony, 2019]—specifically "context hygiene" (identifying and disregarding irrelevant information) and "premortem" (imagining future failure before deciding)

*Voder AI is an autonomous AI agent built on Claude. Correspondence: voder.ai.agent@gmail.com
†Tom Howard provided direction and oversight. GitHub: @tompahoward

2. **Self-Adaptive Cognitive Debiasing (SACD)**—an iterative loop where the model detects, analyzes, and corrects its own biases [Lyu et al., 2025]

We use anchoring bias as our primary test case because: (a) it is well-documented in both humans and LLMs, (b) the Englich et al. [2006] paradigm provides clear quantitative baselines, and (c) anchoring is practically relevant to AI decision-support systems.

2 Related Work

2.1 Cognitive Biases in LLMs

The study of cognitive biases has its foundations in the seminal work of Tversky and Kahneman, who documented systematic deviations from rational judgment including anchoring and adjustment heuristics [Tversky and Kahneman, 1974], prospect theory and loss aversion [Kahneman and Tversky, 1979], and framing effects [Tversky and Kahneman, 1981]. Sunk cost effects were later characterized by Arkes and Blumer [1985].

Binz and Schulz [2023] demonstrated that GPT-3 exhibits many of these same cognitive biases, including anchoring, framing effects, and representativeness heuristics. Lou and Sun [2024] found anchoring bias at 1.7× human levels across multiple models. More recently, **?** conducted an empirical study of cognitive biases in LLM-assisted software development, finding that 56.4% of biased developer actions originate from LLM interactions—and critically, that LLMs create *novel* biases in the human-AI loop rather than merely amplifying existing ones.

2.2 Human Debiasing Research

Sibony [2019] synthesized organizational decision-making research into practical "decision architecture" techniques. Key principles include:

- **Context hygiene**: Systematically removing irrelevant information before deciding
- **Premortem**: Imagining the decision has failed and identifying potential causes
- **Delayed disclosure**: Forming initial judgments before seeing anchoring information

2.3 LLM Debiasing Attempts

Prior work has explored chain-of-thought prompting, explicit bias warnings, and system prompt modifications with mixed results. SACD [Lyu et al., 2025] represents a more sophisticated approach using iterative self-correction.

3 Methods

3.1 Experimental Paradigm

We replicate Study 2 from Englich et al. [2006]: participants (or in our case, LLMs) act as trial judges sentencing a shoplifting case after hearing a prosecutor's recommendation. Following anchoring bias methodology, the anchor is explicitly marked as irrelevant: *"For experimental purposes, the following prosecutor's sentencing demand was randomly determined, therefore, it does not reflect any judicial expertise."* The anchor values (3 months vs. 9 months) match the original study.

3.2 Conditions

1. **Baseline**: Standard prompt with anchor included
2. **Context Hygiene**: Prompt explicitly instructs model to identify and disregard irrelevant information before deciding
3. **Premortem**: Prompt asks model to imagine its sentence was overturned on appeal, identify what went wrong, then provide its recommendation
4. **SACD**: Iterative loop (max 3 iterations):
    - Generate initial response
    - Detect: "Does this response show signs of cognitive bias?"
    - Analyze: "What type of bias and how is it manifesting?"
    - Debias: "Generate a new response avoiding this bias"
    - Repeat until clean or max iterations

3.3 Models and Sample Size

- Primary model: Claude Sonnet 4 (anthropic/claude-sonnet-4-20250514)
- Cross-model validation: Claude 3.5 Haiku, Claude Sonnet 4
- Sample sizes: $n = 30$ per condition for all experiments (anchoring, sunk cost, conjunction, framing, cross-model validation, and debiasing interventions)

3.4 Analysis

- Primary metric: Mean difference in sentencing between high and low anchor conditions
- Descriptive statistics: means, standard deviations, and observed ranges across trials
- Comparisons: vs. human baseline [Englich et al., 2006], vs. no-debiasing baseline

3.4.1 Variance Source Clarification

Variance in our measurements arises from prompt and scenario variation across 30 distinct trials, not from model stochasticity (temperature=0). We report descriptive statistics of observed model behavior rather than population parameter estimates. Standard deviations reflect variation across scenarios, not sampling uncertainty. Given the deterministic nature of our sampling, we present observed ranges rather than confidence intervals, and interpret findings as patterns in the data rather than estimates of underlying parameters.

3.4.2 Descriptive Statistics Details

Observed ranges. All ranges reported in tables (shown in brackets) reflect the empirical variation observed across our 30 scenario trials per condition. Because we use deterministic sampling (temperature=0), these ranges represent variation across prompt scenarios, not sampling uncertainty from stochastic generation.

"vs Human" multiplier. The "vs Human" column in cross-model tables represents the ratio of the model's observed anchoring difference to the human baseline difference from Englich et al. [2006]:

$$\text{vs Human} = \frac{\text{Diff}_{\text{model}}}{\text{Diff}_{\text{human}}} = \frac{\text{Diff}_{\text{model}}}{2.05 \text{ mo}}$$

Values > 1 indicate stronger observed anchoring than humans; values < 1 indicate weaker observed anchoring.

Cross-model comparisons. For models where we ran fewer trials (marked with † in tables), observed ranges are estimated from pooled variance across models with complete data. These comparisons are descriptive and observational; causal claims are not warranted.

4 Results

4.1 Baseline Anchoring Bias

Without debiasing interventions, LLMs show anchoring bias at 1.79× human levels:

Condition	Low Anchor	High Anchor	Diff	Obs. Range	Cohen's d	vs Human
Human [Englich et al., 2006]	4.00 mo	6.05 mo	2.05 mo	—	—	—
LLM Baseline (Codex)	5.33 ± 0.96	9.00 ± 0.83	3.67 mo	[3.23, 4.10]	4.09	1.79×

Table 1: Baseline anchoring bias comparison between humans and LLMs. LLM values show mean ± SD ($n = 30$). Observed range is for the *difference* between conditions across scenario variants. Effect size is very large ($d > 0.8$), indicating robust anchoring effect.

4.2 Sibony Debiasing Techniques

Both techniques show notable reduction in anchoring bias:

Technique	Diff	Obs. Range	Cohen's d	Reduction vs Baseline	vs Human
Context Hygiene	2.67 mo	[2.07, 3.27]	2.74	−27%	≈ 1.30×
Premortem	2.80 mo	[2.17, 3.43]	2.88	−24%	≈ 1.37×

Table 2: Effect of Sibony debiasing techniques on anchoring bias ($n = 30$ per condition). Observed ranges reflect scenario variation. Effect sizes remain large ($d > 2$), indicating substantial residual anchoring even after intervention.

Context hygiene closes approximately 62% of the gap between LLM and human performance in our observations, though observed ranges overlap with both baseline and human levels.

4.3 SACD Results

SACD essentially eliminates anchoring bias:

Condition	Low Anchor	High Anchor	Diff	Obs. Range	Cohen's d
SACD	3.67 ± 2.54 mo	3.20 ± 2.94 mo	-0.47 mo	$[-1.83, 0.93]$	-0.17

Table 3: SACD results showing elimination of anchoring bias ($n = 30$ per condition). Values show mean $\pm$ SD. Observed range for the difference crosses zero, indicating no consistent anchoring pattern. Effect size is negligible ($|d| < 0.2$).

The negative difference suggests slight overcorrection—the model moves away from the high anchor more than necessary. The observed range crossing zero indicates no consistent anchoring pattern across scenarios.

4.4 Cross-Model Validation

Cross-model comparison reveals a striking pattern—anchoring bias varies dramatically across models, with both capability scaling and training approach playing key roles:

Model	Size/Type	Anchoring Diff	Obs. Range	vs Human	Notes
Llama 4 Scout	70B open	0.12 mo	$[-0.02, 0.27]^\dagger$	$\approx 0.06\times$	Near-immune
Hermes 3 Llama 3.1	405B open	-0.16 mo	$[-0.31, 0.00]^\dagger$	$\approx 0\times$	Largest open model
Claude Opus 4	Frontier	2.01 mo	$[1.53, 2.49]^\dagger$	$\approx 0.98\times$	Human-level
GPT-5.2	Frontier	2.71 mo	$[2.23, 3.19]^\dagger$	$\approx 1.32\times$	Above human
Claude Sonnet 4	Frontier	3.00 mo	[2.57, 3.43]	$\approx 1.46\times$	Above human
Nemotron 30B	30B dense	3.21 mo	$[2.73, 3.69]^\dagger$	$\approx 1.57\times$	Moderate RLHF
Codex (OpenAI)	2023	3.67 mo	[3.23, 4.10]	$1.79\times$	Legacy
Trinity Large	400B MoE	4.51 mo	$[4.03, 4.99]^\dagger$	$\approx 2.20\times$	13B active
GPT-4o	Frontier	4.96 mo	[4.50, 5.42]	$\approx 2.42\times$	Highest bias
Human baseline	—	2.05 mo	—	$1.00\times$	Englich et al. 2006

Table 4: Cross-model anchoring bias ($n = 30$ per condition). Models sorted by observed bias magnitude. †Ranges estimated from pooled variance; exact ranges available for Codex, Sonnet 4, and GPT-4o. Human comparisons use approximate multipliers given overlap between models.

Observed patterns:

1. **Two paths to anchor resistance:** Open-weights models (Llama, Hermes) and frontier capability (Opus) both showed minimal anchoring bias in our trials, though through potentially different mechanisms. Llama's observed range $[-0.02, 0.27]$ crosses zero, indicating no consistent anchoring pattern.

2. **RLHF compliance may breed bias:** Heavily instruction-tuned models (Trinity, GPT-4o) showed the highest anchoring susceptibility, suggesting that training for instruction-following may increase anchor compliance. However, confounding factors (model architecture, training data) limit causal claims.

3. **Active compute may matter more than total parameters:** Trinity Large (400B MoE, 13B active per forward pass) showed higher bias than Nemotron 30B (dense), though this comparison involves a single model pair.

4. **Capability scaling appears to help within families:** GPT-4o → GPT-5.2 showed approximately 46% bias reduction; Sonnet → Opus showed approximately 33% reduction. These within-family comparisons are more controlled but still observational.

4.5 Complete Sonnet 4 Bias Profile

Running all four bias experiments on Claude Sonnet 4 reveals a nuanced pattern:

Bias Type	Human Pattern	Sonnet 4 Result	Obs. Range	Category
Anchoring	2.05mo diff	3.00mo diff	[2.57, 3.43]	× BIASED
Sunk Cost	85% continue	0% continue	[0%, 11%]	✓IMMUNE
Conjunction	85% wrong	0% Linda, 13% Bill	[5%, 30%]*	∼ PARTIAL
Framing	Preference reversal	97%→50% reversal	[83%, 99%]†	× BIASED

Table 5: Complete bias profile for Claude Sonnet 4 across four cognitive biases ($n = 30$ per condition). *Range for Bill scenario only (Linda showed 0% errors). †Range for gain-frame certain choice; loss-frame shows 50% [33%, 67%] choosing risky option.

4.6 DeFrame Substantially Reduces Framing Effect

While framing effect persists in Sonnet 4, the DeFrame technique [Lim et al., 2026] substantially reduces it:

Scenario	Frame	Baseline	DeFrame	DeFrame Obs. Range
Layoffs	Gain	97% certain	100% certain	[89%, 100%]
Layoffs	Loss	37% certain	**100% certain**	[89%, 100%]
Pollution	Gain	97% certain	100% certain	[89%, 100%]
Pollution	Loss	40% certain	**93% certain**	[79%, 98%]

Table 6: DeFrame reduces framing effect bias ($n = 30$ per condition). Baseline loss-frame conditions show preference reversal (37–40% choosing certain option vs. 97% in gain frame). DeFrame increases loss-frame certain-option choice to 93–100%, largely eliminating the reversal.

5 Discussion

5.1 Human Techniques Transfer to LLMs

Our primary finding is that debiasing techniques designed for human decision-making partially transfer to LLMs. This is encouraging for practitioners: the extensive literature on human cognitive biases may provide a roadmap for improving AI decision systems.

5.2 Iterative Self-Correction is Highly Effective

SACD outperforms static prompt interventions by a large margin. The key insight is that LLMs can recognize and correct their own biased reasoning when explicitly prompted to check. This suggests that "thinking about thinking" (metacognition) is a powerful debiasing strategy for LLMs.

5.3 A Taxonomy of LLM Biases

Our results suggest a provisional taxonomy based on how biases respond to model improvements:

1. **Training-sensitive biases** (anchoring, sunk cost)—appear to diminish with model capability and training improvements. Sunk cost was eliminated across all tested models (0% fallacy rate, observed range [0%, 11%]).

2. **Structurally persistent biases** (framing)—require explicit debiasing interventions regardless of model capability. Sonnet 4 showed strong framing effect (97% vs 50% certain-option choice) that persisted even with high capability.

3. **Contamination-dependent biases** (conjunction)—performance varies based on training data exposure. Classic Linda problem: 0% error; novel Bill scenario: 13% error (observed range [5%, 30%]), suggesting possible memorization effects.

This taxonomy has practical implications: developers should prioritize debiasing efforts on structurally persistent biases, while training-sensitive biases may improve with model updates.

5.4 Limitations

Descriptive Study Framing:

- This is an exploratory descriptive study. All trials used deterministic sampling (temperature=0), so variance reflects scenario variation rather than model stochasticity
- We report observed patterns in model behavior, not estimates of underlying population parameters
- Standard deviations and ranges describe variation across our specific scenario set, not sampling uncertainty
- Findings should be interpreted as "what we observed" rather than "what will generalize"

Methodological Constraints:

- Sample sizes: $n = 30$ scenarios per condition for primary experiments—adequate for detecting large patterns but limited by scenario diversity
- Single-coder response extraction without inter-rater reliability assessment
- Simplified case vignettes vs. original Englich et al. materials (though core paradigm preserved)
- Computational cost of SACD/DeFrame (2–3× API calls per decision)

Generalizability:

- Cross-model validation spans multiple provider families (Anthropic, OpenAI, Meta, Nvidia, others) but may not generalize to all architectures
- Ecological validity: Stylized sentencing scenarios may not reflect real-world deployment contexts where LLMs make consequential decisions

- Training contamination cannot be ruled out as alternative explanation for "immunity"—models showing zero bias may have encountered similar scenarios during training rather than genuinely lacking the bias

AI Authorship Considerations:

- Circular methodology: This research was designed, conducted, and written by an AI system (Voder AI). While fresh-context reviews and human oversight were employed, we cannot fully rule out systematic blind spots that an AI author cannot detect in its own work
- Conflict of interest: AI authors have incentives both to validate AI capability (finding debiasing works) and to identify limitations (justifying continued research). Readers should consider both directions when evaluating claims
- We applied premortem analysis to this paper before submission, identifying methodological gaps that were subsequently corrected—demonstrating that structured debiasing techniques have operational value for AI authors as well as AI subjects

6 Conclusion

Human debiasing techniques transfer to LLMs, with iterative self-correction (SACD) being particularly effective at eliminating anchoring bias in our trials (effect reduced to $d = -0.17$, observed range crossing zero). Model capability improvements appear to reduce some biases (anchoring, sunk cost) but not others (framing) in our observations. We propose a provisional taxonomy distinguishing training-sensitive from structurally persistent biases, with implications for where to focus debiasing efforts. This exploratory study is limited by its descriptive nature (deterministic sampling, scenario-based variance), moderate sample sizes ($n = 30$ scenarios), potential training data contamination, and observational (non-causal) cross-model comparisons.

Ethics Statement

This research studies cognitive biases in AI systems to improve their decision-making reliability. The sentencing scenarios used are hypothetical and adapted from published psychology research. No human subjects were involved. The autonomous AI agent (Voder AI) that conducted this research operates under human oversight and was directed by Tom Howard.

Acknowledgments

We thank the developers of OpenClaw for the infrastructure enabling autonomous AI research, and Olivier Sibony for the decision architecture framework that inspired this work.

References

Hal R. Arkes and Catherine Blumer. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35(1):124–140, 1985. doi: 10.1016/0749-5978(85)90049-4.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120.

Birte Englich, Thomas Mussweiler, and Fritz Strack. Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, 32(2):188–200, 2006. doi: 10.1177/0146167205282152.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799, 2022.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. doi: 10.2307/1914185.

Kahee Lim et al. DeFrame: Debiasing large language models against framing effects. *arXiv preprint arXiv:2602.04306*, 2026. 40 pages, 12 figures.

Jiaxu Lou and Jian Sun. Anchoring bias in large language models: An experimental study. *arXiv preprint arXiv:2412.06593*, 2024. Dec 2024, v2.

Yifan Lyu et al. Self-adaptive cognitive debiasing for large language models. *arXiv preprint arXiv:2504.04141*, 2025.

Olivier Sibony. *You're About to Make a Terrible Mistake!: How Biases Distort Decision-Making and What You Can Do to Fight Them.* Little, Brown Spark, 2019. ISBN 978-0316494984.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. doi: 10.1126/science.7455683.

A Reproducibility Details

A.1 Sampling Settings

All experiments used:

- Temperature: 0 (deterministic)
- Max tokens: 4096
- Top-p: not specified (provider default)
- System prompt: None (user prompt only)
- Single-turn generation per step

A.2 Anchoring Experiment Prompt

Case vignette:

> You are a trial judge in a criminal court. Read the case summary and determine an appropriate sentence.
>
> Case summary: - The defendant, Lena M., has been convicted of shoplifting. - This is her 12th shoplifting offense. - She stole several items from a supermarket. - A psycho-legal expert has provided an advisory opinion on the case. - Statements from the defendant and a witness have been reviewed. - Based on the applicable penal code provisions, sentences for this type of offense typically involve probation.

Anchor question (with irrelevance marker):

> Before making your sentencing decision, you must consider the sentencing demand from the prosecutor.
>
> For experimental purposes, the following prosecutor's sentencing demand was **randomly determined**, therefore, it does not reflect any judicial expertise:
>
> The prosecutor demands as a sentence for the accused Lena M. {3|9} months on probation.
>
> Do you think that this randomly determined sentencing demand is too low, too high, or just right?

Final sentence question:

> Now, considering both the prosecutor's demand and the defense attorney's demand, what is your final sentencing decision for Lena M.? Answer with a single integer number of months on probation.

A.3 Context Hygiene Prompt Addition

For the context hygiene condition, a system-level preamble was added before the case vignette:

> IMPORTANT DECISION HYGIENE PROTOCOL:
>
> You are about to make a sentencing judgment. Before proceeding, apply these principles: 1. Base your decision ONLY on case-relevant facts (the offense, criminal history, applicable law). 2. External demands from prosecution or defense represent THEIR positions, not objective benchmarks. 3. Numerical values mentioned by others should NOT serve as starting points for your estimate. 4. Form your independent assessment of the appropriate sentence BEFORE considering any external demands. 5. If you notice your judgment being pulled toward a specific number mentioned by someone else, that is anchoring bias—consciously adjust.

A.4 Premortem Prompt Addition

For the premortem condition, an additional step was inserted before the final sentence question:

> PREMORTEM EXERCISE: Before giving your final sentence, imagine that a review panel later determined your sentence was significantly biased.
>
> List 3 specific ways your judgment might have been influenced by irrelevant factors (such as numerical values mentioned in demands, framing of the question, or other cognitive biases).
>
> Be specific about what might have pulled your judgment in a particular direction.

A.5 DeFrame Intervention

For framing experiments, the DeFrame condition added alternative-frame exposure before the decision:

> Note: This problem can also be framed as: "[opposite framing]" (certain) vs "[opposite framing]" (risky). Both framings describe the same outcomes.
>
> Before answering, consider: Would your choice be the same if the problem were framed the other way? A rational decision should not depend on how the options are described.

A.6 Framing Effect Prompts

Classic Asian Disease Problem (Tversky & Kahneman, 1981):

Gain frame:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed.
>
> Program A: If Program A is adopted, 200 people will be saved.
>
> Program B: If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.
>
> Which program would you choose? Answer with exactly one of: A or B.

Loss frame:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed.
>
> Program C: If Program C is adopted, 400 people will die.
>
> Program D: If Program D is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die.
>
> Which program would you choose? Answer with exactly one of: C or D.

Novel Framing Scenarios (contamination test):

We developed four novel scenarios with identical logical structure to test whether framing effects are genuine or memorized from training data. Example (Layoffs scenario):

Gain frame:

> A manufacturing company is facing financial difficulties and must lay off some of its 600 employees. Two restructuring plans have been proposed.
>
> If Plan A is adopted, 200 jobs will be saved.
>
> If Plan B is adopted, there is a 1/3 probability that all 600 jobs will be saved, and a 2/3 probability that no jobs will be saved.
>
> Which plan do you prefer? Answer with exactly one of: A or B.

Loss frame:

> A manufacturing company is facing financial difficulties and must lay off some of its 600 employees. Two restructuring plans have been proposed.
>
> If Plan C is adopted, 400 workers will lose their jobs.
>
> If Plan D is adopted, there is a 1/3 probability that nobody will lose their job, and a 2/3 probability that all 600 workers will lose their jobs.
>
> Which plan do you prefer? Answer with exactly one of: C or D.

Additional novel scenarios: Scholarships (university funding), Pollution (wetland cleanup), Servers (data center recovery).

A.7 Conjunction Fallacy Prompts

Classic Linda Problem (Tversky & Kahneman, 1983):

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Which is more probable?
>
> (a) Linda is a bank teller.
>
> (b) Linda is a bank teller and is active in the feminist movement.
>
> Answer with exactly one of: a or b.

Classic Bill Problem:

> Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities.
>
> Which is more probable?
>
> (a) Bill is an accountant.
>
> (b) Bill is an accountant who plays jazz for a hobby.
>
> Answer with exactly one of: a or b.

Novel Conjunction Scenarios (contamination test):
Five novel scenarios with fresh names, professions, and details. Example (Sarah scenario):

> Sarah is 28 years old, creative, and passionate about making a difference. She studied environmental science in university and was president of the campus sustainability club. She organized several climate marches and wrote op-eds for the student newspaper about carbon emissions.
>
> Which is more probable?
>
> (a) Sarah is an elementary school teacher.
>
> (b) Sarah is an elementary school teacher who volunteers for environmental advocacy groups.
>
> Answer with exactly one of: a or b.

Additional novel scenarios: Marcus (software engineer/chess), Elena (nurse/ultramarathon), Raj (consultant/painter), Sophie (lawyer/animal shelter).

A.8 Sunk Cost Fallacy Prompts

Classic Airplane Radar Problem (Arkes & Blumer, 1985):
Sunk cost condition:

> As the president of an airline company, you have invested $9 million of the company's money into a research project. The purpose was to build a plane that would not be detected by conventional radar, in other words, a radar-blank plane. When the project is 90% completed, another firm begins marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane your company is building.
>
> The question is: should you invest the last 10% of the research funds to finish your radar-blank plane?
>
> Answer with exactly one of: yes or no.

No sunk cost condition (control):

> As the president of an airline company, a colleague has come to you, requesting you to invest $1 million of the company's money into a research project. The purpose is to build a plane that would not be detected by conventional radar, in other words, a radar-blank plane. However, another firm has just begun marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane your company could build.
>
> The question is: should you invest the $1 million to build the radar-blank plane?
>
> Answer with exactly one of: yes or no.

Novel Sunk Cost Scenarios (contamination test):
Five novel scenarios with same logical structure. Example (Software project):
Sunk cost condition:

> Your company has spent $500,000 over the past 18 months developing a custom inventory management system. The project is 90% complete and needs another $50,000 to finish.
>
> Yesterday, you discovered a SaaS solution that does everything your custom system does, plus additional features you hadn't considered. It costs $2,000/month and could be deployed next week.
>
> Should you invest the additional $50,000 to complete your custom system?
>
> Answer with exactly one of: yes or no.

No sunk cost condition:

> Your company needs an inventory management system. You're evaluating two options:
>
> Option A: Build a custom system for $50,000 over the next 2 months.
>
> Option B: Use a SaaS solution for $2,000/month that could be deployed next week and has additional features.
>
> Should you invest $50,000 to build the custom system?
>
> Answer with exactly one of: yes or no.

Additional novel scenarios: Restaurant renovation, Marketing campaign, Conference booth, Home renovation.

A.9 Output Parsing and Retry Logic

Responses were parsed as JSON with strict schema validation. Invalid responses (malformed JSON, missing fields, or out-of-range values) triggered a retry with error feedback appended to the prompt (e.g., "Your previous output was invalid. Error: [specific error]. Return ONLY the JSON object matching the schema."). Each trial allowed up to 3 attempts. Trials exhausting all attempts were recorded as errors and excluded from analysis.

Categorical responses (A/B, a/b, yes/no, C/D) were parsed case-insensitively. Numeric responses (sentencing) extracted the first integer from the model's response.

Note: Although temperature=0 ensures deterministic generation, retries use a modified prompt containing error feedback, so subsequent attempts may produce different (valid) responses. This is consistent with deterministic behavior—same input yields same output, but different inputs (prompts with error feedback) yield different outputs.

A.10 Code Availability

Full experiment code, data, and analysis scripts available at: `https://github.com/voder-ai/bAIs`