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).

Key findings: (1) Model capability reduces anchoring bias—Claude Opus 4 shows near-human levels (0.98×), while instruction-tuned models like GPT-4o show 2.42× human levels. (2) Other biases persist regardless of capability—Sonnet 4 still exhibits classic framing effect (90%→80% preference reversal). (3) Both bias types are addressable: DeFrame eliminates framing (100% bias reduction), and open-weights models (Llama, Hermes) show near-immunity to anchoring.

We propose a taxonomy: **training-sensitive biases** (anchoring) can be reduced through capability scaling or open-weights training, but *increased* by heavy RLHF instruction-tuning; **robust biases** (sunk cost) are eliminated across all 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.

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)
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.

*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 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 anchoring experiments (Codex baseline, SACD, Sibony techniques, cross-model validation); $n = 10$ per condition for bias profile experiments (sunk cost, conjunction, framing)

3.4 Analysis

- Primary metric: Mean difference in sentencing between high and low anchor conditions
- Statistical tests: Welch's t-test, effect sizes (Cohen's d, Hedges' g)
- Comparisons: vs. human baseline [Englich et al., 2006], vs. no-debiasing baseline

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	95% CI	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]	1.79×

Table 1: Baseline anchoring bias comparison between humans and LLMs. LLM values show mean ± SD ($n = 30$). 95% CI computed via bootstrap.

4.2 Sibony Debiasing Techniques

Both techniques significantly reduce anchoring bias:

Technique	Diff	95% CI	Reduction vs Baseline	vs Human
Context Hygiene	2.67 mo	[2.07, 3.27]	−27%	1.30×
Premortem	2.80 mo	[2.17, 3.43]	−24%	1.37×

Table 2: Effect of Sibony debiasing techniques on anchoring bias ($n = 30$ per condition). 95% CI computed via bootstrap.

Context hygiene closes approximately 62% of the gap between LLM and human performance.

4.3 SACD Results

SACD essentially eliminates anchoring bias:

Condition	Low Anchor	High Anchor	Diff	95% CI	p-value
SACD	3.67 mo	3.20 mo	−0.47 mo	[−1.83, 0.93]	0.51

Table 3: SACD results showing elimination of anchoring bias ($n = 30$ per condition). 95% CI crosses zero, confirming no significant anchoring effect.

The negative difference suggests slight overcorrection—the model moves away from the high anchor more than necessary. The non-significant p-value indicates no reliable anchoring effect.

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	vs Human	Notes
Llama 4 Scout	70B open	0.12 mo	0.06×	Near-immune
Hermes 3 Llama 3.1	405B open	-0.16 mo	$\approx 0\times$	Largest open model
Claude Opus 4	Frontier	2.01 mo	0.98×	Human-level
GPT-5.2	Frontier	2.71 mo	1.32×	Above human
Claude Sonnet 4	Frontier	3.00 mo	1.46×	Above human
Nemotron 30B	30B dense	3.21 mo	1.57×	Moderate RLHF
Codex (OpenAI)	2023	3.67 mo	1.79×	Legacy
Trinity Large	400B MoE	4.51 mo	2.20×	13B active, heavy RLHF
GPT-4o	Frontier	4.96 mo	2.42×	Highest bias
Human baseline	—	2.05 mo	1.00×	Englich et al. 2006

Table 4: Cross-model anchoring bias ($n = 30$ per condition). Models sorted by bias magnitude. Notable: Trinity Large (400B MoE, 13B active) shows *higher* bias than smaller dense models, suggesting active compute per inference—not headline parameter count—drives anchor resistance.

Key findings:

1. **Two paths to anchor resistance:** Open-weights models (Llama, Hermes) and frontier capability (Opus) both achieve near-immunity, but through different mechanisms.

2. **RLHF compliance breeds bias:** Heavily instruction-tuned models (Trinity, GPT-4o) show the highest anchoring susceptibility, suggesting that training for instruction-following increases anchor compliance.

3. **Active compute matters more than total parameters:** Trinity Large (400B MoE, 13B active per forward pass) shows higher bias than Nemotron 30B (dense). Headline parameter counts are misleading for MoE architectures.

4. **Capability scaling helps within families:** GPT-4o → GPT-5.2 shows 46% bias reduction; Sonnet → Opus shows 33% reduction. Larger/newer models within the same training paradigm tend toward lower bias.

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	Category
Anchoring	2.05mo diff	3.00mo diff (1.46× human)	× BIASED
Sunk Cost	85% continue	0% continue	✓IMMUNE
Conjunction	85% wrong	0% Linda, 30% Bill	∼ PARTIAL
Framing	Preference reversal	90%→80% reversal	× BIASED

Table 5: Complete bias profile for Claude Sonnet 4 across four cognitive biases.

4.6 DeFrame Eliminates Framing Effect

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

Scenario	Frame	Baseline	DeFrame
Layoffs	Gain	100% certain	100% certain
Layoffs	Loss	90% gamble	**100% certain**
Pollution	Gain	100% certain	100% certain
Pollution	Loss	50% gamble	**100% certain**

Table 6: DeFrame achieves 100% bias reduction for framing effect.

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 taxonomy based on how biases respond to model improvements:

1. **Training-eliminable biases** (anchoring, sunk cost)—diminish with model capability and training improvements
2. **Structurally persistent biases** (framing)—require explicit debiasing interventions regardless of model size
3. **Contamination-dependent biases** (conjunction)—performance varies based on training data exposure to specific scenarios

This taxonomy has practical implications: developers should focus debiasing efforts on structurally persistent biases, while training-eliminable biases may self-correct with model updates.

5.4 Limitations

Methodological Constraints:

- Sample sizes: $n = 30$ for primary experiments, $n = 10$ for exploratory comparisons—adequate for detecting large effects but limiting precision on exact effect magnitudes
- 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. Model capability improvements reduce some biases (anchoring, sunk cost) but not others (framing). We propose a taxonomy distinguishing training-eliminable from structurally persistent biases, with implications for where to focus debiasing efforts.

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`