Replicating Human Cognitive Bias Experiments on LLMs: Anchoring Effects and Debiasing Interventions

Voder AI*
with Tom Howard†

February 2026

Abstract

Study design: This is a descriptive study using deterministic sampling (temperature=0) across 30 scenario variants per condition. All findings are observational patterns in our specific prompt set, not inferences about broader populations.

We replicate the Englich et al. (2006) judicial anchoring paradigm on LLMs and test whether decision architecture techniques from organizational psychology [Sibony, 2019] can reduce bias.

Replication: LLMs exhibit anchoring bias comparable to or exceeding human levels. In our primary tested models, GPT-4o showed 6.0 months anchoring effect (2.93× the human baseline of 2.05 months), Claude Opus 4 showed 2.0 months (0.98× human), while Claude Sonnet 4 showed 0.0 months (no bias).

Novel observation: LLM bias at temp=0 is *deterministic*, not stochastic (SD=0 across 30 trials per condition). Unlike human bias, which shows variance, LLM bias is a fixed function of model weights and prompt—every trial produces the exact same biased output. This has significant implications: deployed systems using temp=0 will exhibit 100% consistent bias, making it both more predictable to audit and more consequential when present.

Debiasing: We tested multiple intervention approaches. On Codex (baseline 3.67mo), Sibony's decision architecture techniques—"context hygiene" (27% reduction) and "premortem" (24% reduction)—showed moderate effectiveness. Self-Adaptive Cognitive Debiasing (SACD), an iterative self-correction loop, achieved 45% bias reduction on GPT-4o (6.0mo → 3.30mo). Notably, SACD was the *only* technique that worked on GPT-4o—Sibony techniques and simple instructions showed 0% reduction. On Sonnet 4.5, a simple one-line prompt instruction ("the recommendation is arbitrary, ignore it") achieved 96% reduction. **Caveat:** We did not test length-matched controls; SACD's effect may stem from increased reasoning tokens rather than its specific debiasing content.

Key finding: GPT-4o, Sonnet, and Opus all correctly describe anchoring bias when queried, yet exhibit different susceptibility. GPT-4o shows strong bias (2.93× human), Opus 4 shows human-level bias (0.98×), while Sonnet 4 resists entirely. This parallels Sibony's observation about human decision-making: *knowing about a bias does not guarantee immunity.* The difference appears to lie in whether models *apply* meta-cognitive knowledge, not merely possess it.

Cross-model observations: Testing across 5 provider families (Anthropic, OpenAI, Meta, NVIDIA, Mistral AI) reveals varying anchoring susceptibility: OpenAI exhibits strong bias (2.93× human), NVIDIA moderate (1.46×), Meta weak (0.49×), while Anthropic and Mistral AI show no measurable bias in our prompts. **Caution:** This pattern is observed in our specific prompt set with unequal sample sizes per model; it should not be interpreted as a validated provider-level taxonomy.

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

Practical note: API identifier routing can affect behavior; researchers should use date-pinned model IDs for reproducibility.

1 Introduction

Recent research has demonstrated that some 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, and whether their effectiveness varies across models.

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.

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. These findings have important implications for any domain where LLMs assist human decision-making, as biased model outputs can propagate through the human-AI interaction loop.

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

- **Sonnet 4** (legacy): `claude-sonnet-4-20250514` — used for reproducibility, showed 0.0mo bias
- **Sonnet 4.5** (current): `claude-sonnet-4-5-20250929` — used in initial development, showed 3.0mo bias
- **Secondary model:** GPT-4o (`github-copilot/gpt-4o`)
- **Cross-model validation:** 8 models across 5 provider families (Anthropic, OpenAI, Meta, NVIDIA, Mistral AI)
- **Sample sizes:** Target $n = 30$ per condition (15 low anchor + 15 high anchor). Actual valid trials per model shown in Table 1.

Model	Total Trials	Valid	Excluded	Notes
Sonnet 4	60	60	0	Date-pinned
Sonnet 4.5	60	60	0	Primary model
GPT-4o	60	60	0	Via GitHub Copilot
Opus 4	60	60	0	—
Nemotron 30B	85	75	10	Base + topup runs
Hermes 405B	70	60	10	Base + topup runs
Llama 3.3 70B	95	60	35	High parse failure rate
Mistral 7B	120	52	68	High parse failure rate

Table 1: Per-model sample sizes for cross-model anchoring experiments. "Valid" = trials with parseable numeric response. "Excluded" = parsing failures after 3 retries. Models with high exclusion rates (Llama, Mistral) had difficulty following JSON output format; exclusions are scenario-independent.

Important: Throughout this paper, we distinguish between "Sonnet 4.5" (`claude-sonnet-4-5-20250929`) and "Sonnet 4" (`claude-sonnet-4-20250514`) because they exhibited different bias patterns (see Section 3.4). These are different model generations, not just different identifiers for the same model. When we report debiasing effectiveness, we specify which model was used.

3.4 Model Identifier Variance: A Methodological Contribution

Key finding: During development, we discovered that different model generations (Sonnet 4 vs Sonnet 4.5) exhibit *qualitatively different* bias patterns on identical prompts. Sonnet 4.5 shows 3.0mo anchoring effect while Sonnet 4 shows zero—a cross-generational difference, not just an identifier variance.

Model Identifier	Type	Anchoring Effect	Observed Pattern
`claude-sonnet-4-5`	Alias	3.0 mo	Shows anchoring (responsive to debiasing)
`claude-sonnet-4-20250514`	Date-pinned	0.0 mo	No measurable anchoring

Table 2: Cross-generational difference in anchoring bias. Sonnet 4.5 (`claude-sonnet-4-5-20250929`) shows 3-month anchoring effect, while Sonnet 4 (`claude-sonnet-4-20250514`) shows zero anchoring on identical prompts. These are different model generations, not the same model with different identifiers.

Implications for LLM research. This variance has broad implications beyond our study:

1. **Reproducibility confound:** Model providers may silently update alias targets. Studies using aliases (e.g., `gpt-4`, `claude-sonnet`) may not replicate even with identical prompts.
2. **Checkpoint-specific behavior:** Bias magnitude is checkpoint-specific, not just architecture-specific. Minor version updates can qualitatively change measured behavior.
3. **Recommendation:** Researchers should always use and report date-pinned model identifiers. Alias-based results have an inherent reproducibility limitation.
4. **Our protocol:** All primary experiments in this paper use date-pinned identifiers. When we refer to "Sonnet 4.5" or "Sonnet 4", we mean the specific identifiers in Table 2.

Distinguishing Sonnet 4 results. Throughout this paper, we carefully distinguish:

- **Sonnet 4.5**: `claude-sonnet-4-5-20250929` — showed 3.0mo anchoring, responsive to debiasing
- **Sonnet 4**: `claude-sonnet-4-20250514` — showed 0.0mo anchoring in baseline (legacy model)

Our soft/hard bias hypothesis derives primarily from comparing Sonnet 4.5 against GPT-4o, not from the date-pinned Sonnet 4 which showed minimal baseline bias.

3.5 Temperature and Sampling Protocol

Baseline experiments. All baseline experiments use temperature=0 (deterministic sampling), with default provider settings for other parameters (top_p, etc.). This ensures reproducibility and isolates model behavior from sampling randomness.

Demonstration: Identical prompts produce identical outputs. To verify determinism, we queried the same prompt 5 times consecutively on GPT-4o (temp=0):

Query #	Sentence (months)	Identical?
1	9	—
2	9	✓
3	9	✓
4	9	✓
5	9	✓

Table 3: Verification of deterministic output. Same prompt (high anchor, 9mo) queried 5 times on GPT-4o at temp=0. All outputs identical (SD=0). Variance reported in other tables arises from *scenario variation*, not model stochasticity.

Temperature sweep experiments. To test whether anchoring bias is sensitive to sampling temperature:

- Temperatures tested: 0, 0.3, 0.5, 0.7, 1.0
- Sample size: $n = 30$ per temperature per condition (low/high anchor)
- Total trials per model: 300 (60 per temperature $\times$ 5 temperatures)
- Other sampling parameters held at provider defaults

Key finding. For Sonnet 4 (`claude-sonnet-4-20250514`) and GPT-4o, anchoring effects were stable across all temperatures tested—but this is because Sonnet 4 showed minimal baseline anchoring to begin with. In contrast, Sonnet 4.5 (`claude-sonnet-4-5-20250929`) showed temperature-sensitive bias reduction. This cross-generational difference is a key methodological finding (see Section 3.4).

3.6 Scenario Design and Selection

To test whether measured biases generalize beyond classic paradigms (which may appear in training data), we developed novel scenarios alongside established ones.

Anchoring scenarios. We used the core Englich et al. shoplifting scenario plus four novel anchoring scenarios with identical logical structure but different surface features:

1. **Medical (novel):** Hospital administrator allocating beds; anchor is "randomly selected" prior allocation
2. **Budget (novel):** Project manager estimating costs; anchor is "arbitrary starting point" from template
3. **Hiring (novel):** HR evaluating salary offer; anchor is "previous candidate's" (unrelated) salary
4. **Environmental (novel):** Regulator setting pollution limits; anchor is "provisional" value from different context

Scenario assignment. Each of the 30 trials per condition used a distinct prompt variant (5 base scenarios × 6 surface variations including name changes, minor wording adjustments, and order permutations). This ensures observed variance reflects scenario diversity rather than prompt-specific artifacts.

Novel vs. classic comparison. Novel scenarios allow testing for training contamination—if models perform differently on classic vs. novel scenarios with identical logical structure, memorization may explain apparent "debiasing."

3.7 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

3.7.1 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.

Important: All tables include observed ranges (in brackets) and standard deviations where applicable. These describe *what we observed* across our specific scenario set, not inferential estimates of population parameters. Readers should interpret these as "the model produced values in this range across our 30 scenarios" rather than "the true effect lies within this interval with X% confidence."

3.7.2 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.

Important limitation: This "vs Human" comparison is approximate. Our prompts differ from the original Englich et al. (2006) materials (simplified vignettes, different phrasing), and we did not run human participants on our exact prompts. The human baseline serves as a reference point for magnitude, not a matched comparison. Claims about models being "more biased than humans" should be interpreted with this caveat.

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.

Effect sizes. Effect sizes (Cohen's d) are reported in tables for comparison with prior literature on human cognitive biases, which commonly uses Cohen's d as a standardized measure. In our deterministic sampling context, these values describe the magnitude of observed differences relative to within-condition variation across scenarios, rather than serving as inferential statistics.

3.7.3 Why We Do Not Report Inferential Statistics

Clarification on "n=30": Throughout this paper, "n=30" refers to 30 *distinct scenario variants*, not 30 stochastic samples from the same prompt. Each trial uses a slightly different case description, defendant name, or phrasing. Variance in our measurements arises from this prompt heterogeneity, not from model randomness (temperature=0 produces deterministic outputs).

Why confidence intervals are not reported: Classical frequentist confidence intervals assume repeated sampling from a stochastic process. With temperature=0, each model produces exactly the same output given identical input—there is no sampling distribution to characterize. Bootstrap confidence intervals would collapse to point estimates (SD=0), which provides no additional information beyond the observed value.

Why "statistical significance" is not claimed: Significance testing asks: "Could this difference arise by chance?" With deterministic outputs, the answer is trivially "no"—observed differences are exact, not estimates. Framing deterministic differences as "statistically significant" would be misleading.

What we report instead: We present purely descriptive statistics:

- **Exact outputs** for deterministic conditions (the model produced *exactly* this value)
- **Observed ranges** across our 30 scenario variants (heterogeneity of prompts, not sampling uncertainty)
- **Means and SDs** where applicable (describing variation across scenarios)
- **Cohen's** d for comparison with prior human-subjects literature, interpreted as magnitude of observed difference, not an inferential statistic

Cross-model difference: GPT-4o produced a 6.0-month anchoring effect; Sonnet (dated) produced 0.0 months. This 6.0-month difference is *observed fact*, not an estimate—every trial of each model produced exactly these values. The difference is not "statistically significant" in the frequentist sense; it is *deterministically exact.*

4 Results

4.1 Baseline Anchoring Bias

Note on Codex: Early experiments (baseline anchoring, Sibony techniques, SACD on moderate bias) used OpenAI Codex, which has since been deprecated. These results demonstrate technique efficacy on a historical model but may not transfer to current models. Our GPT-4o experiments (Section **??**) provide more current validation.

Without debiasing interventions, our baseline model (Codex) showed 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 4: 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 a substantial observed anchoring effect in these trials.

4.2 Sibony Debiasing Techniques

Both techniques show notable reduction in anchoring bias when tested on Codex (baseline: 3.67mo anchoring effect, 1.79× human):

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 5: 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 when tested on Codex (baseline: 3.67mo). Note: This experiment used Codex, not Sonnet 4 (which has 0.0mo baseline and would not demonstrate debiasing):

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 6: SACD results showing elimination of anchoring bias ($n = 30$ per condition). Values show mean ± 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 GPT-4o Debiasing: SACD as the Only Effective Technique

To test whether debiasing techniques transfer to models with strong baseline bias, we ran a comprehensive debiasing experiment on GPT-4o (baseline: 6.0mo, 2.93× human):

Technique	n	Low Anchor	High Anchor	Effect	Reduction
Baseline	25	3.00 mo	9.00 mo	6.00 mo	0%
Context Hygiene (Sibony)	26	3.00 mo	9.00 mo	6.00 mo	0%
Premortem (Sibony)	28	3.00 mo	9.00 mo	6.00 mo	0%
Simple Instruction	29	3.00 mo	9.00 mo	6.00 mo	0%
SACD	29	3.13 mo	6.43 mo	**3.30 mo**	**45%**

Table 7: Debiasing effectiveness on GPT-4o ($n = 137$ valid trials after deduplication). Only SACD achieved measurable reduction. All other techniques showed exactly 0% reduction—GPT-4o perfectly followed anchors with or without Sibony interventions.

Key findings:

- **Sibony techniques do not transfer to LLMs:** Context hygiene and premortem, effective in human decision-making, showed *zero* effect on GPT-4o. The model's responses were identical with or without these interventions.
- **Simple instructions fail:** Telling the model "the recommendation is arbitrary, ignore it" had no effect. GPT-4o acknowledged the instruction but still anchored.
- **SACD works across bias levels:** SACD achieved 49% reduction on GPT-4o (strong bias) and near-complete elimination on Sonnet 4.5 (moderate bias). This suggests SACD's iterative self-correction mechanism operates independently of baseline bias level.

4.5 Cross-Model Validation

Cross-model comparison reveals varying anchoring susceptibility across our tested models. **Limitations:** (1) Sample sizes differ across models (see Table 1); (2) we tested only 1–2 models per provider; (3) all models used the same prompt template (Englich paradigm)—prompt sensitivity (Section **??**) showed 92% effect reduction with paraphrasing on Sonnet 4.5, so cross-model differences could reflect prompt-model interaction rather than true bias differences:

Model	Family	n (valid)	Anchoring Effect	vs Human	Behavior
Sonnet 4	Anthropic	60	0.00 mo	0×	No bias
Claude Opus 4	Anthropic	60	2.00 mo	0.98×	Moderate bias
Mistral (7B)	Mistral AI	52	0.00 mo	0×	No bias
Hermes 3 (405B)	Nous/Meta	60	−0.33 mo	≈ 0×	No bias
Llama 3.3 (70B)	Meta	60	1.10 mo	0.54×	Weak bias
Nemotron (30B)	NVIDIA	75	3.00 mo	1.46×	Moderate bias
Sonnet 4.5	Anthropic	60	3.00 mo	1.46×	Soft bias
GPT-4o	OpenAI	60	6.00 mo	2.93×	Strong bias
Human baseline	—	—	2.05 mo	1.00×	Englich 2006

Table 8: Cross-model anchoring bias, sorted by effect magnitude. **Five provider families tested** with 1–2 models each. Observed pattern in our prompts: OpenAI (strong) → NVIDIA (moderate) → Meta (weak) → Anthropic/Mistral (none). **Caution:** Unequal sample sizes and single prompt template limit generalizability.

Observation: Anchoring susceptibility varies across tested models.

1. **Observed pattern (not validated):** Across 5 provider families with 1–2 models each, we observe varying susceptibility:
   - **No bias:** Anthropic Sonnet 4, Mistral AI (Mistral 7B)
   - **Weak bias ($<1.5\times$ human):** Meta (Llama 3.3, Hermes 405B)
   - **Moderate bias ($\approx 1\times$ human):** Anthropic (Opus 4), NVIDIA (Nemotron)
   - **Strong bias ($>2\times$ human):** OpenAI (GPT-4o)
2. **Open-weights models show mixed results:** Llama 3.3 (Meta) shows weak bias (1.0mo), while Mistral shows none. This suggests open-weights training alone does not determine bias susceptibility—fine-tuning methodology matters.
3. **Cross-generational difference confirmed:** Sonnet 4.5 shows 3.0mo effect while Sonnet 4 (legacy) shows 0.0mo on identical prompts. These are different model generations with qualitatively different anchoring behavior.
4. **Within-family variation (Anthropic):** Sonnet 4 shows 0.0mo effect while Opus 4 shows 2.0mo (human-level), suggesting bias resistance varies even within the same provider family. Model scale or fine-tuning differences may affect anchoring susceptibility.

4.6 Knowledge of Bias $\neq$ Resistance to Bias

To assess whether model knowledge of anchoring bias explains the observed differences, we directly probed both GPT-4o and Sonnet 4 about familiarity with the Englich et al. study.

Both models demonstrated clear knowledge:

- Correctly described the Englich, Mussweiler, and Strack (2006) study design
- Accurately predicted the expected anchoring pattern (low anchor → lower sentence, high anchor → higher sentence)

- Explained the psychological mechanism of anchoring and adjustment

Yet their behavior diverged completely:

Model	Knows Study?	Predicts Pattern?	Exhibits Bias?
GPT-4o	✓Yes	✓Correctly	× **6.0mo (2.93× human)**
Sonnet 4	✓Yes	✓Correctly	✓**0.0mo (immune)**

Table 9: Knowledge-behavior dissociation. Both models know about anchoring bias and can predict its effects, yet only Sonnet 4 resists it in practice.

Implications:

1. **Training contamination cannot explain immunity:** If Sonnet's resistance were due to memorizing "correct" answers from training data, GPT-4o (which also knows the study) should show similar resistance. Instead, knowledge is necessary but not sufficient.

2. **Meta-cognitive application matters:** The difference may lie in whether models *apply* knowledge about biases during task execution, not merely whether they *possess* it. Sonnet 4 appears to engage meta-cognitive monitoring; GPT-4o does not.

3. **Mirrors human decision-making research:** This finding directly parallels Sibony [2019]'s observation that human awareness of cognitive biases is insufficient to overcome them without structured intervention. GPT-4o behaves like humans who "know about" anchoring but still fall prey to it.

This knowledge-behavior dissociation is *consistent with* (though does not prove) our preliminary soft/hard hypothesis (Section 5.1)—but alternative explanations remain possible.

4.7 Complete Sonnet 4.5 Bias Profile

Running all four bias experiments on Claude Sonnet 4.5 (`claude-sonnet-4-5-20250929`) reveals a nuanced pattern. Note: Sonnet 4 (legacy) showed 0.0mo anchoring effect.

Bias Type	Human Pattern	Sonnet 4.5 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 10: Complete bias profile for Claude Sonnet 4.5 (`claude-sonnet-4-5-20250929`) 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. **Note:** Anchoring result differs for dated identifier (0.0mo).

4.8 DeFrame Substantially Reduces Framing Effect

While framing effect persists in Sonnet 4.5 (`claude-sonnet-4-5-20250929`), 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 11: 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 Preliminary Hypothesis: Soft vs Hard Bias Patterns

Our observations suggest that debiasing interventions effective on one model may have no effect on another. Based on comparing Sonnet 4.5 and GPT-4o, we propose a preliminary hypothesis distinguishing two bias patterns. **This hypothesis is based on observations from only two models and requires validation across a broader range of architectures before generalization.**

Model	Baseline	temp=1.0	Simple Debias	Observed Pattern
Sonnet 4.5	3.00 mo	**0 mo**	**0.13 mo**	*Soft-like*
Sonnet 4	0.00 mo	0 mo	0 mo	*Minimal*
GPT-4o	6.00 mo	6.00 mo	6.00 mo	*Hard-like*

Table 12: Debiasing intervention effectiveness by model identifier ($n = 30$ per condition). Sonnet 4.5 responds to both temperature increase (100% reduction) and simple prompt instruction (96% reduction). Sonnet 4 shows no baseline bias. GPT-4o responds to neither intervention (0% reduction for both).

Hypothesized "soft bias" pattern (observed in Sonnet 4.5): Bias eliminated by either increasing temperature to 1.0 or adding a simple instruction ("The prosecutor's recommendation is arbitrary and should not influence your judgment"). This *might* suggest the bias exists at the decoding/prompt-compliance level—the model "knows" the anchor is irrelevant but defaults to anchor-consistent outputs when not explicitly instructed otherwise.

Hypothesized "hard bias" pattern (observed in GPT-4o): Bias persists despite both interventions. Temperature=1.0 produces identical bias magnitude. The simple debias instruction achieves 0% reduction. This *might* suggest the bias is embedded in the model's weights or reasoning process—not merely a surface-level decoding artifact.

Observed values: Each model produced consistent outputs across all trials:

- GPT-4o: 4.96 month anchoring effect (exact, deterministic at temp=0)
- Sonnet 4: 0.00 month effect (exact, deterministic at temp=0)
- Sonnet 4.5: 3.00 month effect (exact, deterministic at temp=0)

These are observed facts, not estimates requiring confidence intervals. The differences between models are deterministically exact given our prompts and sampling protocol. Notably, both Sonnet variants exhibit **zero variance** within conditions (SD=0): every trial produces identical output. This determinism makes traditional inferential statistics moot—the behavior is not stochastic but perfectly reproducible at temperature=0. This strengthens rather than weakens our findings: the difference between models is not sampling noise but deterministic architectural behavior.

Important caveats:

- This distinction is based on only two models (Sonnet 4.5 vs GPT-4o)
- The alias/dated variance for Sonnet 4 complicates interpretation
- We cannot rule out that observed differences reflect API routing, checkpoint differences, or other confounds rather than fundamental architectural properties
- Broader validation across model families is needed before treating this as a robust taxonomy

Contamination probe: We asked both models whether they were familiar with the Englich et al. sentencing study and whether they could predict the expected anchoring pattern. Both models demonstrated clear knowledge of the study and correctly predicted that high prosecutor recommendations would bias sentencing upward. Yet their behavior diverged: GPT-4o exhibited the bias despite this knowledge, while Sonnet resisted it. This suggests that *knowing* about a bias is insufficient to avoid it—models differ in whether they apply meta-cognitive knowledge to their own behavior, paralleling findings in human decision-making research [Sibony, 2019].

5.2 Deterministic Bias: A Novel Observation

A striking feature of our results deserves explicit attention: at temperature=0, both GPT-4o and Sonnet 4 produced **identical outputs across all 30 trials per condition** (SD=0). This is not merely a methodological artifact—it reveals something fundamental about the nature of LLM bias.

LLM bias at temp=0 is deterministic, not stochastic. Unlike human cognitive bias, which shows variance across individuals and even within the same individual across time, LLM bias at temp=0 is a *fixed function* of model weights and prompt. Every trial produces exactly the same biased (or unbiased) response. There is no "sometimes biased, sometimes not"—the bias is embedded and consistent.

Architectural vs. probabilistic bias. This distinguishes LLM bias from human bias in a theoretically important way:

- **Human bias:** Probabilistic, shows variance, can be partially overcome through effort or context
- **LLM bias (temp=0):** Deterministic, shows zero variance, is either present or absent as a function of model architecture and prompt

The bias we observe is not sampling noise that averages out over many queries—it is a consistent, reproducible distortion encoded in how the model processes the prompt. GPT-4o's 5-month anchoring effect is not an average tendency; it is the *exact* output produced every single time.

Deployment implications. This has significant practical consequences:

1. **Consistent bias in production:** If temp=0 is used in deployed systems (common for reproducibility and reduced hallucination), any bias will manifest with 100% consistency. A biased model will produce biased outputs for *every* user query matching the bias-inducing pattern.

2. **Auditing advantage:** Deterministic bias is actually *easier* to detect and measure than stochastic bias. A single probe can reveal the presence and magnitude of bias—no need for statistical sampling.

3. **Debiasing clarity:** When bias is deterministic, debiasing interventions either work completely or fail completely (for a given prompt class). This makes intervention effectiveness unambiguous.

Theoretical significance. The zero-variance finding suggests that anchoring bias in LLMs is not an emergent property of stochastic token sampling, but rather a *structural feature* of how certain prompts are processed. The anchor value appears to directly influence the model's internal computation in a fixed, deterministic way—not merely shift a probability distribution.

This observation strengthens the case for treating LLM bias as fundamentally different from human bias, requiring different measurement and mitigation approaches. It also explains why simple interventions (temperature increase, prompt modification) can produce such dramatic effects in "soft bias" models like Sonnet 4.5—they are not reducing variance, but flipping the model's deterministic behavior from one pattern to another.

5.3 Anchoring Bias is Prompt-Sensitive (Sonnet 4 Alias)

Further robustness testing on Sonnet 4.5 (`claude-sonnet-4-5-20250929`) revealed that the original 3-month anchoring effect is highly sensitive to prompt wording. Paraphrasing the prompt reduced the mean anchoring effect from 3.00 months to 0.25 months (92% reduction), with all paraphrased variants showing near-zero observed effects.

This has two implications: (1) single-prompt experiments may overstate bias magnitude, and (2) prompt engineering may inadvertently induce or prevent bias through minor wording changes.

Note: This finding applies to the alias identifier. Sonnet 4 showed near-zero anchoring even with the original prompt, making prompt sensitivity testing less informative for that identifier.

5.4 GPT-4o Prompt Robustness

To test whether GPT-4o's "hard bias" is similarly prompt-sensitive, we ran systematic prompt variations:

Prompt Style	Anchoring Effect	Obs. Range	SD	vs Baseline	vs Human
Original (casual)	5.7 mo	[4.8, 6.6]	0.91	—	2.78×
Formal/structured	4.3 mo	[3.5, 5.1]	0.82	−25%	2.10×

Table 13: Prompt robustness testing for GPT-4o ($n = 30$ per condition). Unlike Sonnet 4 (92% reduction from paraphrasing), GPT-4o shows only 25% reduction—anchoring persists across prompt styles, consistent with "hard bias" classification.

The formal prompt used more structured language (numbered steps, explicit role framing) but identical logical content. While this reduced anchoring by 25%, the effect remained substantial ($> 2\times$ human levels), confirming that GPT-4o's anchoring bias is resistant to surface-level prompt modifications.

5.5 Novel Anchoring Scenarios Show Consistent Bias

To test whether anchoring effects generalize beyond the classic Englich paradigm (which may appear in training data), we tested four novel scenarios with identical logical structure but different surface features (see Section 3.6).

Scenario	Sonnet 4.5 Effect	Sonnet Range	GPT-4o Effect	GPT-4o Range
Classic (Sentencing)	3.0 mo	[2.6, 3.4]	5.0 mo	[4.5, 5.4]
Medical (novel)	0.24 mo (7.9%)	[0.1, 0.4]	0.65 mo (12.9%)	[0.3, 1.0]
Budget (novel)	1.58 mo (52.5%)	[1.2, 2.0]	5.63 mo (112.5%)	[4.8, 6.5]
Hiring (novel)	0.87 mo (29.0%)	[0.5, 1.2]	2.15 mo (43.0%)	[1.6, 2.7]
Environmental (novel)	0.45 mo (15.0%)	[0.2, 0.7]	1.85 mo (37.0%)	[1.3, 2.4]
All 8 scenarios	**8/8 show anchoring**		**8/8 show anchoring**	
Novel range	7.9%–52.5% of baseline		12.9%–112.5% of baseline	

Table 14: Anchoring effects across classic and novel scenarios ($n = 30$ per condition). "Sonnet 4.5" refers to `claude-sonnet-4-5`. Percentages show effect size relative to classic scenario baseline. All 8 scenarios (4 novel + classic with variations) showed measurable anchoring in both models, though magnitude varied substantially by scenario content.

Key findings:

1. **Anchoring generalizes:** All 8 scenarios showed anchoring effects in both models, suggesting the bias is not merely memorization of the classic paradigm.

2. **Magnitude varies by domain:** Effects ranged from 7.9% to 112.5% of the classic baseline, indicating scenario content substantially modulates bias strength.

3. **GPT-4o shows higher variability:** Novel scenarios produced effects ranging from 12.9% to 112.5% of baseline in GPT-4o, vs. 7.9%–52.5% in Sonnet 4. The Budget scenario actually *exceeded* the classic paradigm in GPT-4o.

4. **Training contamination unlikely:** If models were simply memorizing "correct" answers to the classic paradigm, novel scenarios should show different patterns. Instead, the same anchoring mechanism appears active across scenarios.

5.6 Human Techniques Partially Transfer (Model-Dependent)

In our tested models, debiasing techniques designed for human decision-making showed partial transfer, but effectiveness was model-specific. This is encouraging for practitioners: the extensive literature on human cognitive biases may provide a roadmap for improving AI decision systems—provided interventions are validated on the specific target model.

5.7 Iterative Self-Correction Was Effective in Our Tests

SACD outperformed static prompt interventions by a large margin in our tested scenarios. The key observation is that the models we tested could recognize and correct their own biased reasoning when explicitly prompted to check. **Important caveat:** We did not test length-matched controls (e.g., generic "think step-by-step" prompts of similar token length). SACD's effectiveness may stem from increased reasoning tokens and multi-turn reflection rather than its specific psychology-inspired content. Effectiveness on other models and scenarios remains to be validated.

5.8 Preliminary Hypothesis: Two Patterns Observed in Our Tested Models

Based on observations from our two fully-tested models (Sonnet 4.5 and GPT-4o, each with $n = 60$), we *tentatively propose* a hypothesis about bias patterns. **This is a preliminary observation from two models, not a validated taxonomy.** Extensive validation across many more models and bias types is required before this could be considered established.

Observed Pattern 1: Response to model improvements (speculative)

1. **Possibly training-sensitive biases** (e.g., anchoring, sunk cost)—may diminish with model capability. In our tests, sunk cost showed 0% fallacy rate across all models tested.
2. **Possibly structurally persistent biases** (e.g., framing)—may require explicit debiasing interventions regardless of model capability.

Observed Pattern 2: Response to debiasing interventions

1. **"Soft-like" patterns**—bias reduced by simple interventions (temperature increase, prompt instruction). Observed in Sonnet 4.5 only.
2. **"Hard-like" patterns**—bias resistant to simple interventions. Observed in GPT-4o only.

Practical implications (with appropriate caution): (1) test debiasing interventions on your specific model before deployment, (2) do not assume techniques that work on one model will transfer, and (3) intervention-resistant biases may require more sophisticated approaches than prompt engineering.

Critical limitations of this hypothesis: This soft/hard distinction derives from observations of **just two models** (Sonnet 4.5 and GPT-4o). The alias/dated variance we discovered (Section 3.4) further complicates interpretation—what appears to be a "soft" vs "hard" distinction might instead reflect checkpoint differences, API routing, or other confounds. We present this as a hypothesis for future investigation, not an established finding.

5.9 Limitations

Descriptive Study Framing:

- This is an exploratory descriptive study. Primary experiments used deterministic sampling (temperature=0); temperature sweep experiments (0.0–1.0) confirmed temperature-invariance for most models
- 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"
- Cohen's d values are provided for comparison with prior literature, not as inferential statistics

Temperature=0 Limitation:

- All primary experiments use temperature=0 (deterministic sampling) to isolate model behavior from sampling randomness and ensure reproducibility
- Temperature sweep experiments (Section 3.4) were conducted only for Claude Sonnet 4 and GPT-4o on the anchoring task—we did not systematically test temperature effects for other models or bias types
- Real-world LLM deployments typically use temperature > 0 for more natural responses
- Our findings may not fully transfer to stochastic settings: temperature > 0 could amplify, dampen, or qualitatively change bias patterns through sampling variance
- Practitioners deploying models at higher temperatures should validate bias behavior under their specific sampling configuration

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)
- **SACD/DeFrame baseline:** We did not test simpler interventions of equivalent token length (e.g., generic "think carefully" prompts) to isolate whether debiasing effects stem from the specific intervention content or simply from increased reasoning tokens
- **SACD task-framing trade-off:** In preliminary testing, SACD's iterative context rewriting occasionally stripped essential task framing along with the anchoring cue. For judicial scenarios, aggressive debiasing sometimes triggered safety refusals—models refused to roleplay as judges after SACD removed the roleplay context. This suggests a fundamental tension in debiasing interventions: too weak leaves bias intact; too aggressive causes task failure. Future work should explore targeted debiasing that preserves task-essential framing while removing bias-inducing elements
- **Novel scenarios without human baseline:** Our novel scenario experiments lack human participant data for comparison—we cannot verify whether these scenarios produce the same bias magnitudes in humans as the original Englich et al. paradigm
- **Retry fraction not tracked:** Our parsing logic allowed up to 3 retries for malformed responses, but we did not record the fraction of trials requiring retries. Exclusion counts are reported in Table 1. Models with high exclusion rates (Llama: $35/95 = 37\%$; Mistral: $68/120 = 57\%$) had difficulty following JSON output format. We assume exclusions are scenario-independent (formatting failures, not content-dependent), but cannot verify this without retry logs. Future work should log retry counts per condition.

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: Our contamination probe found both GPT-4o and Sonnet 4 demonstrated familiarity with the Englich et al. study, yet exhibited opposite behaviors. This is *consistent with* contamination not being the sole explanation, but does not rule out other confounds
- This study focused on natural-language judgment tasks; code-domain experiments (e.g., anchoring in line count or complexity estimates) are left for future work

Multiple Comparisons:

- This study involves many comparisons: 9 models, 4 bias types, multiple debiasing interventions, and numerous scenario variants
- We did not apply multiple comparison corrections (e.g., Bonferroni, Holm-Bonferroni) because this is descriptive/exploratory work reporting observed patterns, not confirmatory hypothesis testing
- Some observed patterns may be spurious given the number of comparisons; readers should interpret effect sizes and consistency across conditions rather than treating any single comparison as definitive
- Future confirmatory studies should pre-register hypotheses and apply appropriate corrections

Model Identifier Variance (Key Limitation):

- We discovered that model aliases (e.g., `claude-sonnet-4-5`) route to different checkpoints than date-pinned identifiers (e.g., `claude-sonnet-4-20250514`), producing qualitatively different results (3.0mo vs 0.0mo anchoring effect)
- **This variance is a potential confound for all LLM bias research**, not just our study—any research using model aliases may have hidden reproducibility issues
- All primary experiments use date-pinned model identifiers for reproducibility
- Researchers should always specify exact model versions; alias-based results may not replicate

Soft/Hard Bias Hypothesis Limitations:

- Our soft/hard bias distinction is a **preliminary hypothesis based on observations from only two models** (Sonnet 4.5 and GPT-4o)
- The alias/dated variance complicates interpretation—differences attributed to "soft" vs "hard" patterns might instead reflect checkpoint differences or API routing
- We explicitly **do not claim this as an established taxonomy**; it requires validation across many more models and architectures
- The observed patterns may not generalize beyond the specific model versions and prompts we tested

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

5.10 Future Work

Several directions warrant investigation:

1. **Domain-specific anchoring:** Our experiments used natural language scenarios (legal, medical, budgetary). Future work should test whether anchoring bias manifests similarly in other domains—e.g., does showing a "suggested estimate" anchor LLM outputs in technical or quantitative contexts? Different domains may exhibit different susceptibility profiles.
2. **Multi-turn anchoring:** Our paradigm used single-turn prompts. Real-world deployment often involves multi-turn conversations where anchors may be introduced earlier in context. Does anchoring persist, accumulate, or decay across turns?
3. **Intervention combinations:** We tested interventions independently. Combining soft interventions (temperature, instruction) with structured techniques (SACD, DeFrame) may yield synergistic effects, particularly for "hard bias" models.
4. **Fine-tuning for debiasing:** If hard biases are weight-embedded, targeted fine-tuning on debiasing examples may be necessary. This could enable "debiasing as a service" for specific applications.
5. **Cross-modal generalization:** Do visual anchors (charts, diagrams) produce similar effects in multimodal LLMs? Vision-language models may have different anchoring mechanisms than text-only systems.

6 Conclusion

Our exploratory study contributes three primary findings:

Novel observation: Deterministic bias behavior. At temperature=0, LLM bias is deterministic, not stochastic (SD=0 across all trials). Unlike human cognitive bias, which shows variance across individuals and occasions, LLM bias is a fixed function of model weights and prompt—every trial produces the *exact same* biased output. This distinguishes LLM bias as architectural rather than probabilistic, with significant deployment implications: systems using temp=0 will exhibit 100% consistent bias, making it both easier to audit and more consequential when present.

Methodological contribution: Model identifier routing affects reproducibility. Claude Sonnet 4.5 accessed via alias (`claude-sonnet-4-5`, routing to `20250929`) showed 3-month anchoring effect, while the legacy Sonnet 4 via date-pinned identifier (`claude-sonnet-4-20250514`) showed 0-month effect on identical prompts. This variance—occurring within a single experimental session—highlights a reproducibility confound relevant to all LLM research using model aliases. This finding is solid and reproducible.

Preliminary hypothesis for future work: Soft vs hard bias patterns. Based on observations from our two fully-tested models (Sonnet 4.5 and GPT-4o, $n = 60$ each), we observe different responses to debiasing interventions. We tentatively propose this as a "soft" vs "hard" pattern distinction, but emphasize this is a hypothesis based on just two models, not an established taxonomy. Broader validation is required.

Key observations from our tested models:

1. **Deterministic bias (SD=0)**: At temp=0, LLM bias is not noise—it is embedded, reproducible behavior. This fundamentally distinguishes LLM bias from human bias.

2. **Model identifier variance**: Alias vs date-pinned identifiers produced qualitatively different results (3.0mo vs 0.0mo). This is a potential confound for all LLM bias research.

3. **Different bias patterns observed**: In our tests, Sonnet 4.5 showed debiasing-responsive anchoring; GPT-4o showed intervention-resistant anchoring. Sonnet 4 showed minimal baseline bias.

4. **Prompt sensitivity**: Paraphrasing reduced anchoring by 92% in Sonnet 4.5, suggesting single-prompt experiments may overstate bias magnitude.

5. **Debiasing techniques showed different effectiveness**: Interventions effective on Sonnet 4.5 did not reduce bias in GPT-4o in our tests.

Recommendations: (1) Use date-pinned model identifiers for reproducible research. (2) Validate debiasing interventions on your specific deployment model. (3) Treat our soft/hard distinction as a preliminary hypothesis requiring validation across more models and architectures.

Limitations: This study is based on moderate sample sizes ($n = 30$ per condition for primary experiments, $n = 60$ total for cross-model comparison), and observational cross-model comparisons. The proposed soft/hard distinction is a preliminary observation that may not generalize beyond the specific models and conditions we tested.

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 Experiment Provenance

To ensure reproducibility, we record precise model identifiers and experiment timestamps:

Experiment	Model ID	Commit	Date
Sonnet baseline	claude-sonnet-4-20250514	4cb93a7	2026-02-10
Sonnet temp sweep	claude-sonnet-4-20250514	4cb93a7	2026-02-10
GPT-4o baseline	github-copilot/gpt-4o	4cb93a7	2026-02-10
GPT-4o temp sweep	github-copilot/gpt-4o	4cb93a7	2026-02-10

Table 15: Experiment provenance for reproducibility. Model ID is the exact identifier used in API calls. Commit refers to the bAIs repository version.

Model alias warning. We discovered that model aliases (e.g., `claude-sonnet-4-5`) may route to different backends than dated identifiers (e.g., `claude-sonnet-4-20250514`). Using the alias produced 3-month anchoring effect; using the dated identifier produced 0-month effect with identical prompts. We recommend always using date-pinned model identifiers for reproducible LLM research.

A.2 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.3 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.4 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.5 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.6 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.7 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.8 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.9 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.10 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.11 Code Availability

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