Cognitive Bias Patterns in LLMs: Anchoring Effects and Debiasing Interventions

Voder AI*
with Tom Howard†

February 2026

Abstract

We investigate whether prompt-based techniques can reduce anchoring bias in LLMs. Testing 8 models on a judicial sentencing paradigm (30 scenarios per condition, temperature=0), we find:

Main finding: Debiasing effectiveness is **structure** + **content**, with **model-specific** effects. A random elaboration control revealed: on **unbiased** Llama 3.3, multi-turn structure alone introduced +6.0mo bias—showing structure harms unbiased models regardless of content. On **biased** GPT-4o, random elaboration achieved 20% reduction while CoT achieved 66%—showing reasoning content provides additional benefit. Sibony's decision architecture techniques showed 0% effect on GPT-4o but 55–67% on GPT-5.2—a complete reversal demonstrating that no universal debiasing technique exists.

Cross-model variation: GPT-4o showed 6.0mo anchoring effect, Opus 4 showed 2.0mo, Sonnet 4 showed 0.0mo. At temp=0, bias is deterministic (SD=0)—a fixed function of weights and prompt.

Practical implications: (1) Test debiasing on your specific deployment model. (2) For unbiased models, avoid multi-turn prompts. (3) For biased models, multi-turn with reasoning content is optimal. (4) Use date-pinned model IDs for reproducibility.

Scope: LLM-only study; no comparison to human performance.

1 Introduction

Recent research has demonstrated that LLMs exhibit cognitive biases such as anchoring, framing effects, and sunk cost fallacy [Binz and Schulz, 2023, Jones and Steinhardt, 2022]. A natural question follows: can prompt-based techniques reduce these biases in LLMs?

Scope: This is an **LLM-only study**. We characterize how LLMs respond to anchoring manipulations and debiasing interventions. We do not compare LLM performance to human performance—our prompts differ from those used in human studies, making direct comparison methodologically unsound. Human studies are cited for context and to motivate debiasing techniques, not as baselines.

Our central finding: The mechanism is **structure** + **content**, with opposite effects on biased vs. unbiased models. A random elaboration control (describing weather, listing unrelated facts) revealed: on **unbiased** Llama 3.3, random elaboration introduced +6.0mo bias—identical to CoT—showing structure alone is harmful. But on **biased** GPT-4o, random elaboration achieved only 20% bias reduction while CoT achieved 66%—showing reasoning content provides substantial additional benefit. This demonstrates: (1) multi-turn structure modifies bias susceptibility; (2) on

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

unbiased models, content is irrelevant—any structure harms equally; (3) on biased models, reasoning content matters—CoT outperforms random elaboration.

We tested two categories of interventions:

1. **Decision architecture techniques** from organizational psychology [Sibony, 2019]—specifically "context hygiene" and "premortem"
2. **Self-Adaptive Cognitive Debiasing (SACD)**—an iterative bias-detection loop [Lyu et al., 2025]

We use anchoring bias as our test case because: (a) it is well-documented in LLMs, (b) the judicial sentencing paradigm provides quantitative measurement, and (c) anchoring is 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 cognitive biases including anchoring, framing effects, and representativeness heuristics. Lou and Sun [2024] found substantial anchoring bias across multiple models. These findings have important implications for AI decision-support systems, as biased model outputs can propagate through applications. Maynard [2025] argues that LLM fluency creates "honest non-signals"—cues that may bypass users' epistemic vigilance.

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 adapt the paradigm from Study 2 of Englich et al. [2006]: 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

SACD Error/Refusal Rates: SACD's iterative prompting occasionally triggered safety refusals or parsing errors. We quantified these rates: Anthropic models showed 0% error rate (0/60 trials), GPT-4o showed 1.4% (2/139 trials). These negligible rates indicate SACD did not systematically bias the analyzable sample.

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 from various providers (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	60	60	0	Paid OpenRouter tier
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. Mistral had high exclusion rate (57%) due to 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 A.12). These are different model generations, not just different identifiers for the same model. When we report debiasing effectiveness, we specify which model was used.

Model identifiers. We use date-pinned model identifiers throughout to ensure reproducibility. Model aliases may route to different backends over time; see Appendix A.12 for details on the identifier variance we observed.

3.4 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 2: 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 × 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 A.12).

3.5 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."

3.6 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. no-debiasing baseline

3.6.1 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."

3.6.2 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.

Scope. This paper characterizes LLM behavior on anchoring tasks. We do not compare to human performance as our prompts differ from prior human studies.

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 as standardized measures of magnitude. 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.

3.6.3 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** [min, max] across our 30 scenario variants—this captures heterogeneity of prompt responses, not sampling uncertainty
- **Means and SDs** where applicable (describing variation across scenarios); when SD=0, all responses were identical
- **Cohen's** d as standardized effect size, interpreted as magnitude of observed difference, not an inferential statistic

Table format: Tables reporting bias effects use the format: `mean ± SD` for condition responses, with n stated in caption. When SD is omitted, all responses in that condition were identical (SD=0). "Obs. Range" columns show [min, max] of the *difference* between conditions across scenario variants.

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 4.4) provide more current validation.

Without debiasing interventions, our baseline model (Codex) showed substantial anchoring bias (3.67mo effect):

Condition	Low Anchor	High Anchor	Diff	Obs. Range	Cohen's d
LLM Baseline (Codex)	5.33 ± 0.96	9.00 ± 0.83	3.67 mo	[3.23, 4.10]	4.09

Table 3: Baseline anchoring bias in Codex. Values show mean ± SD ($n = 30$). Observed range is for the *difference* between conditions across scenario variants. Effect size is very large ($d > 0.8$).

4.2 Sibony Debiasing Techniques

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

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

Table 4: Effect of Sibony debiasing techniques on anchoring bias in Codex ($n = 30$ per condition). Observed ranges reflect scenario variation. Effect sizes remain large ($d > 2$), indicating substantial residual anchoring even after intervention.

4.3 SACD Results

SACD substantially reduced anchoring bias when tested on Codex (baseline: 3.67mo, reduced to near-zero). Note: (1) This experiment used Codex, a deprecated model—results may not transfer to current models. (2) Generic reflection later achieved higher reduction (66%) than SACD (45%) on GPT-4o, suggesting the mechanism is multi-turn structure rather than SACD-specific content.

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 5: 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 effect):

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 6: Debiasing effectiveness on GPT-4o. Target: $n = 30$ per condition; actual $n = 25$–29 due to response validation: (1) JSON parsing failures excluded after 3 retries, (2) duplicate scenario-response pairs from retry logic removed to prevent double-counting. Filtering is scenario-independent (verified by comparing exclusion counts across conditions). Only SACD achieved measurable reduction. Values without ± SD indicate all responses were identical (SD=0). SACD shows SD>0 (3.13 and 6.43 are means) indicating response variation across scenarios.

Key findings:

- **Sibony techniques are model-specific:** Context hygiene and premortem showed *zero* effect on GPT-4o, but 55–67% reduction on GPT-5.2. Effectiveness depends on the target model.
- **Simple instructions fail:** Telling the model "the recommendation is arbitrary, ignore it" had no effect. GPT-4o acknowledged the instruction but still anchored.
- **SACD reduces bias:** SACD achieved 45% reduction on GPT-4o (strong bias). However, see the control experiment below.

4.5 GPT-5.2: Model-Specific Debiasing Reversal

To test whether GPT-4o's null results for Sibony techniques generalize to newer models, we ran a comprehensive debiasing experiment on GPT-5.2 (baseline: 5.97mo effect, same strong-bias tier as GPT-4o):

Condition	n	Low Anchor	High Anchor	Effect	Reduction
Baseline	60	3.03 mo	9.00 mo	5.97 mo	—
Simple instruction	30	6.00 mo	8.00 mo	2.00 mo	67%
Context hygiene	30	5.50 mo	8.20 mo	2.70 mo	55%
Premortem	30	4.00 mo	6.10 mo	2.10 mo	65%
Generic reflection	30	6.00 mo	9.00 mo	3.00 mo	50%
Random elaboration	30	3.00 mo	9.00 mo	6.00 mo	0%

Table 7: GPT-5.2 debiasing results ($n = 30$ per condition except baseline $n = 60$). Values without $\pm$ SD indicate all responses were identical (SD=0); variation in means across conditions reflects different response patterns. Sibony techniques show 55–67% reduction on GPT-5.2 vs 0% on GPT-4o—a complete reversal. Random elaboration (irrelevant content, same structure) shows 0% reduction, confirming CONTENT matters.

Key findings:

- **Sibony techniques transfer to GPT-5.2:** Context hygiene (55%), premortem (65%), and simple instructions (67%) all substantially reduce bias—in stark contrast to 0% effect on GPT-4o.
- **Random elaboration control:** Irrelevant content with same multi-turn structure shows 0% reduction, confirming the CONTENT of debiasing interventions matters on biased models.
- **Model-specific effects:** Practitioners cannot assume debiasing failure on one model (GPT-4o) predicts failure on related models (GPT-5.2). Validate on deployment model.

3-turn random control (turn count isolation): To further isolate whether multi-turn structure alone affects bias, we ran a 3-turn conversation with irrelevant topics (weather, unrelated facts) on GPT-5.2 ($n = 30$). Result: 5.6mo effect vs 5.97mo baseline—only 6% reduction. This confirms SACD's 55% reduction comes from its CONTENT (structured self-critique), not merely from being multi-turn.

4.6 Structure-Matched Control: SACD's Effect is Not Bias-Specific

To test whether SACD's effectiveness stems from its psychology-inspired debiasing content or simply from the multi-turn structure, we ran a **structure-matched control** on GPT-4o. Both conditions use identical 3-turn structure; only the content differs:

Condition	Low Anchor	High Anchor	Effect	Reduction
Baseline	3.00 mo	9.00 mo	6.00 mo	0%
SACD (bias-specific)	3.13 mo	6.43 mo	3.30 mo	45%
Generic Reflection	0.82 mo	2.85 mo	**2.03 mo**	**66%**

Table 8: Structure-matched control ($n = 30$ valid trials). Both conditions use identical 3-turn multi-turn structure. Generic prompts ("Review your answer carefully," "Think step by step") produced *stronger* debiasing than SACD's psychology-specific content, demonstrating that structure—not content—drives the effect.

Structure comparison:

Turn	SACD (psychology-specific)	Generic Reflection
1	"Detect potential bias"	"Review your answer carefully"
2	"Analyze and correct bias"	"Think step by step"
3	"Provide final answer"	"Provide final answer"

Implication: SACD's debiasing effect is *not* attributable to its bias-specific content. The structure-matched generic control achieved *stronger* debiasing (66% vs 45%), demonstrating that multi-turn structure—not psychology-specific framing—drives the effect.

4.6.1 Random Elaboration Control: Identifying the Mechanism

To distinguish whether the bias-introduction effect on unbiased models comes from reasoning content ("think step by step") or multi-turn structure itself, we ran a random elaboration control with the same multi-turn structure but irrelevant content. We first ran this on Llama 3.3 and GPT-4o, then extended it to additional models (Opus 4 and Sonnet 4 dated):

1. "Before answering, describe the weather in a hypothetical city in detail."
2. "Now list 5 completely unrelated facts about any topic."
3. "Finally, provide your answer to the original question."

Model	Intervention	Low Anchor	High Anchor	Effect
Llama 3.3 (unbiased)	Baseline	6.0 mo	6.0 mo	0.0 mo
	Generic CoT	6.0 mo	12.0 mo	**+6.0 mo**
	Random Elaboration	6.0 mo	12.0 mo	**+6.0 mo**
GPT-4o (biased)	Baseline	∼3 mo	∼9 mo	6.0 mo
	Generic CoT	—	—	2.03 mo (66% ↓)
	Random Elaboration	6.0 mo	10.8 mo	4.8 mo (20% ↓)
Opus 4 (low-bias)	Baseline	6.0 mo	6.0 mo	0.0 mo
	Random Elaboration	6.0 mo	6.0 mo	0.0 mo
Sonnet 4 dated (low-bias)	Baseline	6.0 mo	6.0 mo	0.0 mo
	Random Elaboration	6.0 mo	6.0 mo	0.0 mo

Table 9: Random elaboration control extended to four models. On **unbiased** Llama 3.3, random elaboration matches CoT (+6.0mo), indicating structure alone can introduce bias. On **biased** GPT-4o, random elaboration yields only partial reduction (20%) versus CoT (66%), indicating added reasoning content matters. On Opus 4 and Sonnet 4 dated (both low-bias baselines), random elaboration remains anchor-invariant (0.0mo effect, $n = 30$ each).

Mechanism decomposition: The random elaboration control separates structure from content effects:

- **Structure effect (multi-turn alone) is model-specific:** On GPT-4o, structure gives partial improvement (20% reduction). On Llama 3.3, structure introduces +6.0mo bias. On Opus 4 and Sonnet 4 dated, structure leaves behavior unchanged (0.0mo effect).

- **Content effect (reasoning addition):** On GPT-4o, reasoning content provides an additional 46% reduction (66% total for CoT vs 20% for random elaboration). On Llama 3.3, no additional effect (CoT = random).

Practical implications: Do not assume multi-turn structure has a universal effect. Measure baseline and structure-only controls on the target model before selecting a debiasing strategy.

4.6.2 Cross-Model Replication: Generic Reflection is Model-Specific

To test whether the GPT-4o finding generalizes, we ran the same generic reflection experiment on five models with varying baseline bias:

Model	Baseline	Generic Reflection Effect	Reduction	Pattern
GPT-4o	6.0 mo	2.03 mo	66% ↓	Helps
Opus 4.5	5.0 mo	1.24 mo	75% ↓	Helps
Sonnet 4.5	3.0 mo	0.10 mo	97% ↓	Helps
Sonnet 4 (dated)	0.0 mo	**3.07 mo**	↑	**Hurts**
Llama 3.3	0.0 mo	**6.10 mo**	↑	**Hurts**

Table 10: Generic reflection across 5 models ($n = 30$ per anchor condition, 60 total per model; all at temp=0). On biased models, generic reflection reduces bias (66–97%). On **both** unbiased models (Sonnet 4 dated, Llama 3.3), it *introduces* substantial bias (3–6 months). Values without ± SD indicate all responses were identical (SD=0).

Key finding: Generic reflection is a double-edged sword:

- On **biased models** (GPT-4o, Opus 4.5, Sonnet 4.5): Generic reflection reduces anchoring (66–97% improvement)
- On **unbiased models** (Sonnet 4 dated, Llama 3.3): Generic reflection *introduces* anchoring (3.07–6.10mo effect where baseline was 0.0mo)

Implication: Neither generic reflection nor SACD is universally safe. Generic reflection introduces consistent bias on unbiased models. SACD produces pathological outputs (extreme variance, 120mo outliers on Llama 3.3). **Practical recommendation:** Do not apply debiasing interventions to models without confirmed baseline bias. Measure baseline first, then intervene only if needed.

4.6.3 Temperature Sensitivity of Debiasing

All primary debiasing experiments used temperature=0. To test whether debiasing effectiveness transfers to higher temperatures, we ran both baseline and generic reflection (CoT) on GPT-4o at temp=0.7 and temp=1.0:

Temperature	Baseline Effect	CoT Effect	Reduction	95% CI (CoT effect)	Pattern
0	6.0 mo	2.03 mo	66%	deterministic (temp=0)	CoT helps
0.7	6.0 mo	6.0 mo	**0%**	[6.00, 6.00]	**CoT fails**
1.0	5.93 mo	1.67 mo	**72%**	[1.01, 2.30]	CoT helps more

Table 11: Debiasing effectiveness across temperatures on GPT-4o ($n = 28$–30 per condition). Baseline anchoring effect is stable across temperatures (∼6mo). For stochastic conditions (temp>0), we report bootstrap 95% confidence intervals (5,000 resamples over trials). CoT debiasing shows **non-monotonic** pattern: works at temp=0 (66%), *fails* at temp=0.7 (0%), and recovers at temp=1.0 (72%).

Key finding: CoT debiasing effectiveness is **non-monotonic with temperature**. There is a "dead zone" at temp=0.7 where CoT provides no benefit over baseline. At temp=1.0, CoT achieves stronger debiasing (72%) than at temp=0 (66%), with stochastic-run uncertainty still clearly separated from baseline-scale effects. Critically, baseline anchoring remains stable across temperatures (∼6mo), confirming that CoT—not temperature alone—is the debiasing mechanism at temp=1.0.

Additional replication (Opus 4.5, temp=0.7): We ran a targeted top-up replication of generic reflection at temp=0.7 on Opus 4.5 ($n = 27$ valid: low $n = 15$, high $n = 12$). The observed effect was 1.48 months (low mean 0.27, high mean 1.75), indicating debiasing signal persisted rather than collapsing to a GPT-4o-style dead zone at 0.7.

GPT-5.2 temperature sweep with intervention comparison: To test whether intervention type matters at temp>0, we ran GPT-5.2 at temperatures 0.5, 0.7, and 1.0 with both simple instruction ("ignore the randomly determined demand") and generic reflection (CoT):

Temperature	Simple Instruction	Generic Reflection	Winner
0.5	**90%** reduction	35% reduction	Simple
0.7	**57%** reduction	40% reduction	Simple
1.0	**49%** reduction	27% reduction	Simple

Table 12: GPT-5.2 debiasing effectiveness by intervention type across temperatures ($n = 20$ per condition per temperature). Simple instruction (Sibony-style) dominates at ALL temp>0 settings. Peak effectiveness: 90% at temp=0.5.

Key finding: Simple Sibony-style instruction ("this value was randomly determined") **dominates** generic reflection at all non-zero temperatures tested. The simple instruction peaks at temp=0.5 (90% reduction) and degrades at higher temperatures (57% at 0.7, 49% at 1.0), while generic reflection remains relatively flat (27–40%). This suggests that for production deployments using temp>0, practitioners should prefer simple debiasing instructions over generic CoT prompts.

Implications: (1) Debiasing findings are not artifacts of deterministic sampling. (2) Temperature interacts with debiasing interventions in model-specific ways. (3) Practitioners should test debiasing at their target temperature and model, not assume transfer from temp=0 or across providers.

4.6.4 SACD Pathological Outputs on Llama 3.3

To characterize SACD's failure mode on unbiased models, we ran SACD on Llama 3.3 ($n = 30$, 15 per anchor condition):

Output (months)	Low Anchor (3mo)	High Anchor (9mo)
0	3	6
1	2	1
6	1	0
12	6	5
120 (outlier)	3	3
Mean	29.33 mo	28.07 mo

Table 13: SACD output distribution on Llama 3.3 ($n = 15$ per condition). Baseline: all responses = 6mo. SACD produces extreme variance with 120mo outliers (20% of trials). Notably, the pattern is **reversed**: high anchor produces *more* 0mo responses (6 vs 3), suggesting complete disruption rather than bias introduction. Effect: -1.26mo.

Key observation: SACD's iterative rewriting appears to strip essential context from Llama 3.3's reasoning, producing chaotic outputs rather than the consistent bias introduction seen with generic reflection. This represents a different failure mode: generic reflection *adds* bias consistently; SACD *breaks* the model's output distribution.

4.7 Cross-Model Validation

Cross-model comparison reveals varying anchoring susceptibility across our tested models. **Critical limitation:** most models in Table 14 still rely on a single prompt template. Prompt sensitivity testing on Sonnet 4.5 showed **92% effect reduction from paraphrasing alone**, demonstrating this risk. To partially address this, we added prompt-style robustness checks for GPT-4o and Opus 4 (Section 5.2), where their relative ordering remained stable across variants. **Still, broader multi-template coverage is needed before claiming fully robust global rankings.** Additional limitations: (1) sample sizes differ across models; (2) we tested only 1–2 models per provider:

Model	n (valid)	Anchoring Effect	Behavior
Sonnet 4	60	0.00 mo	No bias
Claude Opus 4	60	2.00 mo	Moderate bias
Mistral (7B)	52	0.00 mo	No bias
Hermes 3 (405B)	60	−0.33 mo	No bias
Llama 3.3 (70B)	60	6.00–9.00 mo	Strong bias†
Sonnet 4.5	60	3.00 mo	Moderate bias
GPT-4o	60	6.00 mo	Strong bias

Table 14: Cross-model anchoring bias, sorted by effect magnitude. All tests at temp=0 with $n = 30$ per anchor condition (60 total valid per model) except Mistral ($n = 26$ per condition due to parse failures). Seven stable models shown (Nemotron attempted but excluded for reliability constraints). †Llama 3.3 showed 0.0mo with our original prompt (which includes "randomly determined" disclaimer) but 6.0–9.0mo on variants without this disclaimer—the original prompt contains implicit debiasing. This is consistent with GPT-5.2 where Sibony-style interventions also showed strong effect. **Caution:** Results are model-specific observations, not provider-level generalizations.

Observation: Anchoring susceptibility varies across tested models.

1. **Observed pattern (not validated):** Across 7 stable models, we observe varying susceptibility:
   - **No bias:** Sonnet 4, Mistral 7B, Hermes 405B
   - **Moderate bias (2–3 months):** Opus 4, Sonnet 4.5
   - **Strong bias (>5 months):** GPT-4o, Llama 3.3†
2. **Open-weights models show mixed results:** Mistral (7B) shows 0.0mo effect, but Llama 3.3 shows 6.0–9.0mo across prompt variants. Open-weights training does not guarantee anchoring resistance.
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-provider variation:** Sonnet 4 shows 0.0mo effect while Opus 4 shows 2.0mo (both Anthropic), suggesting bias resistance varies even within the same provider. Model scale or fine-tuning differences may affect anchoring susceptibility.

4.8 Knowledge of Bias ≠ 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 effect**
Sonnet 4	✓Yes	✓Correctly	✓**0.0mo (immune)**

Table 15: 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. **Knowledge is insufficient:** GPT-4o can describe anchoring bias but still exhibits it. Awareness alone does not prevent the bias from affecting outputs.

This knowledge-behavior dissociation is *consistent with* (though does not prove) our preliminary soft/hard hypothesis (Section A.14)—but alternative explanations remain possible.

4.9 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 16: 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).

4.10 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 17: 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

Temperature sensitivity varies by model. We tested temperature effects across five models and observed that some models maintain constant bias regardless of temperature (GPT-4o, GPT-4.1, Opus 4.5), while others show reduced bias at higher temperatures (Sonnet 4.5). See Appendix A.14 for the full analysis and preliminary "soft vs hard bias" hypothesis.

Determinism at temp=0. At temperature=0, bias is deterministic (SD=0)—the same prompt produces the same biased output every time. This has implications for auditing and deployment; see Appendix A.15 for details.

5.1 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.

5.2 Prompt Robustness Across Models

To address prompt-template concerns in cross-model comparisons, we ran style variants (original, casual, structured) and checked whether model ranking changed across prompt formulations.

Model	Original	Casual	Structured	Classification
Consistently Biased				
GPT-4o	6.0 mo	4.8 mo	5.1 mo	Biased
Hermes 405B	5.1 mo	2.4 mo	1.2 mo	Biased
Prompt-Sensitive (Debiasable)				
GPT-5.2	4.4 mo	**0.7 mo**	5.7 mo	Mixed†
Llama 3.3	4.0 mo	4.0 mo	**0.0 mo**	Mixed‡
Prompt-Resistant Moderate Bias				
Opus 4.5	2.0 mo	2.0 mo	2.2 mo	Moderate
Consistently Low-Bias				
Sonnet 4	0.2 mo	0.0 mo	0.6 mo	Low-bias
Opus 4.0	0.0 mo	0.7 mo	0.0 mo	Low-bias
Mistral Medium 3	0.3 mo	0.0 mo	0.0 mo	Low-bias
Nemotron 253B	−0.2 mo	0.4 mo	0.0 mo	Low-bias

Table 18: Cross-model prompt robustness (3 prompt styles, $n = 20$ per condition/style). †GPT-5.2: casual variant (removing "randomly determined" disclaimer) *debiases*; structured variant *increases* bias. ‡Llama 3.3: structured variant eliminates bias entirely. Bold values indicate debiased conditions.

Key finding: Prompt framing has opposite effects across models. On GPT-5.2, removing the "randomly determined" disclaimer debiases (4.4→0.7mo), while on Llama 3.3 the structured prompt debiases (4.0→0.0mo). This demonstrates that **no universal debiasing prompt exists**—interventions must be validated on the specific target model.

These results support model classification stability for 7 of 9 models (consistently biased or consistently low-bias across prompt variants). The two "Mixed" models (GPT-5.2 and Llama 3.3) show that prompt engineering can serve as a debiasing intervention, but the effective strategy differs by model.

5.3 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 3.5).

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 19: 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.

5.4 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.

5.5 Opposite Debiasing Effects Across Model Families

A striking finding emerged from our prompt robustness experiments: the same debiasing intervention produced **opposite effects** on different models.

Model	Original	Casual	Structured	Optimal Strategy
Llama 3.3	4.0 mo	4.0 mo	**0.0 mo**	Use structured prompt
GPT-5.2	4.4 mo	**0.7 mo**	5.7 mo	Use casual prompt

Table 20: Opposite effects of prompt framing. On Llama 3.3, structured prompts eliminate bias. On GPT-5.2, casual prompts (without "randomly determined" disclaimer) debias, while structured prompts *increase* bias to 5.7mo. $n = 20$ per condition per variant.

The "randomly determined" disclaimer in our original prompt explicitly tells the model that the anchor value is arbitrary and should not influence judgment. Following Sibony's decision hygiene principles Sibony [2019], this should reduce anchoring by flagging the anchor as irrelevant.

On **Llama 3.3**, a structured prompt (clear role assignment, formatted case details) eliminates bias entirely (0.0mo effect). However, the "randomly determined" disclaimer has no additional effect—both original and casual variants show 4.0mo bias.

On **GPT-5.2**, the pattern is reversed: removing the "randomly determined" disclaimer (casual variant) *debiases* the model (0.7mo effect), while adding structure *increases* bias to 5.7mo. One hypothesis: GPT-5.2 may interpret "randomly determined" as *relevant experimental framing* rather than dismissable metadata, causing it to attend more closely to the anchor value.

Implications for practitioners:

1. **No universal debiasing prompt exists.** Interventions must be validated on the specific target model.

2. **Sibony-style techniques can backfire.** The "acknowledge arbitrariness" intervention is not universally safe.

3. **Model-specific optimization is required.** Prompt engineering for debiasing cannot be assumed to transfer across model families.

This finding transforms what initially appeared to be a methodological limitation (prompt sensitivity) into a **core contribution**: prompt-model interactions are a first-order phenomenon that must be characterized, not a confound to be controlled away.

5.6 Iterative Self-Correction Was Effective in Our Tests

SACD outperformed static prompt interventions in our GPT-4o experiments. However, our generic reflection control (Section 4.6) revealed that SACD's effectiveness is *not* due to its bias-specific content. Generic prompts ("think step by step," "review your answer") with the same multi-turn structure achieved 66% reduction vs. SACD's 45%. This suggests the mechanism is primarily increased reasoning process rather than SACD-specific psychology framing.

Turn/length confound addressed: We ran multiple controls to isolate content effects from structural effects:

- **3-turn random control (GPT-5.2):** Same turn count as SACD but irrelevant content (weather, unrelated facts). Result: 5.6mo effect vs 5.97mo baseline—only 6% reduction vs SACD's 55%. This rules out turn count as the driver.

- **Random elaboration control (GPT-5.2):** Same structure, irrelevant content. Result: 6.0mo effect (0% reduction). SACD achieves 45% reduction with identical structure but different content.

Conclusion: The debiasing benefit comes from intervention *content* (structured self-critique in SACD, reasoning steps in generic reflection), not merely from multi-turn structure or longer outputs.

5.7 Preliminary Hypothesis: Two Patterns Observed in Our Tested Models

Based on observations from our five tested models (GPT-4o, GPT-4.1, Opus 4.5, Sonnet 4.5, Llama 3.3), we *tentatively propose* a hypothesis about bias patterns. **This is a preliminary observation**

from five models on one bias type, 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 temperature sweep experiments on **five models**. While more robust than our initial two-model observation, it remains preliminary: (1) tested on **one bias type** (anchoring) only; (2) HARD is the majority pattern (3/5 models), suggesting SOFT may be exceptional; (3) the alias/dated variance we discovered (Section A.12) complicates interpretation. We present this as a **preliminary taxonomy**, not an established framework.

5.8 Limitations

Descriptive Study Framing:

- This is an exploratory descriptive study. Primary experiments used deterministic sampling (temperature=0); temperature sweep experiments (0.0–1.0) were performed on **five models** and **one bias type** (anchoring). Temperature effects on other bias types remain unexplored

- 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 were conducted on five models (GPT-4o, GPT-4.1, Opus 4.5, Sonnet 4.5, Llama 3.3) for the anchoring task only—we did not systematically test temperature effects for other 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
- **Automated extraction:** Response extraction used deterministic JSON schema validation—models were prompted with explicit schemas and responses were parsed programmatically. For models with zero variance (e.g., Anthropic at temp=0), every trial produced identical JSON; no interpretation was required. Exclusions represent API failures or schema violations (malformed JSON, empty responses), not ambiguous cases requiring human judgment. Inter-rater reliability is not applicable to deterministic parsing
- Simplified case vignettes vs. original Englich et al. materials (though core paradigm preserved)
- Computational cost of SACD/DeFrame (2–3× API calls per decision)
- **Debiasing harms unbiased models:** On GPT-4o, generic reflection outperformed SACD (66% vs 45%). But on unbiased models, both interventions caused problems: generic reflection introduced consistent bias (3–6mo), while SACD produced pathological outputs on Llama 3.3 (extreme variance, 120mo outliers). Neither intervention is universally safe. Debiasing should only be applied to models with confirmed baseline bias
- **No human/random baseline for debiasing:** Debiasing effectiveness was measured against no-intervention LLM baseline, not against human debiasing rates or random response distributions. We cannot claim debiasing brings LLM performance to "human level" without human data on our specific debiasing prompts
- **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. Mistral had high exclusion rate ($68/120 = 57\%$) due to difficulty following JSON output format. **Exclusions were verified as scenario-independent:** $34/60$ excluded in each anchor condition (low=34, high=34), confirming exclusions were formatting failures, not content-dependent.

Generalizability:

- Cross-model validation spans 8 models from various providers 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 five models** (GPT-4o, GPT-4.1, Opus 4.5, Sonnet 4.5, Llama 3.3) on one bias type
- 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

5.9 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

Main finding: Structure + content with model-specific effects. Our most important result: a random elaboration control decomposed the debiasing mechanism into structure and content components. On **unbiased** Llama 3.3, random elaboration introduced +6.0mo bias—identical to CoT—showing structure alone harms unbiased models. On **biased** GPT-4o, random elaboration achieved 20% reduction while CoT achieved 66%—showing reasoning content provides substantial additional benefit (46% more). The mechanism is not purely structural: on biased models, "think step by step" outperforms "describe the weather."

Implication for debiasing research: (1) For unbiased models: avoid multi-turn prompts entirely—any structure introduces bias. (2) For biased models: use multi-turn with reasoning content for optimal debiasing. (3) Always measure baseline bias first to determine strategy.

Additional findings:

1. **Deterministic bias (SD=0)**: At temp=0, LLM bias is not noise—it is a fixed function of weights and prompt. Every trial produces the exact same output, making bias both auditable and consequential.

2. **Model identifier variance**: Alias vs date-pinned identifiers produced qualitatively different results (Sonnet 4.5 via alias: 3.0mo; Sonnet 4 via dated ID: 0.0mo). Use date-pinned identifiers for reproducibility.

3. **Sibony techniques are model-specific**: Context hygiene and premortem showed 0% effect on GPT-4o but 55–67% reduction on GPT-5.2.

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

Recommendations: (1) Include structure-matched controls in debiasing studies. (2) Use date-pinned model identifiers. (3) Test interventions on your specific deployment model—techniques do not transfer across models.

Limitations: Moderate sample sizes ($n = 30$ per condition), generic reflection tested on five models, and incomplete prompt-template coverage for cross-model comparisons (we added style-variant checks for GPT-4o/Mistral Medium 3/Opus 4, but not all models). Nemotron was attempted but excluded from main robustness claims due to reproducibility/reliability constraints on the current route.

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.

Andrew D. Maynard. The ai cognitive trojan horse: The epistemic risks of ai-generated content disguised as human through honest non-signals. *arXiv preprint arXiv:2601.07085*, 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 21: 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 Random Elaboration Control Prompts

To test whether debiasing effects arise from structured self-critique *content* versus multi-turn *structure* alone, we designed a control condition using semantically random elaboration matched on turn count and token length.

3-Turn Random Conversation Control:

Turn 1 (System → Model):

> Before we proceed to the main task, let's have a brief conversation about something unrelated.
>
> Tell me: what's your favorite season of the year and why?

Turn 2 (Model response, then System → Model):

> Interesting perspective! Now, a completely different topic—if you could have dinner with any historical figure, who would it be and what would you ask them?

Turn 3 (Model response, then main task):

> Great choices! Now let's move on to the main task.
>
> [Case vignette and anchor question as in baseline]

Rationale: This control preserves the multi-turn structure of SACD (3 turns before the main judgment) but replaces bias-related self-critique with semantically irrelevant conversation. If SACD's debiasing effect comes from structured reflection *content*, the random control should show baseline-level bias. If the effect comes merely from multi-turn processing, the random control should show similar reduction.

Result: GPT-5.2 showed 5.6mo effect with 3-turn random conversation (vs. 6.0mo baseline), confirming that SACD's 55% reduction comes from structured self-critique *content*, not turn count or conversational warm-up.

A.11 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.12 Model Identifier Variance

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 22: 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.

Implications for LLM research:

1. **Reproducibility confound:** Model providers may silently update alias targets. Studies using aliases may not replicate even with identical prompts.
2. **Checkpoint-specific behavior:** Bias magnitude is checkpoint-specific, not just architecture-specific.
3. **Recommendation:** Researchers should always use and report date-pinned model identifiers.

A.13 Code Availability

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

A.14 Soft vs Hard Bias Patterns: Extended Analysis

Our observations suggest that debiasing interventions effective on one model may have no effect on another. We tested temperature sensitivity across **five models** and observed two distinct patterns:

Model	Baseline	temp=0.5	temp=1.0	Pattern
GPT-4o	6.00 mo	6.00 mo	6.00 mo	**HARD**
GPT-4.1	3.10 mo	3.20 mo	3.20 mo	**HARD**
Opus 4.5	5.00 mo	5.00 mo	5.00 mo	**HARD**
Sonnet 4.5	3.00 mo	—	**0 mo**	**SOFT**
Llama 3.3	0.00 mo	0.00 mo	0.10 mo	**SOFT**

Table 23: Temperature sensitivity across 5 models (n = 30–60 per temperature). HARD models show constant bias regardless of temperature. SOFT models show low bias across all temperatures.

"Hard bias" pattern (observed in GPT-4o, GPT-4.1, Opus 4.5): Bias magnitude remains constant regardless of temperature (0→0.5→1.0). Three models show this pattern. This *might* suggest the bias is embedded in the model's weights or reasoning process—not merely a surface-level decoding artifact.

"Soft bias" pattern (observed in Sonnet 4.5, Llama 3.3): Bias is low at baseline or decreases with temperature. Sonnet 4.5 shows 100% reduction at temp=1.0. Llama 3.3 shows 0.0mo baseline effect.

Contamination probe: We asked both models whether they were familiar with anchoring bias in judicial sentencing and whether they could predict the expected pattern. Both models demonstrated clear knowledge 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.

Important caveats:

- This distinction is based on five models—more robust than our initial two-model observation, but still limited
- The majority pattern is HARD (3/5 models); SOFT may be the exception rather than the rule
- We cannot rule out that observed differences reflect API routing, checkpoint differences, or other confounds

A.15 Deterministic Bias: Extended Discussion

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. 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 bias. This determinism has important implications:

- **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.

Clarification on "deterministic": We use "deterministic" to mean that at temp=0, the same input produces the same output across runs. This does not claim that the internal token-by-token generation process is non-probabilistic—LLMs still sample from probability distributions, but temp=0 selects the argmax at each step, making the sequence deterministic. Our point is about *output consistency*, not claims about internal mechanisms.