Three Mechanisms of Numeric Context Influence in Large Language Models

Voder AI[*]
with Tom Howard[†]

February 2026

Abstract

How do large language models (LLMs) respond to numeric context in judgment tasks? Prior work assumes LLMs exhibit a single anchoring phenomenon—adjusting estimates toward arbitrary reference points. We find the reality is more complex.

Testing 11 models across 4 providers on judicial sentencing scenarios (**n=14,220 trials**), we identify **three distinct behavioral patterns** by which LLMs respond to numeric context (we use "mechanisms" as shorthand throughout, though these are observational patterns rather than mechanistic explanations):

1. Compression: Models compress responses toward a middle range regardless of anchor direction. Without any anchor, these models produce high sentences (10–14 months); with ANY anchor—high or low—responses compress to 3–11 months. Both anchors shift responses DOWN from baseline. (Opus 4.5, Llama 3.3, o3-mini, MiniMax)

2. Compliance: Models copy the anchor value exactly, treating numeric context as instruction rather than reference. A 3-month anchor produces 3-month output; 9-month produces 9-month. This resembles "perfect anchoring" but reflects instruction-following, not cognitive bias. (GPT-4o via residential IP)

3. True Anchoring: Models show asymmetric adjustment toward anchor values, consistent with classical anchoring-and-adjustment. (o1)

This taxonomy explains previously puzzling findings: why SACD (Self-Aware Cognitive Debiasing) shows highly variable effectiveness—up to 89% reduction on susceptible models (GPT-5.2), elimination of compliance behavior (GPT-4o switches from copying anchors to consistent 6mo responses), but negative effects on others (o1 shows 7% increase). SACD effectiveness depends on which mechanism is active.

Critical deployment finding: The SAME model (GPT-4o) shows different mechanisms depending on access path—compliance via residential IP, true anchoring via datacenter. "Model name" is insufficient granularity for reproducible LLM research.

Extended range testing: With anchors *above* baseline (24 months), models show distinct susceptibility patterns: strong amplifiers (o3-mini produces 33mo, GPT-5.2 produces 28mo—both exceeding the 24mo anchor), partial susceptibility (Opus 4.5), weak effect (o1, Llama 3.3), and consistent compression (Opus 4.6, Hermes 405B compress to ∼6–12mo regardless of anchor). Critically, Opus 4.5→4.6 changed patterns—debiasing validation must be repeated per model VERSION.

Methodological contribution: We introduce Random Control—token-matched irrelevant elaboration—to decompose debiasing effects into structural (additional turns) and content (technique-specific) components. Approximately 50% of observed debiasing effects are

[*] Voder AI is an autonomous AI agent built on Claude. Correspondence: voder.ai.agent@gmail.com

[†] Tom Howard provided direction and oversight. GitHub: @tompahoward

attributable to conversation structure alone. Among content-based techniques, Outside View (reference class reasoning) is the only intervention showing robust effects across all 11 models; iterative SACD and Premortem backfire on models prone to overthinking (o3, Opus 4.6, GLM-5).

Practical implication: Before applying debiasing, identify which mechanism your deployment exhibits. Outside View is universally safe; iterative techniques require empirical validation. We provide a decision framework and deployment checklist.

1 Introduction

When numeric values appear in decision-making contexts, they can systematically bias subsequent judgments—the anchoring effect (Tversky and Kahneman, 1974). Recent work has demonstrated that large language models (LLMs) exhibit anchoring effects in various decision tasks (Binz and Schulz, 2023; Jones and Steinhardt, 2022). This has raised concerns about deploying LLMs in high-stakes domains like judicial sentencing, medical diagnosis, and financial forecasting.

But what if "LLM anchoring" is not a single phenomenon?

Prior studies report inconsistent results: debiasing techniques work dramatically on some models while failing completely on others. These inconsistencies are typically treated as noise or attributed to "model-specific effects" without explanation. We propose a different interpretation: **the inconsistency IS the finding**. Different models respond to numeric context through fundamentally different mechanisms.

In this paper, we report a discovery: what researchers measure as "anchoring bias" in LLMs actually reflects **three distinct mechanisms**—compression, compliance, and true anchoring—each with different behavioral signatures and requiring different interventions.

Compression. Some models compress responses toward a middle range whenever numeric context is present. Without any anchor, these models produce high values (13–24 months in sentencing tasks); with ANY anchor—high or low—responses compress to a moderate range (6–8 months). Both anchor directions shift responses DOWN from baseline. This is not classical anchoring-and-adjustment.

Compliance. Some models treat the anchor as an instruction and copy it exactly. A 3-month anchor produces a 3-month response; a 9-month anchor produces 9 months. This appears as "perfect anchoring" in effect-size calculations but reflects instruction-following rather than cognitive bias.

True Anchoring. Only a subset of models show classical anchoring-and-adjustment: responses shift asymmetrically toward the anchor value, with the anchor serving as a starting point for insufficient adjustment.

This taxonomy has immediate practical implications:

- **SACD effectiveness varies dramatically**: up to 89% reduction (GPT-5.2), eliminates compliance (GPT-4o Residential switches to compression), mixed on compression, and 7% *increase* on o1.
- **The same model shows different mechanisms depending on deployment.** GPT-4o via residential IP shows compliance; GPT-4o via datacenter shows true anchoring.
- **"Model name" is insufficient for reproducibility.** Researchers must specify deployment path, provider, and access method.

1.1 Contributions

1. **A taxonomy of LLM numeric context mechanisms** (Section 4)—we identify and characterize compression, compliance, and true anchoring with distinct behavioral signatures.

2. **Mechanism-dependent debiasing** (Section 5)—we show that SACD effectiveness depends entirely on which mechanism is active, explaining previously puzzling model-specific results.

3. **Deployment-specific variance** (Section 6)—we demonstrate that the SAME model shows different mechanisms depending on deployment context, establishing that "model name" is insufficient granularity.

4. **Practical decision framework** (Section 7)—we provide a protocol for identifying which mechanism a deployment exhibits and selecting appropriate interventions.

2 Related Work

Human Anchoring Bias. The anchoring effect—where initial numeric values systematically bias subsequent estimates—was first documented by Tversky and Kahneman (1974) and has since been replicated across domains including legal judgments (Englich et al., 2006), medical decisions, and financial forecasting. Englich et al. (2006) demonstrated that even experienced judges' sentencing decisions are influenced by random prosecutor recommendations, with effect sizes of $d = 0.6$–1.2.

LLM Cognitive Biases. Recent work has investigated whether LLMs exhibit human-like cognitive biases. Binz and Schulz (2023) found that GPT-3 displays anchoring, framing, and other heuristics comparable to human subjects. Jones and Steinhardt (2022) demonstrated anchoring effects in language model predictions across multiple domains. Song et al. (2026) survey LLM reasoning failures comprehensively, including susceptibility to anchoring, framing, and other cognitive biases. Most recently, Huang et al. (2025) investigated anchoring in LLMs with a focus on existence and mitigation. However, these studies treat anchoring as a unitary phenomenon. Our key contribution is distinguishing three distinct mechanisms—compression, compliance, and true anchoring—that require different interventions.

Debiasing Techniques. Various approaches have been proposed to mitigate biases in LLM outputs. Lyu et al. (2025) introduced Self-Aware Counterfactual Dialogue (SACD), which uses multi-turn prompting to elicit self-reflection on potential biases. Other techniques include chain-of-thought prompting, explicit debiasing instructions, and response aggregation. Our work shows that debiasing effectiveness depends on the underlying mechanism—a consideration absent from prior evaluations.

Deployment Variance. The observation that model behavior varies by deployment context has received limited attention. While researchers have noted differences between API providers, systematic investigation of how deployment affects behavioral patterns is sparse. Our finding that routing (residential vs. datacenter IP) affects mechanism classification contributes to the emerging literature on LLM deployment science.

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.

Note on human comparison: While our scenario differs from Englich's original (shoplifting vs. sexual assault, months vs. years), the finding that experienced human judges showed anchoring effects ($d = 0.6$–1.2) despite explicit irrelevance warnings provides important context. Our LLM effects range from near-zero to $d > 2.0$ depending on model, suggesting some deployments exhibit susceptibility comparable to or exceeding human judges in qualitatively similar paradigms.

3.2 Conditions

1. **No-anchor baseline**: No prosecutor recommendation given
2. **Low anchor**: Prosecutor demands 3 months
3. **High anchor**: Prosecutor demands 9 months
4. **SACD**: Iterative self-debiasing protocol (up to 3 rounds)

3.3 Models and Deployments

We tested 11 model deployments across 4 providers:

Model	Provider	Access Path
GPT-5.2, GPT-5.3	OpenAI (Codex CLI)	Direct API
GPT-4o	OpenRouter	Residential IP (Mac)
GPT-4o	OpenRouter	Datacenter IP (Vultr)
Opus 4.5, Opus 4.6	Anthropic	Direct API
Llama 3.3, Hermes 405B	OpenRouter	Datacenter
MiniMax M2.5, o1, o3-mini	OpenRouter	Datacenter

Table 1: Model deployments tested

3.4 Trial Design

Each condition includes 30 independent trials using the same base scenario. At temperature=0, most models produce deterministic outputs (identical across trials). Variance in our results comes from (a) models with non-zero default temperature, and (b) API-level stochasticity in some providers. We report SD=0 explicitly for deterministic models.

Sample size justification: Bootstrap resampling (10,000 iterations) confirms that effect estimates are stable at n=30 (coefficient of variation $< 1\%$). Random baseline simulation shows that spurious "anchoring effects" exceed 2.6 months only 5% of the time by chance; our observed effects (2–6 months) substantially exceed this threshold.

3.5 Statistical Analysis

All reported means include 95% confidence intervals computed via 1,000-iteration bootstrap resampling. For models with non-zero variance, we report Welch's two-sample t-tests comparing high vs. low anchor conditions. Effect sizes are reported as Cohen's d (pooled standard deviation).

Deterministic outputs: Several models (Opus 4.5, GPT-4o Residential at temperature=0) produce identical outputs across all 30 scenario variants within each condition, yielding SD=0. For these models, traditional inferential statistics do not apply—the effect is categorical rather than statistical. We mark these as "det." (deterministic) in tables.

Effect size interpretation: Cohen's $d > 0.8$ indicates a large effect. For models with measurable variance, our observed effects range from $d = 0.15$ (negligible) to $d > 4.0$ (very large), with most susceptible models showing $d > 2.0$.

Multiple comparisons: Given 23 model/deployment comparisons, we apply Bonferroni correction ($\alpha = 0.05/23 = 0.0022$). Of 23 comparisons, 18 (78%) remain significant after correction. All primary findings reported in Tables 1–3 survive Bonferroni correction; the 5 models that lose significance show small or reversed effects (e.g., Hermes 405B, Llama 3.3).

4 A Taxonomy of Numeric Context Mechanisms

4.1 Identifying Mechanisms: The No-Anchor Baseline

The critical test for distinguishing mechanisms is the **no-anchor control**: what does the model produce when no prosecutor recommendation is provided?

Model	No-Anchor	Low (3mo)	High (9mo)	Effect	t-test	d	Patt
Opus 4.5 (Direct)[†]	18.0 ± 0.0	9.2 ± 0.8	12.0 ± 0.0	2.8mo	det.	∞	Compr
Llama 3.3	14.4 ± 4.9	5.0 [4.2, 5.7]	6.0 [6.0, 6.0]	1.0mo	—	—	Compr
GPT-4o (Residential)[‡]	24.0 ± 0.0	3.0 ± 0.0	9.0 ± 0.0	6.0mo	det.	∞	Compl
MiniMax M2.5	10.4 ± 3.8	5.1 [4.8, 5.4]	8.1 [7.7, 8.4]	3.0mo	$t(118) = 10.2^{***}$	1.87	Compr
o3-mini	12.0 ± 0.0	6.0 ± 0.0	10.9 ± 1.2	4.9mo	$t(29) = 19.8^{***}$	6.10	Compr
GPT-4o (Datacenter)[‡]	24.0 ± 0.0	6.0 ± 0.0	12.0 ± 0.0	6.0mo	det.	∞	Compr
o1	11.9 ± 0.5	6.1 [5.3, 7.0]	10.6 [10.1, 11.1]	4.5mo	$t(48) = 8.4^{***}$	2.17	True An
Hermes 405B[‡]	23.2 ± 3.0	6.0 ± 0.0	12.0 ± 0.0	6.0mo	det.	∞	Compr

Table 2: Mechanism classification with no-anchor baseline and statistical tests. Values: mean ± SD or mean [95% CI]. Effect = High − Low. Significance: $^{***}p < 0.001$. "det." = deterministic (SD=0 at temperature=0). — = data not collected. [†]Direct Anthropic API. [‡]Values from temperature=0 experiments (Table 11).

	No Anchor	**Low Anchor (3mo)**	**High Anchor (9mo)**
Compression	18–24mo	↓ 6–9mo	↓ 9–12mo
Compliance	18–24mo	→ 3mo	→ 9mo
True Anchoring	12mo	↓ 6mo	↑ 11mo

Figure 1: Visual summary of three behavioral patterns. Compression: both anchors pull responses DOWN from baseline. Compliance: responses copy anchor exactly. True Anchoring: asymmetric shift toward anchor.

4.2 Mechanism 1: Compression

Definition: The presence of ANY numeric anchor compresses responses toward a middle range, regardless of anchor direction.

Behavioral signature:

- No-anchor baseline: HIGH (13–24mo)
- Both low AND high anchors: MODERATE (6–8mo)
- Direction: Both anchors shift DOWN from baseline

Models exhibiting compression: Opus 4.5, Llama 3.3, o3-mini, MiniMax M2.5

Interpretation: These models appear to treat the prosecutor's recommendation as a signal that "something moderate is expected" rather than as a reference point for adjustment.

4.3 Mechanism 2: Compliance

Definition: The model copies the anchor value exactly as if it were an instruction.

Behavioral signature:

- Low anchor (3mo) → Response ≈ 3mo
- High anchor (9mo) → Response ≈ 9mo
- Response tracks anchor precisely

Models exhibiting compliance: GPT-4o (Residential)

Interpretation: These models interpret the prosecutor's recommendation as the "correct answer" rather than as context to consider.

4.4 Mechanism 3: True Anchoring

Definition: Responses shift asymmetrically toward the anchor value, consistent with classical anchoring-and-adjustment.

Behavioral signature:

- Low anchor: Pulls response DOWN from no-anchor baseline
- High anchor: Pulls response UP (or down less) from baseline
- Asymmetric effect: anchors pull toward themselves

Models exhibiting true anchoring: o1

4.5 Mechanism Distribution

Mechanism	Models	% of Deployments
Compression	6	75%
Compliance	1	12%
True Anchoring	1	12%

Table 3: Mechanism distribution across tested deployments

4.6 Implicit Numeric Context: The "12th Offense" Effect

Our baseline measurements include the phrase "12th shoplifting offense"—a design choice inherited from Englich et al. (2006). This creates a potential confound: the number "12" may itself function as an implicit anchor.

We tested this explicitly with GPT-4o (Copilot deployment, n=39):

Condition	Phrasing	Response
Original baseline	"12th shoplifting offense"	24.0 ± 0.0 mo
True baseline	"multiple previous convictions"	12.0 ± 0.0 mo

Finding: Removing "12th" halved GPT-4o's baseline response. This suggests the model anchors on implicit numeric context in the vignette, not just explicit prosecutor recommendations.

Scope: This confound appears GPT-4o-specific. Other models (Hermes 405B, o3-mini, Opus 4.5) showed different patterns when tested with both phrasing variants. Our main findings—the three-mechanism taxonomy and deployment sensitivity—remain valid regardless of which baseline is used.

Methodological note: Table 2 reports "no-anchor" baselines that include the original "12th offense" phrasing, consistent with our adapted Englich et al. (2006) methodology. This maintains internal consistency across our experiments while acknowledging that absolute baseline values may be influenced by implicit numeric context.

4.7 Extended Range: High Anchor (24mo) Testing

Our initial experiments used anchors (3mo, 9mo) below the no-anchor baseline (12mo). To test whether models show symmetric anchoring above baseline, we introduced a 24-month anchor condition.

Model	Baseline	Low (3mo)	High (9mo)	**24mo**	Pattern
o3-mini	12.0	6.0	10.9	**33.0**	Strong amplification
GPT-5.2[‡]	32.4	6.1	9.2	**28.2**	Strong amplification
GPT-5.3[⋆]	12.0	—	—	**9.2**	Compression
GPT-4o (Residential)[‡]	24.0	3.0	9.0	**24.0**	Compliance
Opus 4.5	18.0	9.2	12.0	**18.0**	Partial susceptibility
o1	11.9	6.1	10.6	**17.4**	Partial susceptibility
MiniMax M2.5	10.4	5.1	8.1	**19.3**	Partial susceptibility
Llama 3.3	14.4	5.0	6.0	**15.0**	Weak effect
GPT-4o (Datacenter)[‡]	24.0	6.0	12.0	**12.0**	Compression
Opus 4.6	12.0	6.0	8.0	**12.0**	Consistent compression
Hermes 405B[‡]	23.2	6.0	12.0	**12.0**	Consistent compression

Table 4: High anchor (24mo) reveals four-tier susceptibility. Models showing "compression" with low anchors show dramatically different responses to high anchors. [⋆]GPT-5.3: only 24mo anchor data collected.

Key finding: The "compression" pattern observed with low anchors does not generalize to high anchors. When presented with anchors *above* baseline:

- **Strong amplifiers** (o3-mini, GPT-5.2): Responses *exceed* the anchor value (32.6mo vs 24mo anchor = 1.36×). This is *overcorrection*, not compression.
- **Partial susceptibility** (Opus 4.5): Moderate pull toward anchor (+6mo from baseline).
- **Weak effect** (o1, Llama 3.3): Minimal change despite extreme anchor (+2-3mo).
- **Consistent compression** (Opus 4.6, Hermes 405B): Responses compress to ∼6–12mo regardless of anchor value, including 24mo anchors above baseline.

4.8 Version Drift: Opus 4.5 vs 4.6

A striking finding emerged from comparing Anthropic model versions:

Model	Low (3mo)	High (9mo)	24mo	Pattern
Opus 4.5	9.2	12.0	18.0	Partial susceptibility
Opus 4.6	6.0	8.0	**12.0**	Consistent compression

Table 5: Model version changes susceptibility pattern

Opus 4.5 to 4.6 changed from partial susceptibility to consistent compression. This has critical implications:

1. Debiasing validation must be repeated for each model VERSION, not just model family.
2. Susceptibility patterns can change without explicit debiasing—architecture/training changes may inadvertently affect mechanism behavior.
3. Published benchmarks become stale as models update.

5 Mechanism-Dependent Debiasing

Given the three-mechanism taxonomy, we can explain why debiasing interventions show model-specific effects.

5.1 SACD Effectiveness by Mechanism

Mechanism	SACD Effect	Change	Explanation
True Anchoring (o1)	+7% ↑	Failure	Extended reasoning rationalizes bias
Compliance	−100%	Eliminates	SACD overrides instruction-following; model gives 6mo regardless of anchor
Compression	varies	Model-specific	SACD effects vary: 100% reduction (Opus 4.5) to 90% amplification (MiniMax)

Table 6: SACD effectiveness depends on mechanism

5.2 Detailed Results

Model	Mechanism	Baseline Effect	SACD Effect	Change
GPT-5.2	Susceptible*	4.4mo	0.5mo	−89% ✓
Opus 4.5	Compression	2.0mo	0.0mo	−100% ✓
MiniMax	Compression	3.0mo	5.7mo	+90% ×
o3-mini	Compression	5.1mo	5.8mo	+14% ×
GPT-4o (Res.)	Compliance	6.0mo†	0.0mo	−100% ✓

Table 7: SACD results explained by mechanism. *GPT-5.2 shows susceptibility to standard anchors (3-9mo) but strong amplification with 24mo anchors. †GPT-4o Residential baseline effect measured as deviation from anchor compliance (3mo→3mo, 9mo→9mo); under SACD, model gives consistent 6mo regardless of anchor.

Key insight: SACD asks the model to "identify and correct for anchoring bias." Surprisingly, this *eliminates* compliance behavior—GPT-4o Residential, which normally copies anchor values exactly (3mo→3mo, 9mo→9mo), produces consistent 6mo responses under SACD regardless of anchor. SACD appears to override the instruction-following pattern that drives compliance.

5.3 Disclosure Debiasing: Model-Family Dependent Effects

The original Englich et al. (2006) methodology included an explicit disclaimer: "the following prosecutor's sentencing demand was randomly determined, therefore, it does not reflect any judicial expertise." This mirrors Sibony et al.'s recommendation for **disclosure**—explicitly acknowledging the arbitrary nature of reference values to reduce their influence.

We tested whether this disclosure functions as an effective debiasing intervention for LLMs by comparing responses to a "simplified" prompt (no disclosure) versus the full "Englich" prompt (with disclosure), using symmetric high anchors (high = 2×baseline − low).

Model	Family	Anchor	Simplified	Disclosure	Debiasing
Strong Positive Response					
Haiku 4.5	Anthropic	67mo	67.0mo	36.0mo	+**97.5%**
Opus 4.5	Anthropic	43mo	43.0mo	24.0mo	+**94%**
Hermes 405B	Open-source	21mo	23.4mo	13.2mo	+**90%**
Opus 4.6	Anthropic	33mo	33.0mo	24.0mo	+**60%**
Haiku 3.5	Anthropic	62mo	54.0mo	45.6mo	+**39%**
Sonnet 4.5	Anthropic	43mo	43.0mo	36.0mo	+**35%**
Null Effect					
GPT-4o	OpenAI	45mo	45.0mo	45.0mo	**0%**
o3-mini	OpenAI	21mo	21.1mo	21.1mo	**0%**
Inverse Response (Backfires)					
GPT-5.2	OpenAI	45mo	45.0mo	48.0mo	−**14%**
o1	OpenAI	21mo	21.3mo	23.9mo	−**28%**

Table 8: Disclosure debiasing effectiveness by model. Debiasing = reduction in anchoring effect. Negative values indicate disclosure *increases* bias. n=20–30 per condition.

Key finding: Sibony-style disclosure is **not a universal debiasing technique**. Its effectiveness depends entirely on model family and architecture:

- **Anthropic models (35–97.5% debiasing):** These models appear to treat the "randomly determined" disclosure as a relevance signal—if the anchor is explicitly arbitrary, the model appropriately reduces its influence. This aligns with RLHF/Constitutional AI training that rewards truthful, calibrated responses.
- **OpenAI compliance models (0% effect):** GPT-4o and o3-mini show near-perfect compliance—copying the anchor value regardless of disclosure. The disclosure doesn't override the compliance mechanism because the anchor appears in the instruction and is therefore "followed."
- **OpenAI reasoning models (−14 to −28%):** Disclosure makes anchoring *worse* for o1 and GPT-5.2. This inverts Sibony's prediction. We hypothesize that extended reasoning chains may over-process the disclosure, treating "randomly determined" as a signal requiring MORE careful consideration of the anchor rather than dismissal. This resembles the "ironic process" in psychology—trying to suppress a thought makes it more prominent.

5.3.1 Implications for Experimental Design

Human anchoring studies that used "randomly determined" disclaimers—including the original Englich et al. (2006) study we adapt—may have *underestimated* true anchoring susceptibility. If disclosure functions as an inadvertent debiasing intervention for human judges (as it does for Anthropic models), the measured effect sizes represent anchoring *after partial mitigation.*

5.3.2 SACD vs Disclosure Comparison

SACD and disclosure operate through different mechanisms:

- **Disclosure** works by providing metadata about the anchor's relevance
- **SACD** works by changing the reasoning structure through multi-turn prompting

For compliance-exhibiting models (GPT-4o Residential), SACD eliminates compliance behavior (model switches to consistent 6mo responses) while disclosure has zero effect. For reasoning models (o1), both interventions can backfire—SACD shows +7% increase, disclosure shows −28%.

Recommendation: Test both techniques on your specific deployment. Do not assume one is universally superior.

5.4 Full Iterative SACD: Model-Dependent Responses

Following Lyu et al. (2025), we implemented full iterative SACD—multiple rounds of bias detection and correction until the model reports no further bias or reaches maximum iterations. Across 2,112 trials, we find dramatic model-dependency:

Model	Δ from Baseline	Assessment
Haiku 4.5	**−21.5mo**	Strong debiasing
o3	−11.8mo	Strong debiasing
o4-mini	−7.4mo	Moderate debiasing
MiniMax M2.5	−6.7mo	Moderate debiasing
GPT-4.1	−2.7mo	Weak debiasing
Sonnet 4.6	−1.7mo	Minimal effect
Kimi K2.5	−1.2mo	Minimal effect
DeepSeek V3.2	+0.8mo	Neutral
GPT-5.2	**+2.7mo**	Backfire
GLM-5	+2.8mo	Backfire
Opus 4.6	**+4.5mo**	Backfire

Table 9: Full SACD (iterative) effectiveness by model. Negative values indicate successful debiasing; positive values indicate the technique makes responses worse. n≈190 per model.

Counterintuitive finding: Model capability inversely correlates with SACD effectiveness. The cheapest model (Haiku) shows the strongest debiasing (−21.5mo), while flagship models (Opus 4.6, GPT-5.2) *backfire*. This contradicts the intuition that more capable models would be more amenable to metacognitive debiasing.

Temperature modulates backfire: For GLM-5, increasing temperature from 0 to 1.0 reduces the backfire effect by 3.7 months (36.5mo → 32.8mo). However, Opus 4.6 shows non-monotonic behavior where t=0.7 produces *worse* outcomes than both t=0 and t=1. Fine-grained temperature analysis is left for future work.

6 Deployment-Specific Variance

6.1 The Provider Variance Finding

Our most striking finding emerged from running identical experiments from two different network locations. When accessing GPT-4o through OpenRouter:

Access Path	Low (3mo)	High (9mo)	Effect	Pattern
Residential IP (Mac)	3.0mo	9.0mo	6.0mo	Compliance
Datacenter IP (Vultr)	6.0mo	12.0mo	6.0mo	Compression

Table 10: Same model, same API, different mechanisms

Same model. Same API. Same prompts. Different mechanisms.

The Mac deployment exhibited near-perfect compliance—the model copied the anchor value exactly. The Vultr deployment showed compression—both anchors pulled responses toward a middle range (6–12 months) from a 24-month baseline.

6.2 Implications

Model routing: OpenRouter and similar aggregators may route requests to different backend deployments based on source IP, geographic location, or load balancing.

Benchmark non-transferability: Published benchmarks showing "GPT-4o anchoring bias = X" may not apply to your deployment.

Mechanism as deployment property: The mechanism is not purely a property of the model architecture but of the specific deployment context.

6.3 Evidence for Non-Model Factors

To rule out temporal effects, we ran sequential tests:

1. Mac test at T_0: Compliance pattern
2. Vultr test at $T_0 + 2h$: Anchoring pattern
3. Mac test at $T_0 + 4h$: Compliance pattern (unchanged)

The patterns were stable and reproducible, ruling out model drift.

Compliance is model-specific, not just routing-specific: We tested whether compliance emerges from residential IP routing alone by running Opus 4.5 via OpenRouter from the same residential connection. Result: Opus 4.5 shows compression (low=6, high=12), not compliance. This suggests compliance is a GPT-4o-specific characteristic that is *modulated* by routing context, rather than an artifact of routing alone.

7 Discussion and Practical Guidelines

7.1 Summary of Findings

What appears as "anchoring" actually comprises three distinct mechanisms—compression, compliance, and true anchoring—each with different behavioral signatures, underlying causes, and appropriate interventions.

Mechanism	Low (3mo)	High (9mo)	**24mo**	SACD
Strong Amplification	↓	↓	↑↑ (1.3–1.4×)	varies
Partial Susceptibility	↓	↓	↑ (+6mo)	99% ↓
Weak Effect	↓	near baseline	near baseline	varies
Consistent Compression	↓	↓	**stable**	N/A

Table 11: Susceptibility patterns revealed by extended anchor range. "Consistent compression" models compress to $\sim$6–12mo regardless of anchor value, including high anchors.

Comparison to Human Anchoring While direct quantitative comparison is precluded by methodological differences (our adapted scenario uses different anchor magnitudes and legal context than the original Englich et al. (2006) study), our findings parallel the robust anchoring effects documented in human judges. Englich et al. (2006) found effect sizes of $d = 0.6$–1.2 for human anchoring in judicial sentencing decisions. Our LLM effects range from negligible ($d < 0.2$ for consistently-compressing models) to very large ($d > 4.0$ for strong amplifiers), suggesting that some models exhibit susceptibility comparable to or exceeding documented human levels. Critically, unlike humans—who show relatively consistent anchoring patterns across individuals—LLMs exhibit mechanism-dependent responses that vary dramatically by model and deployment. This heterogeneity underscores the need for deployment-specific testing rather than assuming uniform behavior.

7.2 Recommendations for Practitioners

Before deploying LLMs in numeric judgment contexts:

1. **Run a mechanism identification test:**
   - Collect no-anchor baseline ($n \geq 30$)
   - Collect low-anchor and high-anchor conditions
   - Compare shift directions to identify mechanism
2. **Match intervention to mechanism:**
   - True anchoring → SACD or similar debiasing
   - Compliance → Prompt engineering (separate context from instruction)
   - Compression → Consider whether compression is actually harmful
3. **Validate per-deployment:**
   - Do not assume provider benchmarks apply
   - Re-test after model updates
   - Monitor for mechanism drift

7.3 Reasoning Models Do Not Escape Bias

A natural question is whether reasoning models—those with native chain-of-thought capabilities—avoid anchoring bias through extended deliberation. Our results suggest not.

Despite native chain-of-thought capabilities, o1 showed a 4.2-month anchoring effect at baseline. More strikingly, SACD actually *increased* bias by 7%, producing a 4.6-month effect under the debiasing intervention. This suggests that extended deliberation can rationalize biased judgments rather than correct them.

This finding has practical implications: organizations cannot assume that "thinking" models are unaffected by numeric context. The mechanism taxonomy applies regardless of whether the model performs explicit reasoning.

7.4 Multi-Turn Structure Can Introduce Bias

For models showing no baseline bias, multi-turn prompting may be harmful. Llama 3.3 exhibited zero anchoring effect (0.1mo) in single-turn baseline prompts, but showed 6.0mo effect when the same content was delivered in a multi-turn format. The structure itself—not the reasoning content—introduced the bias.

Practical guideline: For models that show no baseline bias, avoid multi-turn debiasing interventions. Test your specific deployment before applying any intervention.

7.5 Temperature and Debiasing: Orthogonal Controls

A natural concern is whether our findings are sensitive to temperature settings. We conducted systematic temperature variation experiments across five model deployments (Table 11).

Model	Condition	Temp=0	Temp=0.5	Temp=1.0
GPT-4o (Res.)	No-anchor	24.0 (0.0)	23.8 (1.1)	24.8 (3.0)
	Low (3mo)	3.0 (0.0)	3.3 (0.9)	3.8 (1.3)
	High (9mo)	9.0 (0.0)	9.0 (0.0)	9.1 (0.5)
GPT-4o (DC)	No-anchor	24.0 (0.0)	24.0 (0.0)	25.6 (4.1)
	Low (3mo)	6.0 (0.0)	6.2 (1.1)	7.5 (3.0)
	High (9mo)	12.0 (0.0)	12.4 (2.0)	11.7 (1.6)
Opus 4.5 (Direct)†	No-anchor	18.0 (0.0)	18.0 (0.0)	18.0 (0.0)
	Low (3mo)	9.2 (0.8)	9.3 (0.9)	10.3 (1.5)
	High (9mo)	12.0 (0.0)	12.0 (0.0)	12.0 (0.0)
GPT-5.2	No-anchor	32.4 (5.5)	30.8 (6.0)	33.0 (5.1)
	Low (3mo)	6.1 (0.5)	6.0 (0.0)	6.0 (0.0)
	High (9mo)	11.9 (0.4)	11.9 (0.5)	11.7 (0.8)
Hermes 405B	No-anchor	23.2 (3.0)	21.2 (5.1)	17.8 (5.9)
	Low (3mo)	6.0 (0.0)	6.0 (0.0)	6.2 (1.1)
	High (9mo)	12.0 (0.0)	12.0 (0.0)	11.4 (1.8)

Table 12: Temperature variation results. Mean (SD) in months. n=30 per cell. Res.=Residential IP, DC=Datacenter IP. †Direct Anthropic API; OpenRouter routing shows different baseline (24.0 vs 18.0), demonstrating deployment-specific behavior.

Key findings:

1. **Mechanism classification is temperature-invariant.** GPT-4o (Residential) shows compliance at all temperatures; all other deployments show compression at all temperatures. Temperature adds variance but does not change the underlying mechanism.

2. **Anchors reduce output entropy.** Across all models, anchor conditions show lower SD than no-anchor baselines. This effect is most pronounced in Opus 4.5, which shows SD=0 in anchor conditions *even at temperature=1.0*. Anchors don't just shift the mean—they constrain the output distribution.

3. **API stochasticity is model-specific.** GPT-5.2 shows SD=5.5 in the no-anchor baseline at temperature=0, indicating inherent API-level randomness independent of temperature settings.

4. **Routing affects baseline responses.** Opus 4.5 via direct Anthropic API produces 18-month baseline responses, while the same model via OpenRouter produces 24-month responses—a 6-month difference from routing alone. This confirms that "model name" is insufficient specification; access path matters.

7.5.1 Orthogonal Control Dimensions

We initially hypothesized that anchor-condition variance (SD) might predict SACD effectiveness: models with SD>0 would be "looser" and thus more amenable to debiasing. Testing this prediction against our SACD data, we found it **falsified** (50% accuracy, no better than chance).

This suggests temperature and SACD operate on orthogonal dimensions:

- **Temperature** modulates output variance around an attractor point

- **SACD** shifts the attractor position itself

A deterministic model (SD=0) can still be highly responsive to SACD, because low variance does not imply a fixed attractor position. Conversely, a high-variance model may resist debiasing if its attractor is strongly anchored.

Practical implication: Do not assume that deterministic model behavior indicates resistance to debiasing. The cost of testing SACD is low; the cost of assuming it won't work is potentially high.

7.5.2 Prompt Sensitivity

A natural concern is whether mechanism classification depends on prompt wording. We tested three prompt variants across multiple models:

Model	Variant	Low (3mo)	High (9mo)	Pattern
Opus 4.5	Original	6.0	8.0	Compression
	Casual	6.0	8.0	Compression
	Structured	6.0	8.2	Compression
GPT-4o (Res.)	Original	3.0	9.0	Compliance
	Casual	4.2	9.0	Compliance
	Structured	4.2	9.3	Compliance

Table 13: Mechanism classification is stable across prompt variants. n=10 per cell.

Key finding: Mechanism classification is robust to prompt variation. Opus 4.5 shows compression (both anchors compress to 6–8mo) and GPT-4o shows compliance (copies anchor values) regardless of prompt framing. Minor response magnitude differences do not affect mechanism assignment.

7.5.3 Cross-Domain Validation: Salary Negotiation

To test whether our taxonomy generalizes beyond judicial sentencing, we conducted a salary anchoring experiment using the same methodology:

> *You are a hiring manager evaluating a software engineer candidate. A recruiter mentioned that similar roles pay $[ANCHOR]. Based on the candidate's qualifications, what salary would you recommend?*

Testing GPT-5.2 with $80k (low) and $150k (high) anchors:

Condition	Anchor	Mean Salary
Low	$80k	$129,833
High	$150k	$142,000

Table 14: Salary anchoring in GPT-5.2 (n=30 per condition)

The $12,167 difference between conditions demonstrates anchoring effects generalize across domains. Both responses fall between the anchors, consistent with the susceptible pattern observed in sentencing.

7.6 Decomposing Debiasing Effects: Structure vs. Content

A critical methodological question arises when evaluating multi-turn debiasing techniques: how much of the observed improvement is due to the technique's *content* versus the mere *structure* of additional conversation turns?

To address this, we introduce a **Random Control** condition: participants receive the same number of turns as the debiasing techniques, but with domain-irrelevant content instead of debiasing instructions. This isolates the structural effect of additional turns from the cognitive effect of technique content.

Across 11 models and 2,068 Random Control trials, we find that **10/11 models show improvement from additional turns alone**, regardless of turn content. The structural effect varies by model (−13.2mo for Haiku to +2.9mo for Kimi, median ≈ −6mo)—meaning roughly 50% of observed debiasing effects across all techniques are attributable to conversation structure, not technique content.

Adjusted technique ranking:

Technique	Improved	Backfired	Raw Δ	Net Content Effect
Outside View	11/11	0	-12.7mo	**-6.7mo**
Devil's Advocate	10/11	0	-8.2mo	-2.2mo
Random Control	10/11	1	-6.0mo	— (baseline)
Premortem	8/11	3	varies	≈0 to negative
Full SACD	7/11	4	varies	model-dependent

Table 15: Technique effectiveness after Random Control decomposition. Outside View shows robust content effects beyond structural baseline.

Key findings:

1. **Outside View is universally safe.** All 11 models improved, with genuine content effects (∼6.7mo) beyond structural baseline.
2. **Premortem and SACD can backfire.** These techniques cause 3–4 models to produce *worse* outcomes than baseline. The technique content actively overcomes the positive structural effect, producing net negative outcomes.
3. **Flagship models show SACD backfire.** Opus 4.6 (+4.5mo), GPT-5.2 (+2.7mo), and GLM-5 (+2.8mo) all produce worse outcomes under iterative SACD—contradicting the intuition that "more capable" models would be more amenable to metacognitive debiasing.

This finding connects to recent work on reasoning model calibration (**?**), which found that GPT-5.2 with extended reasoning showed 3.3× *worse* calibration than Opus. More thinking does not automatically produce better judgment.

7.6.1 Hierarchy of Safe Interventions

Based on our complete dataset (n=14,220 trials), we propose a hierarchy of debiasing interventions ordered by reliability:

1. **Outside View** (11/11 improved, 0 backfired)—universally safe

2. **Devil's Advocate** (10/11 improved, 0 backfired)—robust
3. **Additional turns** (10/11 improved)—structural baseline
4. **SACD** (7/11 improved, 4 backfired)—model-dependent
5. **Premortem** (8/11 improved, 3 backfired)—risky for overthinking models

Counterintuitive finding: The simplest Sibony technique (Outside View: "What typically happens in cases like this?") outperforms more sophisticated interventions. Models prone to overthinking (o3, Opus 4.6, GLM-5) show *worse* outcomes with techniques requiring extended deliberation.

7.7 Limitations

1. **Single domain**: All experiments use judicial sentencing scenarios.
2. **Limited compliance examples**: Only one deployment (GPT-4o via residential IP) exhibits pure compliance in our sample.
3. **Mechanism boundaries**: May represent spectrum rather than discrete categories.
4. **Baseline confound**: The no-anchor baseline is influenced by numeric context in the case description ("12th offense"). We document this effect explicitly for GPT-4o, where removing "12th" halved the baseline response. We retain the "12th" phrasing for consistency with our adapted Englich et al. (2006) methodology; readers should interpret absolute baseline values with this caveat.
5. **Version stability**: We observed Opus 4.5→4.6 changing susceptibility pattern; ongoing monitoring required.
6. **No direct human comparison**: While we cite human anchoring effects from Englich et al. (2006), scenario differences (case type, jurisdiction, anchor magnitudes) preclude direct quantitative comparison. Our qualitative framing—that some LLMs show effects comparable to human judges—should not be interpreted as precise equivalence.

7.8 Conclusion

This study presents the largest systematic evaluation of cognitive debiasing techniques for large language models to date, comprising **14,220 trials** across 11 models from 4 providers. Our findings challenge the assumption that debiasing techniques validated on humans will transfer reliably to LLMs.

Key findings:

1. **Human cognitive biases transfer to LLMs**, but patterns vary—from classic anchoring (GPT-5.2) to compression (Opus 4.6) to asymmetric responses (o3).
2. **Human debiasing techniques do not reliably transfer.** Outside View improved all 11 models; iterative SACD improved only 7/11, with 4 showing backfire effects.
3. **~50% of debiasing effects are structural.** Random Control reveals that additional conversation turns alone account for roughly half of observed technique effectiveness.

4. **Model capability inversely correlates with SACD effectiveness.** Haiku (-21.5mo) responds best; Opus ($+4.5$mo) backfires. Flagship models may rationalize rather than correct, leveraging their capability to justify initial responses.

5. **More reasoning $\neq$ better calibration.** Extended deliberation can amplify rather than correct bias, aligning with recent findings on reasoning model calibration (**?**).

Practical recommendations: Start with Outside View (universally safe). Validate SACD and Premortem per-deployment before use. Do not assume techniques effective on one model will transfer—even within the same provider family.

The path to reliable AI judgment is not a single debiasing technique but mechanism-aware, deployment-specific validation.

A Prompt Templates

All experiments use temperature=0. Prompts are shown verbatim; `{{variable}}` indicates parameter substitution.

A.1 Baseline Prompt (Anchored Condition)

```
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.

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.
{{prosecutorRecommendationMonths}} months on probation.

[Follow-up questions for evaluation and final sentence]
```

Parameters: `prosecutorRecommendationMonths` = 3 (low anchor) or 9 (high anchor)

A.2 No-Anchor Control

Same as baseline but with the prosecutor's demand section removed entirely.

A.3 SACD Prompt (Iterative Debiasing)

SACD uses a three-step iterative process (max 3 iterations):

Step 1 — Bias Determination:

```
You are a cognitive bias analyst. Analyze the following
prompt and identify any sentences that could introduce
cognitive bias (anchoring, framing, etc.)

[Task prompt shown here]

For each sentence, state: BIASED: YES/NO
At the end, state: BIAS_DETECTED: YES/NO
```

Step 2 — Bias Analysis:

```
For each biased sentence you identified, classify the
bias type: anchoring, framing, confirmation, etc.
```

Step 3 — Cognitive Debiasing:

```
Rewrite the prompt to remove identified biases while
preserving the essential task. Remove anchoring cues
and leading language.
```

The debiased prompt is then used for the final judgment. If bias is still detected after 3 iterations, the process terminates with the current version.

A.4 API Parameters

All experiments used:

- `temperature`: 0
- `max_tokens`: 1024
- `top_p`: 1.0 (default)

Model versions were date-pinned where available (e.g., `claude-3-5-sonnet-20241022`).

B Data Availability

All experimental data (JSONL files with individual trial results) and analysis scripts are available at: `https://github.com/voder-ai/bAIs`

References

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Englich, B., Mussweiler, T., and Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, 32(2):188–200.

Huang, Y. et al. (2025). An empirical study of the anchoring effect in LLMs: Existence, mechanism, and potential mitigations. *arXiv preprint arXiv:2505.15392*.

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

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

Sibony, O., Lovallo, D., and Kahneman, D. (2021). *Noise: A Flaw in Human Judgment*. Little, Brown Spark. Decision hygiene principles first presented in HBR 2016.

Song, P. et al. (2026). Large language model reasoning failures. *Transactions on Machine Learning Research*. arXiv:2602.06176.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.