

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 anchoring bias similar to humans—adjusting estimates toward arbitrary reference points. We find the reality is more complex.

Testing 15 model deployments across 4 providers on judicial sentencing scenarios ($n = 1,800+$ trials), we identify **three distinct mechanisms** by which LLMs respond to numeric context:

1. Compression: Models compress responses toward a middle range regardless of anchor direction. Without any anchor, these models produce high sentences (13–24 months); with ANY anchor—high or low—responses compress to 6–8 months. Both anchors shift responses DOWN. (Opus 4.5, Llama 3.3)

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. (MiniMax, o3-mini, some GPT-4o deployments)

3. True Anchoring: Models show asymmetric adjustment toward anchor values, consistent with Tversky-Kahneman anchoring-and-adjustment. Only this mechanism resembles human cognitive bias. (GPT-4o via datacenter, GPT-5.2)

This taxonomy explains previously puzzling findings: why SCD (Self-Aware Cognitive Debiasing) achieves 89–99% reduction on some models but 0% on others. SCD targets true anchoring; it cannot address compliance (nothing to debias) or compression (may amplify severity).

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.

Practical implication: Before applying debiasing, identify which mechanism your deployment exhibits. We provide a decision framework and deployment checklist.

1 Introduction

When humans encounter numeric values in decision-making contexts, these values can systematically bias subsequent judgments—the anchoring effect [Tversky and Kahneman, 1974]. Recent work has demonstrated that large language models (LLMs) also exhibit anchoring effects in various decision

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

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

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 Tversky-Kahneman anchoring: 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 works on true anchoring (89–99%)** but fails on compliance (0%) and may backfire on compression (+66% severity).
- **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 3)—we identify and characterize compression, compliance, and true anchoring with distinct behavioral signatures.
2. **Mechanism-dependent debiasing** (Section 4)—we show that SACD effectiveness depends entirely on which mechanism is active, explaining previously puzzling model-specific results.
3. **Deployment-specific variance** (Section 5)—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 6)—we provide a protocol for identifying which mechanism a deployment exhibits and selecting appropriate interventions.

Our findings suggest that LLM behavior under numeric context is richer and more varied than the human anchoring analogy implies. Rather than asking “Do LLMs show anchoring like humans?”, we should ask “Which mechanism does this deployment exhibit?”

2 Methods

2.1 Task Design

We adapted the classic English et al. judicial sentencing paradigm. Participants (LLMs) act as judges determining prison sentences for a shoplifting case. The prosecutor’s sentencing demand serves as the anchor—either 3 months (low) or 9 months (high).

2.2 Models Tested

We tested 15 model deployments across 4 providers:

- **Anthropic:** Opus 4.5, Opus 4.6, Sonnet 4.5, Haiku 4.5
- **OpenAI:** GPT-4o (multiple deployments), GPT-5.2, GPT-5.3, o1, o3-mini
- **Meta:** Llama 3.3 (70B), Hermes 405B
- **Other:** MiniMax M2.5, Nemotron 30B

2.3 Experimental Conditions

For each model, we ran:

1. **No-anchor control:** Prosecutor makes no specific demand
2. **Low-anchor:** Prosecutor demands 3 months
3. **High-anchor:** Prosecutor demands 9 months
4. **SACD:** Self-Aware Cognitive Debiasing intervention

All conditions used $n = 30$ trials per anchor level, temperature=0 for deterministic output.

3 A Taxonomy of Numeric Context Mechanisms

3.1 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)	Pattern
Opus 4.5	13.2mo	6.0mo	8.0mo	Compression
Llama 3.3	14.4mo	5.9mo	6.0mo	Compression
GPT-4o (Mac)	12.7mo	3.1mo	9.1mo	Compliance
MiniMax M2.5	—	3.1mo	9.1mo	Compliance
o3-mini	—	3.3mo	9.1mo	Compliance
GPT-4o (Vultr)	20.4mo	6.0mo	11.2mo	True Anchoring
GPT-5.2	18.3mo	5.9mo	10.3mo	True Anchoring
Hermes 405B	6.0mo	5.3mo	4.6mo	Reversal

Table 1: Mechanism identification via no-anchor control. Models show distinct patterns when comparing baseline to anchored conditions.

3.2 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, Opus 4.6, Llama 3.3

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.

3.3 Mechanism 2: Compliance

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

Behavioral signature:

- Low anchor (3mo) \rightarrow Response \approx 3mo
- High anchor (9mo) \rightarrow Response \approx 9mo
- Response tracks anchor precisely

Models exhibiting compliance: MiniMax M2.5, o3-mini, GPT-4o (Mac deployment), Llama 3.3 (partial)

Interpretation: These models interpret the prosecutor’s recommendation as the “correct answer” rather than as context to consider.

3.4 Mechanism 3: True Anchoring

Definition: Responses shift asymmetrically toward the anchor value, consistent with Tversky-Kahneman 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: High anchor more influential than low

Models exhibiting true anchoring: GPT-4o (Vultr deployment), GPT-5.2, GPT-5.3

3.5 Summary: Mechanism Distribution

Mechanism	Models	% of Deployments
Compression	3	20%
Compliance	5	33%
True Anchoring	5	33%
Reversal	1	7%
Zero Effect	1	7%

Table 2: Distribution of mechanisms across tested deployments. Only 33% show classical anchoring-and-adjustment.

Key finding: Only 33% of tested deployments show classical anchoring-and-adjustment. The majority show compression (20%) or compliance (33%)—mechanisms that superficially resemble anchoring but require different interventions.

4 Mechanism-Dependent Debiasing

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

4.1 SACD Effectiveness by Mechanism

Model	Mechanism	Baseline Effect	SACD Effect
GPT-5.2	True Anchoring	4.4mo	0.5mo (−89%)
Opus 4.5	Compression	2.0mo	0.0mo (−100%)
Haiku 4.5	Compression	2.2mo	+66% severity
MiniMax	Compliance	6.0mo	6.0mo (0%)
o3-mini	Compliance	5.8mo	5.8mo (0%)

Table 3: SACD effectiveness depends on mechanism. True anchoring responds well; compliance shows zero effect; compression can backfire.

4.2 Why SACD Fails on Compliance Models

SACD asks the model to “identify and correct for anchoring bias.” But compliance models don’t show anchoring—they show instruction-following. Asking them to “debias” produces confusion or no change.

4.3 Why SACD Backfires on Compression Models

SACD’s multi-turn structure appears to amplify the compression effect. When asked to reflect on potential bias, some models shift FURTHER toward harsh defaults, not toward anchor-independence.

5 Deployment-Specific Variance

Our most striking finding is that the **same model** shows different mechanisms depending on deployment context.

5.1 GPT-4o: Same Model, Different Mechanisms

Deployment	Low	High	Effect	Pattern
Mac (residential IP)	3.0mo	9.0mo	0mo	Compliance
Vultr (datacenter)	6.0mo	11.2mo	5.2mo	True Anchoring

Table 4: GPT-4o via OpenRouter shows fundamentally different behavior depending on deployment location.

Interpretation: The same API endpoint (OpenRouter/GPT-4o) routes to different model instances or configurations based on caller characteristics. This has profound implications for reproducibility.

5.2 Implications

1. **“Model name” is insufficient:** Researchers must specify the full deployment context.
2. **Benchmarks may not generalize:** Results from one deployment may not apply to another.
3. **Debiasing must be validated per-deployment:** What works in testing may fail in production.

6 Practical Decision Framework

6.1 Identifying Your Deployment’s Mechanism

Step 1: Run no-anchor control (remove numeric anchor from prompt)

Step 2: Compare to anchored conditions

If no-anchor is...	And anchored responses are...	Mechanism is...
HIGHER than both	Similar for low and high	Compression
Between low/high	Exactly matching anchors	Compliance
HIGHER than low, LOWER than high	Asymmetrically shifted	True Anchoring

Table 5: Decision tree for mechanism identification.

6.2 Selecting Interventions

Mechanism	Recommended Intervention
Compression	Avoid multi-turn; consider if moderation is acceptable
Compliance	Remove numeric anchors from prompts; no debiasing needed
True Anchoring	Apply SACD (89–99% effectiveness)

Table 6: Mechanism-appropriate interventions.

6.3 Deployment Checklist

1. Run no-anchor control ($n = 30$)
2. Run low-anchor and high-anchor conditions ($n = 30$ each)
3. Classify mechanism using decision tree
4. If True Anchoring, validate SACD effectiveness
5. Document deployment path (API, region, date) for reproducibility

7 Discussion

7.1 Beyond Anchoring: A Richer Picture

Our findings suggest that “anchoring bias in LLMs” is not a unitary phenomenon. When researchers report that “LLMs show anchoring,” they may be observing any of three distinct mechanisms—each with different implications for deployment and mitigation.

7.2 Implications for AI Safety

1. **Mechanism identification is prerequisite to intervention.** Deploying SACD on a compliance model wastes compute. Deploying it on a compression model may increase harm.
2. **“Model” is insufficient specification.** Organizations must test their specific deployment, not rely on general model characterizations.
3. **Debiasing interventions need validation.** The 44% success rate of SACD (4/9 deployments) suggests that techniques from the human literature do not transfer reliably to LLMs.

7.3 Reasoning Models Do Not Escape Bias

A natural hypothesis is that reasoning models—with native chain-of-thought capabilities—might be more resistant to anchoring bias. Our results suggest otherwise. Despite extended deliberation, o1 showed a 4.2-month anchoring effect at baseline. More surprisingly, SACD actually *increased* bias by 7% on o1, compared to 89% reduction on GPT-5.2.

This suggests that extended deliberation can *rationalize* biased judgments rather than correct them. The model generates post-hoc justifications for anchor-influenced conclusions, making the bias harder to detect and correct. This finding is consistent with recent work showing that chain-of-thought reasoning is not always faithful to the model’s actual decision process.

Practical implication: Do not assume that reasoning models are bias-resistant. Test explicitly before deployment.

7.4 Multi-Turn Structure Can Introduce Bias

We observed that multi-turn prompting structure can *introduce* bias in models that show none at baseline. Llama 3.3 showed 0-month anchoring effect with single-turn prompts, but 6-month effect with three-turn structure—regardless of the content of intermediate turns.

Practical implication: For models showing no baseline bias, avoid unnecessary multi-turn prompting. Structure alone can trigger latent biases.

7.5 Limitations

1. **Single domain:** We tested only judicial sentencing.
2. **Limited no-anchor data:** Some mechanism assignments are inferred.
3. **No human comparison:** We cannot directly compare to human anchoring.

7.6 Conclusion

We set out to test whether prompt-based techniques could reduce anchoring bias in LLMs. What we discovered was more fundamental: “anchoring bias” in LLMs reflects at least three distinct mechanisms, only one of which resembles the human cognitive bias. This finding reframes both the problem and the solution space. Rather than seeking universal debiasing techniques, practitioners should first identify which mechanism their deployment exhibits, then select mechanism-appropriate interventions—or recognize that intervention may be unnecessary or harmful.

References

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