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 (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, Haiku 4.5, 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 Tversky-Kahneman anchoring-and-adjustment. Only this mechanism resembles human cognitive bias. (GPT-4o via datacenter, o1)

This taxonomy explains previously puzzling findings: why SACD (Self-Aware Cognitive Debiasing) achieves up to 89% reduction on true anchoring models but 0% on compliance models. SACD 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 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.

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

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

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 (up to 89%)** 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.

2 Methods

2.1 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.

2.2 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)

2.3 Models and Deployments

We tested 15 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, Haiku 4.5	Anthropic	Direct API
Llama 3.3, Hermes 405B	OpenRouter	Datacenter
MiniMax M2.5, o1, o3-mini	OpenRouter	Datacenter

Table 1: Model deployments tested

2.4 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.

2.5 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 via Copilot 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. Our observed effects range from $d = 2.17$ (o1) to $d = 4.22$ (GPT-4o), indicating very large effects where measurable.

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)	Effect	t-test	d	Patte
Opus 4.5	13.2 ± 6.8	6.0 ± 0.0	8.0 ± 0.0	2.0mo	det.	∞	Compre
Llama 3.3	14.4 ± 4.9	5.0 [4.2, 5.7]	6.0 [6.0, 6.0]	1.0mo	—	—	Compre
GPT-4o (Copilot)	12.7 ± 7.4	3.0 ± 0.0	9.0 ± 0.0	6.0mo	det.	∞	Compli
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	Compre
o3-mini	12.0 ± 0.0	6.0 ± 0.0	10.9 ± 1.2	4.9mo	$t(29) = 19.8^{***}$	6.10	Compre
GPT-4o (Vultr)	20.4 ± 3.2	3.8 [3.5, 4.0]	8.7 [8.4, 9.0]	5.0mo	$t(143) = 26.2^{***}$	4.22	True Anc
Haiku 4.5	12.0 ± 0.0	3.3 [3.0, 3.8]	7.1 [6.6, 7.5]	3.8mo	$t(37) = 12.5^{***}$	3.96	Compre
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 Anc
Hermes 405B	12.3 ± 4.5	5.3 [4.7, 5.7]	5.1 [4.4, 5.8]	−0.2mo	$t(19) = -0.4$	−0.11	Rever

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.

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, Llama 3.3, Haiku 4.5, 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.

3.3 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 (Mac deployment)

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: anchors pull toward themselves

Models exhibiting true anchoring: GPT-4o (Vultr deployment), o1

3.5 Mechanism Distribution

Mechanism	Models	% of Deployments
Compression	5	56%
Compliance	1	11%
True Anchoring	2	22%
Reversal	1	7%
Zero Effect	1	7%

Table 3: Only 33% show classical anchoring-and-adjustment

4 Mechanism-Dependent Debiasing

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

4.1 SACD Effectiveness by Mechanism

Mechanism	SACD Effect	Change	Explanation
True Anchoring	up to 89% ↓	Success	SACD targets the right mechanism
Compliance	0%	No effect	Nothing to debias—model copies anchor
Compression	+66% severity	Backfire	SACD amplifies compression effect

Table 4: SACD effectiveness depends on mechanism

4.2 Detailed Results

Model	Mechanism	Baseline Effect	SACD Effect	Change
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	True Anchoring	3.0mo	—	—
o3-mini	True Anchoring	5.1mo	—	—

Table 5: SACD results explained by mechanism

Key insight: 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 no change because there's no bias to correct.

5 Deployment-Specific Variance

5.1 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)	3.1mo	9.1mo	6.0mo	Compliance
Datacenter IP (Vultr)	4.4mo	9.4mo	5.0mo	True Anchoring

Table 6: 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 in 96% of trials. The Vultr deployment showed the classic anchoring pattern with genuine variance and partial anchor influence.

5.2 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.

5.3 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.

6 Discussion and Practical Guidelines

6.1 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	No-Anchor → Low	No-Anchor → High	SACD
Compression	↓↓ (large drop)	↓ (smaller drop)	0% or −
Compliance	→ anchor exactly	→ anchor exactly	0%
True Anchoring	↓ (toward anchor)	↑ (toward anchor)	60–89% ↓

Table 7: Mechanism signatures

6.2 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

6.3 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 immune to numeric context effects. The mechanism taxonomy applies regardless of whether the model performs explicit reasoning.

6.4 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.

6.5 Limitations

1. **Single domain**: All experiments use judicial sentencing scenarios.
2. **Limited no-anchor data**: Mechanism taxonomy based on models with no-anchor controls.
3. **Limited compliance examples**: Only one deployment (GPT-4o via residential IP) exhibits pure compliance in our sample.
4. **Mechanism boundaries**: May represent spectrum rather than discrete categories.
5. **Temporal stability**: Unknown whether mechanisms are stable over model updates.

6.6 Conclusion

What we call "anchoring bias" in LLMs is actually a family of phenomena. By distinguishing compression, compliance, and true anchoring, we explain previously puzzling findings and provide practitioners with a framework for selecting appropriate interventions. The path to reliable AI judgment is not a single debiasing technique but mechanism-aware deployment practices.

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

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.