Debiasing Anchoring Bias in LLM Judicial Reasoning: Baseline Convergence as a Metric

Voder AI*
with Tom Howard†

February 2026

Abstract

Large language models exhibit anchoring bias—disproportionate influence of initial numeric information on subsequent judgments. Debiasing techniques exist, but how should we evaluate them? Standard methodology compares responses under high vs. low anchor conditions; a technique "works" if it reduces this gap. We identify a critical limitation: this metric misses **overcorrection**, where techniques move responses away from anchors but past the unbiased answer.

We introduce **baseline convergence** as a complementary evaluation metric. By collecting baseline responses without explicit anchors (n=909 across 10 models), we measure whether techniques bring outputs closer to the model's considered judgment. This metric reveals *overcorrection*—when techniques move responses away from anchors but past where the model would naturally respond. While there is no objective "correct" sentence, the baseline provides a meaningful reference for measuring technique effects. Using this metric across 13,369 trials, we find that technique effectiveness varies substantially:

- **Full SACD** (Self-Administered Cognitive Debiasing; iterative self-reflection): +24% improvement ($d = 0.41$, $p < .001$)
- **Premortem**: +10% improvement ($p < .001$, $d = 0.17$)
- **Random Control**: +9% improvement ($p < .001$, $d = 0.15$)
- **Devil's Advocate**: +2% (not significant, $p = 0.33$)

(Our Outside View implementation produced confounded results and is discussed separately in Section 5.3.)

Iterative self-reflection (Full SACD) is the most effective technique, but with high model variance: 5/10 models significantly improve, while Claude Opus 4.6 shows 68% *worse* convergence ($p < .001$). Devil's Advocate shows no significant effect ($p = 0.33$).

Without baseline collection, overcorrection would be invisible under standard susceptibility metrics. We propose baseline convergence as a complementary metric for LLM debiasing research, particularly useful for detecting overcorrection.

1 Introduction

When large language models make judgments, do debiasing techniques actually help—or do they just move errors in a different direction?

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

We report findings from a large systematic evaluation of LLM debiasing techniques (13,369 trials across 10 models). Our core contribution is methodological: by collecting baseline responses without explicit anchors, we can measure not just whether techniques *reduce susceptibility* to anchors, but whether they bring outputs *closer to the model's considered judgment*.

Study design note: We analyze each model independently, measuring how debiasing techniques affect that model's responses at different temperatures. Anchors use *constant relative scaling*: low anchor = baseline × 0.5, high anchor = baseline × 1.5. This ensures anchors are equally strong relative to each model's natural judgment—a fixed 3-month anchor would be a strong pull for a model with an 18-month baseline but negligible for one with a 36-month baseline. This design answers: "How do techniques affect *this* model?" rather than "Which model is least susceptible?"

This distinction matters. Standard anchoring studies compare high-anchor and low-anchor conditions—if the gap shrinks, the technique "works." But this metric misses a critical failure mode: **overcorrection**. A technique that moves every response to 15 months, regardless of whether the unbiased answer is 30 months or 6 months, would show "reduced susceptibility" while actually *increasing* distance from truth.

1.1 The Baseline Convergence Metric

We introduce a complementary evaluation metric: **baseline convergence**.

- **Susceptibility** (standard): $|\bar{R}_{high} - \bar{R}_{low}|$
- **Convergence** (ours): $|R_{technique} - R_{baseline}|$

A technique succeeds on convergence if it brings the response *closer* to what the model would say without any anchor present.

1.2 Findings Preview

Using this metric, we observe technique rankings with clear statistical separation:

Convergence metric: A hierarchy emerges:

1. **Full SACD** (+24%, $p < .001$, $d = 0.41$)—iterative self-reflection
2. **Premortem** (+10%, $p < .001$)—imagine failure mode
3. **Random Control** (+9%, $p < .001$)—extra turns, no debiasing content
4. **Devil's Advocate** (+2%, $p = 0.33$, not significant)—argumentation

Simple structural interventions (extra turns) produced meaningful improvements with minimal prompt complexity. Our Outside View implementation showed worse convergence (−22%), but this result is confounded and discussed separately in Section 5.3.

1.3 Why This Matters

This has immediate practical implications:

1. **Practitioners don't need complex debiasing prompts.** Simply adding conversation turns helps more than specific debiasing instructions.

2. **Reference class reasoning (Outside View) may introduce secondary anchors.** In our implementation, specifying jurisdiction to avoid model refusals may have anchored responses to that jurisdiction's typical sentences.

3. **Baseline collection enables overcorrection detection.** Without baselines, techniques that overcorrect would appear effective under susceptibility metrics.

4. **The standard evaluation metric would have misled us completely.** Direction-based analysis showed Outside View as universally effective; calibration analysis reveals it as worst.

1.4 Contributions

1. **A baseline convergence metric for debiasing evaluation** that catches overcorrection invisible to susceptibility measures.

2. **Technique rankings differ between metrics**: Under susceptibility (spread reduction), most techniques appear effective. Under convergence, Full SACD leads (+24%) while our Outside View implementation showed −22% worse convergence (confounded; see Section 5.3). Effect sizes are small ($d \leq 0.41$).

3. **High model variance**: 5/10 models significantly improve with SACD, but Opus 4.6 shows 68% *worse* convergence.

4. **13,369 trials across 10 models** with Bonferroni-corrected statistics and effect sizes.

2 Related Work

2.1 Anchoring Bias in Human Judgment

Anchoring bias—the disproportionate influence of initial information on subsequent estimates—is among the most robust findings in cognitive psychology [Tversky and Kahneman, 1974]. Even experts are susceptible: Englich et al. [2006] demonstrated that experienced judges' sentencing decisions were influenced by random numbers generated by dice rolls. Effect sizes of $d = 0.6$–1.2 persist regardless of anchor source or participant awareness. Our experimental paradigm adapts this judicial sentencing design.

2.2 Cognitive Biases in LLMs

Recent work has shown that LLMs exhibit human-like cognitive biases [Binz and Schulz, 2023, Jones and Steinhardt, 2022]. Anchoring effects have been documented across multiple model families [Huang et al., 2025], with susceptibility varying by model architecture and size. Song et al. [2026] survey LLM reasoning failures comprehensively, including susceptibility to anchoring and framing effects. Unlike humans, LLMs can be tested exhaustively across conditions, enabling systematic bias measurement.

2.3 Debiasing Techniques

Several techniques have been proposed for mitigating anchoring:

Outside View / Reference Class Forecasting: Prompting models to consider what typically happens in similar cases [Sibony, 2019]. Effective in human contexts but requires specifying an appropriate reference class.

Self-Administered Cognitive Debiasing (SACD): Iterative prompting that guides models through bias detection and correction [Lyu et al., 2025]. Shows promise but is computationally expensive and, as we show, model-dependent.

Devil's Advocate: Prompting models to argue against their initial response. Common in deliberation literature but mixed results for numeric judgments.

Premortem Analysis: Asking models to imagine the decision failed and explain why. Drawn from project management practice [Klein, 2007].

2.4 Evaluation Methodology

Standard anchoring evaluation compares high-anchor and low-anchor conditions [Englich et al., 2006, Huang et al., 2025]:

$$\text{Susceptibility} = |\bar{R}_{high} - \bar{R}_{low}|$$

A technique "works" if it reduces this gap. This methodology does not require ground truth—it measures susceptibility to anchors, not accuracy of outputs. This is a valid and important metric.

We extend this by introducing **baseline convergence**:

$$\text{Convergence Error} = |R_{technique} - R_{baseline}|$$

This requires collecting baseline responses but enables detection of **overcorrection**—a failure mode invisible to susceptibility-only evaluation. To our knowledge, no prior work on LLM anchoring has systematically collected baselines without explicit anchors for convergence evaluation.

3 Methodology

3.1 Evaluation Metrics

We distinguish two evaluation approaches for debiasing techniques:

3.1.1 Standard Metric: Anchor Susceptibility

The conventional approach compares responses under high vs. low anchor conditions:

$$\text{Susceptibility} = |\bar{R}_{high} - \bar{R}_{low}|$$

A technique "works" if it reduces this gap. This metric answers: *Does the technique reduce the anchor's influence?*

3.1.2 Our Metric: Baseline Convergence

We collected baseline responses without explicit anchors—model outputs with no prosecutor demand anchor present. This enables a second metric:

$$\text{Convergence Error} = |\bar{R}_{technique} - \bar{R}_{baseline}|$$

A technique succeeds if it reduces convergence error relative to the anchored (no-technique) condition:

$$\text{Improved} = |R_{technique} - R_{baseline}| < |R_{anchored} - R_{baseline}|$$

This metric answers: *Does the technique bring the response closer to the model's unprompted judgment?*

3.1.3 Why Both Metrics Matter

These metrics can diverge. Consider:

- Baseline: 30mo
- High-anchor response: 50mo (convergence error = 20mo)
- Technique response: 12mo (convergence error = 18mo... but overcorrected)

Under susceptibility, the technique "worked" (moved away from anchor). Under convergence, it marginally helped—but a different technique might achieve 28mo (convergence error = 2mo).

3.2 Experimental Design

3.2.1 Models

We evaluated 10 models across 4 providers:

Provider	Models
Anthropic	Claude Haiku 4.5, Sonnet 4.6, Opus 4.6
OpenAI	GPT-4.1, GPT-5.2, o3, o4-mini
DeepSeek	DeepSeek-v3.2
Others	Kimi-k2.5 (Moonshot), GLM-5 (Zhipu)

3.2.2 Conditions

1. **Baseline**: Sentencing prompt with no anchor
2. **Low anchor**: 3-month anchor in prosecutor demand
3. **High anchor**: 36–60 month anchor in prosecutor demand
4. **Techniques**: Applied to high-anchor condition

3.2.3 Techniques Evaluated

Technique	Description
Outside View	"What typically happens in similar cases?" (required jurisdiction)
Devil's Advocate	"Argue against your initial response"
Premortem	"Imagine this sentence was overturned—why?"
Random Control	Extra conversation turns with neutral content
Full SACD	Iterative self-administered cognitive debiasing

3.2.4 Temperature Conditions

Each technique was tested at three temperatures: t=0 (deterministic), t=0.7 (moderate variance), and t=1.0 (high variance). Baseline responses were collected at all three temperatures. Results are

aggregated across temperatures; preliminary analysis showed no significant temperature×technique interaction effects ($p > 0.1$ for all comparisons).

3.2.5 Trial Counts and Procedure

- **Total trials**: 13,369
- **Per model-technique-temperature**: 30–90 trials (target $n \geq 30$; range reflects iterative data collection with some conditions receiving additional trials for robustness)
- **Baseline trials**: 909 total (approximately 90 per model across all temperatures)
- **Response extraction**: Final numeric response extracted via regex pattern matching for integer month values
- **Trial assignment**: Trials run in batches by model and technique; order randomized within batches
- **Anchor values**: Constant relative scaling—low anchor = baseline × 0.5, high anchor = baseline × 1.5. This ensures equal anchor strength across models with different baselines.

Table 1: Trial distribution by condition. Each comparison uses the sample sizes shown; trials are assigned to conditions without overlap. All model-technique-temperature cells achieved $n \geq 30$.

Condition	n
Debiasing Techniques (vs. anchored control)	
Full SACD	2,391
Outside View	2,423
Random Control	2,215
Premortem	2,186
Devil's Advocate	2,166
Control Conditions	
Anchored (no technique)	1,509
Unanchored Baseline	909
Total unique trials	**13,369**

3.3 Confounds and Limitations

3.3.1 Outside View Jurisdiction Context

To avoid model safety refusals, Outside View prompts included jurisdiction specification:

> "In German federal courts, what is the TYPICAL probation sentence..."

This may have introduced a secondary anchor toward German sentencing norms (∼12–18 months for probation). Other techniques did not require this modification.

4 Results

4.1 Baseline Responses

Unanchored baseline responses varied substantially across models:

Model	Baseline Mean	SD
o4-mini	35.7mo	4.7
o3	33.7mo	5.6
GLM-5	31.9mo	5.7
GPT-5.2	31.8mo	5.7
Kimi-k2.5	30.6mo	7.4
DeepSeek-v3.2	29.6mo	8.0
Haiku 4.5	29.1mo	11.2
GPT-4.1	25.1mo	3.4
Sonnet 4.6	24.1mo	1.3
Opus 4.6	18.0mo	0.0

Table 2: Model baselines range from 18.0mo (Opus) to 35.7mo (o4-mini)—a 17.7mo spread. Opus 4.6 shows zero variance at all temperatures, consistently responding with exactly 18 months—suggesting highly deterministic reasoning for this prompt type.

4.2 High-Anchor Responses (No Technique)

Under high-anchor conditions without intervention, two anchor response patterns emerge:

1. **Compression**: Response pulled below baseline (Anthropic models, GPT-4.1)
2. **Inflation**: Response pulled above baseline (GPT-5.2, GLM-5, o3)

4.3 Technique Effectiveness: Baseline Convergence

Technique	n	Mean Dist	95% CI	Improvement	p (Bonf)	d
Anchored baseline	1509	12.4mo	[12.0, 12.7]	—	—	—
Full SACD	2391	9.4mo	[9.1, 9.8]	**+24%**	$< .001$	0.41
Premortem	2186	11.1mo	[10.8, 11.5]	+10%	$< .001$	0.17
Random Control	2215	11.3mo	[11.0, 11.6]	+9%	$< .001$	0.15
Devil's Advocate	2166	12.1mo	[11.8, 12.4]	+2% (ns)	1.000	0.03
Outside View†	2423	15.1mo	[14.8, 15.4]	−22%	$< .001$	−0.38

Table 3: Technique effectiveness with 95% confidence intervals and Bonferroni-corrected p-values. Effect sizes are small by Cohen's conventions ($d < 0.5$); statistical significance does not imply practical significance. †Outside View result confounded by required jurisdiction specification; included for transparency but excluded from primary conclusions.

4.4 Model-Specific Results: Full SACD

Full SACD shows high variance across models (Bonferroni-corrected, 10 tests):

Model	Improvement	p (adj)	Result
o3	+51%	$< .001$	Significant improvement
GPT-4.1	+48%	$< .001$	Significant improvement
Sonnet 4.6	+46%	$< .001$	Significant improvement
DeepSeek-v3.2	+30%	$< .001$	Significant improvement
GPT-5.2	+20%	0.022	Significant improvement
o4-mini	+12%	0.210	Not significant
Haiku 4.5	−2%	1.000	Not significant
Kimi-k2.5	−3%	1.000	Not significant
GLM-5	−4%	1.000	Not significant
Opus 4.6	**−68%**	$< .001$	**Significant backfire**

Table 4: Full SACD model-specific results. 5/10 significantly improve, 1/10 significantly worsens (Opus 4.6).

Key findings:

1. **5/10 models significantly improve** after Bonferroni correction
2. **Opus 4.6 shows severe backfire** (−68%, $p < .001$)—the technique makes it *worse*
3. **Effect sizes remain small** even for best performers ($d \leq 0.41$)

4.5 Why Baseline Collection Matters

Consider a technique that reduces all responses to the same value regardless of anchor. Under susceptibility ($|R_{high} - R_{low}|$), this appears perfect—zero spread. Under convergence ($|R - R_{baseline}|$), the technique may perform poorly if that fixed value diverges from the baseline.

Our Outside View implementation (as confounded by jurisdiction specification) exemplifies this: it produces consistent responses that diverge from model baselines by 22%. Without baseline collection, this overcorrection would be invisible.

5 Discussion

5.1 Why Full SACD Works (and Fails)

Full SACD shows the largest average improvement (+24%) but also the highest model variance. We propose:

Hypothesis 1: Iterative reflection enables genuine reconsideration. Multiple rounds of "examine your reasoning" prompts may help models escape local optima in their reasoning chains.

Hypothesis 2: Some models perform "debiasing theater." Opus 4.6's severe backfire (−68%) suggests the technique can activate surface compliance without genuine reconsideration—the model may be optimizing for *appearing* to reconsider rather than actually doing so.

Hypothesis 3: Baseline proximity matters. Opus 4.6 has the lowest baseline (18mo), meaning SACD may be pulling it *away* from its natural judgment toward a perceived "expected answer."

5.2 Why Random Control Works

Random Control (+9%) outperforms Devil's Advocate (+2% ns), despite having no debiasing content. Possible mechanisms:

Attention redistribution. Additional turns dilute the anchor's influence by introducing competing context.

Implicit reconsideration. Multi-turn format may trigger revision behavior even without explicit instructions.

5.3 The Outside View Confound

Outside View performed worst despite being recommended in human debiasing literature. Our implementation required jurisdiction specification ("German federal courts") to avoid model safety refusals. This may have introduced a secondary anchor:

- German probation for repeat shoplifting: ∼12–18 months
- Our model baselines (without explicit anchor): 18–36 months
- Outside View consistently pulled toward ∼15 months

Implication for practitioners: When using Outside View, ensure the reference class matches your actual decision context. Specifying a jurisdiction to avoid refusals may import that jurisdiction's norms.

5.4 Limitations

1. **Single domain.** All experiments use judicial sentencing vignettes. Replication across other anchoring domains (estimation, forecasting, negotiation) is needed before generalizing findings.
2. **Outside View confound.** Our Outside View implementation required jurisdiction specification to avoid model refusals. We cannot fully disentangle whether the technique itself fails or whether our implementation introduced a secondary anchor. Future work should test jurisdiction-neutral Outside View prompts.
3. **Baseline interpretation.** Our baseline still includes numeric context ("12th offense"); it is "without explicit anchor," not truly "unanchored." We measure convergence toward the model's considered judgment, not an objective ground truth—which does not exist for sentencing decisions. The baseline represents the model's response absent *explicit prosecutor demand anchoring*, not an "unbiased" state.
4. **Model coverage.** 10 models from 4 providers is substantial but not exhaustive. Results may not apply to other model families.
5. **Prompt disclosure.** Complete prompt templates are available at our repository; we acknowledge that prompt engineering choices may influence results.

5.5 Practical Recommendations

Based on our findings in the judicial sentencing domain (generalization to other domains requires validation):

1. **Consider structural interventions.** Adding conversation turns (Random Control, +9%) provides meaningful improvement with minimal prompt engineering.

2. **Test per-model.** Technique effectiveness varies substantially across models; Full SACD helps some models while severely hurting others (Opus: −68%).

3. **Collect baselines.** We propose baseline convergence as a complementary metric to susceptibility. Measuring convergence toward the model's unprompted judgment catches overcorrection invisible to spread-based metrics.

4. **Be cautious with reference class prompts.** Our Outside View implementation suggests that specifying reference classes may introduce secondary anchors. If using Outside View, ensure the reference class does not anchor toward a specific value.

6 Conclusion

We introduced baseline convergence as a metric for evaluating LLM debiasing techniques. This metric catches overcorrection—a failure mode invisible to standard susceptibility measures.

Our key findings from 13,369 trials across 10 models:

1. **Full SACD leads, but with high variance.** +24% average improvement ($d = 0.41$), but Opus 4.6 shows −68% backfire. Practitioners must test per-model.

2. **Our Outside View implementation produced confounded results.** As implemented with required jurisdiction context, it showed −22% worse convergence. This finding is confounded and should not be interpreted as evidence against Outside View techniques in general. Future work should test jurisdiction-neutral implementations.

3. **Effect sizes are small.** Even the best technique achieves only $d = 0.41$ ("small" by Cohen's conventions). Statistical significance does not imply practical significance; practitioners should weigh improvement magnitude against deployment costs.

4. **Baseline collection is essential.** Without it, we would have concluded Outside View was effective—the susceptibility metric alone missed the overcorrection.

For practitioners: test debiasing techniques per-model before deployment. Full SACD is effective for most models but can severely backfire. Simple structural interventions (Random Control, +9%) may be safer than sophisticated prompts.

For researchers: collect baselines without explicit anchors. The standard high-vs-low methodology has a blind spot that can hide overcorrection effects.

A Prompt Templates

A.1 Case Vignette (All Conditions)

> 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.

A.2 Anchor Introduction (High/Low Anchor Conditions)

> 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. `[ANCHOR]` months on probation.

A.3 Outside View (As Implemented)

> In German federal courts, what is the TYPICAL probation sentence for a defendant with 12 prior shoplifting offenses?

Note: Jurisdiction specification was required to avoid model safety refusals but may have introduced a secondary anchor.

A.4 Full SACD (Iterative Self-Administered Cognitive Debiasing)

Following Lyu et al. [2025], Full SACD implements three iterative steps:

1. **Bias Determination**: "Analyze the following prompt... For each sentence, determine if it contains a cognitive bias"
2. **Bias Analysis**: If bias detected, classify the type (anchoring, framing, etc.)
3. **Cognitive Debiasing**: "Rewrite the flagged sentences to remove the bias"

Steps repeat until no bias is detected or maximum iterations (5) reached. Average iterations to convergence: 2.3.

A.5 Random Control

Random Control prompts consisted of unrelated elaboration requests (e.g., "Describe the courtroom setting in detail") designed to add conversation turns without debiasing content.

Data and Code Availability

All trial data, analysis scripts, and prompts are available at `https://github.com/voder-ai/bAIs`. The repository includes raw JSONL trial data for all 13,369 trials, statistical analysis scripts reproducible from raw data, and complete prompts for all debiasing techniques.

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.

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.

Yucheng Huang et al. Anchoring bias in large language models: An empirical study. *arXiv preprint*, 2025. Preprint.

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.

Gary Klein. *The Power of Intuition: How to Use Your Gut Feelings to Make Better Decisions at Work*. Currency, 2007. ISBN 978-0385502894.

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

Olivier Sibony. *You're About to Make a Terrible Mistake!: How Biases Distort Decision-Making and What You Can Do to Fight Them*. Little, Brown Spark, 2019. ISBN 978-0316494984.

Yang Song et al. A survey of reasoning failures in large language models. *arXiv preprint*, 2026. Preprint.

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