Baseline Convergence, Not Susceptibility: Evaluating LLM Debiasing with Unanchored Baselines

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 metric. By collecting unanchored responses (n=909 across 10 models), we can measure whether techniques bring outputs closer to the model's unprompted judgment, not just away from anchors. Using this metric across 13,369 trials, we discover rankings that invert conventional wisdom:

- **Full SACD** (iterative self-reflection): +24% improvement ($d = 0.41$, $p < .001$)
- **Premortem / Random Control**: +9–10% improvement ($p < .001$)
- **Outside View** (reference class reasoning): −22%—significantly *worsens* convergence

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, we would have concluded Outside View was universally effective—a finding completely inverted by proper convergence measurement. We argue baseline collection should become standard practice in LLM debiasing research.

1 Introduction

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

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 unanchored baseline responses, we can measure not just whether techniques *reduce susceptibility* to anchors, but whether they bring outputs *closer to the model's unprompted judgment.*

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

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

unbiased 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 discover rankings that invert conventional wisdom:

Standard metric (susceptibility): All techniques appear roughly equivalent—most reduce the high-low gap.

Convergence metric: Clear hierarchy emerges with statistical significance:

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
5. **Outside View** (−22%, $p < .001$)—reference class reasoning *backfires*

The counterintuitive finding: **Outside View, often recommended in human debiasing literature, significantly worsens model performance**. Meanwhile, simple structural interventions (extra turns) help nearly as much as sophisticated techniques.

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. **Temperature interacts with technique type.** Deterministic responses (t=0) work best for structural interventions; moderate variance (t=0.7) helps self-reflection.
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. **Inverted technique rankings**: Outside View, recommended in human literature, *backfires* (-22%) while Full SACD leads ($+24\%$).
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 unanchored baselines 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 unanchored baseline responses—model outputs with no 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).

3.2.5 Trial Counts

- **Total trials**: 13,369
- **Per model-technique-temperature**: 30–90 trials (target $n \geq 30$)
- **Baseline trials**: 909 total across all models

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 1: Model baselines range from 18.0mo (Opus) to 35.7mo (o4-mini)—a 17.7mo spread. Opus shows zero variance (deterministic).

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	Improvement	p (Bonf)	Effect Size
Anchored baseline	1509	12.4mo	—	—	—
Full SACD	2391	9.4mo	**+24%**	$< .001$	$d = 0.41$
Premortem	2186	11.1mo	+10%	$< .001$	$d = 0.17$
Random Control	2215	11.3mo	+9%	$< .001$	$d = 0.15$
Devil's Advocate	2166	12.1mo	+2% (ns)	1.000	$d = 0.03$
Outside View	2423	15.1mo	−22%	$< .001$	$d = -0.38$

Table 2: Technique effectiveness with Bonferroni-corrected p-values (5 tests). Full SACD shows largest improvement; Outside View significantly *worsens* convergence. Effect sizes are small by Cohen's conventions.

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 3: 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 Comparison: Susceptibility vs. Convergence Metrics

Under the standard susceptibility metric, Outside View appeared to "improve" models by reducing the high-low gap. Under convergence:

Metric	Outside View	Full SACD
Susceptibility ($\|high - low\|$)	Appears effective	Appears effective
Convergence ($\|response - baseline\|$)	-22% **(backfires)**	$+$**24% (best)**

Table 4: Rankings invert between metrics. Without baseline collection, Outside View appears to "work."

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: $\sim$12–18 months
- Our unanchored baselines: 18–36 months (model-dependent)
- Outside View consistently pulled toward $\sim$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. Results may not generalize.
2. **Outside View confound.** We cannot fully disentangle technique failure from implementation choice.
3. **Baseline validity.** Our "unanchored" baseline still includes numeric context ("12th offense").
4. **Model coverage.** 10 models from 4 providers is substantial but not exhaustive.

5.5 Practical Recommendations

Based on our findings:

1. **Start with structure, not content.** Adding conversation turns is simpler and more effective than crafting debiasing prompts.
2. **Match temperature to technique.** Use t=0 for structural interventions, t=0.7 for self-reflection.
3. **Validate with calibration metric.** Don't just measure susceptibility—measure whether outputs land closer to baseline.
4. **Test per-model.** Technique effectiveness varies substantially across models.

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.
2. **Outside View backfires.** Despite recommendations in human debiasing literature, it shows −22% worse convergence ($p < .001$).
3. **Effect sizes are small.** Even the best technique achieves only $d = 0.41$—practitioners should calibrate expectations.
4. **Baseline collection is essential.** Without it, we would have concluded Outside View was effective.

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 unanchored baselines. The standard high-vs-low methodology has a blind spot that inverted our technique rankings.

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.

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

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*. Crown Business, 2007.

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.

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

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