Calibration, 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 **calibration to baseline** as a complementary metric. By collecting unanchored responses (n=1,001 across 11 models), we can measure whether techniques bring outputs closer to ground truth, not just away from anchors. Using this metric across 14,220 trials, we discover rankings that invert conventional wisdom:

- **Random Control** (extra turns, no debiasing content): 91% of models improved
- **Self-reflection techniques** (Premortem, SACD): 82%
- **Outside View** (reference class reasoning): **36%**—worst performer

The simplest structural intervention outperforms sophisticated prompt engineering. Temperature interacts with technique type: deterministic sampling (t=0) optimizes structural interventions; moderate variance (t=0.7) aids self-reflection.

Without baseline collection, we would have concluded Outside View was universally effective—a finding completely inverted by proper calibration 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 the largest systematic evaluation of LLM debiasing techniques to date (14,220 trials across 11 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 ground truth*.

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 Calibration Metric

We introduce a complementary evaluation metric: **calibration to baseline**.

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

A technique succeeds on calibration 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.

Calibration metric: Clear hierarchy emerges:

1. **Random Control** (10/11 models calibrated)—extra conversation turns with no debiasing content
2. **Premortem / Full SACD** (9/11)—self-reflection techniques
3. **Devil's Advocate** (7/11)—argumentation
4. **Outside View** (4/11)—reference class reasoning

The counterintuitive finding: **the simplest intervention beats the most sophisticated**. Extra turns with irrelevant content outperform carefully crafted debiasing prompts.

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 calibration metric for debiasing evaluation** that catches overcorrection invisible to susceptibility measures.

2. **Inverted technique rankings** showing structure (conversation turns) beats content (debiasing instructions).

3. **Temperature × technique interaction effects**—first systematic analysis of temperature's role in debiasing.

4. **14,220 trials across 11 models**—the largest LLM debiasing evaluation to date.

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 [**?**]. 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 [**?**].

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 **calibration to unanchored baselines**:

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

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 calibration 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 Calibration

We collected unanchored baseline responses—model outputs with no anchor present. This enables a second metric:

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

A technique succeeds if it reduces calibration 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 ground truth?*

3.1.3 Why Both Metrics Matter

These metrics can diverge. Consider:

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

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

3.2 Experimental Design

3.2.1 Models

We evaluated 11 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, GLM-5, MiniMax-m2.5

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**: ∼14,100
- **Per model-technique-temperature**: 20–50 trials
- **Baseline trials per model**: 91

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
o4-mini	35.7mo
o3	33.7mo
GLM-5	31.8mo
GPT-5.2	31.8mo
Kimi-k2.5	30.7mo
DeepSeek-v3.2	29.6mo
Haiku 4.5	29.1mo
GPT-4.1	25.1mo
Sonnet 4.6	24.1mo
MiniMax-m2.5	24.1mo
Opus 4.6	18.0mo

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

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: Calibration Metric

4.3.1 High-Anchor Conditions

Technique	Improved	Success Rate
Random Control	10/11	**91%**
Premortem	9/11	82%
Full SACD	9/11	82%
Devil's Advocate	7/11	64%
Outside View	4/11	**36%**

Table 2: Random Control—which adds conversation turns without debiasing content—outperforms all content-based techniques.

4.3.2 Low-Anchor Conditions

Technique	Improved	Success Rate
Full SACD	11/11	**100%**
Premortem	9/11	82%
Random Control	7/11	64%
Outside View	5/11	45%
Devil's Advocate	4/11	36%

Table 3: Full SACD achieves perfect calibration under low anchors. Rankings shift between anchor conditions.

4.4 Temperature × Technique Interaction

Technique	t=0	t=0.7	t=1	Optimal
Random Control	**100%**	80%	91%	**t=0**
Premortem	70%	**80%**	64%	t=0.7
Full SACD	64%	**73%**	64%	t=0.7
Devil's Advocate	60%	60%	64%	t=1
Outside View	30%	30%	36%	t=1

Table 4: Temperature effects on calibration success (high-anchor conditions).

Key findings:

1. **Random Control at t=0 achieves 100% success**—deterministic extra turns are optimal

2. **Self-reflection techniques (SACD, Premortem) prefer t=0.7**—moderate variance aids deliberation

3. **Outside View fails at all temperatures**—the technique itself is problematic, not the sampling

4.5 Comparison: Susceptibility vs. Calibration Metrics

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

Metric	Outside View Ranking
Susceptibility ($\|high - low\|$)	Best (11/11 "improved")
Calibration ($\|response - baseline\|$)	**Worst** (4/11 improved)

Table 5: This inversion demonstrates why baseline collection matters.

5 Discussion

5.1 Why Structure Beats Content

Our most surprising finding is that Random Control—conversation turns with irrelevant content—outperforms deliberate debiasing techniques. We propose several explanations:

Hypothesis 1: Attention redistribution. Additional turns may dilute the anchor's influence by introducing competing context. The model's attention becomes distributed across more tokens, reducing the relative weight of the anchoring value.

Hypothesis 2: Implicit reconsideration. Multi-turn format may trigger different inference patterns than single-shot prompts. The model may treat subsequent turns as opportunities to revise rather than defend prior responses.

Hypothesis 3: Debiasing content backfires. Explicit debiasing instructions may activate "debiasing theater"—surface compliance without genuine reconsideration. Structure avoids this because there's nothing to perform.

The temperature findings support Hypothesis 2: Random Control works best at t=0 (deterministic), suggesting the structural effect is robust and doesn't require sampling variance.

5.2 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 unanchored baselines: 18–36 months (model-dependent)
- 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.

5.3 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.** 11 models from 4 providers is substantial but not exhaustive.

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

Our key findings:

1. **Structure beats content.** Random Control (extra turns, no debiasing content) achieves 91% calibration improvement vs. 36% for Outside View.

2. **Temperature matters.** Structural interventions prefer t=0; self-reflection prefers t=0.7.

3. **Baseline collection is essential.** Without it, we would have published inverted rankings.

For practitioners: start with structure. Add conversation turns before crafting complex debiasing prompts. Validate with calibration metrics, not just susceptibility.

For researchers: collect unanchored baselines. The standard high-vs-low methodology has a blind spot. Ground truth matters.

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.