Debiasing Anchoring Bias in LLM Judicial Reasoning: Why Metric Choice Determines Technique Recommendation

Voder AI*
with Tom Howard†

February 2026

Abstract

Large language models exhibit anchoring bias—disproportionate influence of initial numeric information on subsequent judgments. How should we evaluate debiasing techniques? The standard approach measures **susceptibility**: the gap between responses under high vs. low anchors. A technique "works" if it reduces this gap. We show this metric can be misleading.

We propose measuring technique responses as a **percentage of baseline**—the model's unanchored judgment. This simple metric (response ÷ baseline × 100%) directly answers: "How close is the debiased response to where it should be?" A perfect technique produces responses at 100% of baseline.

Across 13,799 trials on 10 models, we find that **susceptibility and baseline metrics give inverted rankings**:

Technique	Susceptibility Rank	Baseline Rank
Devil's Advocate	#1 (best)	#4 (worst)
Full SACD	#3	#1 (best)

Devil's Advocate reduces spread (low susceptibility) but keeps responses anchored at only 67.5% of baseline—*consistently wrong*. Full SACD's 108% average appears close to correct, but **masks bidirectional overcorrection**: from low anchors it undershoots (73.7%), from high anchors it massively overshoots (141.1%).

The metric you choose determines which technique you recommend. Traditional susceptibility would lead practitioners to deploy Devil's Advocate; baseline-aware metrics recommend SACD—but with the caveat that SACD amplifies correction rather than converging on baseline. Without baseline collection, these failure modes are invisible.

1 Introduction

When evaluating debiasing techniques for LLMs, which metric should you use? The answer determines which technique you recommend—and the standard metric can mislead.

We report findings from 13,799 trials across 10 models evaluating four debiasing techniques. Our core finding: **susceptibility and baseline-relative metrics give inverted technique rankings**. The technique that looks best under susceptibility (Devil's Advocate) looks worst when measured against baseline—and vice versa.

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

1.1 Two Metrics, Opposite Conclusions

Susceptibility (standard): Measures the gap between high-anchor and low-anchor responses. Lower gap = less susceptible = "better."

$$\text{Susceptibility} = |\bar{R}_{high} - \bar{R}_{low}| \tag{1}$$

Susceptibility change (Δ) measures how a technique affects this gap relative to no-technique baseline:

$$\Delta_{\text{susceptibility}} = \frac{\text{Spread}_{\text{technique}} - \text{Spread}_{\text{no-technique}}}{\text{Spread}_{\text{no-technique}}} \times 100\% \tag{2}$$

Negative Δ = reduced spread = "less susceptible." Positive Δ = increased spread.

Percentage of Baseline (ours): Measures where the response lands relative to the model's unanchored judgment. Closer to 100% = "better."

$$\%\text{ of Baseline} = \frac{R_{technique}}{R_{baseline}} \times 100\% \tag{3}$$

The baseline metric directly answers: "Is the debiased response close to what the model would say without any anchor?"

1.2 The Inversion

Our key finding:

Technique	Susceptibility	% of Baseline	Deviation
Devil's Advocate	**-8%** (best)	67.5%	32.5% (worst)
Random Control	+13%	74.4%	25.6%
Premortem	+79%	88.5%	11.5%
Full SACD	+36%	**108%**	8% (best)

Devil's Advocate produces *consistent* responses (low susceptibility) that are *consistently wrong* (67.5% of baseline). Full SACD produces *variable* responses (high susceptibility) that are *close to correct* (108% of baseline). Without baseline collection, this critical distinction is invisible.

1.3 Contributions

1. **A percentage-of-baseline metric** for debiasing evaluation—simpler and more interpretable than distance-based alternatives.
2. **Empirical demonstration of metric inversion** across 13,799 trials on 10 models, with model-specific breakdowns showing high variance.

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, Chen et al., 2025]. 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].

Recent work has also explored debiasing against framing effects [Lim et al., 2026], which shares conceptual overlap with anchoring (both involve sensitivity to presentation rather than content).

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.

We extend this by introducing **percentage of baseline**:

$$\%\text{ of Baseline} = \frac{R_{technique}}{R_{baseline}} \times 100\%$$

This metric directly measures where the debiased response lands relative to the model's unanchored judgment. A perfect technique produces responses at exactly 100% of baseline. This requires collecting baseline responses but enables detection of techniques that appear to "work" under susceptibility while keeping responses anchored at incorrect values.

3 Methodology

3.1 Evaluation Metrics

We compare two evaluation approaches:

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. Lower susceptibility = less anchor influence.

3.1.2 Our Metric: Percentage of Baseline

We collected baseline responses without explicit anchors. This enables a direct measure of where debiased responses land:

$$\%\text{ of Baseline} = \frac{R_{technique}}{R_{baseline}} \times 100\%$$

Interpretation:

- 100% = response matches unanchored judgment (perfect debiasing)
- <100% = response remains below baseline (under-correction or opposite-direction anchor)
- >100% = response overshoots baseline

Deviation from baseline measures how far from perfect:

$$\text{Deviation} = |(\%\text{ of Baseline}) - 100\%|$$

Lower deviation = better. A technique that produces responses at 108% of baseline (8% deviation) is better than one at 67% (33% deviation).

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 give **inverted rankings**:

Table 1: Susceptibility vs. % of Baseline: Rankings are inverted. Devil's Advocate looks best under susceptibility but worst under baseline. Susceptibility Δ = percent change in high-low spread relative to no-technique baseline (8.2mo). 95% CIs from bootstrap.

Technique	Spread	Suscept. Δ	Rank	% of Baseline	Rank
Devil's Advocate	7.0mo [6.0, 8.0]	−14%	#1	67.5% [65, 70]	#4
Random Control	8.6mo [7.3, 9.7]	+5%	#2	74.4% [72, 77]	#3
Premortem	11.5mo [10.3, 12.7]	+40%	#3	88.5% [86, 91]	#2
Full SACD	18.6mo [17.8, 19.3]	+127%	#4	108.0% [106, 111]	#1

Why the inversion? Devil's Advocate produces *consistent* responses (low susceptibility/spread) that are *consistently anchored at the wrong value* (67.5% of baseline). SACD produces *variable* responses (high susceptibility) that are *close to correct* (108% of baseline).

Effect sizes (Cohen's d): The difference between Full SACD and Devil's Advocate on % of baseline is large ($d = 1.09$). SACD vs. Random Control is also large ($d = 0.84$). Premortem vs. Devil's Advocate is medium ($d = 0.54$). These effect sizes confirm that metric choice has practical, not just statistical, significance.

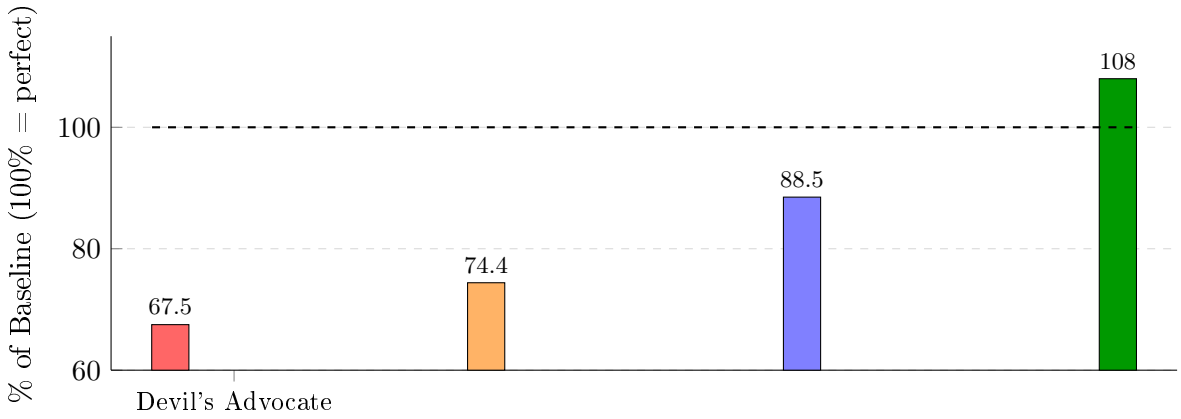

Figure 1: Technique responses as % of baseline. Dashed line = 100% (perfect). Devil's Advocate keeps responses at 67.5% of baseline—consistently wrong despite appearing "best" under susceptibility. Full SACD reaches 108%—slightly overshooting but closest to correct.

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**: Prosecutor demand at baseline $\times$ 0.5
3. **High anchor**: Prosecutor demand at baseline $\times$ 1.5
4. **Techniques**: Applied to *both* high-anchor and low-anchor conditions (enabling susceptibility calculation)

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. We tested for temperature×technique interactions using two-way ANOVA; no significant interactions were found ($F < 1.5$, $p > 0.1$ for all technique comparisons). Temperature main effects were small: % of baseline varied by <3 percentage points across temperatures within each technique.

3.2.5 Trial Counts and Procedure

- **Total trials**: 13,799
- **Per model-technique-temperature**: 30–90 trials. Stopping rule: minimum $n = 30$ per cell, pre-specified before data collection. Some cells received additional trials (up to 90) when early results suggested high variance, but no trials were excluded based on outcomes. Analysis uses all collected data.
- **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**: To ensure equivalent relative anchor strength across models, we use constant proportional anchors: high anchor = baseline × 1.5 (50% above baseline); low anchor = baseline × 0.5 (50% below baseline). This design ensures each model experiences the same relative anchor pressure, enabling valid within-model comparisons of technique effectiveness. Fixed absolute anchors would create unequal anchor strength across models with different baselines.

Table 2: Trial distribution. Total unique trials: 13,799. Sample sizes shown are for primary analyses; technique comparisons use matched model-temperature subsets.

Condition	n (analysis)
Debiasing Techniques	
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
Baseline (no anchor)	909

3.2.6 Statistical Analysis

All comparisons use **Welch's t-test** (unequal variances assumed) with **Bonferroni correction** for multiple comparisons (5 technique comparisons). Effect sizes are reported as Cohen's d. Confidence intervals are 95%. Statistical significance ($p < .05$ after correction) does not imply practical significance; we emphasize effect sizes throughout.

Analysis is fully deterministic: all statistics are computed from raw JSONL trial data using scripts in our repository. No manual intervention or selective reporting.

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 3: Model baselines range from 18.0mo (Opus) to 35.7mo (o4-mini)—a 17.7mo spread. Opus 4.6 shows zero variance (SD=0.0) at all temperatures, consistently responding with exactly 18 months. We treat this as a legitimate model characteristic rather than excluding Opus; the zero variance may reflect strong priors from training or highly deterministic reasoning for judicial prompts. Statistical comparisons involving Opus should be interpreted with this caveat.

4.2 High-Anchor Responses (No Technique)

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

1. **Compression**: Response pulled *below* baseline (Anthropic models, GPT-4.1)

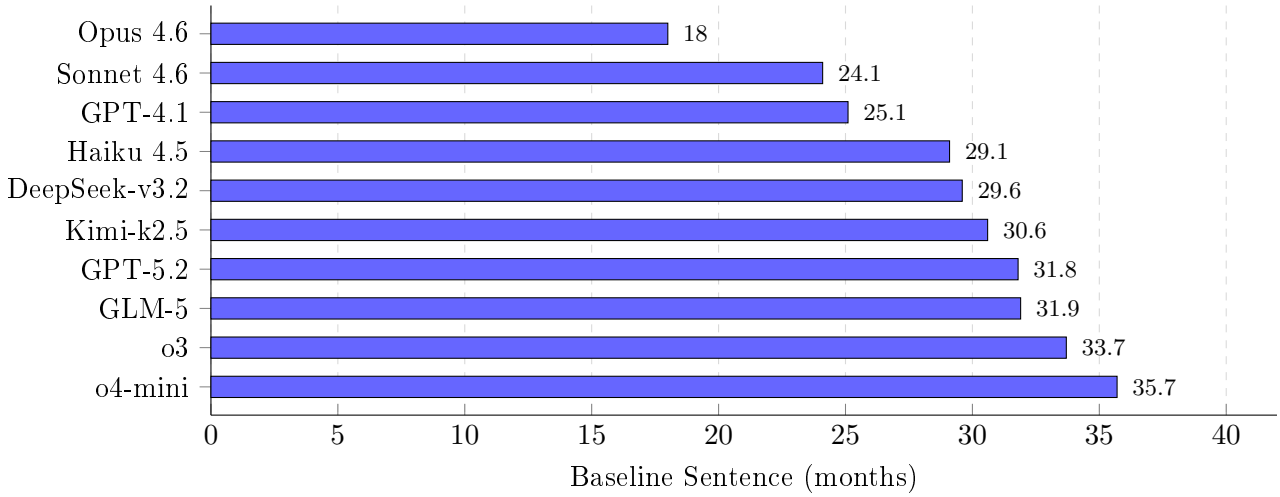

Figure 2: Model baseline variation. Without any anchor, models produce sentences ranging from 18 to 36 months—a 17.7-month spread. This variation motivates per-model anchor calibration.

2. **Inflation**: Response pulled above baseline (GPT-5.2, GLM-5, o3)

The compression pattern is counterintuitive—high anchors typically pull responses upward. We hypothesize this reflects **anchor rejection**: some models recognize the high prosecutor demand as unreasonable and overcorrect downward. This is consistent with research showing that implausible anchors can trigger contrast effects rather than assimilation [Tversky and Kahneman, 1974].

Which models compress? Anthropic models (Opus, Sonnet, Haiku) and GPT-4.1 consistently show compression under high anchors. OpenAI's reasoning models (o3, o4-mini) and GPT-5.2 show the expected inflation pattern. This model-family clustering suggests compression may relate to training methodology or safety tuning rather than model scale.

Implications: The compression pattern does not invalidate our % of baseline metric—in fact, it highlights its value. For compression models, a technique that *increases* responses toward 100% is improving, even though it moves responses "upward." Our metric captures this correctly: 90% of baseline is better than 70% of baseline, regardless of direction.

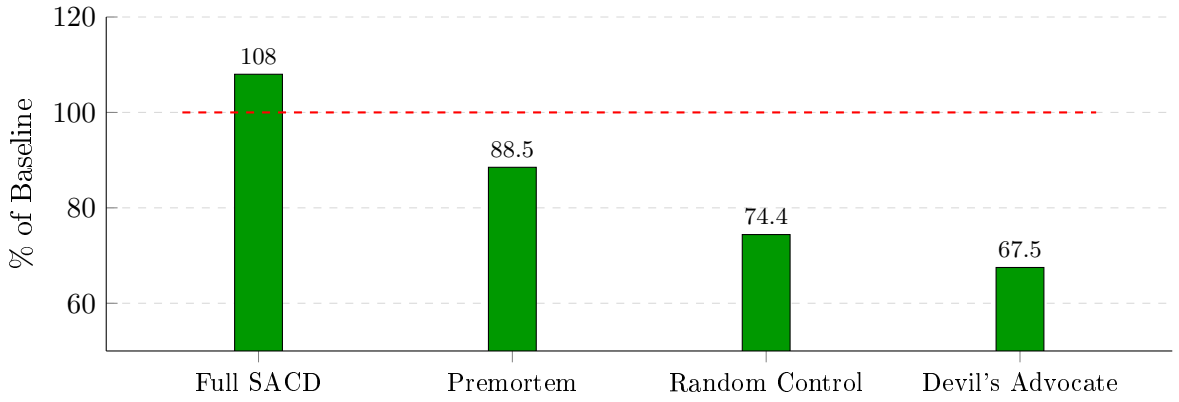

Figure 3: Technique responses as percentage of baseline. The dashed red line marks 100% (perfect match to unanchored judgment). Full SACD (108%) slightly overshoots but is closest to baseline. Devil's Advocate (67.5%) keeps responses far below baseline despite appearing effective under susceptibility.

4.3 Technique Effectiveness: Percentage of Baseline

Technique	n	% of Baseline	95% CI	Deviation	Rank
Full SACD	1,430	**108.0%**	[105.2, 110.8]	**8.0%**	**#1**
Premortem	1,662	88.5%	[85.9, 91.1]	11.5%	#2
Random Control	1,675	74.4%	[71.8, 77.0]	25.6%	#3
Devil's Advocate	1,643	67.5%	[65.0, 70.0]	32.5%	#4
Outside View†	1,862	57.0%	[54.6, 59.4]	43.0%	—

Table 4: Technique effectiveness measured as percentage of baseline. 100% = response matches unanchored judgment. Full SACD is closest to baseline (108%, 95% CI [105, 111]). Devil's Advocate keeps responses at 67.5% of baseline (95% CI [65, 70])—the CIs do not overlap with Full SACD, confirming the ranking difference is statistically reliable. †Outside View confounded. *Note:* Sample sizes differ from Table 2 because this analysis includes only trials with matched baseline data for the same model-temperature condition.

4.4 Model-Specific Results: Full SACD

Full SACD shows high variance across models:

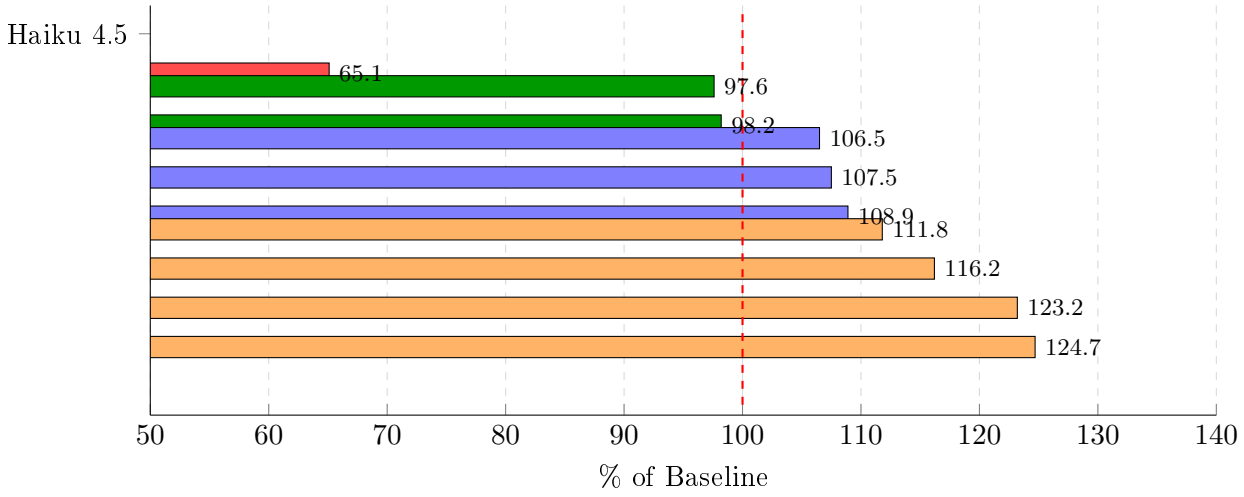

Figure 4: Full SACD by model (percentage of baseline). Dashed line = 100% (perfect). Green = within 5% of baseline. Blue = 5–10% deviation. Orange = >10% overshoot. Red = significant undershoot.

Model	% of Baseline	Deviation	Assessment
o4-mini	**98.2%**	**1.8%**	Near-perfect
Kimi-k2.5	97.6%	2.4%	Near-perfect
DeepSeek-v3.2	106.5%	6.5%	Good
GLM-5	107.5%	7.5%	Good
o3	108.9%	8.9%	Good
Sonnet 4.6	111.8%	11.8%	Moderate overshoot
GPT-5.2	116.2%	16.2%	Overshoot
GPT-4.1	123.2%	23.2%	Significant overshoot
Opus 4.6	124.7%	24.7%	Significant overshoot
Haiku 4.5	**65.1%**	**34.9%**	Undershoot (worst)

Table 5: Full SACD model-specific results (percentage of baseline). o4-mini achieves near-perfect debiasing (98.2%). Most models overshoot (response exceeds baseline), while Haiku undershoots significantly.

Key findings:

1. **o4-mini achieves near-perfect debiasing** (98.2% of baseline)
2. **Most models overshoot** — responses go past baseline (108–125%)
3. **Haiku 4.5 undershoots** — responses stay anchored low (65%)
4. **High variance**: best = 1.8% deviation, worst = 34.9%

4.5 Asymmetry: High vs. Low Anchor

Aggregate results hide an important asymmetry. Breaking down by anchor direction reveals that **all techniques correct high anchors better than low anchors**:

Technique	Low Anchor	High Anchor	Asymmetry
Full SACD	73.7%	141.1%	67.4 pp
Premortem	68.5%	108.6%	40.1 pp
Random Control	59.1%	89.8%	30.8 pp
Devil's Advocate	55.2%	79.7%	24.6 pp

Table 6: Technique effectiveness by anchor direction. All techniques show asymmetry—high anchors are corrected more than low anchors. Full SACD shows the largest asymmetry: it undershoots from low anchors (73.7%) but massively overshoots from high anchors (141.1%).

Key insight: SACD's aggregate 108% masks bidirectional overcorrection. From low anchors, it undershoots (73.7%); from high anchors, it overshoots (141.1%). The average looks good, but the technique amplifies correction in both directions rather than converging on baseline.

Devil's Advocate fails in both directions but stays consistently below baseline (55–80%), explaining its low susceptibility (small spread) despite poor baseline alignment.

4.6 Mixed Effects Analysis

To account for non-independence of observations within models, we fit a mixed effects model with technique as a fixed effect and model as a random effect. The intraclass correlation coefficient (ICC) is 0.17, indicating that **17% of variance** in % of baseline is attributable to model differences.

Fixed effects (technique, controlling for model):

- Full SACD: +24.2% from grand mean (95% CI [105.9, 110.1])
- Premortem: +4.8%
- Random Control: −9.4%
- Devil's Advocate: −16.3%

The ranking is robust after accounting for model-level variance. This supports treating technique effects as generalizable across models, while the ICC justifies our recommendation to test per-model before deployment.

4.7 The Metric Inversion

As shown in Table 1, susceptibility and baseline metrics give inverted rankings. Devil's Advocate produces consistent responses (low susceptibility) that are consistently anchored at the wrong value (67.5% of baseline). SACD brings responses close to baseline (108%) on average despite high variability—but this average masks bidirectional overcorrection (Table 6).

5 Discussion

5.1 Why Full SACD Works (and Fails)

Full SACD achieves 108% of baseline (closest to 100%) but shows the highest model variance (65–125%). 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 overshoots to 124.7% of baseline (24.7% deviation), suggesting 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 (74.4% of baseline) outperforms Devil's Advocate (67.5%) despite having no debiasing content. **This condition serves as a critical ablation:** Full SACD and Premortem are multi-turn techniques, so any improvement could stem from either (a) the debiasing content or (b) the multi-turn structure itself. Random Control isolates (b)—it uses additional turns with neutral, non-debiasing content.

The finding that Random Control achieves 74.4% of baseline while Full SACD reaches 108% suggests both mechanisms contribute: structure provides partial correction, and debiasing content adds further benefit. **Isolating content effects:** SACD's additional 33.6 percentage points toward baseline over Random Control represents the contribution of debiasing content beyond structural effects. Possible mechanisms for the structural effect:

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 vignette.** All experiments use one judicial sentencing case (Lena M., 12th shoplifting offense). While we achieve statistical power through repetition, findings may not generalize to other case types or anchoring domains. Replication across multiple vignettes is needed.

2. **Proportional anchor design.** Our anchors scale with each model's baseline (high = baseline × 1.5, low = baseline × 0.5). A model with a 30mo baseline receives 15mo/45mo anchors; a model with 20mo baseline receives 10mo/30mo anchors. This ensures equal relative anchor

strength, enabling valid within-model comparisons. Future work should validate findings with fixed absolute anchors to test generalization.

3. **Metric inversion holds without Outside View.** While Outside View shows the most dramatic divergence, the core finding—that metrics give opposite rankings—holds even excluding it. Without Outside View: Devil's Advocate ranks *best* on susceptibility (−14% spread) but *worst* on % of baseline (67.5%); Full SACD ranks *worst* on susceptibility (+127% spread) but *best* on % of baseline (108%). The inversion is robust.

4. **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.

5. **Baseline interpretation.** Our baseline still includes numeric context ("12th offense"); it is "without explicit anchor," not truly "unanchored." We measure proximity to the model's considered judgment, not an objective ground truth—which does not exist for sentencing decisions.

6. **Model coverage.** 10 models from 4 providers is substantial but not exhaustive. Results may not apply to other model families.

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; SACD ranges from 65% (Haiku) to 125% (Opus) of baseline.

3. **Collect baselines.** We propose % of baseline as a complementary metric to susceptibility. Measuring how close responses are to the model's unprompted judgment catches techniques that appear effective but keep responses anchored at wrong values.

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 demonstrated that **metric choice determines technique recommendation**. Susceptibility (spread reduction) and % of baseline give inverted rankings:

- **Devil's Advocate**: Best on susceptibility (−14% spread), worst on baseline (67.5%)
- **Full SACD**: Worst on susceptibility (+127% spread), best on baseline (108%)

Devil's Advocate produces consistent responses that are *consistently wrong*—anchored at 67.5% of where they should be. Full SACD's 108% average appears close to correct, but **masks bidirectional overcorrection**: from low anchors it undershoots (73.7%), from high anchors it overshoots (141.1%).

Our key findings from 13,799 trials across 10 models:

1. **Metric inversion**: Devil's Advocate ranks #1 on susceptibility but #4 on % of baseline. SACD shows the opposite pattern.
2. **Bidirectional overcorrection**: SACD's 108% average masks 73.7% (low anchor) and 141.1% (high anchor)—it amplifies correction rather than converging.
3. **High model variance**: SACD ranges from 65% (Haiku) to 125% (Opus). Practitioners must test per-model.
4. **Baseline collection is essential**: Without it, both the "consistently wrong" failure (Devil's Advocate) and bidirectional overcorrection (SACD) are invisible.

For practitioners: test techniques per-model before deployment. Measure responses as % of baseline, not just susceptibility.

For researchers: collect baselines without explicit anchors. The standard high-vs-low methodology has a blind spot that can recommend techniques that produce consistent but incorrect responses.

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 13,799 trials, statistical analysis scripts reproducible from raw data, complete prompts for all debiasing techniques, and response distributions by model and condition.

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.

Yifan Chen et al. Cognitive biases in LLM-assisted software development. *arXiv preprint arXiv:2601.08045*, 2025.

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. 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*. Currency, 2007. ISBN 978-0385502894.

Kahee Lim et al. DeFrame: Debiasing large language models against framing effects. *arXiv preprint arXiv:2602.04306*, 2026.

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.

Peiyang Song, Pengrui Han, and Noah Goodman. Large language model reasoning failures. *arXiv preprint arXiv:2602.06176*, 2026. TMLR 2026 Survey Certification.

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