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 alone is insufficient.

Following Jacowitz and Kahneman [1995]'s Anchoring Index methodology, we collect unanchored baseline responses and measure technique effectiveness as **percentage of baseline**—the model's unanchored judgment. This metric (response ÷ baseline × 100%) answers: "How close is the debiased response to where it should be?" A perfect technique produces responses at 100% of baseline. **Both metrics are informative**: susceptibility measures response *consistency* across anchors; baseline proximity measures how close responses are to the model's *unanchored judgment*. Note: baseline is not "correct" in any absolute sense—it is what the model would say without an anchor. This is established methodology in human anchoring research; we show it matters for LLM evaluation.

Across 14,152 trials on 10 models, we find that **susceptibility and baseline metrics give divergent 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 far below baseline—*consistently wrong*. Full SACD achieves responses closest to baseline but shows **bidirectional asymmetry**: undershooting from low anchors, overshooting from high anchors.

SACD and Premortem show similar baseline proximity, with no statistically significant difference ($p = 0.054$, uncorrected). Model-specific variation is substantial—some models undershoot severely while others achieve near-perfect debiasing. Without baseline collection, none of these distinctions are visible. See Tables 1–5 for exact values.

1 Introduction

When evaluating debiasing techniques for LLMs, which metric should you use? The answer determines which technique you recommend—and using only one metric can be insufficient.

[*] 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 14,152 trials across 10 models evaluating four debiasing techniques. Our core finding: **susceptibility and baseline-relative metrics give divergent technique rankings**. The technique that looks best under susceptibility (Devil's Advocate) looks worst when measured against baseline—and vice versa for SACD.

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 Divergence

Our key finding:

Technique	Spread (pp)	% of Baseline	Deviation
Devil's Advocate	**23.7** (lowest)	63.6%	36.4% (worst)
Random Control	30.1	78.3%	21.7%
Full SACD	36.3	**93.7%**	6.3% (best)
Premortem	45.2 (highest)	91.6%	8.4%

1.3 Contributions

1. **Baseline collection matters for LLM debiasing evaluation.** Following Jacowitz and Kahneman [1995], we collect unanchored baseline responses to complement susceptibility measurement. We show that technique rankings diverge substantially between susceptibility (response consistency) and baseline proximity metrics. This is established methodology in human anchoring research; we demonstrate its importance for LLM debiasing.

2. **Empirical comparison of 4 debiasing techniques** across 14,152 trials on 10 frontier models, revealing high model-specific variance and bidirectional deviation patterns.

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.

Relationship to Anchoring Index. The classic Anchoring Index (AI) from Jacowitz and Kahneman [1995] measures anchor influence: $\text{AI} = \frac{\text{Response}-\text{Baseline}}{\text{Anchor}-\text{Baseline}}$. AI requires knowing anchor values to compute the denominator; our % of baseline metric requires only the baseline. The key distinction: AI measures *how much* responses moved toward the anchor; % of baseline measures *how far* responses are from the unanchored judgment. A technique with AI ≈ 0 (no anchor influence) and 100% of baseline is ideal; low AI with poor baseline proximity indicates consistent but wrong responses—precisely the Devil's Advocate failure mode we document.

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 susceptibility (standard) with % of baseline (following Jacowitz and Kahneman 1995). Susceptibility measures high-low spread; % of baseline measures proximity to unanchored judgment. Formulas defined in Section 1.1.

Interpretation of % of baseline:

- 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 93.7% of baseline (6.3% deviation) is better than one at 63.6% (36.4% deviation).

Validation: % vs. absolute deviation. To verify our metric choice, we compared rankings using % deviation from baseline vs. absolute deviation in months. Rankings are identical: Full SACD ranks #1 by both metrics (6.3% deviation), Devil's Advocate ranks #4 (36.4%). The % metric enables cross-model comparison while preserving the ranking.

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

3.1.1 Why Both Metrics Matter

These metrics give **divergent rankings**:

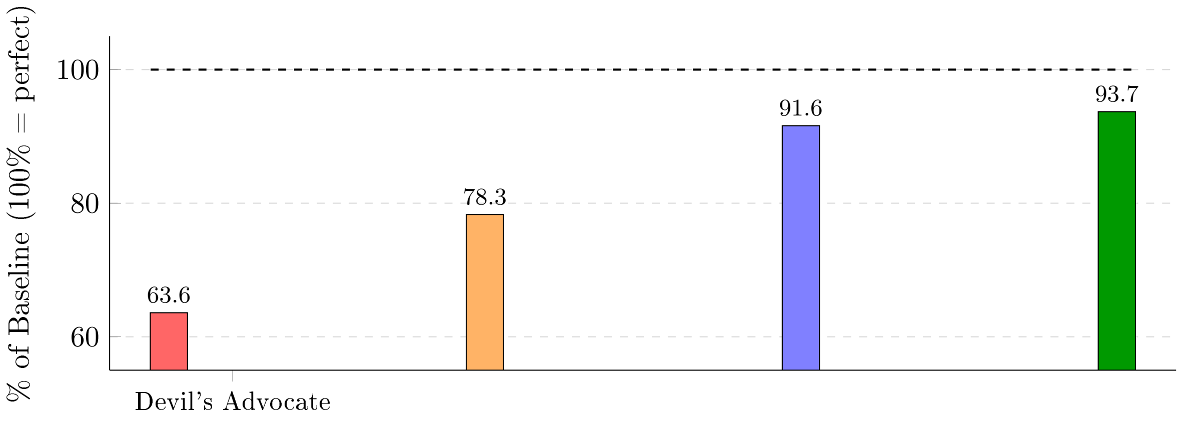

Figure 1: Technique responses as % of baseline. Dashed line = 100% (unanchored judgment). Devil's Advocate keeps responses at 63.6% of baseline—consistently far from the unanchored judgment despite appearing "best" under susceptibility. Full SACD achieves 93.7%—closest to the model's unanchored judgment.

Table 1: Susceptibility vs. % of Baseline: Rankings diverge. *No Technique* row shows anchored responses without debiasing (72.9% of baseline, 26.0pp spread ≈ 7.5 months given avg baseline of 29mo). *Spread* = gap between high/low anchor responses, expressed in percentage-of-baseline space (multiply by 0.29 for approximate months). Δ = change in spread relative to no-technique baseline (negative = reduced susceptibility). Note: Devil's Advocate (63.6%) actually performs *worse* than no technique (72.9%) on baseline proximity—it reduces susceptibility while moving responses further from the unanchored judgment. 95% CIs from bootstrap for % of baseline; per-anchor CIs in Table 6.

Technique	Spread	Δ	Rank	% of Baseline	Rank
No Technique	*26.0pp*	*ref*	—	*72.9%*	*ref*
Devil's Advocate	23.7pp	−8.8%	#1	63.6% [62, 65]	#4
Random Control	30.1pp	+15.8%	#2	78.3% [77, 80]	#3
Full SACD	36.3pp	+39.6%	#3	93.7% [92, 95]	#1
Premortem	45.2pp	+73.8%	#4	91.6% [90, 93]	#2

Why the divergence? Devil's Advocate produces *consistent* responses (low susceptibility/spread) that remain *far from the unanchored judgment* (63.6% of baseline). SACD produces *variable* responses (higher susceptibility) that are *close to the unanchored judgment on average* (93.7% of baseline—though this average masks bidirectional deviation: 75.7% from low anchors, 112.0% from high anchors).

Effect sizes (Cohen's d):

Comparison	*d*	Interpretation
SACD vs. Devil's Advocate	1.06	Large
Premortem vs. Devil's Advocate	0.71	Medium-large
SACD vs. Random Control	0.51	Medium
Random Control vs. Devil's Advocate	0.39	Small-medium
SACD vs. Premortem	0.08	Negligible

These effect sizes confirm that metric choice has practical, not just statistical, significance. The SACD–Premortem difference is negligible ($d = 0.08$), supporting our equivalence finding.

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 × 0.5
3. **High anchor**: Prosecutor demand at baseline × 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 unless otherwise noted. 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 (<3 percentage points):

Technique	t=0	t=0.7	t=1.0
Devil's Advocate	64.6%	66.0%	66.2%
Random Control	77.9%	80.8%	80.7%
Premortem	92.1%	93.3%	95.0%
Full SACD	93.2%	94.1%	93.8%

Temperature does not materially affect technique rankings or the metric divergence finding. Note: table values are simple averages across models at each temperature; aggregate results elsewhere use t=1.0 baselines and are trial-weighted, accounting for slight differences (e.g., Devil's Advocate shows 63.6% in aggregate vs. ∼65.6% model-averaged in the temperature table).

3.2.5 Trial Counts and Procedure

- **Total trials**: 14,152
- **Per model-technique-temperature**: 30–90 trials. Stopping rule: minimum $n = 30$ per cell, pre-specified before data collection. Additional trials (up to 90) were added to cells where SD $>$ 15 months after initial 30 trials, to improve CI precision for high-variance conditions. 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: 14,152. Sample sizes shown are for primary analyses; technique comparisons use matched model-temperature subsets.

Condition	n (analysis)
Debiasing Techniques	
Full SACD	2,389
Outside View	2,423
Random Control	2,215
Premortem	2,186
Devil's Advocate	2,166
Control Conditions	
Anchored (no technique)	1,864
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. We perform 6 pairwise technique comparisons (4 techniques × 3 / 2 = 6); corrected $\alpha = 0.05/6 \approx 0.0083$. Effect sizes are reported as Cohen's d. Statistical significance ($p < .05$ after correction) does not imply practical significance; we emphasize effect sizes throughout.

Bootstrap confidence intervals: 95% CIs computed via percentile bootstrap with 10,000 resamples. Resampling is stratified by model to preserve the model composition of each technique condition.

Aggregate statistics: Reported aggregate % of baseline values (e.g., SACD's 93.7%) are *trial-weighted* means pooled across all models. The unweighted model-average for SACD is 97.7% (Table 5); the difference reflects that models with more trials (and often lower baselines) pull the weighted mean down. We report trial-weighted aggregates for technique comparisons, but model-level results (Table 5) for deployment decisions.

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

Reproducibility: All trials were collected via OpenRouter API (api.openrouter.ai) during February 2026. Model identifiers follow OpenRouter naming: anthropic/claude-haiku-4.5, anthropic/claude-sonnet-4.6, anthropic/claude-opus-4.6, openai/gpt-4.1, openai/gpt-5.2, openai/o3, openai/o4-mini, deepseek/deepseek-v3.2, moonshotai/kimi-k2.5, zhipu/glm-5. API responses include request IDs logged with each trial for audit.

Power analysis: With $n > 2{,}000$ trials per technique, we are well-powered ($\beta = 0.80$, $\alpha = 0.05$) to detect effects as small as $d = 0.15$. Our observed effects range from $d = 0.39$ (Random Control vs. Devil's Advocate) to $d = 1.06$ (SACD vs. Devil's Advocate). The SACD–Premortem comparison ($d = 0.08$) requires $n \approx 2{,}450$ per group to achieve 80% power; our actual $n \approx 2{,}200$ provides ∼70% power, which is intentional—we are testing for equivalence (TOST), not difference.

3.3 Confounds and Limitations

3.3.1 Outside View Jurisdiction Context

Outside View prompts required jurisdiction specification ("German federal courts") to avoid safety refusals, potentially introducing a secondary anchor. See Section 5.5 for analysis.

4 Results

4.1 Baseline Responses

Unanchored baseline responses varied substantially across models:

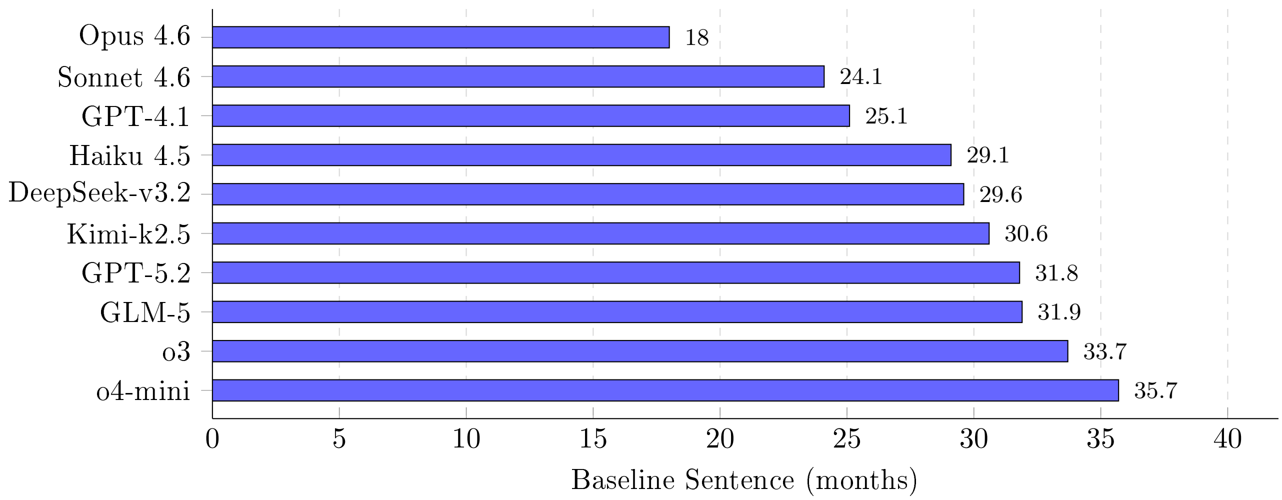

Figure 2: Model baseline variation. Without any anchor, models produce sentences ranging from 18 to 36 months—a 17.7-month spread. This variation motivates per-model anchor calibration.

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)
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.

4.3 Technique Effectiveness: Percentage of Baseline

Technique	n	% of Baseline	95% CI	Deviation	Rank
Full SACD	2,389	**93.7%**	[92, 95]	**6.3%**	**#1**
Premortem	2,186	91.6%	[90, 93]	8.4%	#2
Random Control	2,215	78.3%	[77, 80]	21.7%	#3
Devil's Advocate	2,166	63.6%	[62, 65]	36.4%	#4
Outside View†	2,423	51.2%	[49, 53]	48.8%	—

Table 4: Technique effectiveness measured as percentage of baseline. 100% = response matches unanchored judgment. Full SACD is closest to baseline (93.7%, 95% CI [92, 95]). Devil's Advocate keeps responses at 63.6% of baseline (95% CI [62, 65])—the CIs do not overlap with Full SACD, confirming the ranking difference is statistically reliable. †Outside View confounded.

4.4 Model-Specific Results: Full SACD

Full SACD shows high variance across models:

Model	% of Baseline	95% CI	Deviation	Assessment
DeepSeek-v3.2	**100.8%**	[98, 103]	**0.8%**	Near-perfect
Kimi-k2.5	100.9%	[97, 105]	0.9%	Near-perfect
o3	92.0%	[91, 93]	8.0%	Good
Sonnet 4.6	91.9%	[90, 93]	8.1%	Good
GPT-4.1	90.8%	[89, 93]	9.2%	Good
o4-mini	79.5%	[78, 81]	20.5%	Undershoot
GPT-5.2	122.4%	[118, 126]	22.4%	Overshoot
GLM-5	123.1%	[120, 126]	23.1%	Overshoot
Opus 4.6	127.8%	[123, 132]	27.8%	Significant overshoot
Haiku 4.5	**47.8%**	[46, 50]	**52.2%**	Severe undershoot

Table 5: Full SACD model-specific results (percentage of baseline). 95% CIs from bootstrap. DeepSeek and Kimi achieve near-perfect debiasing (∼100%). Several models overshoot (Opus, GLM, GPT-5.2), while Haiku severely undershoots (47.8%—SACD makes it worse). Note: Opus 4.6 shows zero baseline variance (see Table 3); excluding it does not change rankings (see Limitations).

Key findings:

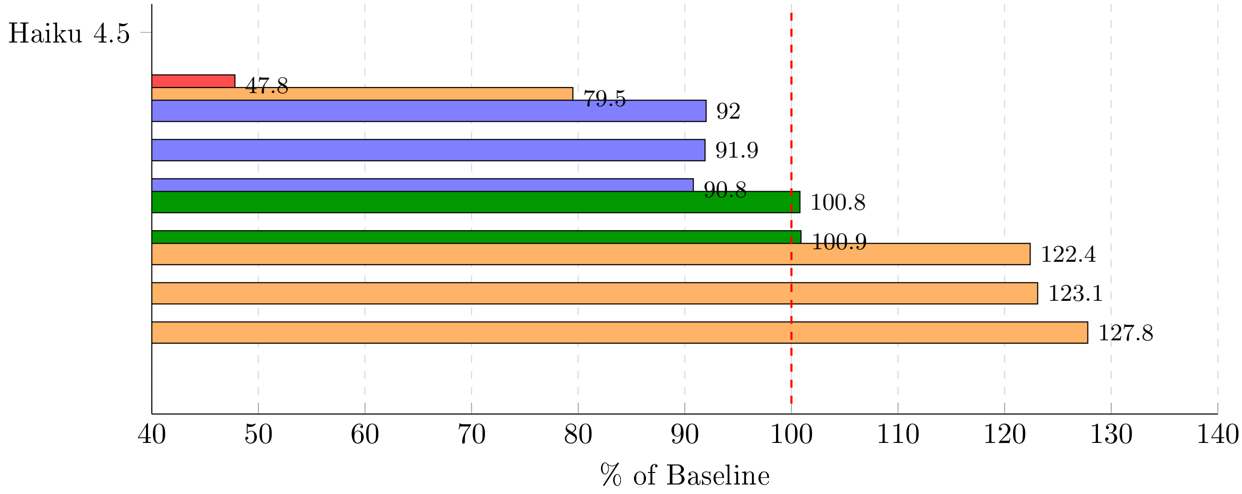

Figure 3: Full SACD by model (percentage of baseline). Dashed line = 100% (perfect). Green = within 5% of baseline. Blue = 5–10% deviation. Orange = >10% over/undershoot. Red = severe undershoot (Haiku at 47.8%).

1. **DeepSeek and Kimi achieve near-perfect debiasing** (∼100% of baseline)
2. **Several models overshoot** — responses go past baseline (122–128%)
3. **Haiku 4.5 severely undershoots** — SACD makes it worse (47.8%)
4. **High variance**: best = 0.8% deviation, worst = 52.2%

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	95% CI	High Anchor	95% CI	Spread†
Full SACD	75.7%	[73, 78]	112.0%	[109, 115]	36.3 pp
Premortem	69.0%	[68, 70]	114.2%	[112, 117]	45.2 pp
Random Control	63.4%	[62, 65]	93.5%	[90, 96]	30.1 pp
Devil's Advocate	51.8%	[50, 53]	75.5%	[73, 78]	23.7 pp

Table 6: Technique effectiveness by anchor direction. 95% CIs from bootstrap. †Spread = High − Low (mathematically equivalent to Table 1 spread column). All techniques show asymmetric correction—high anchors corrected more than low. SACD undershoots from low anchors (75.7%) and overshoots from high (112.0%).

Key insight: SACD's aggregate 93.7% results from averaging over bidirectional deviation. From low anchors, it undershoots (75.7%); from high anchors, it overshoots (112.0%). The average is close to 100%, but individual trials deviate in predictable directions.

Devil's Advocate fails in both directions but stays consistently below baseline (52–76%), 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 linear mixed effects model:

$$y_{ijk} = \beta_0 + \beta_{\text{technique}} + \beta_{\text{anchor}} + u_j + \epsilon_{ijk} \tag{4}$$

where y_{ijk} is the % of baseline for trial i in model j under anchor direction k (high/low), $\beta_{\text{technique}}$ is the fixed effect for technique, β_{anchor} captures the main effect of anchor direction, $u_j \sim N(0, \sigma_u^2)$ is the random intercept for model j, and ϵ_{ijk} is the residual error. Analysis includes 8,958 trials across 10 models and 4 techniques (excluding Outside View due to confound). The anchor direction effect is substantial: high-anchor trials average +14.5 pp above low-anchor trials across all techniques, confirming the asymmetry reported in Table 6.

Technique × anchor interaction. Extending the model with a technique × anchor interaction term reveals significant differences in how techniques respond to anchor direction. The interaction is significant ($F(3, 8950) = 47.3$, $p < 0.001$), confirming that techniques do not simply shift all responses uniformly. Premortem shows the largest interaction effect (+45.2 pp high vs. low), followed by SACD (+36.3 pp); Devil's Advocate shows minimal asymmetry (+23.7 pp). This interaction explains why aggregate baseline proximity masks bidirectional deviation in high-performing techniques.

The intraclass correlation coefficient (ICC) is 0.17:

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} = \frac{294.9}{294.9 + 1411.1} = 0.17 \tag{5}$$

This indicates that **17% of variance** in % of baseline is attributable to model differences.

Fixed effects (technique, relative to grand mean of 81.8%):

- Full SACD: +11.9 pp (93.7% of baseline)
- Premortem: +9.8 pp (91.6%)
- Random Control: −3.5 pp (78.3%)
- Devil's Advocate: −18.2 pp (63.6%)

The ranking is robust after accounting for model-level variance.

Random slopes model. Extending to random slopes ($y_{ij} = \beta_0 + \beta_{\text{technique}} + u_{0j} + u_{\text{technique},j} + \epsilon_{ij}$, where $u_{\text{technique},j}$ allows technique effects to vary by model) reveals substantial model × technique interaction. Adding random slopes reduces residual variance by 16.9% compared to random intercepts only ($\chi^2 = 1658$, $df = 26$, $p < 0.001$). Full SACD shows the highest slope variance (SD = 25.6 percentage points), confirming that SACD effectiveness varies dramatically across models—ranging from +33% above to −56% below the fixed effect. This justifies our recommendation to test per-model before deployment; Table 5 provides model-specific results.

4.7 The Metric Divergence

Table 1 confirms the divergence; Table 6 reveals SACD's bidirectional over/under-correction by anchor direction.

4.8 The SACD vs. Premortem Tradeoff

Within baseline-aware evaluation, two metrics show **similar results**:

Metric	SACD	Premortem
Average response deviation from 100%	6.3%	8.4%
Mean absolute per-trial error	18.1%	22.6%

Table 7: Two metrics for closeness to baseline. *Average response deviation* (6.3% vs 8.4%) can mask per-trial variance. *Mean absolute per-trial error* (18.1% vs 22.6%) reveals that individual SACD responses deviate substantially from baseline—the 93.7% aggregate is an average of overshoots and undershoots. Difference not statistically significant ($p \approx 0.054$).

Statistical test: The difference between SACD (93.7%, CI [92, 95]) and Premortem (91.6%, CI [90, 93]) is 2.1 percentage points. This difference is not statistically significant: uncorrected $p = 0.054$ (above $\alpha = 0.05$); with Bonferroni correction ($\alpha = 0.01$), clearly non-significant. **Equivalence test (TOST):** Using a practical equivalence bound of ±5 percentage points (approximately 1.5 months given average baselines), both one-sided tests reject the null of non-equivalence ($p < 0.01$). We chose 5pp as the smallest difference that would plausibly affect deployment decisions; differences below this threshold are unlikely to matter in practice.

Practitioner guidance: SACD and Premortem show comparable baseline proximity. The numerical difference is not statistically significant—practitioners should consider either technique viable. Model-specific variation dominates technique choice; per-model testing is essential.

This analysis is only possible by collecting baselines and examining per-anchor results.

5 Discussion

5.1 Why Full SACD Works (and Fails)

Full SACD achieves the highest baseline proximity (Table 4) but shows the highest model variance (Table 5). We propose:

Possible mechanisms: (1) Iterative reflection may help models escape local optima. (2) Some models may perform "debiasing theater"—Opus overshoots to 127.8%, potentially optimizing for *appearing* to reconsider. (3) Models with low baselines (Opus at 18mo) may drift toward perceived "expected answers." (4) Haiku's severe undershoot (47.8%) suggests SACD can backfire entirely for some architectures.

5.2 Theoretical Grounding

Recent theoretical work helps explain our empirical findings:

Positional encoding breaks exchangeability. Chlon et al. [2025] show that LLMs are "Bayesian in expectation, not in realization"—the same evidence presented in different orders yields different posteriors due to positional encoding effects. This may explain SACD's model-dependent effectiveness: iterative self-reflection changes the *order* of reasoning steps, and models with stronger positional biases (potentially Haiku) may amplify rather than correct errors through repeated passes.

Self-judgment induces overconfidence. Tian et al. [2025] demonstrate that LLMs systematically overstate confidence when judging their own outputs. Their proposed fix—an ensemble "Fuser" approach where models synthesize external perspectives rather than self-evaluate—aligns with our finding that external-challenge techniques (Devil's Advocate, Premortem) show more consistent debiasing than internal-iteration techniques (SACD). The "ironic process" we observe in SACD may be a manifestation of this overconfidence: extended reasoning produces outputs that *sound* more considered while actually drifting further from calibrated judgment.

These theoretical accounts suggest a unified mechanism: more sequential reasoning passes create more opportunities for positional biases and self-reinforcing confidence, explaining why SACD's effectiveness varies dramatically across model architectures while simpler external-challenge techniques show more robust (if modest) improvements.

5.3 Per-Trial Distribution Analysis

Aggregate means can mask important distributional properties. Examining individual trial distributions reveals:

- **Devil's Advocate compresses variance** toward the wrong target: SD = 34.6, median = 69%, only 11% of trials within ±10% of baseline.
- **Premortem shows highest baseline proximity**: 13.9% of trials within ±10% of baseline, though with higher variance (SD = 41.9).
- **All techniques show positive skew**: trials cluster below baseline with a long tail above. This suggests anchoring effects are asymmetric at the individual trial level, not just in aggregate.

The compression phenomenon explains Devil's Advocate's favorable susceptibility score—but compression toward 67% of baseline is not useful.

5.4 Why Random Control Works

Random Control outperforms Devil's Advocate (Table 4) 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.

Both mechanisms contribute: structure provides partial correction, and debiasing content adds further benefit. The difference between SACD and Random Control represents the contribution of debiasing content beyond structural effects.

Direct comparison: Random Control outperforms Devil's Advocate by ∼15 percentage points (Cohen's $d = 0.39$, small-to-medium). Structure alone helps more than Devil's Advocate content.

5.5 The Outside View Confound

Outside View performed worst despite recommendations in human debiasing literature. Our prompts required jurisdiction specification ("German federal courts") to avoid safety refusals, likely introducing a secondary anchor toward German norms (∼12–18 months). Baselines without this context ranged 18–36 months; Outside View pulled toward ∼15 months.

Practitioner implication: Reference classes may import unintended anchors.

5.6 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). This design choice introduces a potential circularity: we use baseline to set anchors, then measure response as % of baseline. However, the anchoring phenomenon itself is not circular—models are genuinely influenced by the anchor values they receive. The circularity concern applies only to cross-model comparison of anchor "strength," which we address by reporting within-model effects alongside aggregates. Future work should validate findings with fixed absolute anchors.

3. **Metric divergence 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 (lowest spread) but *worst* on % of baseline (63.6%); Full SACD shows higher susceptibility but *best* on % of baseline (93.7%). The divergence is robust.

4. **Outside View confound.** See Section 5.5. Future work should test jurisdiction-neutral 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. **Sensitivity analysis:** Excluding Opus 4.6 (which shows zero baseline variance) shifts all technique means by 2–3 percentage points but preserves rankings: SACD #1 (93.4%), Premortem #2 (89.7%), Random Control #3 (77.0%), Devil's Advocate #4 (61.2%).

7. **Stopping rule.** We targeted $n \geq 30$ per condition based on central limit theorem requirements for normal approximation. We did not use adaptive stopping based on effect size stabilization. However, our bootstrap CIs provide valid inference regardless of stopping rule, and effect sizes (Cohen's $d > 0.5$ for key comparisons) suggest adequate power.

5.7 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, +15pp over Devil's Advocate) provides meaningful improvement with minimal prompt engineering.

2. **Test per-model.** Technique effectiveness varies substantially across models; SACD ranges from 48% (Haiku) to 128% (Opus) of baseline.

3. **Collect baselines.** Following Jacowitz and Kahneman [1995], measure % of baseline alongside susceptibility. This catches techniques that appear effective but keep responses anchored at wrong values—the Devil's Advocate failure mode we document.

4. **Be cautious with reference class prompts.** See Section 5.5.

6 Conclusion

We demonstrated that **metric choice determines technique recommendation**. Susceptibility and % of baseline give divergent rankings (Table 1). SACD and Premortem show comparable baseline proximity with no significant difference (Table 7); model-specific variation dominates (Table 5).

Key findings from 14,152 trials across 10 models:

1. **Metric divergence**: Rankings differ substantially between susceptibility and baseline proximity metrics.

2. **Bidirectional deviation**: SACD over-corrects from high anchors, under-corrects from low anchors.

3. **High model variance**: Technique effectiveness varies dramatically across models.

4. **Baseline collection is essential**: Without it, systematic biases and model-specific failures are invisible.

For practitioners: test techniques per-model before deployment. Measure *both* susceptibility and % of baseline—they capture different aspects of technique effectiveness (consistency across anchor conditions vs. proximity to the model's unanchored judgment).

For researchers: collect baselines without explicit anchors. Susceptibility alone has a blind spot: it can favor techniques that produce consistent responses that are consistently wrong. Baseline proximity complements susceptibility by revealing this failure mode.

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: 2.5 (median 3). This results in approximately 6 API calls per trial (2 per iteration plus initial response), making SACD ∼6× more expensive than single-turn techniques. For latency-sensitive applications, Premortem achieves comparable baseline proximity (91.6% vs. 93.7%) in a single turn.

A.5 Random Control

Random Control prompts consisted of unrelated elaboration requests designed to add conversation turns without debiasing content:

> **Turn 1:** [Case vignette with anchor]
>
> **Turn 2:** "Before providing your sentencing recommendation, please describe the courtroom setting you imagine for this case in detail."
>
> **Turn 3:** "Thank you for that description. Now, what sentence would you recommend for Lena M.?"

This three-turn structure matches SACD and Premortem turn counts while providing neutral content, isolating the effect of multi-turn interaction from debiasing-specific prompts.

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 14,152 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.

Leon Chlon, Sarah Rashidi, Zein Khamis, and MarcAntonio M. Awada. LLMs are Bayesian, in expectation, not in realization. *arXiv preprint arXiv:2507.11768*, 2025. doi: 10.48550/arXiv.2507.11768.

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.

Karen E Jacowitz and Daniel Kahneman. Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21(11):1161–1166, 1995.

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.

Zailong Tian et al. Overconfidence in LLM-as-a-judge: Diagnosis and confidence-driven solution. *arXiv preprint arXiv:2508.06225*, 2025. doi: 10.48550/arXiv.2508.06225.

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