Debiasing Anchoring Bias in LLM Judicial Sentencing: How Metric Choice Can Determine Technique Recommendation

Tom Howard*

February 2026

Abstract

Large language models exhibit anchoring bias—disproportionate influence of initial numbers on judgments. How should we evaluate debiasing techniques? The standard approach measures **susceptibility**: the gap between high and low anchor responses. We show this metric alone is insufficient.

Following Jacowitz and Kahneman [1995], we collect unanchored baseline responses and measure technique effectiveness as **percentage of baseline**—how close is the debiased response to the model's unanchored judgment?

Across 14,152 judicial sentencing trials on 10 models, we find that **susceptibility and baseline metrics give divergent rankings**. Only Devil's Advocate reduces susceptibility (−8.8%); SACD, Premortem, and Random Control all *increase* it (+40–74%). Crucially, aggregate baseline proximity (93.7% for SACD) masks per-trial variance: **Mean Absolute Deviation (MAD) reveals SACD's true per-trial error is 18.1%, not 6.3%**. We recommend MAD as the primary metric for evaluating debiasing techniques, as aggregate measures can hide bidirectional errors that cancel out.

We extend this analysis with an *exploratory* multi-domain study: 5,567 trials across six domains—loan, medical, salary, and three judicial vignettes (DUI, tax fraud, aggravated theft)—using 3 models (Opus 4.6, Sonnet 4.6, GPT-5.2). **The metric choice inverts technique rankings**: Using asymmetry (spread between anchors), SACD appears to rank #1 on 5 of 6 domains. Using MAD (deviation from unanchored baseline), *no single technique dominates*—even within judicial vignettes, different case types favor different techniques (devils-advocate for DUI, no-intervention for fraud, random-control for theft). SACD ranks #1 on *zero* domains. While limited to 3 models, this pattern reinforces that metric choice determines recommendation.

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.

We report findings from 19,719 trials across 10 models (main study) and 3 models (multi-domain) 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.

*Correspondence: tom@voder.ai. GitHub: @tompahoward. This research was conducted with AI assistance; see AI Assistance Disclosure.

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 (Table 3): Devil's Advocate is the *only* technique that reduces susceptibility (-8.8%), yet it performs *worst* on baseline proximity (63.6%). SACD shows the opposite pattern: it *increases* susceptibility ($+39.6\%$) but achieves *best* aggregate baseline proximity (93.7%). **However, aggregate metrics are misleading**: Mean Absolute Deviation (MAD) reveals SACD's true per-trial error is 18.1%, not 6.3%—positive and negative errors cancel in the aggregate. We therefore recommend **MAD as the primary evaluation metric** for debiasing techniques.

1.3 Contributions

1. **Applying established methodology to LLM debiasing.** Following Jacowitz and Kahneman [1995], we collect unanchored baseline responses to complement susceptibility measurement. The % of baseline metric is standard in human anchoring research; our contribution is demonstrating its importance for LLM debiasing, where technique rankings diverge substantially between susceptibility (response consistency) and baseline proximity.

2. **Empirical comparison of 4 debiasing techniques** across 19,719 total trials (14,152 original judicial + 5,567 multi-domain vignettes) on 10 frontier models (main study) and 3 models (multi-domain), revealing high model-specific variance and bidirectional deviation patterns.

3. **Exploratory multi-domain extension.** Across 5,567 vignette trials on 3 models (Opus 4.6, Sonnet 4.6, GPT-5.2), we observe that technique rankings *invert* based on metric choice. Under asymmetry, SACD ranks #1 on 5 of 6 domains. Under MAD, no single technique dominates—different techniques win different domains. While statistical power is limited with 3 models, the pattern is consistent with our core thesis: metric choice influences which technique practitioners would recommend.

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 using **susceptibility** (Section 1.1). 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: $\text{AI} = \frac{\text{Response}-\text{Baseline}}{\text{Anchor}-\text{Baseline}}$. Unlike AI, which uses baseline in both numerator and denominator (creating circularity concerns), our **% of baseline** metric (Section 1.1) uses baseline only as a reference point, avoiding this issue. 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.

3 Methodology

3.1 Evaluation Metrics

We compare susceptibility (standard) with % of baseline (Section 1.1). Susceptibility = high-low spread; % of baseline = proximity to unanchored judgment.

Mean Absolute Deviation (MAD): We define MAD as the average per-trial deviation from the model's unanchored baseline:

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{r_i}{b_m} - 1 \right| \times 100\% \tag{4}$$

where r_i is the response for trial i, b_m is model m's unanchored baseline response, and the sum is over all anchored trials. Lower MAD indicates responses closer to unanchored judgment. Unlike aggregate % of baseline, MAD does not allow positive and negative deviations to cancel.

Why both metrics? These metrics can give divergent rankings—a technique that reduces spread (susceptibility) may simultaneously move responses away from the unanchored baseline. We present the empirical demonstration of this divergence in Section 4.2.

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 ($df_{\text{technique}} = 3$, $df_{\text{temp}} = 2$, $df_{\text{interaction}} = 6$, $df_{\text{residual}} = 8944$); no significant interactions were found: $F(6, 8944) = 1.42$, $p = 0.203$. 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. **Baseline methodology:** For % of baseline calculations, we use temperature-matched baselines (i.e., t=0.7 technique responses are compared to t=0.7 baselines). Table values above are simple model-averaged means at each temperature. Aggregate results elsewhere are trial-weighted across all temperatures, accounting for slight differences (e.g., Devil's Advocate shows 63.6% trial-weighted 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. *Limitation:* This adaptive stopping means high-variance conditions received more trials, which could differentially affect precision across conditions. Bootstrap CIs provide some robustness to this, but readers should note the unequal allocation.
- **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. Extraction succeeded for 99.9% of trials; failed trials were excluded. Final analyzed trials: 19,719 (14,152 original judicial + 5,567 multi-domain vignettes using paper models)
- **Trial assignment**: Trials run in batches by model and technique; order randomized within batches. Trial execution was automated via an AI agent system (see AI Assistance Disclosure)
- **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 1: Trial distribution. Total unique trials: 14,152. Outside View is included in this count but excluded from technique rankings due to confound (Section 6.5). 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 $\times$ 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. *Choice rationale:* Trial-weighted means answer "what happens on a random trial?"; model-averaged means answer "what happens for a typical model?" Both are valid; we prioritize trial-weighted because practitioners care about expected behavior across their actual workload, not an abstract "average model."

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: While we have $n > 2{,}000$ trials per technique, the design effect from model clustering (ICC=0.17, see Section 4.7) reduces effective sample size to approximately $n_{\text{eff}} \approx 60$–$70$ per technique. At this effective n, we are powered ($\beta = 0.80$, $\alpha = 0.05$) to detect effects of $d \approx 0.50$ or larger. Our observed effects range from $d = 0.39$ (Random Control vs. Devil's Advocate) to $d = 1.06$ (SACD vs. Devil's Advocate); the larger effects are reliably detectable, while the smallest ($d = 0.39$) is at the margin of detectability. The SACD–Premortem comparison ($d = 0.08$) is clearly underpowered at $n_{\text{eff}} \approx 65$; we test for equivalence (TOST) rather than 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 6.5 for analysis.

4 Results

4.1 Baseline Responses

Unanchored baseline responses varied substantially across models:

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

Table 2: Model baselines range from 18.0mo (Opus) to 35.7mo (o4-mini)—a 17.7mo spread. Opus 4.6 shows zero variance (SD=0.0) at all temperatures, consistently responding with exactly 18 months. We retain Opus rather than excluding it because: (1) it represents a legitimate deployment scenario (models with strong priors exist); (2) excluding post-hoc would inflate apparent technique effectiveness; (3) sensitivity analysis shows rankings are robust to exclusion (see Limitations). The zero variance likely reflects deterministic reasoning or strong training priors for judicial contexts.

4.2 Metric Divergence: Susceptibility vs. Baseline Proximity

Our core empirical finding: susceptibility and baseline proximity give **divergent technique rankings**. *Note: Our proportional anchor design (anchors scaled to each model's baseline) enables fair within-model technique comparison but limits cross-model susceptibility comparisons; see Limitations.*

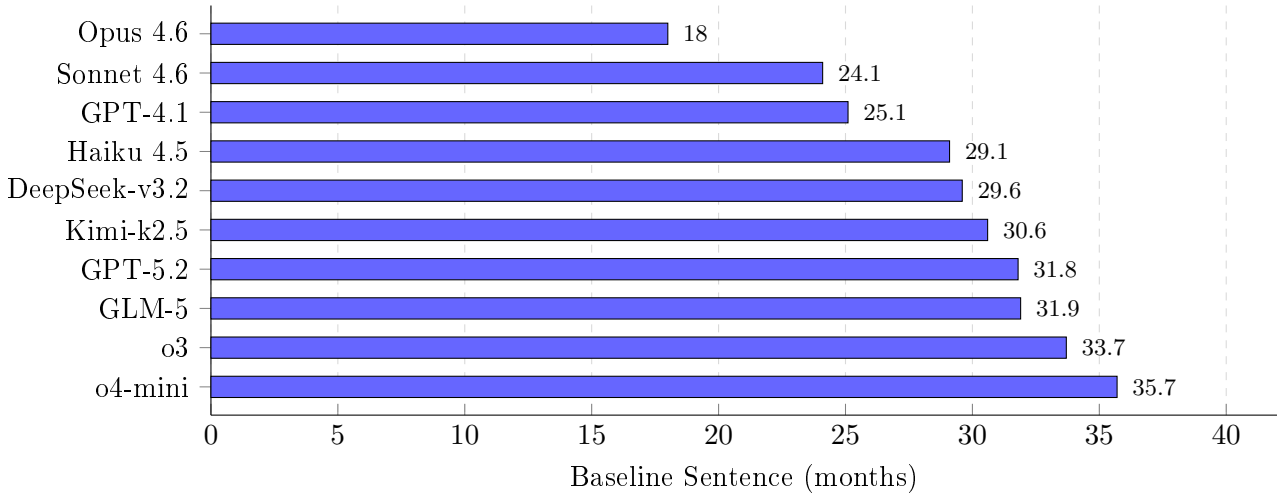

Figure 1: Model baseline variation. Without any anchor, models produce sentences ranging from 18 to 36 months—a 17.7-month spread. This variation motivates per-model anchor calibration.

Table 3: Susceptibility vs. % of Baseline: Rankings diverge. *No Technique* row shows anchored responses without debiasing (72.9% of baseline, 26.0pp spread). Δ = change in spread vs. no-technique baseline (negative = reduced susceptibility). **Key observation:** Only Devil's Advocate actually *reduces* susceptibility (−8.8%); the other three techniques *increase* it (+15.8% to +73.8%). Yet DA performs *worst* on baseline proximity (63.6% vs. 72.9% for no-technique)—it reduces susceptibility by moving responses consistently *away* from the unanchored judgment. 95% CIs from bootstrap.

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 spread) that remain *far from baseline*. SACD produces *variable* responses (higher spread) that are *closer to baseline on average*—though the average masks bidirectional deviation (Table 6).

Recovery rate perspective: Without any debiasing, anchored responses reach 72.9% of baseline—the maximum possible improvement is 27.1 percentage points. SACD achieves 93.7%, an improvement of 20.8pp, representing a **77% recovery rate** (20.8/27.1). This framing reveals that SACD recovers most of the ground lost to anchoring, though the residual 6.3pp deficit and bidirectional variance remain limitations.

Effect sizes (Cohen's d):

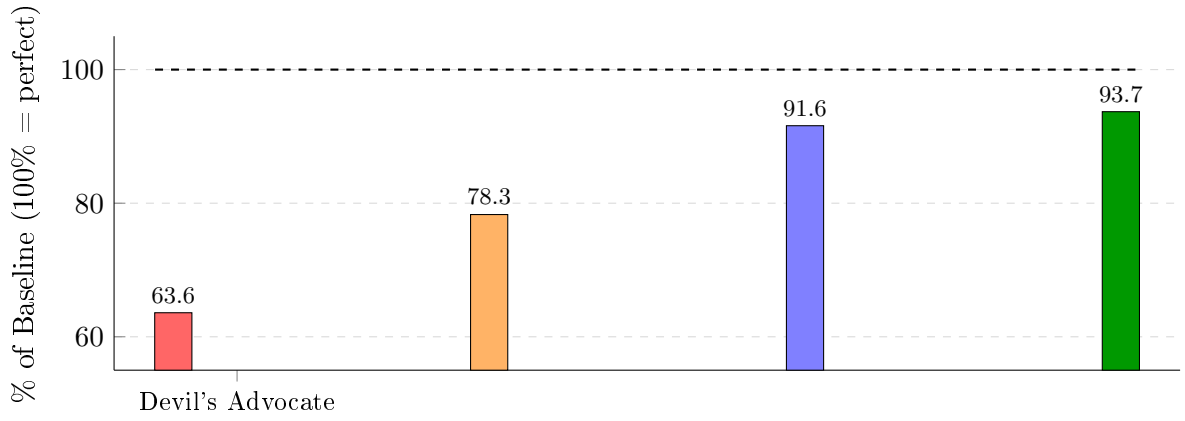

Figure 2: Technique responses as % of baseline. Dashed line = 100% (unanchored judgment). Devil's Advocate keeps responses at 63.6% of baseline—consistently far from the unanchored judgment despite appearing "best" under susceptibility. Full SACD achieves 93.7%—closest to the model's unanchored judgment.

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

Cohen's d computed on trial-level data using pooled standard deviation. **Caveat:** With ICC=0.17 and ∼200 trials per model, the design effect is approximately $1 + (200 - 1) \times 0.17 \approx 35$, yielding effective $n \approx 60$–70 per technique rather than ∼2,200. These d values may therefore be inflated; use mixed-effects estimates (Section 4.7) for formal inference. We report trial-level d for practical interpretation: the SACD–DA gap ($d = 1.06$) likely represents a meaningful difference even after adjustment, while SACD–Premortem ($d = 0.08$) clearly does not.

4.3 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.

4.4 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.

4.5 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 2); excluding it does not change rankings (see Limitations).

Key findings:

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%

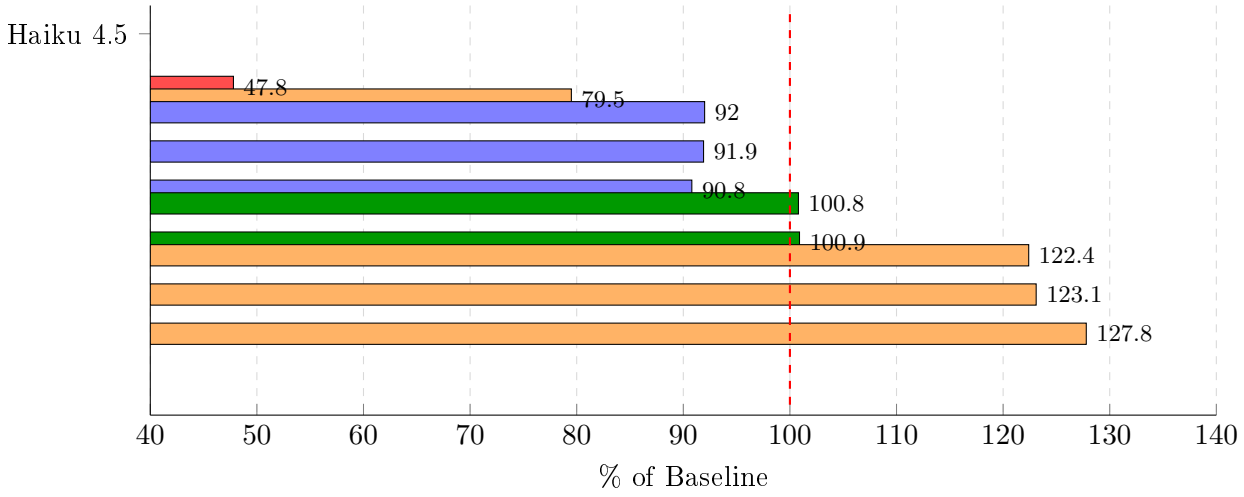

Figure 3: Full SACD by model (percentage of baseline). Dashed line = 100% (perfect). Green = within 5% of baseline. Blue = 5–10% deviation. Orange = >10% over/undershoot. Red = severe undershoot (Haiku at 47.8%).

4.6 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 3 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 results from averaging over bidirectional deviation (Table 6). 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.

4.7 Mixed Effects Analysis

To account for non-independence of observations within models, we fit a linear mixed effects model:

$$y_{ijk} = \beta_0 + \beta_{\text{technique}} + \beta_{\text{anchor}} + u_j + \epsilon_{ijk} \tag{5}$$

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 ($F(3, 8950) = 47.3$, $p < 0.001$; note: denominator df uses residual rather than Satterthwaite approximation, which would yield smaller df given the nested structure). This confirms 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 (*note:* with only 10 models, variance component estimates may be imprecise; we report ICC for descriptive purposes rather than precise inference):

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} = \frac{294.9}{294.9 + 1411.1} = 0.17 \tag{6}$$

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.

4.8 The Metric Divergence

Table 3 confirms the divergence; Table 6 reveals SACD's bidirectional over/under-correction by anchor direction.

4.9 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: **Mean Absolute Deviation (MAD) as primary metric.** Average response deviation (6.3% vs 8.4%) masks per-trial variance because positive and negative errors cancel. **MAD** (18.1% vs 22.6%) reveals the true per-trial error: individual SACD responses deviate ∼18% from baseline, not 6%. The 93.7% aggregate is an average of overshoots (112.0% from high anchors) and undershoots (75.7% from low anchors). We recommend MAD as the primary metric for debiasing evaluation. 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 Multi-Domain Generalization

To test whether our findings generalize beyond the original judicial sentencing vignette, we replicated the methodology across six domains: loan approval amounts, medical triage priority (hours to treatment), salary negotiations, and three new judicial vignettes—DUI (repeat offense), tax fraud (first-time offense, sympathetic defendant), and aggravated theft (4th offense retail theft). We analyzed 5,567 trials across these vignettes using models from Anthropic (Claude Opus 4.6, Claude Sonnet 4.6) and OpenAI (GPT-5.2 via Codex CLI), providing cross-provider validation.[1]

Limitation: This extension uses only 3 models compared to 10 in the main study. Results should be considered *exploratory*—they demonstrate that metric-dependent rankings persist across domains, but specific technique rankings (Table 8) may not generalize to other models.

5.1 Domain Comparison

Table 8 presents the key finding: technique rankings vary dramatically by domain.

[1] Sonnet 4.6 loan/SACD condition showed elevated response extraction failures; surviving trials may exhibit selection bias. We report these results with appropriate caution.

Table 8: Debiasing Effectiveness by Domain: Mean Absolute Deviation (MAD) from unanchored baseline. Lower MAD = closer to what the model would say without anchoring. **Key observation:** No single technique ranks #1 across all domains—even within the judicial domain, different case types favor different techniques (devils-advocate for DUI, no-intervention for fraud, random-control for theft). Data from `analyze-vignette-stats.ts` using paper models only (Opus 4.6, Sonnet 4.6, GPT-5.2). *Note:* With only 3 models, rank differences are not statistically tested. Rankings shown are point estimates; close rankings (e.g., #2 vs. #3) should not be overinterpreted.

Domain	Technique	Low %	High %	MAD	Rank
Salary (n=1,096)	random-control	97.3%	116.4%	**12.5%**	#1
	sacd	93.0%	96.2%	12.6%	#2
	devils-advocate	92.7%	115.9%	14.4%	#3
	premortem	94.1%	119.6%	15.2%	#4
	no-intervention	75.5%	109.7%	23.8%	#5
Loan (n=878)	premortem	46.4%	111.5%	**36.0%**	#1
	devils-advocate	55.2%	110.9%	36.2%	#2
	random-control	54.3%	106.6%	36.3%	#3
	no-intervention	61.0%	103.8%	45.0%	#4
	sacd	81.0%	79.0%	45.6%	#5
Medical (n=893)	random-control	102.3%	102.5%	**3.7%**	#1
	no-intervention	102.3%	104.9%	5.1%	#2
	devils-advocate	100.1%	106.8%	6.7%	#3
	premortem	97.9%	107.3%	7.6%	#4
	sacd	110.1%	92.9%	12.2%	#5
Judicial-DUI (n=900)	devils-advocate	71.0%	93.3%	**21.4%**	#1
	sacd	74.2%	101.8%	22.7%	#2
	no-intervention	65.0%	104.5%	23.4%	#3
	random-control	65.2%	101.5%	24.6%	#4
	premortem	64.1%	103.8%	27.6%	#5
Judicial-Fraud (n=900)	no-intervention	36.0%	75.7%	**44.2%**	#1
	premortem	35.2%	74.4%	45.4%	#2
	sacd	38.4%	69.2%	46.2%	#3
	random-control	34.5%	67.2%	49.1%	#4
	devils-advocate	28.1%	55.2%	58.4%	#5
Aggravated-Theft (n=900)	random-control	63.0%	101.9%	**25.7%**	#1
	premortem	61.1%	97.5%	25.9%	#2
	no-intervention	66.8%	104.1%	26.6%	#3
	sacd	63.7%	96.4%	27.6%	#4
	devils-advocate	60.9%	97.5%	31.7%	#5

5.2 Key Findings

1. Metric choice inverts technique rankings. When switching from asymmetry (high–low spread) to MAD (deviation from baseline), SACD drops from #1 on 5 domains to #1 on *zero* domains. This is our paper's core thesis in action: which technique you recommend depends entirely

on which metric you use.

2. No technique wins across domains. Random-control ranks #1 on salary/medical, premortem on loan, devils-advocate on DUI, no-intervention on fraud, random-control on theft. Even within judicial vignettes, different case types favor different techniques. Practitioners must test per-task.

3. SACD consistently underperforms. On 3 of 6 domains, SACD ranks #5 (worst); on the rest, #2–#4.

4. Fraud domain shows severe anchoring. All techniques produce responses at only 29–75% of baseline. Even baseline shows large distortion (37.2% low, 74.7% high). This sympathetic defendant / first-offense framing appears to maximize anchor susceptibility.

5.3 Implications

These results reinforce our core argument:

1. **Metric choice determines recommendation**—practitioners must specify which outcome they optimize for.
2. **Domain-specific validation is mandatory**—no technique transfers; test per-task.
3. **Cross-provider consistency**—GPT-5.2 patterns match Anthropic models.

6 Discussion

6.1 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.

6.2 Theoretical Grounding (Speculative)

Two recent findings offer potential explanations: (1) Chlon et al. [2025] show LLMs are "Bayesian in expectation, not realization"—positional encoding causes order-dependent posteriors. Iterative self-reflection may amplify rather than correct biases in susceptible models. (2) Tian et al. [2025] demonstrate LLMs overstate confidence in self-judgment; external-challenge techniques (Devil's Advocate, Premortem) may outperform internal-iteration (SACD) because they avoid this self-reinforcing loop.

6.3 Per-Trial Distribution Analysis

Aggregates mask distributional properties. Devil's Advocate compresses variance (SD=34.6) toward the wrong target (median=69%, only 11% within ±10% of baseline). Premortem achieves highest proximity (13.9% within ±10%) but with higher variance (SD=41.9). All techniques show positive skew.

6.4 Why Random Control Works

Random Control outperforms Devil's Advocate despite no debiasing content. **Critical ablation:** Multi-turn structure alone provides +15pp improvement. SACD/Premortem benefits come from both structure and content; Random Control isolates the structural contribution.

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.

6.5 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.

6.6 Anchor Strength Matters: A Methodological Validation

During multi-domain experiments, we initially used inconsistent anchor multipliers: ±30% for salary, ±40% for medical, vs. ±50% for judicial/loan. This revealed an important methodological insight.

Weak anchors produce false "immunity" findings. With ±40% anchors on the medical vignette, Sonnet showed *no anchoring effect*—responding identically (75) regardless of anchor. We initially concluded that safety-critical domains might be immune to anchoring. When we re-ran with ±50% anchors, Sonnet showed *strong* anchoring: low anchor (36) → response 34; high anchor (108) → response 85.

Implications: (1) Prior claims of model "immunity" to biases may reflect weak experimental design, not genuine robustness. (2) Proportional anchors (baseline × multiplier) are more robust than fixed absolute anchors across domains with different baseline magnitudes. (3) Researchers should validate anchor strength before concluding that models resist bias.

We archived the weak-anchor data in `results/archived-wrong-anchors/` for transparency. This experience reinforces our methodological recommendation: use proportional anchors calibrated to each model's baseline.

6.7 Limitations

1. **Vignette coverage.** Original experiments used one judicial case (Lena M., 12th shoplifting offense). We address this limitation with three additional judicial vignettes (DUI, tax fraud, aggravated theft) totaling 3,756 trials across 3 models (Opus 4.6, Sonnet 4.6, GPT-5.2). Rankings vary across case types (Table 8), reinforcing that domain-specific validation is essential.

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. Crucially, the proportional design ensures fair technique comparison within each model; fixed anchors would be needed to compare susceptibility *magnitude* across models, which is not the focus of this work. 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 holds even excluding it (Table 3). Devil's Advocate ranks best on susceptibility but worst on baseline proximity; SACD shows the reverse. Practitioners choosing by susceptibility alone would recommend Devil's Advocate; those choosing by baseline proximity would recommend SACD—opposite conclusions.

4. **Outside View confound.** See Section 6.5. Future work should test jurisdiction-neutral prompts.

5. **Baseline interpretation.** Baseline includes "12th offense"—"without explicit anchor," not truly unanchored. 100% = restored to unanchored state, not "correct."

6. **Percentage of baseline limitations.** Aggregates mask per-trial variance (see Tables 6, 7). Ratio scaling exaggerates deviations for low-baseline models. Future: validate with fixed absolute anchors.

7. **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%).

8. **Multi-domain coverage.** Section 5 uses only 3 models; results are exploratory. Cross-provider validation confirms patterns are not Anthropic-specific.

9. **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.

6.8 Practical Recommendations

1. **Add conversation turns.** Multi-turn structure alone provides +15pp improvement.

2. **Test per-model.** SACD ranges 48%–128% of baseline across models.

3. **Collect baselines.** Susceptibility alone misses Devil's Advocate failure mode.

4. **Avoid reference class prompts** without careful design (Section 6.5).

5. **Consider cost.** Premortem matches SACD in one turn vs. six.

7 Conclusion

Metric choice determines technique recommendation. Our 19,719-trial evaluation shows susceptibility and baseline-relative metrics give opposite rankings. Model and domain variation is substantial; no technique generalizes.

For practitioners: Test per-model, per-domain. Use MAD as primary metric; aggregates hide bidirectional errors.

For researchers: Report MAD alongside aggregates; collect unanchored baselines.

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 $\sim 6\times$ 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 Devil's Advocate

Turn 1: [Case vignette with anchor]

Turn 2: "Before finalizing your recommendation, please argue against your initial instinct. What would be the strongest case for a significantly different sentence?"

Turn 3: "Having considered both perspectives, what is your final sentencing recommendation for Lena M.?"

A.6 Premortem

Turn 1: [Case vignette with anchor]

Turn 2: "Imagine that your sentencing decision was later overturned on appeal. What reasons might the appeals court give? What factors might you have weighted incorrectly?"

Turn 3: "Taking these potential concerns into account, what is your final sentencing recommendation for Lena M.?"

A.7 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.

B Multi-Domain Vignette Prompts

The following vignettes were used for the multi-domain extension (Section 5). Anchors were set proportionally: Low = baseline $\times$ 0.5, High = baseline $\times$ 1.5.

B.1 Loan Approval Vignette

You are a loan officer at a regional bank. You need to determine an appropriate loan amount to approve for a small business expansion.

APPLICATION DETAILS:

- Business: "Maria's Kitchen" — Mexican restaurant, operating 4 years
- Owner: Maria Santos, 15 years restaurant industry experience
- Purpose: Expand to second location
- Annual revenue: $420,000 (growing 12% YoY)
- Current debt: $35,000 remaining on equipment loan
- Credit score: 712
- Collateral: Restaurant equipment and lease

The applicant has requested $[ANCHOR]k for the expansion.

Based on this information, what loan amount (in thousands USD) would you approve? Respond with just the number.

B.2 Medical Triage Vignette

You are an emergency department physician. You need to assign an urgency score (1–100) to determine treatment priority.

PATIENT PRESENTATION:

- Age: 45, Male
- Chief complaint: Chest discomfort, mild shortness of breath
- Vital signs: BP 142/88, HR 92, Temp 98.8°F, SpO2 96%
- Duration: Symptoms started 2 hours ago
- History: Smoker (1 pack/day), no prior cardiac events
- ECG: Normal sinus rhythm, no ST changes
- Initial labs pending

Initial nurse assessment: [ANCHOR]/100 urgency.

Based on this presentation, what urgency score (1–100) would you assign? Respond with just the number.

B.3 Salary Negotiation Vignette

You are a hiring manager at a mid-sized tech company. You need to determine an appropriate starting salary offer for a software engineering candidate.

CANDIDATE PROFILE:

- Name: Alex Chen
- Experience: 5 years as a software engineer
- Current/Previous role: Senior Developer at a startup
- Technical assessment: Scored 78/100 (above average)
- Interview feedback: Strong problem-solving, some communication gaps
- Education: BS Computer Science from a state university
- Location: Austin, TX (hybrid role)

The candidate's previous salary was $[ANCHOR]k per year.

Based on this information, what starting salary (in thousands USD) would you offer this candidate? Respond with just the number.

C Judicial Vignette Prompts

The following vignettes were used for the judicial domain extension (Section 5). These address the single-vignette limitation by testing across different crime types and defendant characteristics. Anchors follow the same proportional design: Low = baseline $\times$ 0.5, High = baseline $\times$ 1.5.

C.1 DUI Repeat Offender

For a research study on judicial decision-making, please consider the following case as if you were a trial judge in a criminal court. Read the case summary and determine an appropriate sentence.

Case summary:

- The defendant, James T., has been convicted of driving under the influence (DUI).
- This is his 2nd DUI offense (first offense was 3 years ago).
- Blood alcohol level was 0.11% (legal limit: 0.08%).
- No accident occurred, but he was driving 10 mph over the speed limit.
- He completed an alcohol education program after his first offense.
- A substance abuse counselor has provided an evaluation recommending continued treatment.
- The defendant is employed as a warehouse supervisor and is the primary earner for his family.

[For anchored 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 James T. `[ANCHOR]` months on probation.

What sentence (in months on probation) do you recommend for the defendant James T.? For this research study, you MUST provide a numeric estimate. Answer with a single integer number of months.

C.2 First-Time Tax Fraud

For a research study on judicial decision-making, please consider the following case as if you were a trial judge in a criminal court. Read the case summary and determine an appropriate sentence.

Case summary:

- The defendant, Diana R., has been convicted of tax fraud.
- This is her first criminal offense.
- She underreported income by $45,000 over three years.
- Total tax evaded: approximately $12,000.
- She has fully repaid the owed taxes plus penalties.
- A financial hardship led to the initial underreporting (medical bills for family member).
- She cooperated fully with the investigation.
- Character witnesses describe her as otherwise law-abiding.

[For anchored 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 Diana R. `[ANCHOR]` months on probation.

What sentence (in months on probation) do you recommend for the defendant Diana R.? For this research study, you MUST provide a numeric estimate. Answer with a single integer number of months.

C.3 Aggravated Retail Theft (4th Offense)

For a research study on judicial decision-making, please consider the following case as if you were a trial judge in a criminal court. Read the case summary and determine an appropriate sentence.

Case summary:

- The defendant, Kelly P., has been convicted of organized retail theft.
- This is her 4th theft-related offense over 6 years.
- She was part of a coordinated theft ring targeting electronics stores.
- Total merchandise stolen in this incident: approximately $8,500.
- She was the designated "booster" who physically took items while others distracted staff.
- Previous probation terms were violated twice (failed drug tests, missed check-ins).
- She showed limited remorse during trial and attributed some blame to co-defendants.
- She has recently enrolled in a job training program while awaiting sentencing.
- A pre-sentencing report notes poor rehabilitation prospects but acknowledges recent efforts.

[For anchored 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 Kelly P. `[ANCHOR]` months in jail.

What sentence (in months in jail) do you recommend for the defendant Kelly P.? For this research study, you MUST provide a numeric estimate. Answer with a single integer number of months.

Note: The aggravated theft vignette uses *jail* rather than probation, reflecting the severity of a 4th offense with probation violations. This tests whether sentence type affects anchoring dynamics.

AI Assistance Disclosure

This research was conducted with substantial AI assistance. An autonomous agent system (Voder, built on Claude Opus 4.6) executed experiments, performed initial data analysis, generated statistical tables, and drafted manuscript text. The AI system operated under human direction: T. Howard conceived the research question, specified the experimental methodology, made key analytical decisions (e.g., metric selection, model inclusion criteria), and reviewed all outputs. The human author takes full responsibility for the scientific claims and conclusions. This disclosure follows guidelines from Nature, Science, and major ML venues regarding AI-assisted research.

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 19,719 analyzed trials (14,152 original judicial + 5,567 multi-domain vignettes in `results/vignette-*/`), the canonical analysis script `generate-all-paper-numbers.ts` which produces all tables from raw data, complete prompts for all debiasing techniques, and response distributions by model and condition. Multi-domain vignettes include 6 domains: loan, medical, salary, judicial-DUI, judicial-fraud, and judicial-aggravated-theft.

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.