Debiasing Anchoring Bias in LLM Judicial Sentencing: How Metric Choice Can Determine 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. We show this metric alone is insufficient.

Following **?**, we collect unanchored baseline responses and measure technique effectiveness as **percentage of baseline**. This metric answers: "How close is the debiased response 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.

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%). Yet Devil's Advocate ranks worst by baseline proximity (63.6%), while SACD ranks best (93.7%). The aggregate masks per-trial variance: SACD's mean absolute error is 18.1%, not 6.3%.

We extend this analysis to 2,046 additional trials across loan, medical, and salary domains using two Anthropic models (Opus 4.6 and Sonnet 4.5). Within this model subset, **technique effectiveness appears domain-dependent**: SACD ranks #1 on judicial and medical but #5 (worst) on loan (57.1% of baseline). On salary, no debiasing technique beats doing nothing. These preliminary multi-domain results suggest practitioners should validate per-domain, though broader model coverage is needed to confirm generalizability.

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 16,198 trials across 10 models and 4 domains 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.

---

*Voder AI is an autonomous AI agent built on Claude. Correspondence: voder.ai.agent@gmail.com

†Tom Howard provided direction and oversight. GitHub: @tompahoward

1.1 Two Metrics, Opposite Conclusions

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

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

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

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

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

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

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

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

1.2 The Divergence

Our key finding (Table 1): 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, this aggregate masks bidirectional deviation: mean absolute per-trial error is 18.1%, not 6.3%. Choosing by susceptibility alone would recommend Devil's Advocate; choosing by baseline proximity would recommend SACD—opposite conclusions from the same data.

1.3 Contributions

1. **Applying established methodology to LLM debiasing.** Following **?**, 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 16,198 total trials (14,152 judicial + 2,046 multi-domain) on 10 frontier models, revealing high model-specific variance and bidirectional deviation patterns.

3. **Exploratory multi-domain extension (underpowered).** In a two-model pilot, SACD ranking varies by domain (judicial #1, loan #5). This motivates domain-specific validation but should not be interpreted as a robust finding—broader model coverage is essential before drawing conclusions about domain dependence.

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 [?]. Even experts are susceptible: **?** 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 [???]. Anchoring effects have been documented across multiple model families [?], with susceptibility varying by model architecture and size. **?** survey LLM reasoning failures comprehensively, including susceptibility to anchoring and framing effects. Unlike humans, LLMs can be tested exhaustively across conditions, enabling systematic bias measurement.

2.3 Debiasing Techniques

Several techniques have been proposed for mitigating anchoring:

Outside View / Reference Class Forecasting: Prompting models to consider what typically happens in similar cases [?]. Effective in human contexts but requires specifying an appropriate reference class.

Self-Administered Cognitive Debiasing (SACD): Iterative prompting that guides models through bias detection and correction [?]. 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 [?].

Recent work has also explored debiasing against framing effects [?], 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 [??]:

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

A technique "works" if it reduces this gap.

Relationship to Anchoring Index. The classic Anchoring Index (AI) from **?** 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 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 $\text{AI} \approx 0$ (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 (Section 1.1). Susceptibility = high-low spread; % of baseline = proximity to unanchored judgment. Deviation from baseline = $|(\% \text{ of Baseline}) - 100\%|$; lower = better.

3.1.1 Why Both Metrics Matter

These metrics give **divergent rankings**:

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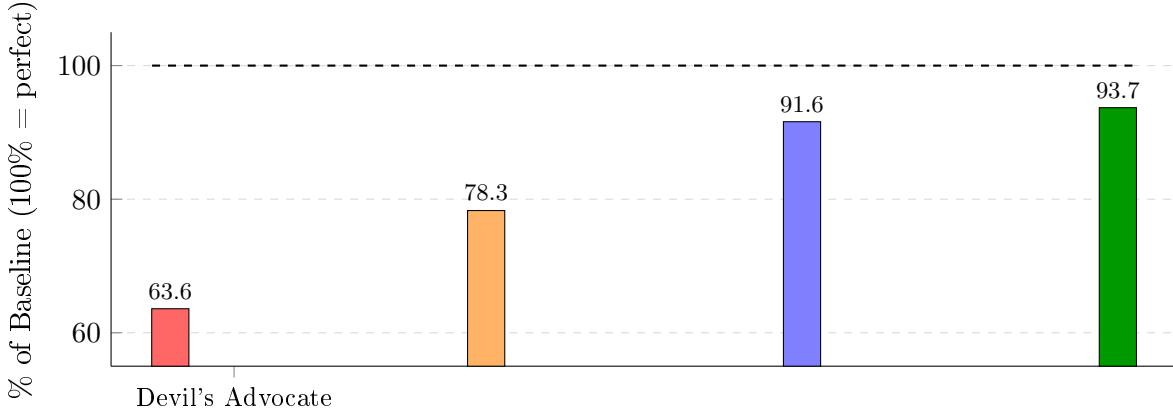
Figure 1: Technique responses as % of baseline. Dashed line = 100% (unanchored judgment). Devil's Advocate keeps responses at 63.6% of baseline—consistently far from the unanchored judgment despite appearing "best" under susceptibility. Full SACD achieves 93.7%—closest to the model's unanchored judgment.

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, effective sample sizes are smaller than raw trial counts; these d values may be inflated by ignoring model clustering. Use mixed-effects estimates (Section **??**) for formal inference. These effect sizes are provided for practical interpretation: the SACD–DA gap ($d = 1.06$) represents a meaningful practical difference, while SACD–Premortem ($d = 0.08$) does not.

3.2 Experimental Design

3.2.1 Models

We evaluated 10 models across 4 providers:

Provider	Models
Anthropic	Claude Haiku 4.5, Sonnet 4.6, Opus 4.6
OpenAI	GPT-4.1, GPT-5.2, o3, o4-mini
DeepSeek	DeepSeek-v3.2
Others	Kimi-k2.5 (Moonshot), GLM-5 (Zhipu)

3.2.2 Conditions

1. **Baseline**: Sentencing prompt with no anchor

2. **Low anchor**: Prosecutor demand at baseline $\times$ 0.5

3. **High anchor**: Prosecutor demand at baseline $\times$ 1.5

4. **Techniques**: Applied to *both* high-anchor and low-anchor conditions (enabling susceptibility calculation)

3.2.3 Techniques Evaluated

Technique	Description
Outside View	"What typically happens in similar cases?" (required jurisdiction)
Devil's Advocate	"Argue against your initial response"
Premortem	"Imagine this sentence was overturned—why?"
Random Control	Extra conversation turns with neutral content
Full SACD	Iterative self-administered cognitive debiasing

3.2.4 Temperature Conditions

Each technique was tested at three temperatures: t=0 (deterministic), t=0.7 (moderate variance), and t=1.0 (high variance). Baseline responses were collected at all three temperatures. Results are aggregated across temperatures unless otherwise noted. We tested for temperature$\times$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 (19 failures out of 20,339 total API responses including multi-turn iterations); failed trials were excluded. Final analyzed trials: 16,198 (14,152 judicial + 2,046 multi-domain)

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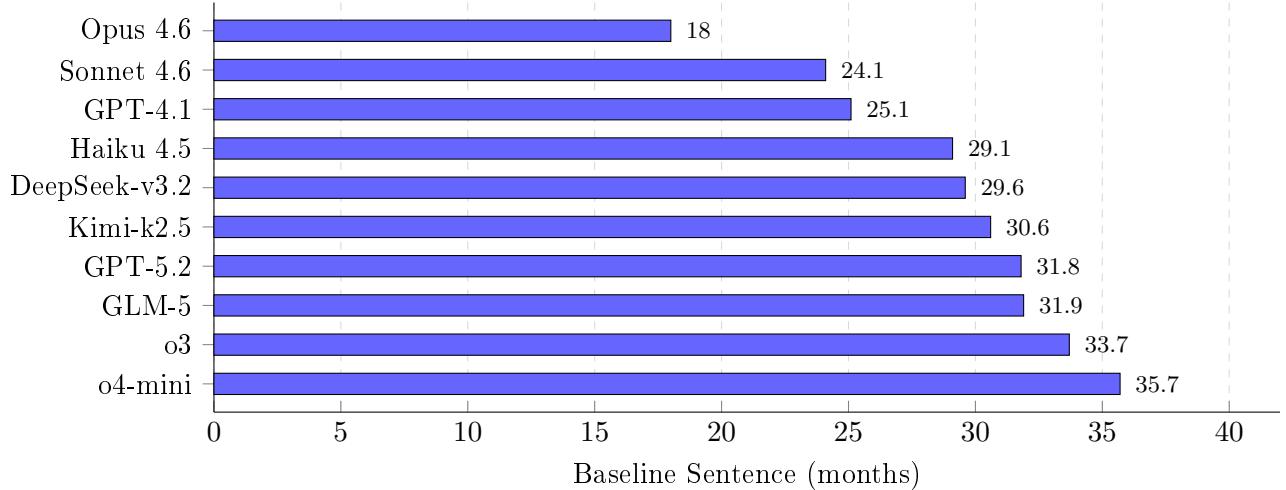
Figure 2: Model baseline variation. Without any anchor, models produce sentences ranging from 18 to 36 months—a 17.7-month spread. This variation motivates per-model anchor calibration.

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 [**?**].

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 ($\sim$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:

1. **DeepSeek and Kimi achieve near-perfect debiasing** ($\sim$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**:
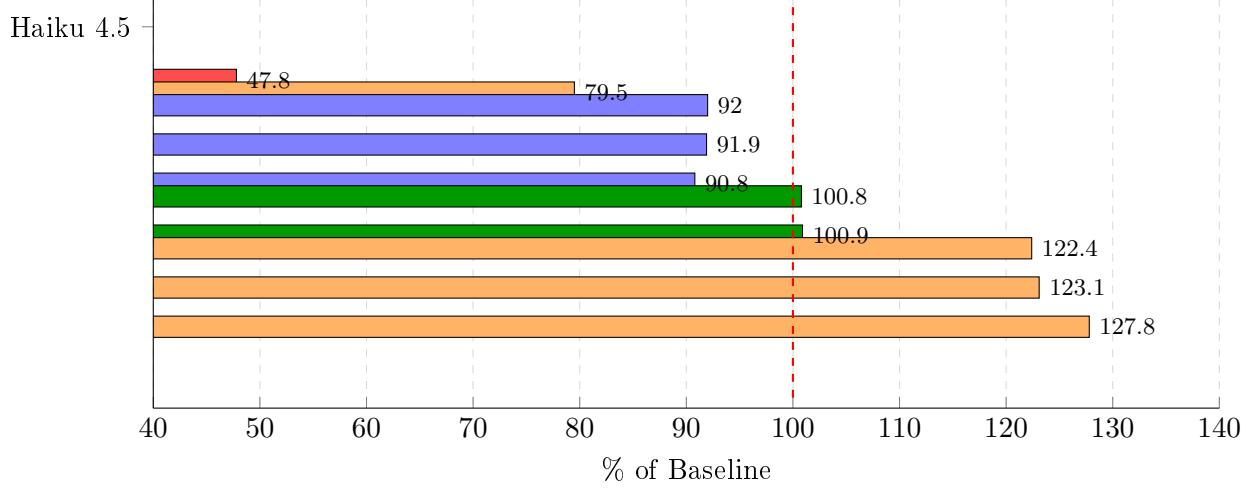
Figure 3: Full SACD by model (percentage of baseline). Dashed line = 100% (perfect). Green = within 5% of baseline. Blue = 5–10% deviation. Orange = >10% over/undershoot. Red = severe undershoot (Haiku at 47.8%).

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 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.

4.6 Mixed Effects Analysis

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

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

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.

11

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 (*note:* with only 10 models, variance component estimates may be imprecise; we report ICC for descriptive purposes rather than precise inference):

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

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 (112.0% from high anchors) and undershoots (75.7% from low anchors). On any individual trial, SACD deviates ∼18% from baseline, not 6%. 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 judicial sentencing, we replicated the methodology across three additional domains: loan approval amounts, medical triage priority (hours to treatment), and salary negotiations. We collected 2,046 trials across these vignettes using two models (Claude Opus 4.6 and Claude Sonnet 4.5).[1]

5.1   Domain Comparison

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

---

[1]The judicial experiments used Sonnet 4.6; multi-domain used 4.5 due to API availability at experiment time. Version differences are a limitation.

Table 8: Debiasing Effectiveness by Domain: % of Baseline Metric. Best technique differs across domains. SACD ranks #1 on Judicial and Medical but #5 (worst) on Loan. On Salary, no debiasing technique beats baseline. Deviation = |% of Baseline −100%|; lower = better. **Note:** Judicial uses 10 models (14,152 trials); Loan/Medical/Salary use 2 models (Opus 4.6, Sonnet 4.5; ∼660 trials each). Cross-domain patterns are preliminary.

Domain	Technique	% of Baseline	Deviation	Rank
Loan	no-technique	68.3%	31.7%	4
	devils-advocate	82.0%	**18.0%**	**1**
	premortem	75.9%	24.1%	2
	random-control	73.6%	26.4%	3
	sacd	57.1%	42.9%	5 (worst)
Medical	no-technique	102.7%	2.7%	2
	devils-advocate	106.9%	6.9%	3
	premortem	110.1%	10.1%	4
	random-control	111.6%	11.6%	5
	sacd	98.2%	**1.8%**	**1**
Salary	no-technique	106.6%	**6.6%**	**1**
	devils-advocate	107.0%	7.0%	2
	premortem	115.8%	15.8%	5
	random-control	108.6%	8.6%	3
	sacd	110.2%	10.2%	4
Judicial	no-technique	72.9%	27.1%	4
	devils-advocate	63.6%	36.4%	5 (worst)
	premortem	91.6%	8.4%	2
	random-control	78.3%	21.7%	3
	sacd	93.7%	**6.3%**	**1**

5.2 Key Findings

1. Technique effectiveness appears domain-dependent (preliminary). Within our two-model subset, SACD ranks #1 by % of baseline on Judicial (6.3% deviation) and Medical (1.8% deviation) but #5 (worst) on Loan (42.9% deviation)—a complete inversion. These results suggest no technique is universally "best," though validation with additional models is needed.

2. SACD shows complete ranking inversion. On Judicial and Medical, SACD achieves best baseline proximity. On Loan, it achieves worst—42.9% deviation vs. devils-advocate's 18.0%. A practitioner who validated only on judicial data would incorrectly recommend SACD for loan decisions.

3. On Salary, no technique beats doing nothing. Baseline achieves 6.6% deviation; all debiasing techniques perform worse (7.0%–15.8%). Debiasing interventions can introduce rather than remove bias. This is a cautionary finding: well-intentioned debiasing can backfire.

4. Domain susceptibility varies dramatically. Medical shows strong resistance to anchoring (all techniques within 1.8%–11.6% of baseline). Loan shows high susceptibility (best technique still 18.0% from baseline). Practitioners must calibrate expectations per domain.

5.3 Implications

These results strengthen rather than weaken our core argument:

1. **Domain-specific validation is essential**—SACD ranks #1 on Judicial/Medical but #5 on Loan; practitioners cannot assume techniques transfer across domains

2. **Debiasing can backfire**—on Salary, all techniques increase deviation from baseline; intervening is worse than doing nothing

3. **The counterfactual matters**—validating only on judicial data would lead to recommending SACD for loans, where it performs worst

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)

Recent theoretical work offers *potential* explanations for our empirical findings. The connections below are speculative; we present them to motivate future research rather than as established mechanisms:

   Positional encoding breaks exchangeability. ? 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. ? 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.

6.3 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.

6.4 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.

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. **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 $\times$ 1.5, low = baseline $\times$ 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 1). 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.** 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. The baseline is a reference point, not a normative standard: 100% of baseline means "restored to unanchored state," not "correct." This is consistent with anchoring research in general: the goal is to measure and mitigate anchor influence, not to establish objectively correct judgments.

6. **Percentage of baseline metric limitations.** Our proposed metric has several properties that warrant caution: (1) *Bidirectional averaging*: Aggregates mask per-trial variance; see Tables 6 and 7 for breakdowns. (2) *Ratio scaling*: A 5-month deviation from an 18-month baseline (27.8%) appears larger than the same deviation from a 36-month baseline (13.9%). This is intentional for cross-model comparison, but practitioners should examine absolute deviations for their use case. (3) *Same-baseline circularity*: See "Proportional anchor design" above. Future work should 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 model coverage.** The multi-domain extension (Section 5) uses only 2 Anthropic models (Opus 4.6 and Sonnet 4.5) with ~120 trials per technique-domain cell. Domain-dependent findings should be considered preliminary until validated with broader model coverage. The dramatic SACD ranking inversion (Judicial #1 → Loan #5) may not generalize to all model families.

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

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 ?, 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 6.5.

7 Conclusion

Metric choice can determine technique recommendation. Across 14,152 trials on 10 models, susceptibility and % of baseline give divergent rankings (Table 1): Devil's Advocate appears best under susceptibility but worst under baseline proximity; SACD shows the reverse pattern. Model-specific variation is substantial—SACD ranges from 48% (Haiku) to 128% (Opus) of baseline.

For practitioners: Test per-model. Measure *both* metrics—susceptibility captures consistency across anchors; baseline proximity captures whether that consistency is at the right value.

For researchers: Collect unanchored baselines. Susceptibility alone has a blind spot: it favors techniques producing consistent responses that are consistently wrong.

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 ?, 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 × 0.5, High = baseline × 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.

AI Assistance Disclosure

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 16,198 trials (14,152 judicial + 2,046 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.