Anchoring Depends on Confidence and Post-Training in Language Models

Hillary N. Owusu[1] **and Naomi H. Feldman**[2]
[1]Department of Computer Science
[2]Department of Linguistics and UMIACS
University of Maryland, College Park
hnyowusu@cs.umd.edu; nhf@umd.edu

Abstract

Anchoring bias causes large language models (LLMs) to shift quantitative judgments in response to irrelevant numerical primes. We analyze this bias as a function of model confidence and accuracy in base, instruction-tuned, and distilled variants of Llama and Qwen models. We find that anchoring susceptibility is negatively correlated with model confidence without regard to accuracy: confidently incorrect models resist anchoring as effectively as accurate ones, provided their internal priors are sufficiently strong. We further show that post-training impacts the strength of this relationship, and that models are more susceptible to high anchors than to low anchors. Our findings suggest anchoring resistance is a structural property of distributional concentration (certainty) rather than knowledge correctness (factual accuracy), with implications for deploying LLMs in numerical reasoning tasks.

1 Introduction

While post-training techniques such as instruction tuning have successfully reduced many social biases in Large Language Models (LLMs), these systems remain vulnerable to cognitive biases inherited from human-generated data (Echterhoff et al., 2024). A prominent example is the anchoring effect (Tversky and Kahneman, 1974), in which quantitative judgments are disproportionately influenced by an irrelevant initial value. Existing work has demonstrated the presence of anchoring bias in LLMs (Nguyen, 2024; Lou and Sun, 2024; Huang et al., 2025; Valencia-Clavijo, 2025) and has established that anchoring susceptibility varies substantially across model families and training paradigms (Hagendorff et al., 2023; Lou and Sun, 2024; Nguyen, 2024). However, there is not yet a robust way of predicting when a model will be most susceptible to anchoring bias.

One open question is whether a model's confidence can serve as a practical signal of robustness to anchoring, i.e., whether high-confidence predictions are systematically more resistant to anchors than low- or mid-confidence predictions. Confidence is often treated as a practical indicator of reliability in deployed language models. Work on confidence calibration shows that, when confidence is elicited appropriately, it can meaningfully predict correctness and support human oversight on uncertain cases (Tian et al., 2023). If model confidence also predicted robustness to anchoring, practitioners could use confidence-stratified strategies to flag potentially anchor-sensitive numerical judgments. A second critical question is whether robustness is grounded in factual correctness, specifically, whether possessing an accurate internal representation of a ground-truth value provides a baseline of resistance to numerical primes.

In this paper, we provide evidence that a model's internal certainty, the concentration of its baseline probability distribution, serves as a robust diagnostic signal of anchoring resistance. Through a series of regression analyses across the Llama and Qwen families, we find that susceptibility is primarily a function of this distributional confidence rather than factual accuracy, uncovering a decoupling between what a model knows and how easily it is influenced by numeric context. We further show that post-training strengthens the certainty-robustness coupling: while base models exhibit heterogeneous certainty effects that are weak in aggregate, instruction-tuned and distilled models demonstrate more consistent and stronger relationships between confidence and resistance to anchoring. Finally, we establish that anchoring is asymmetric, with high anchors exerting a reliably stronger influence than low anchors.

By characterizing these patterns, we provide an empirical account of when LLMs are most vulnerable to numerical manipulation and how anchoring susceptibility varies across post-training paradigms.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 174–180

2 Experiments

Our experiments examine the extent to which model confidence and accuracy can predict susceptibility to anchoring.

Models We evaluate six models across two architectural families, Llama and Qwen, representing three developmental stages: *base*, *instruction-tuned*, and *distilled*. The Llama-8B variants follow a consistent general-purpose `Llama-3.1-8B` lineage across all stages (Base, Instruct, and DeepSeek-R1-Distill). The Qwen-7B family includes general-purpose `Base` and `Instruct` versions alongside `DeepSeek-R1-Distill-Qwen-7B`, which is derived from the math-specialized `Qwen2.5-Math-7B`. Models were executed locally using the Hugging Face `transformers` library (Wolf et al., 2020).

Anchoring Task We evaluate anchoring susceptibility using 50 questions sampled randomly from the **Open Anchoring Dataset (OpAQ)** (Röseler et al., 2022). This dataset provides factual numeric questions and anchor values derived from empirical human trials.

For each question, we define three prompt conditions, denoted by the variable x: a control condition (x_{ctl}), a low-anchor condition (x_{low}), and a high-anchor condition (x_{high}). We adopt the slot-generator prompt template from Huang et al. (2025) to test for numerical priming. In this setup, an explicitly irrelevant numerical statement is prepended to the question, ensuring the anchor holds no semantic relationship to the query. Evaluation was conducted in a zero-shot setting.

Table 1 illustrates these conditions and their corresponding mathematical notation using an example stimulus. For every question, we provide a fixed candidate set $\mathcal{Y} = \{y_1, \ldots, y_n\}$ of nine discrete numeric answer strings which remains constant across all three prompt conditions.

Model Properties We characterize each model's internal state using two primary metrics: certainty and factual accuracy.

Certainty. We operationalize certainty as the maximum probability assigned to any candidate in the control distribution: $C(x_{\text{ctl}}) = \max_{y_i \in \mathcal{Y}} P(y_i \mid x_{\text{ctl}})$. This reflects the sharpness of the model's preference prior to anchoring.

Accuracy. We operationalize factual accuracy using the mean absolute error (MAE) under control conditions: $MAE(x_{\text{ctl}}) = |\mathbb{E}[y \mid x_{\text{ctl}}] - y^*|$, where y is the numerical value associated with a response in the candidate set and y^* is the ground-truth value. Lower values of MAE indicate higher factual accuracy (closer proximity to the ground truth), whereas higher values represent significant factual error. This reflects the model's correctness prior to introducing an anchor.

For regression analyses, we utilize the scale-free relative error, $MAE_{rel} = MAE/|y^*|$, and apply a $\log(1 + MAE_{rel})$ transformation.

Condition	Notation	Prompt Template
Control	x_{ctl}	What was the age of Mahatma Gandhi when he died?
Low Anchor	x_{low}	The slot machine stopped on **64**. What was the age of Mahatma Gandhi when he died?
High Anchor	x_{high}	The slot machine stopped on **79**. What was the age of Mahatma Gandhi when he died?
Candidates	$\mathcal{Y}$	{64, 67, 70, 72, 74, 76, 77, 78, 79}

Table 1: Example stimulus set with corresponding notation (Source: Röseler et al., 2022; Huang et al., 2025).

Measuring Anchoring Effects To quantify the impact of numerical priming, we evaluate model preferences over a fixed candidate set by computing the resulting probability distribution over nine candidate completions and measuring how that distribution shifts under anchoring. This protocol is standard in closed-set language model assessment and avoids the variability introduced by stochastic decoding or sampling.

Candidate scoring and normalization. Let $x \in \{x_{\text{ctl}}, x_{\text{low}}, x_{\text{high}}\}$ represent the prompt tokens and $\mathcal{Y} = \{y_1, \ldots, y_n\}$ be the set of discrete candidate answer strings. Because our candidate set includes values of varying token lengths, we score each candidate y_i using its length-normalized log-likelihood to prevent length bias toward shorter sequences:

$$s(y_i \mid x) = \frac{1}{L_i} \sum_{j=1}^{L_i} \log P(w_j \mid x, w_{<j}) \quad (1)$$

where L_i is the number of tokens w in completion y_i. These normalized scores are then converted into a probability distribution via a softmax:

$$P(y_i \mid x) = \frac{\exp(s(y_i \mid x))}{\sum_{j=1}^{n} \exp(s(y_j \mid x))} \quad (2)$$

Distributional shift. To measure the degree to which the anchor causes a shift in the model's

probability distribution over the candidate set $\mathcal{Y}$, we compute the Total Variation Distance (TVD) between the control and anchored distributions:

$$\text{TVD}(x_{\text{ctl}}, x_{\text{anc}}) = \frac{1}{2} \sum_{i=1}^{n} |P(y_i \mid x_{\text{ctl}}) - P(y_i \mid x_{\text{anc}})| \quad (3)$$

In the Appendix A we show that this shift in probability mass is predominantly in the direction of the anchor.

Statistical Analysis. We analyze anchoring susceptibility (N = 600 observations) using Total Variation Distance (TVD, Eq. 3).[1] Our analysis follows a three-tiered approach: (1) pooled Ordinary Least Squares (OLS) to estimate the global associations between certainty, accuracy, and anchoring susceptibility across models; (2) Linear Mixed-Effects (LME) models with crossed random intercepts for Model and Question ID to better handle structural variation across models and data (Barr et al., 2013); and (3) interaction terms between Paradigm (Base, Instruct, Distill) and Certainty to evaluate how post-training modulates cognitive defenses. Continuous predictors are z-scored to normalize effect sizes, and categorical variables are dummy-coded with the Base paradigm, Llama family, and Low anchor serving as reference levels. Finally, per-model OLS with question fixed effects (λ_q) and HC3 robust standard errors are used to isolate model-specific effects that may be obscured in pooled analyses (Table 2).

3 Results

The Decoupling of Accuracy and Anchoring Resistance Anchoring susceptibility (TVD) is predicted by model control certainty, but not by factual accuracy (Fig 2). In pooled OLS with question-clustered standard errors, certainty has a significant negative association with TVD ($\beta = -0.037$, $p = .001$), and this effect is robust in a mixed-effects model with random intercepts for Model and Question ID ($\beta = -0.052$, $p < .001$). Because predictors are standardized, the pooled estimate implies that a one-standard-deviation increase in certainty corresponds to an average 0.04 decrease in TVD.

In contrast, as illustrated in Fig 2, factual accuracy (operationalized as previously described) is non-significant once certainty is included ($p = .48$; mixed-effects: $p = .74$). While internal certainty acts as a robust diagnostic of anchoring resistance, a model's proximity to the ground truth does not.

Post-training Amplifies Confidence-Driven Resistance. The relationship between certainty and anchoring resistance is modulated by the training stage. Using the Base paradigm as a reference level, our interaction analysis reveals that while Base models show a negligible pooled relationship between certainty and resistance ($\beta = -0.012, p = .54$), post-trained models exhibit significantly steeper certainty–resistance slopes: Instruct ($\beta_{\text{int}} = -0.072, p = .008$) and Distill ($\beta_{\text{int}} = -0.054, p = .021$). These interaction effects are robust in the LME specification, confirming that the shift in anchoring resistance is a structural byproduct of the training paradigm rather than individual model variance.

However, the seemingly "weak" relationship in Base models may reflect cross-model differences in calibration rather than weak within-model relationships between certainty and anchoring. Per-model regressions (Table 2) reveal that certainty is a predictor of anchoring resistance within individual architectures when evaluated across questions (Llama-Base: $\beta = -0.055, p < .001$; Qwen-Base: $\beta = -0.023, p < .1$). This discrepancy suggests that while Base models possess an internal certainty–resistance link, their confidence scales are not aligned with one another, leading to a "washed out" effect when pooled. In contrast, instruction-tuned and distilled models show a more consistent relationship between certainty and anchoring resistance in the pooled analysis, making certainty a more reliable diagnostic of anchoring resistance after post-training.

General Susceptibility and Asymmetry To quantify the directional influence of anchors, we define an *Asymmetry Score* as the difference in Total Variation Distance (TVD) between high and low anchor conditions: Asym $= TVD_{high} - TVD_{low}$.

As shown in Figure 1, the Asymmetry Score remains positive across nearly all model families and certainty regimes. Directional analysis (Appendix A, Table 3) confirms these shifts reflect genuine anchoring: base models show $\geq$92% directional consistency for both anchor types, but instruction-tuned models degrade asymmetrically, maintaining 78–86% consistency for high anchors while dropping to 50–64% for low anchors. This selective destabilization of downward shifts is consistent with the observed high-anchor dominance.

[1] All core findings replicate on an expanded set of 99 OpAQ items (1,188 total observations), with all main effects remaining significant at $p < .001$.

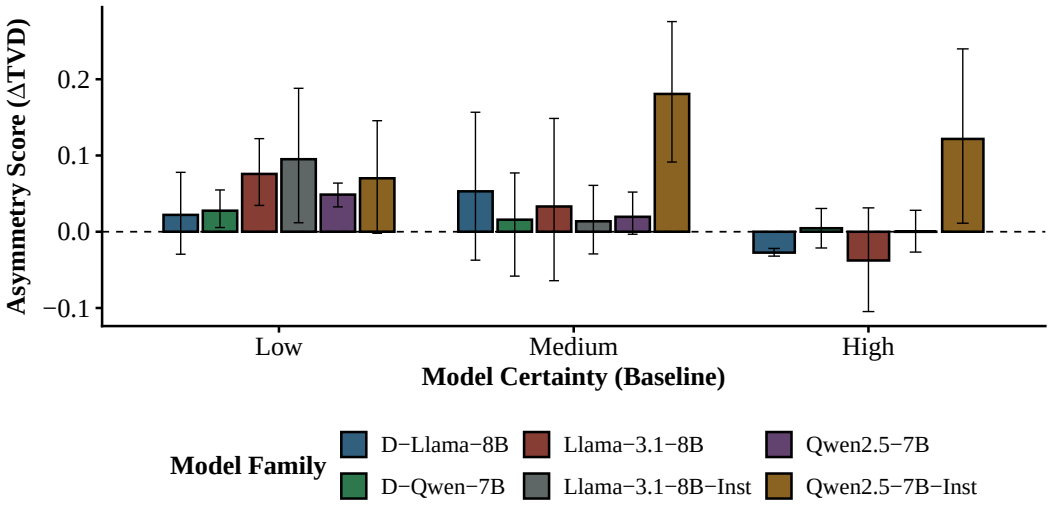

Figure 1: Directional Asymmetry Score across model lineages and certainty regimes. The score is defined as $TVD_{high} - TVD_{low}$. Positive values indicate that high anchors exert a stronger influence on the internal probability distribution than low anchors. Error bars represent 95% bootstrap confidence intervals.

The asymmetry peaks at medium certainty, where anchors exert maximal influence. For instruction-tuned models like `Qwen2.5-7B-Instruct`, high-anchor effects persist even at high certainty (>0.8), though with substantial variability across questions. Distilled models partially recover stability (`DS-Llama-8B`: 92% for low anchors), suggesting distillation buffers against instruction-tuning instability.

Model	Cert. (C)	Accuracy	High	R^2
Base Models				
Llama-3.1-8B	-0.055***	0.020	0.056*	0.850
Qwen-2.5-7B	-0.023†	-0.005	0.046***	0.799
Instruction-Tuned				
Llama-3.1-8B	-0.053***	0.021†	0.035	0.731
Qwen-2.5-7B	-0.084***	0.059*	0.128**	0.810
Distilled (DeepSeek)				
DS-Llama-8B	-0.066***	0.015	0.023	0.812
DS-Qwen-7B	-0.017***	-0.002	0.025*	0.863

Table 2: Per-model OLS with question fixed effects (λ_q) and HC3 standard errors. Models belong to either the Llama or Qwen family and to one training paradigm (Base/Instruct/Distill). Certainty and Accuracy are standardized within each model; Accuracy serves as our operationalized measure of factual accuracy. Significance levels ***p < .001, **p < .01, *p < .05. † indicates marginal significance (p < .1).

4 Discussion

Our regression analyses provide an answer to the question of whether anchoring vulnerability can be anticipated: internal certainty is a robust diagnostic signal of anchoring resistance, whereas baseline factual accuracy is not (Fig. 2). We find that the concentration of a model's baseline distribution consistently forecasts its susceptibility to numeric context, regardless of the accuracy of its initial estimate. This establishes that internal confidence is the primary predictor of anchoring, enabling the anticipation of model vulnerability even in the absence of ground-truth knowledge. One possible interpretation is that LLM anchoring reflects a trade-off between uncertainty and computation, analogous to resource-rational theories of human cognition (Lieder et al., 2018). These patterns also echo classic human findings that anchoring is stronger under lower confidence and uncertainty, suggesting a broader parallel between human and model susceptibility to numerical primes (Jacowitz and Kahneman, 1995).

The certainty-robustness relationship is consistent across model lineages, but the strength of this coupling varies. Notably, Table 2 shows that DeepSeek-R1-Distill-Qwen-7B, derived from the math-specialized Qwen2.5-Math-7B, has a weaker certainty effect ($\beta = -0.017$) compared to other models, though still statistically significant. Because our design does not allow us to isolate training corpus effects from architectural differences, we cannot make strong causal claims about math specialization, but our results suggest that further research into the impact of math-specialized pretraining on model certainty and robustness could be fruitful.

We also find that high anchors exert a reliably stronger influence than low anchors, even after con-

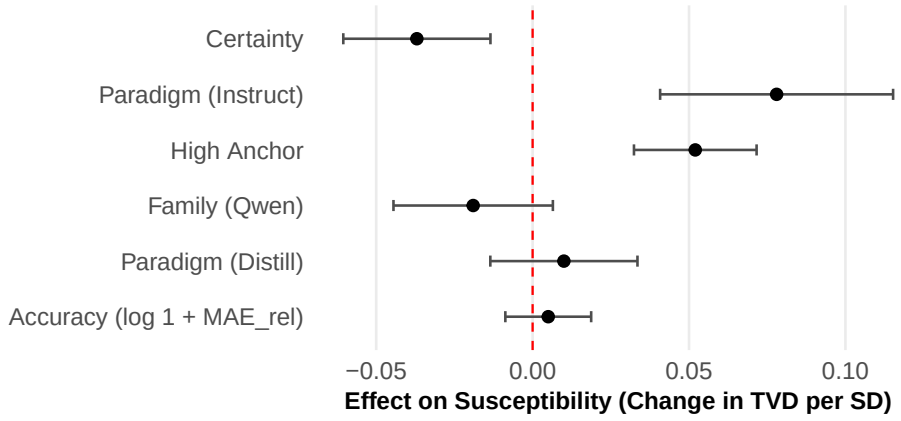

Figure 2: Main effects on anchoring susceptibility (TVD) from a pooled OLS regression ($N = 600$, 50 questions × 2 anchor conditions × 6 models). Point estimates represent standardized coefficients (z-scores) with 95% confidence intervals. Internal certainty significantly reduces susceptibility, while factual accuracy shows a null effect (crossing the red dashed zero-line).

trolling for baseline certainty, accuracy, and model attributes (Fig. 2). This directional asymmetry is not a byproduct of specific model families or training stages, but a persistent property of how numerical context perturbs the distribution in our setup, suggesting LLMs are generally more sensitive to the upward expansion of the response space.

Finally, interaction analyses indicate that post-training significantly alters the relationship between certainty and robustness. While base models show a weak coupling, instruction-tuned and distilled variants exhibit a much stronger correlation: the same increase in baseline concentration is more predictive of anchoring resistance after post-training. This suggests that while alignment and distillation may not directly target anchoring, these processes may nevertheless impact how the model utilizes its internal confidence.

5 Limitations

Our analysis is based on 50 questions sampled from the Open Anchoring dataset, which limits statistical power and prevents strong claims about population-level prevalence of anchoring effects. The trends we report across models, certainty regimes, and anchor directions should therefore be interpreted as indicative patterns. At the same time, our core findings remain stable in an expanded replication on 99 OpAQ items, making it less likely that the main relationships we identify are artifacts of the smaller sample. Broader coverage would be required to assess anchoring behavior across a wider range of domains, numeric scales, and question formats.

Second, our evaluation measures model preferences over a fixed candidate set using log-probabilities under teacher forcing, rather than sampling free-form generations. This design reduces decoding variability and enables controlled comparisons across prompt conditions, but it may not fully reflect anchoring effects in open-ended settings where models can produce out-of-set answers or vary their response formats. Relatedly, distributional shift measures (e.g., TVD) depend on the provided candidate set and its granularity.

Finally, we focus on diagnostic characterization rather than mitigation. We do not test interventions (e.g., prompting strategies or fine-tuning) that could reduce anchoring, nor do we provide mechanistic explanations for the observed family-dependent differences. These are important directions for future work.

6 Acknowledgments

We thank Shramay Palta, Rachel Rudinger, and Sarah Wiegreffe for helpful discussion.

References

Dale J. Barr, R. Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68 3.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with llms. *Preprint*, arXiv:2403.00811.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Yiming Huang, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. 2025. An empirical study of the anchoring effect in llms: Existence, mechanism, and potential mitigations. *Preprint*, arXiv:2505.15392.

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

Falk Lieder, Thomas L. Griffiths, and Quentin J. M. Huys. 2018. The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, 25(1):322–349.

Jiaxu Lou and Yifan Sun. 2024. Anchoring bias in large language models: An experimental study. *Preprint*, arXiv:2412.06593.

Jeremy K. Nguyen. 2024. Human bias in ai models? anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance*, 43(C):None.

Lukas Röseler, Lucia Weber, Ena P. B. Stijović, Katharina A. K. Jaekel, J. F. (Janne)ke M. T. (Janneke) G. (Gijsbers) Gijsbers, and Nir Milstein. 2022. The Open Anchoring Quest Dataset: Anchored Estimates from 96 Studies on Anchoring Effects. *Journal of Open Psychology Data*, 10(1):16.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

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

Felipe Valencia-Clavijo. 2025. Anchors in the machine: Behavioral and attributional evidence of anchoring bias in llms. *Preprint*, arXiv:2511.05766.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

A Validation of Distributional Metrics

To ensure that Total Variation Distance (TVD) reflects specific anchoring susceptibility rather than general model instability, we perform two tests. Both tests rely on Δ, defined below, which measures the average shift of the model's numerical estimate toward the anchor.

Each candidate string y_i corresponds to a unique numeric value $v_i \in \mathbb{R}$. To measure the extent to which the distributional shift moves probability mass toward the anchor value, we compute the percentage change in the model's estimate in each anchoring condition relative to the control baseline:

$$\Delta_{\text{low}} = \frac{\mathbb{E}[v \mid x_{\text{ctl}}] - \mathbb{E}[v \mid x_{\text{low}}]}{\mathbb{E}[v \mid x_{\text{ctl}}]} \tag{4}$$

$$\Delta_{\text{high}} = \frac{\mathbb{E}[v \mid x_{\text{high}}] - \mathbb{E}[v \mid x_{\text{ctl}}]}{\mathbb{E}[v \mid x_{\text{ctl}}]} \tag{5}$$

where $\mathbb{E}[v \mid x]$ is the expected value over the candidate set:

$$\mathbb{E}[v \mid x] = \sum_{i=1}^{n} P(y_i \mid x)\, v_i. \tag{6}$$

A positive value for either Δ_{low} or Δ_{high} indicates that the model's estimate was pulled in the direction of the provided anchor.

To quantify the linear relationship between distributional volatility (TVD) and the magnitude of the resulting anchoring effect ($|\Delta|$), we calculate the Pearson correlation coefficient between these two variables. We find a significant positive correlation ($r = 0.63$, $p < .001$), indicating that larger reallocations of internal probability mass correspond to larger changes in the model's expected value. This moderate-to-strong positive correlation validates TVD as a meaningful proxy for behavioral anchoring effects.

Model	Corr. (r)	Dir. (Low)	Dir. (High)
Base Models			
Llama-3.1-8B	0.566***	92.0%***	92.0%***
Qwen-2.5-7B	0.500***	92.0%***	96.0%***
Instruction-Tuned			
Llama-3.1-8B	0.603***	64.0%**	78.0%***
Qwen-2.5-7B	0.526***	50.0%	86.0%***
Distilled (DeepSeek)			
DS-Llama-8B	0.702***	92.0%***	84.0%***
DS-Qwen-7B	0.419***	74.0%***	90.0%***

Table 3: Validation of distributional metrics. Correlation (r) represents the Pearson correlation between TVD and $|\Delta|$. Directional percentages represent anchor adherence, tested against chance (50%) via exact binomial test. Significance levels: ***p < .001, **p < .01, *p < .05. † indicates marginal significance (p < .1).

We also compute the percentage of trials where the expected value moved toward the provided anchor (Table 3). Across all models, 82.5% of shifts were directional, reflecting an anchoring bias. However, in some cases, particularly for the low anchor

values in the instruction-tuned models, distributional shifts were less consistently in the direction of the anchor. In those cases, TVD may be capturing general instability in the models' estimates rather than anchoring biases per se.