Anchors in the Machine: Behavioral and Attributional Evidence of Anchoring Bias in LLMs

Felipe Valencia-Clavijo
Dataplicada
feval@dataplicada.com

Abstract

Large language models (LLMs) are increasingly examined as both behavioral subjects and decision systems, yet it remains unclear whether observed cognitive biases reflect surface imitation or deeper probability shifts. *Anchoring bias*, a classic human judgment bias, offers a critical test case. While prior work shows LLMs exhibit anchoring, most evidence relies on surface-level outputs, leaving internal mechanisms and attributional contributions unexplored. This paper advances the study of anchoring in LLMs through three contributions: (1) a log-probability-based behavioral analysis showing that anchors shift entire output distributions, with controls for training-data contamination; (2) exact Shapley-value attribution over structured prompt fields to quantify anchor influence on model log-probabilities; and (3) a unified Anchoring Bias Sensitivity Score integrating behavioral and attributional evidence across six open-source models. Results reveal robust anchoring effects in Gemma-2B, Phi-2, and Llama-2-7B, with attribution signaling that the anchors influence reweighting. Smaller models such as GPT-2, Falcon-RW-1B, and GPT-Neo-125M show variability, suggesting scale may modulate sensitivity. Attributional effects, however, vary across prompt designs, underscoring fragility in treating LLMs as human substitutes. The findings demonstrate that anchoring bias in LLMs is robust, measurable, and interpretable, while highlighting risks in applied domains. More broadly, the framework bridges behavioral science, LLM safety, and interpretability, offering a reproducible path for evaluating other cognitive biases in LLMs.

***K*eywords** Large Language Models (LLMs) · Anchoring bias · Cognitive bias · Interpretability · Explainable AI (XAI) · Shapley values

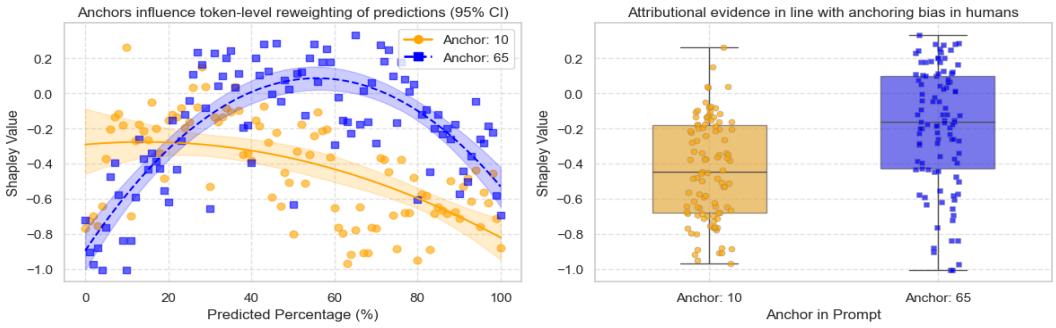

Figure 1: Replication of Tversky and Kahneman's anchoring experiment [1] in GPT-2 with Shapley values.

1 Introduction

Two complementary research streams have emerged at the intersection of decision sciences and artificial intelligence. One treats large language models (LLMs) as *subjects*, probing whether they reproduce human cognitive phenomena with the purpose of using them as substitutes for real participants in human-subject research [2, 3]. The other leverages theories and tools from neuroscience, behavioral economics, and psychology

to interrogate the capabilities and limits of LLMs as decision systems in their own right, with the aim of anticipating failure modes and improving safety [4, 5]. Both perspectives are increasingly relevant as LLMs are delegated to consequential tasks in industry and government via autonomous and semi-autonomous agents.

A persistent challenge for both streams is that LLMs are often described as "black boxes": their predictions are easy to observe but difficult to trace. Research in interpretable and explainable AI has responded with a range of approaches, from behavioral input–output analysis and probing, to attributional analyses of token influence, to concept-level probing of internal representations, and finally to mechanistic investigations of neural circuits [6, 7, 8]. While these paradigms differ in depth, they share a common goal: to make model decisions more transparent and, ultimately, more reliable. Yet a wide gap remains between surface-level behavioral observation studies and deeper, interpretable approaches.

Cognitive biases and heuristics provide a principled lens through which to bridge this gap, and the discovery of such biases in LLMs can also be beneficial for AI alignment. Decades of behavioral economics have demonstrated that human judgments systematically deviate from rational choice in predictable ways [9, 1, 10]. Mirroring this tradition, recent studies show that LLMs reproduce psychology-style effects [2, 4]. Beyond human parallels, there is also growing interest in identifying failure modes of large language models that resemble cognitive biases [11, 12, 13]. Understanding whether anchoring in LLMs is merely a surface imitation of human behavior or reflects deeper probability shifts matters for both research streams: it informs whether LLMs can credibly stand in for human participants, and it highlights potential bias-driven failure modes that could undermine safety in real-world deployments.

In this paper, I focus on *anchoring*, a cognitive bias where exposure to an initial number (the "anchor") systematically adjusts their judgments upward or downward, even when the anchor is irrelevant. For example, when asked to estimate the percentage of African countries in the United Nations, human participants give higher estimates after seeing a high anchor (e.g., 65) and lower estimates after seeing a low anchor (e.g., 10) [1]. Prior work shows that LLMs display a similar pattern [14, 15, 16]. However, most of the evidence relies only on analyzing chat-style outputs. It does not analyze internal probability distributions, nor does it test whether the anchor itself contributes to the log-probability of the prediction. I only found one study that analyzed anchoring-like behavior at the mechanistic level in the GPT-2 family of models [17].

My contribution is to strengthen both behavioral and attributional approaches. On the behavioral side, I go beyond surface-level experimental designs by analyzing sequence *log-probabilities* of candidate answers, which allows me to detect systematic shifts in the entire output distribution rather than relying only on repetition of chat-generated answers. On the attributional side, I introduce a Shapley-value framework over structured prompt fields to quantify how much the *anchor* contributes to the model's log-probability of candidate answers, as seen in Figure 1 for GPT-2. Shapley values, originally developed in cooperative game theory [18], have been adapted to LLMs as a principled way of assigning influence to tokens or fields [19, 20, 21]. To my knowledge, no previous study has combined robust log-probability analysis with Shapley attribution to study anchoring bias in LLMs. By integrating these two levels of evidence, I provide a methodological extension to prior behavioral studies with robust interpretability.

The purpose of this paper is therefore twofold: to advance the study of anchoring in LLMs by providing both log–probability–based behavioral evidence and Shapley-value attributional evidence, and to contribute methodologically by demonstrating how exact Shapley attribution can be applied across open log-prob models. Empirically, I compare six open-source LLMs (GPT-2, GPT-Neo-125M, Falcon-RW-1B, Gemma-2B, Phi-2, and Llama-2-7B) using a controlled set of prompts, and I summarize results with an Anchoring Bias Sensitivity Score that integrates behavioral and attributional evidence. This combined analysis shows that anchoring bias in LLMs is not merely a surface imitation of human behavior but is often accompanied by measurable internal reweighting, offering a clearer, more interpretable account of how biases emerge in model predictions and also a step toward more transparent, explainable, and ultimately safer AI systems.

2 Related Work

2.1 Interpretable AI Paradigms.

Bereska and Gavves [8] outline four main paradigms for interpreting model behavior:

1. **Behavioral interpretability** — treating models as black boxes and studying input–output patterns.
2. **Attributional interpretability** — tracing predictions back to the influence of input features.

3. **Concept-based interpretability** — probing internal representations for higher-level abstractions governing behavior.
4. **Mechanistic interpretability** — mapping neurons, layers, and circuits to specific causal relationships.

This study is positioned between the first two paradigms. By combining log-probability analysis with Shapley-value attribution, it moves beyond surface-level output comparisons to capture systematic distributional shifts, while also quantifying how anchors directly contribute to predictions.

2.2 Cognitive biases.

Cognitive biases are systematic patterns of deviation from rational judgment, first formalized in the pioneering work of Tversky and Kahneman. Their early study on the availability heuristic showed that people often judge frequencies or probabilities based on the ease with which examples come to mind, leading to systematic errors when availability is distorted [9]. This was followed by their famous article on heuristics and biases, which identified representativeness, availability, and anchoring as core mechanisms through which intuitive judgments depart from rational choice [1]. Later, they demonstrated the power of framing, showing that logically equivalent outcomes are perceived differently depending on how they are presented, producing predictable shifts in preference [10]. Tversky and Kahneman, along with other researchers, laid the foundations of behavioral economics by revealing that human decision-making systematically departs from rational choice theory in predictable ways. This line of research on human decision-making has also sparked growing interest in examining large language models, both to test whether they exhibit human-like cognitive biases and to explore the possibility that they may display an entirely distinct set of biases unique to their architecture.

2.3 Cognitive bias in LLMs.

Cognitive biases in LLMs are increasingly examined through psychology-inspired replications. Cui et al. [2] show that models reproduce many classic effects, though often with inflated magnitudes or spurious significance. Similarly, Binz and Schulz [4] found GPT-3 to appear human-like in decision-making yet fragile under perturbations or causal reasoning tasks. Taken together, these studies suggest that while LLMs convincingly resemble human judgments, their cognitive biases differ from humans in scale and stability, making careful interpretation essential. It is important to note, however, that both studies are confined to behavioral outputs and do not reveal attribution level analysis, nor probe the internal processes underlying model decisions.

2.4 Anchoring bias in LLMs.

Among the cognitive biases observed in large language models, anchoring has received particular attention. Experimental studies consistently show that LLMs, like humans, adjust their judgments upward or downward depending on irrelevant numeric cues. Suri et al. [14] demonstrate this by replicating a classic anchoring task with ChatGPT-3.5, ChatGPT-4, and human participants, finding statistically robust shifts between high- and low-anchor conditions that mirror human behavior. Similarly, Lou and Sun [15] evaluate anchoring across GPT-3.5, GPT-4, and GPT-4o, showing that stronger models are more consistently biased by numeric hints, while weaker models introduce more variability. They also find that simple prompt-level mitigation strategies are largely ineffective, indicating that anchoring is a robust feature of model behavior. Stureborg et al. [16] extend this line of work by analyzing anchoring in multi-attribute evaluation tasks. When GPT-4 generated scores for several text attributes in sequence, later ratings were disproportionately biased by earlier ones, reflecting the autoregressive or sequential dependency nature of model outputs. The authors argue that such dependencies undermine LLM reliability as evaluators, particularly in multi-criteria settings.

2.5 Other related cognitive biases in LLMs.

Recent experimental work has documented several cognitive biases in LLMs that are closely related to anchoring. Wang et al. [22] show that LLM-based evaluators display strong positional bias, with judgments easily manipulated by the order of candidate responses. Chen et al. [23] identify threshold priming bias in information retrieval assessments, where prior relevance scores systematically influence subsequent judgments across GPT-3.5, GPT-4, and LLaMa2 models. Sumita et al. [24] provide a broader survey, experimentally confirming six cognitive biases, including order effects. Notably, Li and Gao [17] go beyond behavioral evidence by applying mechanistic interpretability to multiple-choice question answering in GPT-2 models,

uncovering an internal preference for the first option ("A"). This remains one of the only studies to probe an anchoring-like cognitive bias as an internal mechanism.

2.6 Shapley values for attribution in LLMs.

With the exception of Li and Gao [17], existing studies rely exclusively on experimental designs, providing behavioral or mimicking evidence that anchoring and anchoring-related biases are stable properties of LLMs across tasks and models. However, they do not probe or attribute internal log-probabilities, leaving a gap between purely behavioral findings and mechanistic interpretability. A natural tool for bridging this gap is the Shapley value, originally introduced in cooperative game theory to fairly allocate payoffs among players based on their marginal contributions [18]. In the context of LLMs, Shapley values can be adapted to attribute the influence of individual prompt tokens or fields to the model's log-probability of specific outputs, providing a principled way to move from behavioral observations to more interpretable evidence of internal scoring dynamics. Therefore, attribution-based analysis offers a promising middle ground, as it can reveal how anchors shape token-level reweighting within model predictions.

The idea of applying Shapley values to interpret LLM behavior has recently gained traction. Mohammadi [19] proposed a Shapley-based framework that treats prompt components as players in a cooperative game, quantifying their marginal contributions to choice probabilities and exposing the "token noise" phenomenon, where seemingly irrelevant tokens exert disproportionate influence on decisions. In parallel, Horovicz and Goldshmidt [20] introduced *TokenSHAP*, which estimates Shapley values at the token level via Monte Carlo sampling and uses semantic similarity between generated responses as the payoff function. Beyond these research contributions, the official SHAP library itself provides demonstration notebooks for GPT-2, where Shapley values are computed over open-ended text generation tasks to explain which input tokens drive the log-probability of generating specific outputs [21].

These examples show the potential of combining token-level Shapley attribution with logit-based teacher forcing to interpret LLM predictions. However, there is a trade-off: using raw log-probabilities enables exact attribution but is computationally expensive, while Monte Carlo sampling with log-probabilities or semantic similarity reduces cost at the expense of approximate Shapley values. These lines of work illustrate the versatility but also the challenges of Shapley-based approaches for probing LLM decision-making at the level of discrete choice experiments, fine-grained token importance, or practical debugging of generation behavior.

3 Methods

I study anchoring in large language models with a simple and strict design. I look for two kinds of evidence based on Bereska and Gavves interpretability paradigms [8]: (i) behavioral (B) shifts in the distribution over numeric targets when I swap a low anchor for a high anchor, and (ii) attributional (A) changes showing how much the *anchor* field contributes to the model's log-probability for those same targets. I do not attempt a concept-based (C) or mechanistic (M) analysis here; rather, the middle ground attribution-based Shapley values to reveal how anchors shape token-level reweighting within model predictions.

3.1 Design, stimuli, and hypotheses

Positive-control replication (V0; Tversky & Kahneman). As a positive control I replicate the classic "African countries in the UN" anchoring experiment reported by Tversky and Kahneman [1]. Inspired by the prompt template approach of Mohammadi [19], the prompt I used has three sentences and a single number shown to the model (the anchor). V0 is kept for reference in figures but is *excluded* from the model-level aggregate score (to avoid contamination or, in other words, inflating the results with a canonical item likely seen during pre-training). All other stimuli are listed in Appendix A.

Template (fixed structure). Prompts are rendered with a Jinja template with four fields `{scene, comparative, absolute, anchor}`:

```
{{ scene }}{{ anchor }}.

{{ comparative }}{{ anchor }}?

{{ absolute }}
```

V0 example (rendered).

- **Low anchor (10):**

  The roulette wheel landed on 10.
  Is the percentage of African countries in the United Nations larger or smaller than 10?
  What is your best guess of the percentage of African countries in the UN?

- **High anchor (65):**

  The roulette wheel landed on 65.
  Is the percentage of African countries in the United Nations larger or smaller than 65?
  What is your best guess of the percentage of African countries in the UN?

Question families and anchor regimes. Beyond V0 I include *different questions* that keep the *same structure* to measure anchoring across content while controlling the form of the input (e.g., Asian, South American, English-speaking, EU, French-speaking countries in the UN). I run two regimes: (i) a *standard-anchors* set that reuses the pair (10, 65) across questions; and (ii) a *different-anchors* set that moves the pair while keeping the *same 55-point gap* (15–70, 20–75, . . . , 35–90). Keeping the gap constant keeps effect sizes comparable; moving the absolute numerals reduces the chance of training-set memorization of a particular pair. Full lists are in Appendix A.

Hypotheses and direction calls.

- **Behavior (B):**

$$H_0^B : \mathbb{P}(Y \mid \text{high}) = \mathbb{P}(Y \mid \text{low}), \qquad H_1^B : \mathbb{P}(Y \mid \text{high}) \neq \mathbb{P}(Y \mid \text{low}).$$

  I interpret **B+** as behavioral evidence *aligned with anchoring bias* (the soft expectation increases under the high anchor), **B**– as evidence aligned in the opposite direction (it decreases), and **B0** as no directional shift. I call **B+**/**B**–/**B0** and report p-values; I denote statistical significance using symbols ($*$, $**$, $***$) at conventional thresholds (e.g., p<.10, .05, .01).

- **Attribution (A):**

$$H_0^A : \mathbb{E}[\phi(\text{anchor}) \mid \text{high}] = \mathbb{E}[\phi(\text{anchor}) \mid \text{low}], \qquad H_1^A : \text{ a difference exists.}$$

  I interpret **A+** as attributional evidence *aligned with anchoring bias* (mean Shapley(anchor) is larger under the high anchor, i.e., the anchor field contributes more), **A**– as the reverse, and **A0** as no difference. I call **A+**/**A**–/**A0** and report p-values; I denote statistical significance using symbols ($*$, $**$, $***$). Log units allow odds interpretation; a difference of ≈ 0.69 nats corresponds to $\times 2$ odds.

3.2 End-to-end procedure

1. **Setup.** I load the model and tokenizer (Hugging Face), fix RNG seeds, and evaluate in teacher-forced mode (no sampling).[1]
2. **Render prompts.** For each variation v and each anchor $a \in \{\text{low}, \text{high}\}$ I render the three-sentence prompt with the template above (Appendix A lists all non-V0 variations).
3. **Score all targets.** Using the prompts from Step 2, I restrict outputs to the fixed set $\{0, 1, \ldots, 100\}$, rendered as strings "i%". For each target y_i I compute the sequence log-probability

$$\ell_i = \log P(y_i \mid \text{prompt}),$$

   summing across the target's sub-tokens.[1]
4. **Normalize to a categorical.** From the log-probabilities $\{\ell_i\}$ in Step 3, I form probabilities with log-sum-exp:

$$p_i = \exp\big(\ell_i - \text{logsumexp}_j\, \ell_j\big), \qquad \sum_i p_i = 1.$$

   I use standard numerical guards (e.g., `logsumexp` for stability), if a quantized backend returns logits as `uint8`, I cast to `float16` before the softmax.

[1] **Decoding.** I score `prompt + " " + target` by summing token-level log-probabilities for the target string; causal alignment uses logits at position $t-1$ to score token t.

5. **Behavioral summary (SoftEV).** From the probabilities $\{p_i\}$ in Step 4, I summarize the distribution with

$$\text{SoftEV} = \sum_{i=0}^{100} i \cdot p_i.$$

In plots I show a *95% parametric predictive interval* for the mean of n draws from p (I use $n = 100$ with $B = 5000$ bootstrap resamples).[2]

6. **Primary behavioral or mimicking test (B).** Using the ℓ_i from Step 3, for each target y_i I compute

$$\ell_i^{(\text{high})} = \log P(y_i \mid \text{prompt}_{\text{high}}), \qquad \ell_i^{(\text{low})} = \log P(y_i \mid \text{prompt}_{\text{low}}),$$

and form the paired *log-probability* differences

$$d_i = \ell_i^{(\text{high})} - \ell_i^{(\text{low})}.$$

I apply a two-sided *paired t-test on the log-probability differences* $\{d_i\}$. This is evidence that the anchor *reweights* the model's distribution over the same fixed targets (compositional dependence), not evidence about iid human sampling. I then call **B+/B–/B0** based on the SoftEV difference from Step 5 and report p-values, denoting statistical significance using symbols ($*$, $**$, $***$).

7. **Robustness for B.** On the same differences $\{d_i\}$ from Step 6, I add (i) a two-sided *Wilcoxon signed-rank* test with Pratt zeros, and (ii) a *permutation* test using random sign-flips with the mean as statistic.[3][4]

8. **Attribution (A).** In contrast to stochastic approaches, I compute Shapley(anchor) *exactly* over the four structured prompt fields. For each (v, a, y_i), I evaluate the log-probability payoff

$$v(S) = \log P(y_i \mid \text{prompt with fields } S)$$

for all 2^4 possible subsets S, and compute

$$\phi_{\text{anchor}}(i) = \text{mean}_{S:\,\text{anchor}\notin S}\big[v(S \cup \{\text{anchor}\}) - v(S)\big].$$

This full enumeration is computationally feasible with four fields and avoids the variance introduced by sampling. Compared to prior work methods, my approach is deterministic and log-probability based: Mohammadi [19] and the SHAP GPT-2 demo [21] also ground Shapley values in log-probs but reduce computation through Monte Carlo or masking, while TokenSHAP [20] instead defines payoff via semantic similarity with Monte Carlo sampling. By reducing the feature space to structured fields, I make exact Shapley computation tractable and directly test attribution shifts with the anchors.

9. **Aggregation across variations and models: Anchoring Bias Sensitivity Score (ABSS).** I compute a model-level *ABSS* that combines: the behavioral summary from Step 5 with significance from Step 6 (and robustness from Step 7), and the attribution shift with significance from Step 8. For each variation (excluding V0[5]), I define

$$S_B = \text{sign}(\Delta\text{EV}/100) \cdot \big|\Delta\text{EV}/100\big|, \qquad S_A = \text{sign}(\Delta\phi) \cdot \tanh\big(|\Delta\phi|\big),$$

where ΔEV is the SoftEV gap and $\Delta\phi$ is the mean Shapley(anchor) gap. I map p-values to weights with $w(p) = \text{clip}(-\log_{10} p/3,\, 0,\, 1)$; I build a robustness factor $\rho = 0.5 + 0.5 \cdot \text{mean}\big(w(p_{\text{Wil}}),\, w(p_{\text{Perm}})\big)$; and I add a small concordance bonus $c \in \{-1, 0, +1\}$ when both sides have weight and the signs agree/disagree. With $\alpha = \beta = 1$ and $\lambda_{\text{conc}} = 0.15$:

$$\text{ABSS} = \rho\big(\alpha\, S_B\, w(p_{\log}) + \beta\, S_A\, w(p_{\text{shap}})\big) + \lambda_{\text{conc}}\, c.$$

I report per-variation ABSS and then sum/average per model with predefined tie-breakers. V0 does not enter this aggregation.

4 Results

For reference, the full set of results for all six models is reported in Table B that can be found in Appendix B.

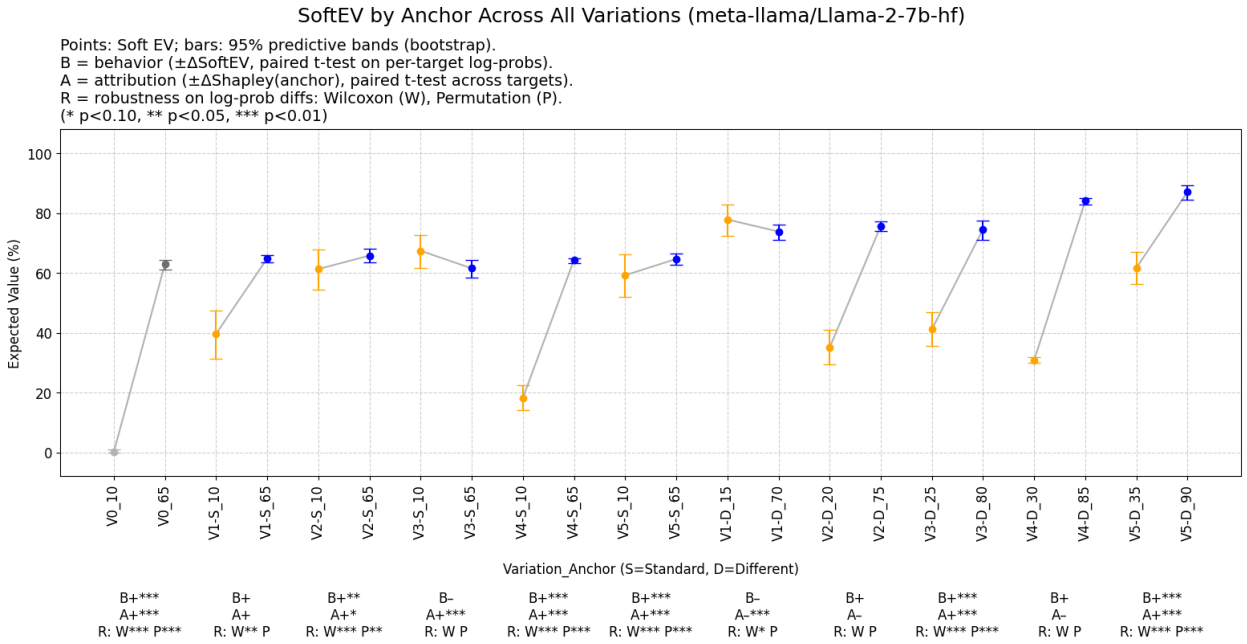

Figure 2: SoftEV by anchor across all variations (Llama-2-7b-hf).

4.1 Llama-2-7b-hf

In the positive-control replication (V0), the higher anchor produced a very large shift (+62.84), significant both at mimicking behavioral shifts (**B+*****) and attributional anchor changes (**A+*****) with robustness support (**W*****, **P*****). Nonetheless, this may be the result of contamination, which is very likely given the excessively large shift in behavior, suggesting it might have been pretrained or finetuned for the classic experiment. For such reason, this replication is excluded from aggregate scores.

Among the other variations, the strongest behavioral effects appear in V2-S (**B+****), V3-D (**B+*****), V4-S (**B+*****), V5-S (**B+*****), and V5-D (**B+*****). All of these have robustness confirmed by Wilcoxon and permutation tests (with V2-S at **W*****, **P****, and the rest at **W*****, **P*****).

Attribution alignment is also frequent and strong: V3-S (**A+*****), V3-D (**A+*****), V4-S (**A+*****), V5-S (**A+*****), and V5-D (**A+*****) all show highly significant positive contributions, with effect sizes ranging from +0.80 to +1.05 nats (×2.2–×2.9 odds multipliers).

Two notable reversals occur in behavior: V1-D (**B**–) and V3-S (**B**–), though neither reaches ** or *** significance. On the attribution side, strong negative signals appear in V1-D (**A**–*****, −0.63 nats, ×0.53) and, not strongly, in V4-D (**A**–, −0.42 nats, ×0.65).

Overall, Llama-2-7b-hf demonstrates multiple highly significant **B+** and **A+** effects with robustness support, alongside a small number of discordant cases in the different-anchor regime. Full attribution distributions are shown in Appendix 9 and Appendix 10.

4.2 falcon-rw-1b

In the positive-control replication (V0), the higher anchor produced a moderate shift (+11.86) with a **B+** call and aligned attribution (**A+**), though without strong robustness support. As with other models, V0 is excluded from aggregation due to possible pre-training contamination.

[2]**Predictive band.** The SoftEV interval is the 2.5–97.5 percentile of means of n simulated draws from the implied categorical p; it is predictive, not a confidence interval for a human population parameter.

[3]**Wilcoxon.** Two-sided signed-rank with Pratt handling of zeros.

[4]**Permutation.** Rademacher sign-flips on $\{d_i\}$; statistic is mean(d); default 10,000 permutations.

[5]**VO exclusion.** V0 does not enter this aggregation, although it can be added for experimentation and identification of contamination.

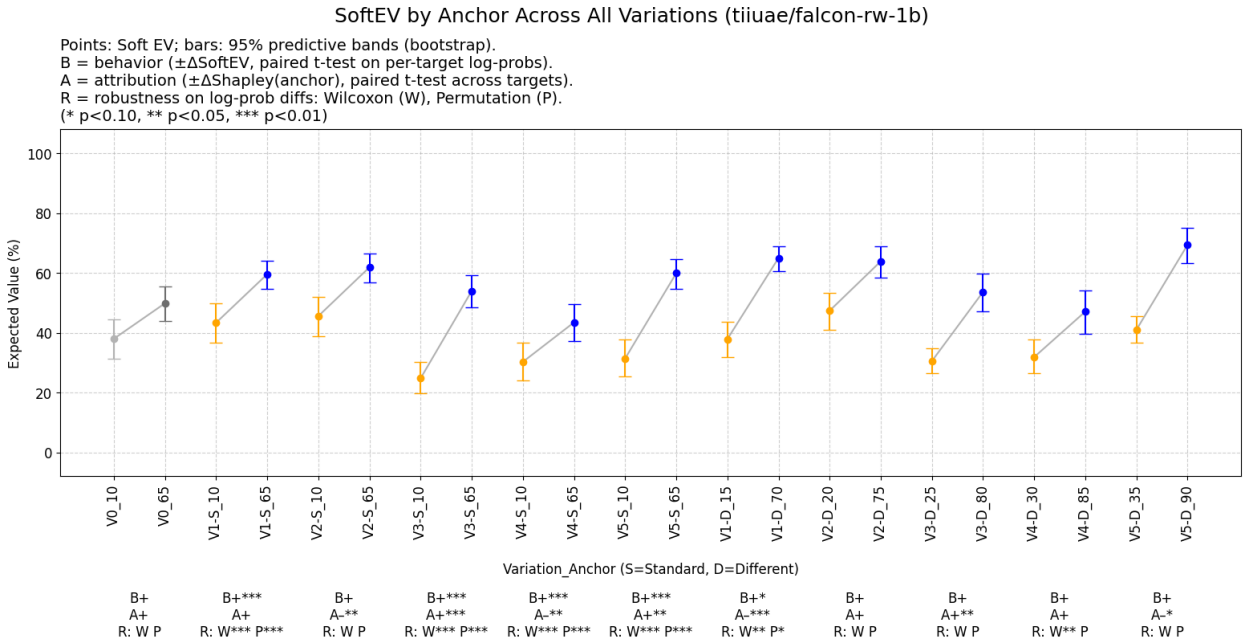

Figure 3: SoftEV by anchor across all variations (falcon-rw-1b).

Among the other variations, the strongest behavioral effects are seen in V1-S (**B+*****), V3-S (**B+*****), V4-S (**B+*****), and V5-S (**B+*****), each confirmed by robustness tests (**W*****, P*****). These variations show shifts in the range of +13.2 to +29.2 points. Additional significant but weaker effects include V1-D (**B+***, **W** P***).

Attribution shows a more mixed picture. Strong positive contributions occur in V3-S (**A+*****), V3-D (**A+****), and V5-S (**A+****), with effect sizes of +0.19–0.27 nats (×1.2–×1.3 odds multipliers). In contrast, several cases display significant negative attribution: V1-D (**A–*****), V2-S (**A–****), V4-S (**A–****), and V5-D (**A–***). These reversals suggest that while the higher anchor raises the expected value, the Shapley decomposition sometimes credits the low anchor with stronger marginal influence.

Overall, falcon-rw-1b demonstrates reliable and often significant behavioral anchoring, but attribution alignment is inconsistent, with both **A+** and **A−** outcomes across variations. Full attribution distributions are shown in Appendix 11 and Appendix 12.

4.3 gemma-2b

In the positive-control replication (V0), gemma-2b shows a strong shift of +26.75 with clear evidence of anchoring both behaviorally (**B+*****) and attributionally (**A+*****, +1.78 nats, ×5.91), supported by robustness tests (**W*****, P*****). This item is excluded from aggregation due to possible pre-training contamination.

Among the other variations, gemma-2b demonstrates one of the most coherent profiles across all models. Strong behavioral effects are observed in V1-S (**B+*****), V1-D (**B+*****), V2-S (**B+*****), V2-D (**B+*****), V3-D (**B+*****), V4-D (**B+*****), V5-S (**B+*****), and V5-D (**B+*****), all with robustness confirmed (**W*****, P*****). Particularly large shifts are seen in V2-S (+28.70), V2-D (+28.28), V4-D (+31.99), V5-S (+33.72), and V5-D (+35.68).

Attribution alignment is consistently strong. Highly significant **A+*** calls appear in V1-S, V1-D, V2-S, V2-D, V3-S, V3-D, V4-S, V4-D, and V5-S, with effect sizes often exceeding +1.0 nats and reaching up to +2.45 nats (×11.5 odds) in V2-S. Only V5-D departs from this pattern, registering **A−** with essentially no difference (−0.00 nats).

Overall, gemma-2b exhibits widespread, highly significant behavioral anchoring coupled with strong and large-magnitude positive attribution effects. The single discordant case (V5-D) does not alter the overall

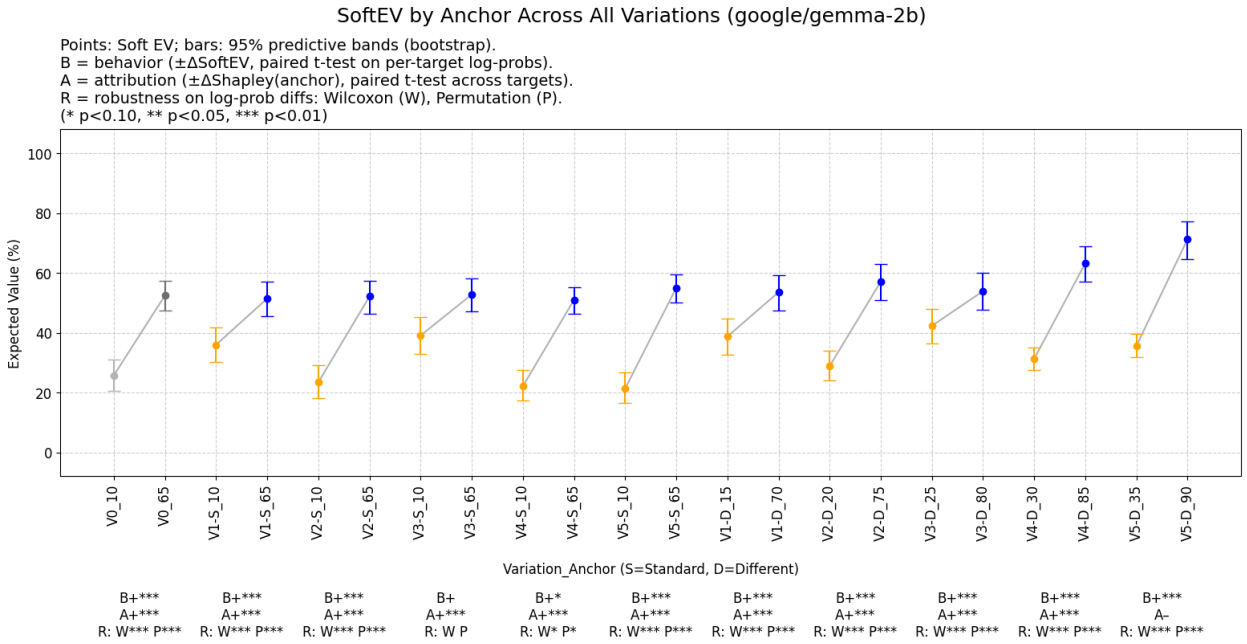

Figure 4: SoftEV by anchor across all variations (gemma-2b).

conclusion that gemma-2b presents a consistent joint B+ and A+ profile across models. Full attribution distributions are shown in Appendix 13 and Appendix 14.

4.4 gpt-neo-125M

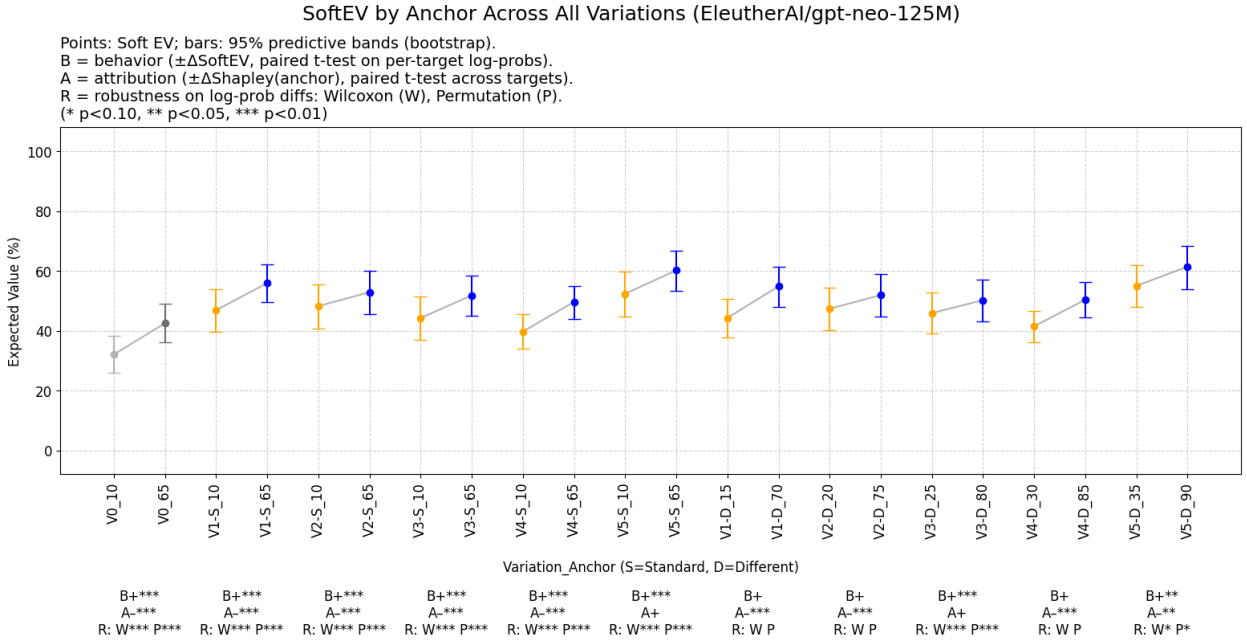

Figure 5: SoftEV by anchor across all variations (gpt-neo-125M).

In the positive-control replication (V0), gpt-neo-125M shows a moderate shift of +10.45 with strong behavioral evidence (**B+*****) supported by robustness tests (**W*****, **P*****). However, attribution is significantly

negative (**A−*****, −0.35 nats, ×0.70), indicating discordance between behavior and attribution. As with other models, this item is excluded from aggregation due to possible pre-training contamination.

Among the other variations, gpt-neo-125M exhibits reliable behavioral anchoring. Strong effects are observed in V1-S (**B+*****), V2-S (**B+*****), V3-S (**B+*****), V3-D (**B+*****), V4-S (**B+*****), and V5-S (**B+*****), each with robustness confirmed (**W*****, **P*****). Additional significant effects occur in V5-D (**B+****, **W* P***). Shifts are generally in the +4 to +10 point range.

Attribution, in contrast, is dominated by negative signals. Highly significant **A−*** calls appear in V1-S, V1-D, V2-S, V2-D, V3-S, V4-S, and V4-D, with effect sizes ranging from −0.40 to −0.21 nats (×0.67–×0.81 odds multipliers). One case, V5-D, registers **A−****. Only two variations, V3-D and V5-S, show non-significant **A+** outcomes, with negligible effect sizes (+0.04 and +0.01 nats).

Overall, gpt-neo-125M displays consistent and statistically robust behavioral anchoring but systematic negative attributional alignment, leading to frequent discordant outcomes where **B+** coincides with **A−**. Full attribution distributions are shown in Appendix 15 and Appendix 16.

4.5 gpt2

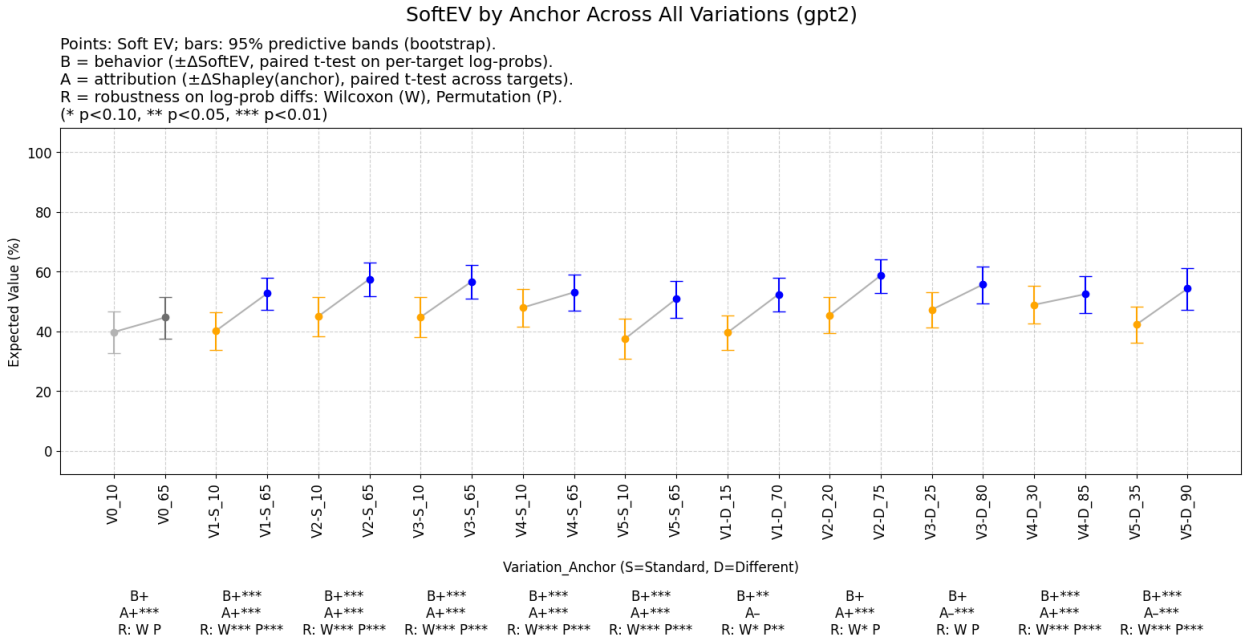

Figure 6: SoftEV by anchor across all variations (gpt2).

In the positive-control replication (V0), gpt2 shows a modest shift of +4.87 with behavioral anchoring (**B+**) but without strong robustness, while attribution is highly significant and positive (**A+*****, +0.24 nats, ×1.27). As usual, this item is excluded from model-level aggregation due to possible pre-training contamination.

Among the other variations, gpt2 demonstrates consistent behavioral anchoring. Strong effects are observed in V1-S (**B+*****), V2-S (**B+*****), V3-S (**B+*****), V4-S (**B+*****), V4-D (**B+*****), V5-S (**B+*****), and V5-D (**B+*****), each confirmed by robustness tests (**W*****, **P*****). Two additional cases, V1-D (**B+****, **W* P****) and V2-D (**B+**, **W* P**), also show upward shifts. Magnitudes range from +3.55 to +13.35 points, indicating systematic anchoring across both standard and different anchors.

Attribution is largely positive. Highly significant **A+*** calls appear in V1-S, V2-S, V2-D, V3-S, V4-S, V4-D, and V5-S, with effect sizes from +0.17 to +0.35 nats (×1.19–×1.42 odds multipliers). However, two different-anchor cases, V3-D and V5-D, show strongly negative attribution (**A−*****, −0.20 and −0.43 nats, ×0.82 and ×0.65). V1-D also registers **A−** without significance.

Overall, gpt2 exhibits broad and highly significant behavioral anchoring supported by robustness, paired with mostly positive attribution alignment. The few different-anchor reversals underscore sensitivity to

the absolute anchor values but do not alter the general conclusion of consistent anchoring behavior. Full attribution distributions are shown in Appendix 17 and Appendix 18.

4.6 phi-2

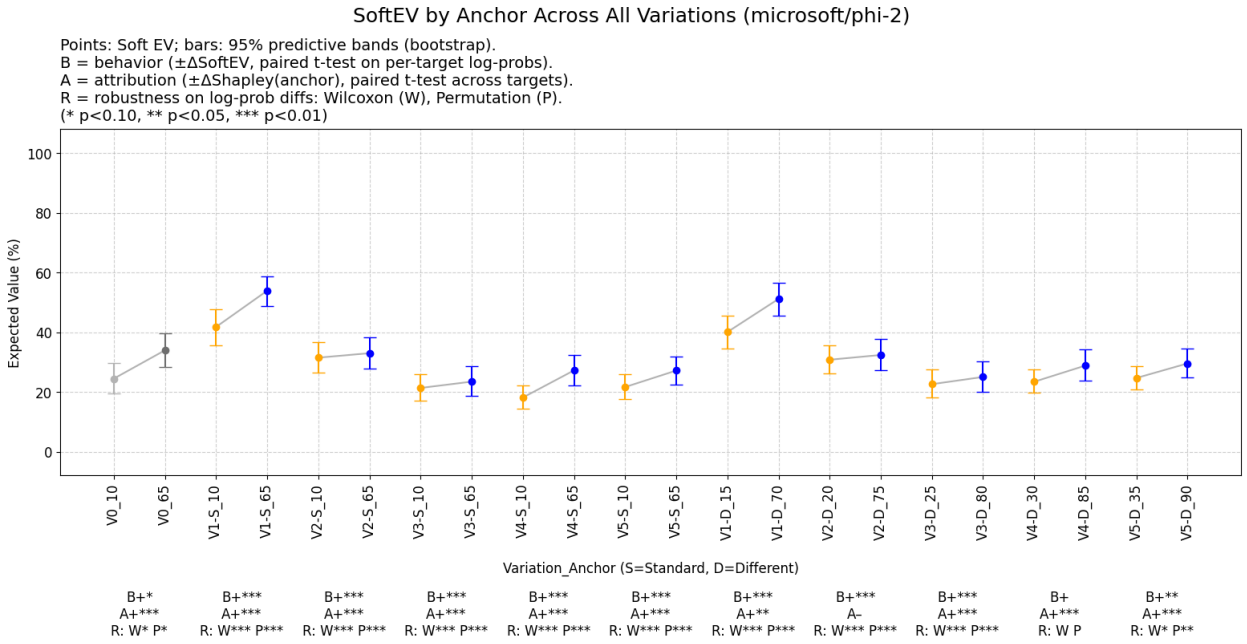

Figure 7: SoftEV by anchor across all variations (phi-2).

In the positive-control replication (V0), phi-2 shows a shift of +9.49 with behavioral evidence (**B+***) supported by weaker robustness (**W* , P***) and strongly positive attribution (**A+*****, +0.52 nats, ×1.69). As elsewhere, this item is excluded from aggregation due to possible pre-training contamination.

Among the other variations, phi-2 exhibits consistently positive behavioral anchoring. Strong effects are observed in V1-S (**B+*****), V1-D (**B+*****), V2-S (**B+*****), V2-D (**B+*****), V3-S (**B+*****), V3-D (**B+*****), V4-S (**B+*****), and V5-S (**B+*****), each with robustness confirmed (**W*****, P*****). Additional significant effects occur in V5-D (**B+****, **W* P****). V4-D shows **B+** without strong robustness. Although several ΔEV magnitudes are small (e.g., +1.5–+2.4 pts in V2–V3), the per-target design yields high power.

Attribution alignment is pervasive. Highly significant **A+***** calls appear in V1-S, V2-S, V3-S, V3-D, V4-S, V4-D, V5-S, and V5-D, with effect sizes ranging from +0.22 to +1.27 nats (×1.25–×3.56 odds multipliers); V1-D also shows **A+**** (+0.14 nats, ×1.15). The only departure is V2-D, which registers **A−** with essentially zero magnitude (−0.00 nats), i.e., not a meaningful difference.

Overall, phi-2 presents reliable, highly powered behavioral anchoring with strong and frequent positive attributional shifts, yielding one of the most coherent joint **B+**/**A+** profiles. Full attribution distributions are shown in Appendix 19 and Appendix 20.

4.7 Standard vs. different anchors

Across models, moving the anchor pair from the standard (S) regime (10–65) to the different (D) regime (15–70, . . . , 35–90) generally preserved strong behavioral anchoring (**B+**), but attributional alignment showed more variation. *gemma-2b* and *phi-2* were the most stable: both retained widespread **B+** in D with consistently positive attribution (**A+**); Gemma displayed **A+***** across nearly all S and D variations with a single negligible exception (V5-D, **A−** at ≈ 0 nats), while phi-2 mirrored this pattern with predominantly **A+***** in both regimes (with only V2-D, **A−**). *gpt2* was stable behaviorally but sensitive attributionally: S variations were uniformly **A+*****, while D introduced inversions (V3-D, V5-D, **A−*****) despite **B+** holding. *Llama-2-7b-hf* also maintained strong **B+** under D, often with very large ΔEV (e.g., V4-D), but D surfaced mixed attribution (V1-D, V2-D, V4-D, **A−**), in contrast to **A+** under all S. *falcon-rw-1b* showed reliable **B+**

in both regimes, yet attribution remained heterogeneous: S already mixed **A+** and **A−**, and D continued this split (e.g., V1-D, **A−***** vs. V3-D, **A+****). Finally, *gpt-neo-125M* was the outlier: it produced frequent, robust **B+** in both S and D, but attribution was predominantly negative (**A−*****) across regimes, indicating persistent discordance. In short, behavioral sensitivity transferred from S to D for all models, but attributional stability under D ranged from high (*gemma-2b*, *phi-2*) to fragile (*gpt2*, *Llama-2-7b-hf*, *falcon-rw-1b*), with *gpt-neo-125M* consistently discordant in both.

4.8 Ranking: Anchoring Bias Sensitivity Score (ABSS)

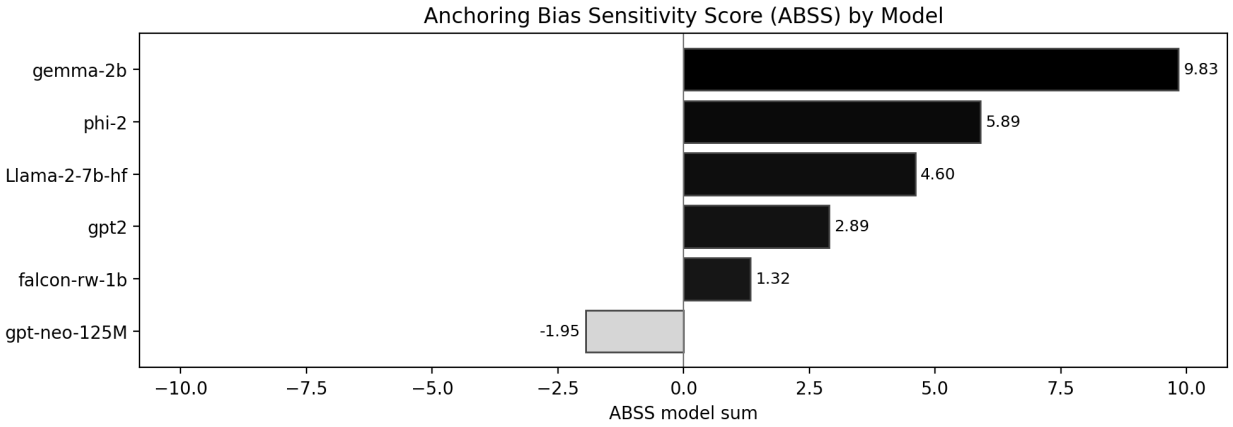

Figure 8: Anchoring Bias Sensitivity Score (ABSS) by model.

The Anchoring Bias Sensitivity Score (ABSS) aggregates behavioral and attribution-based evidence across all variations. As shown in Figure 8, the most biased models in the pool are *gemma-2b*, *phi-2*, and *Llama-2-7b-hf*, which lead the ranking with high positive ABSS values. In contrast, the lower-ranked models are *gpt2*, *falcon-rw-1b*, and *gpt-neo-125M*, with the latter displaying a negative ABSS, indicating systematic attribution shifts in the opposite direction of anchoring. Together, these results highlight that anchoring bias is strongest in the top three models, while the bottom three show weaker or reversed forms of anchoring bias sensitivity. Per-variation rankings are provided in Appendix 21.

5 Limitations

The present design has three main limitations. First, discretization: outputs were conditioned on fixed numeric strings, which may not capture the full variability of free-form generations. Second, I do not make mechanistic claims because internal parameters and circuits were not probed, limiting the conclusions to behavioral and attribution-based evidence. Third, the method only works for models with open log-probabilities of outputs.

6 Discussion and Conclusion

The results demonstrate that large language models systematically mimic the anchoring effect: higher anchors reliably shift the distribution of predicted numeric values upward (**B+**), often with attribution-based evidence that the anchor field itself is influencing reweighting (**A+**). The presence of strong and robust effects across multiple models, particularly *gemma-2b*, *phi-2*, and *Llama-2-7b-hf*, indicates that anchoring is not an incidental phenomenon but rather a consistent property of LLMs. At the same time, the mixed attribution calls observed in some variations suggest that the influence of anchors is not always uniform across targets, raising questions about how anchoring operates at a more granular level. Interestingly, the other three models: *gpt2*, *falcon-rw-1b*, and *gpt-neo-125M* appear comparatively more resistant to anchoring bias and also happen to be smaller in parameter size, suggesting (without claiming causality) that scale may shape the expression of anchoring sensitivity as Lou and Sun [15] similarly found.

My method demonstrates that employing exact Shapley computation over structured prompt fields with log-probability as the payoff allows for deterministic attribution tied directly to model scoring. By coupling these attributional results with paired tests on log-probability shifts across controlled anchor manipulations,

the approach extends beyond interpretability to deliver an explainable anchoring bias sensitivity score that integrates both behavioral and attributional evidence. Unlike SHAP-style demo notebooks that illustrate token-level attributions in open-ended generation, this design formalizes Shapley attribution into a controlled experimental framework for bias measurement. In doing so, the study provides a structured middle ground between behavioral observation and deeper causal inquiry.

The findings carry implications for both streams of research highlighted in the introduction. For the *LLMs-as-subjects* perspective, anchoring sensitivity suggests that models can reproduce psychology-style effects in controlled settings, but attributional fragility across variations warns that they should not be treated as straightforward substitutes for human participants. For the *LLMs-as-decision-systems* perspective, the presence of anchoring bias means that outputs can be systematically shifted by arbitrary cues, raising concerns for governance and safety in high-stakes domains such as healthcare, finance, and law. In this context, attribution analysis functions as a useful proxy and complement to mechanistic interpretability: whereas mechanistic studies probe internal circuits, Shapley-value attribution with log-probability as a payoff provides a principled account of which input fields influence model predictions and by how much.

Risks follow directly from these implications, since biased outputs can cascade into downstream decisions, influencing users or automated systems in subtle but consequential ways. Recognizing and quantifying such biases is therefore essential for responsible deployment. Potential mitigation strategies include filtering layers that remove arbitrary numeric elements (which I tested successfully in controlled settings, though without broad generalization), as well as broader strategies such as chain-of-thought prompting. More generally, bias-aware governance frameworks will be needed as LLMs become embedded into workflows where anchoring effects and other cognitive biases may otherwise propagate unchecked. Taken together, these implications provide a fuller understanding of where LLMs align with or deviate from human-like biases, and how such biases can be managed responsibly.

Several extensions follow naturally from this study. First, expanding to a broader set of models, including instruction-tuned, RLHF-trained, and frontier systems with log-probability access, would enable systematic benchmarking of anchoring across architectures and training paradigms. Second, integrating attributional evidence with concept-based and mechanistic interpretability could clarify when anchoring reflects surface-level probability reweightings versus deeper representational dynamics. Third, extending the methodology to other cognitive biases (e.g., framing, availability, endowment) would help construct a comparative map of LLM capabilities and limitations in behavioral terms and build a taxonomy of LLM behavioral biases, while linking them to richer interpretability frameworks beyond chat-level output analysis, provided that log-probability access remains available. Taken together, these directions would not only deepen understanding of where LLMs align with or diverge from human-like biases but also reinforce the role of attributional methods as a structured middle ground between behavioral observation and causal inquiry, offering methodological clarity, improved explainability, and a foundation for more interaction between behavioral science and LLMs research under more comprehensive interpretability paradigms.

In summary, this study shows that anchoring bias in LLMs is both robust and measurable: it shifts output distributions in predictable ways and is often accompanied by attributional signals that the anchor field itself influences these shifts. Anchoring sensitivity demonstrates that models can reproduce psychology-style effects, but also reveals fragility that limits their use as substitutes for human participants. At the same time, the fact that arbitrary cues can systematically shift outputs highlights risks for LLMs as decision systems in high-stakes domains. By combining log–probability–based behavioral analysis with exact Shapley attribution over structured fields, I demonstrate a methodological contribution that goes beyond surface-level output studies, offering a reproducible framework for behavioral measurement of open log-probability LLMs. This framework strengthens the bridge between behavioral science, LLM safety, and interpretability, establishes attributional methods as a middle ground for explaining model behavior, and underscores the need for continued research into how cognitive biases emerge, how they can be mitigated and evaluated, and how they shape the safe and responsible deployment of large language models.

7 Data Availability

To facilitate transparency and replication, the GitHub repository at `Anchoring-LLMs` contains the code, processed datasets, prompt templates, and supplementary documentation.

References

[1] Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974.

[2] Ziyan Cui, Ning Li, and Huaikang Zhou. A Large-Scale Replication of Scenario-Based Experiments in Psychology and Management Using Large Language Models. *Nature Computational Science*, 5(7):627–634, 2025.

[3] James Mooney, Josef Woldense, Zheng Robert Jia, Shirley Anugrah Hayati, My Ha Nguyen, Vipul Raheja, and Dongyeop Kang. Are LLM Agents Behaviorally Coherent? Latent Profiles for Social Simulation, 2025.

[4] Marcel Binz and Eric Schulz. Using Cognitive Psychology to Understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

[5] Shazeda Ahmed, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. Field-Building and the Epistemic Culture of AI Safety. *First Monday*, Apr 2024.

[6] Christoph Molnar. *Interpretable Machine Learning*. 3 edition, 2025.

[7] Naomi Saphra and Sarah Wiegreffe. Mechanistic?, 2024.

[8] Leonard Bereska and Efstratios Gavves. Mechanistic Interpretability for AI Safety – A Review, 2024.

[9] Amos Tversky and Daniel Kahneman. Availability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology*, 5(2):207–232, 1973.

[10] Amos Tversky and Daniel Kahneman. The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481):453–458, 1981.

[11] Erik Jones and Jacob Steinhardt. Capturing Failures of Large Language Models via Human Cognitive Biases, 2022.

[12] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What Makes Good In-Context Examples for GPT-3?, 2021.

[13] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate Before Use: Improving Few-Shot Performance of Language Models, 2021.

[14] Gaurav Suri, Lily R. Slater, Ali Ziaee, and Morgan Nguyen. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5, 2023.

[15] Jiaxu Lou and Yifan Sun. Anchoring Bias in Large Language Models: An Experimental Study, 2024.

[16] Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large Language Models are Inconsistent and Biased Evaluators, 2024.

[17] Ruizhe Li and Yanjun Gao. Anchored Answers: Unravelling Positional Bias in GPT-2's Multiple-Choice Questions, 2025.

[18] Lloyd S. Shapley. A Value for n-Person Games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

[19] Behnam Mohammadi. Explaining Large Language Models Decisions Using Shapley Values, 2024.

[20] Roni Goldshmidt and Miriam Horovicz. TokenSHAP: Interpreting Large Language Models with Monte Carlo Shapley Value Estimation, 2024.

[21] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

[22] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large Language Models are not Fair Evaluators, 2023.

[23] Nuo Chen, Jiqun Liu, Xiaoyu Dong, Qijiong Liu, Tetsuya Sakai, and Xiao-Ming Wu. AI Can Be Cognitively Biased: An Exploratory Study on Threshold Priming in LLM-Based Batch Relevance Assessment, 2024.

[24] Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments, 2024.

A Prompt Templates & Anchor Lists

Standard-anchors set (pair 10–65; different questions, same structure)

- **S1.** *The spinner stopped at* | *Do you think the percentage of Asian countries in the UN is higher or lower than* | *Guess the percentage of Asian countries in the UN.* Anchors: 10, 65.
- **S2.** *You saw a wheel display* | *Would you say the percentage of South American countries in the UN is more or less than* | *Estimate the percentage of South American countries in the UN.* Anchors: 10, 65.
- **S3.** *The number you received was* | *Is the share of English-speaking countries in the UN larger or smaller than* | *Provide your best estimate of their percentage.* Anchors: 10, 65.
- **S4.** *The game spun a number:* | *Is the proportion of EU member countries in the UN greater or lesser than* | *What do you think is their percentage in the UN?* Anchors: 10, 65.
- **S5.** *Imagine you just saw the number* | *Would you say the percentage of French-speaking countries in the UN is above or below* | *Estimate the percentage of French-speaking countries in the UN.* Anchors: 10, 65.

Different-anchors set (constant 55-point distance; different questions, same structure)

- **D1.** Asian countries in the UN. Anchors: 15, 70.
- **D2.** South American countries in the UN. Anchors: 20, 75.
- **D3.** English-speaking countries in the UN. Anchors: 25, 80.
- **D4.** EU member countries in the UN. Anchors: 30, 85.
- **D5.** French-speaking countries in the UN. Anchors: 35, 90.

Design note. Keeping the 55-point gap constant allows effect-size comparability, while moving the absolute numerals reduces risks of training-set contamination effects. V0 is retained for reference and comparison but excluded from the aggregate ABSS.

B Anchoring Results Grouped by Model

Table 1: Anchoring results grouped by model, showing all variations per LLM.
SoftEV: point + 95% predictive bands (bootstrap).
ΔEV: SoftEV gap.
B: ΔSoftEV direction with paired t-test on per-target log-probs.
R: robustness (Wilcoxon=W, Permutation=P).
A: ΔShapley(anchor) with paired t-test across targets.
ΔShapley: in nats; multiplier = e^{Δ}.

Model	Var	Anchors	SoftEV (Low) [95% pred]	SoftEV (High) [95% pred]	ΔEV	B	R	A	ΔShapley [nats] (×mult)
Llama-2-7b-hf	V0	10→65	0.11 [0.00-0.90]	62.95 [61.24-64.35]	+62.84	B+***	W*** P***	A+***	0.50 (×1.65)
	V1-S	10→65	39.61 [31.40-47.43]	64.85 [63.44-65.87]	+25.24	B+	W** P	A+	0.04 (×1.04)
	V1-D	15→70	77.86 [72.39-82.75]	73.84 [71.17-76.27]	-4.01	B–	W* P	A–***	-0.63 (×0.53)
	V2-S	10→65	61.30 [54.28-67.94]	65.81 [63.49-68.02]	+4.51	B+**	W*** P**	A+*	0.37 (×1.45)
	V2-D	20→75	35.05 [29.49-40.98]	75.74 [73.91-77.34]	+40.69	B+	W P	A–	-0.11 (×0.90)
	V3-S	10→65	67.43 [61.77-72.59]	61.53 [58.49-64.22]	-5.89	B–	W P	A+***	0.80 (×2.22)
	V3-D	25→80	41.15 [35.70-46.90]	74.65 [70.98-77.62]	+33.49	B+***	W*** P***	A+***	0.89 (×2.44)
	V4-S	10→65	18.19 [14.15-22.57]	64.33 [63.25-65.01]	+46.15	B+***	W*** P***	A+***	0.85 (×2.35)

Continued on next page

Model	Var	Anchors	SoftEV (Low) [95% pred]	SoftEV (High) [95% pred]	ΔEV	B	R	A	ΔShapley [nats] (×mult)
	V4-D	30→85	30.88 [30.05-31.80]	84.16 [82.83-85.04]	+53.29	B+	W P	A–	-0.42 (×0.65)
	V5-S	10→65	59.25 [51.96-66.11]	64.68 [62.71-66.36]	+5.43	B+***	W*** P***	A+***	1.05 (×2.85)
	V5-D	35→90	61.76 [56.17-67.09]	87.14 [84.58-89.18]	+25.38	B+***	W*** P***	A+***	0.52 (×1.69)
falcon-rw-1b	V0	10→65	37.97 [31.40-44.54]	49.84 [43.95-55.43]	+11.86	B+	W P	A+	0.03 (×1.03)
	V1-S	10→65	43.34 [36.74-49.80]	59.57 [54.67-63.94]	+16.23	B+***	W*** P***	A+	0.09 (×1.10)
	V1-D	15→70	37.89 [31.97-43.72]	65.00 [60.64-68.82]	+27.11	B+*	W** P*	A–***	-0.22 (×0.80)
	V2-S	10→65	45.57 [38.79-52.09]	61.90 [56.82-66.53]	+16.33	B+	W P	A–**	-0.14 (×0.87)
	V2-D	20→75	47.38 [41.00-53.45]	63.90 [58.47-68.86]	+16.51	B+	W P	A+	0.00 (×1.00)
	V3-S	10→65	24.76 [19.70-30.34]	53.99 [48.45-59.21]	+29.24	B+***	W*** P***	A+***	0.27 (×1.31)
	V3-D	25→80	30.54 [26.45-34.88]	53.67 [47.16-59.80]	+23.13	B+	W P	A+**	0.22 (×1.25)
	V4-S	10→65	30.26 [24.07-36.66]	43.45 [37.18-49.65]	+13.19	B+***	W*** P***	A–**	-0.14 (×0.87)
	V4-D	30→85	31.87 [26.42-37.66]	47.09 [39.70-54.17]	+15.22	B+	W** P	A+	0.10 (×1.10)
	V5-S	10→65	31.38 [25.31-37.64]	59.92 [54.64-64.66]	+28.53	B+***	W*** P***	A+**	0.19 (×1.21)
	V5-D	35→90	41.03 [36.56-45.62]	69.44 [63.21-75.08]	+28.41	B+	W P	A–*	-0.21 (×0.81)
gemma-2b	V0	10→65	25.70 [20.68-31.04]	52.45 [47.30-57.25]	+26.75	B+***	W*** P***	A+***	1.78 (×5.91)
	V1-S	10→65	35.97 [30.19-41.85]	51.44 [45.61-57.05]	+15.47	B+***	W*** P***	A+***	1.16 (×3.18)
	V1-D	15→70	38.74 [32.78-44.63]	53.56 [47.50-59.35]	+14.82	B+***	W*** P***	A+***	0.46 (×1.58)
	V2-S	10→65	23.43 [18.24-29.16]	52.13 [46.48-57.46]	+28.70	B+***	W*** P***	A+***	2.45 (×11.54)
	V2-D	20→75	28.85 [23.98-34.07]	57.13 [50.89-62.88]	+28.28	B+***	W*** P***	A+***	1.00 (×2.71)
	V3-S	10→65	39.07 [32.89-45.17]	52.78 [47.16-58.16]	+13.71	B+	W P	A+***	1.50 (×4.50)
	V3-D	25→80	42.33 [36.51-47.99]	53.91 [47.63-59.92]	+11.58	B+***	W*** P***	A+***	0.67 (×1.94)
	V4-S	10→65	22.12 [17.33-27.48]	51.00 [46.34-55.27]	+28.88	B+*	W* P*	A+***	1.76 (×5.82)
	V4-D	30→85	31.28 [27.68-35.06]	63.27 [57.21-68.79]	+31.99	B+***	W*** P***	A+***	1.54 (×4.65)
	V5-S	10→65	21.31 [16.42-26.66]	55.03 [50.13-59.48]	+33.72	B+***	W*** P***	A+***	1.90 (×6.70)
	V5-D	35→90	35.66 [31.84-39.61]	71.34 [64.73-77.19]	+35.68	B+***	W*** P***	A–	-0.00 (×1.00)
gpt-neo-125M	V0	10→65	32.05 [25.99-38.29]	42.50 [36.02-48.97]	+10.45	B+***	W*** P***	A–***	-0.35 (×0.70)
	V1-S	10→65	46.86 [39.72-53.86]	55.93 [49.52-62.08]	+9.07	B+***	W*** P***	A–***	-0.32 (×0.73)
	V1-D	15→70	44.30 [37.66-50.70]	54.87 [48.09-61.41]	+10.57	B+	W P	A–***	-0.21 (×0.81)
	V2-S	10→65	48.24 [40.67-55.55]	52.88 [45.63-59.92]	+4.64	B+***	W*** P***	A–***	-0.31 (×0.73)
	V2-D	20→75	47.36 [40.04-54.41]	51.94 [44.61-59.08]	+4.58	B+	W P	A–***	-0.35 (×0.71)
	V3-S	10→65	44.25 [36.94-51.36]	51.84 [44.93-58.48]	+7.59	B+***	W*** P***	A–***	-0.22 (×0.81)
	V3-D	25→80	45.92 [39.04-52.69]	50.24 [43.16-57.10]	+4.32	B+***	W*** P***	A+	0.04 (×1.04)

Continued on next page

Model	Var	Anchors	SoftEV (Low) [95% pred]	SoftEV (High) [95% pred]	ΔEV	B	R	A	ΔShapley [nats] (×mult)
	V4-S	10→65	39.74 [33.89-45.53]	49.64 [44.05-55.03]	+9.89	B+***	W*** P***	A–***	-0.30 (×0.74)
	V4-D	30→85	41.49 [36.14-46.74]	50.40 [44.40-56.19]	+8.91	B+	W P	A–***	-0.40 (×0.67)
	V5-S	10→65	52.37 [44.74-59.67]	60.21 [53.24-66.78]	+7.84	B+***	W*** P***	A+	0.01 (×1.01)
	V5-D	35→90	55.03 [47.90-61.87]	61.43 [53.96-68.45]	+6.40	B+**	W* P*	A–**	-0.09 (×0.91)
gpt2	V0	10→65	39.73 [32.56-46.75]	44.60 [37.49-51.49]	+4.87	B+	W P	A+***	0.24 (×1.27)
	V1-S	10→65	40.18 [33.86-46.37]	52.67 [47.07-57.91]	+12.49	B+***	W*** P***	A+***	0.29 (×1.34)
	V1-D	15→70	39.52 [33.76-45.17]	52.36 [46.50-57.96]	+12.84	B+**	W* P**	A–	-0.04 (×0.96)
	V2-S	10→65	45.01 [38.42-51.43]	57.48 [51.75-62.94]	+12.47	B+***	W*** P***	A+***	0.21 (×1.23)
	V2-D	20→75	45.40 [39.41-51.33]	58.69 [52.84-64.20]	+13.29	B+	W* P	A+***	0.17 (×1.19)
	V3-S	10→65	44.73 [38.00-51.32]	56.67 [50.83-62.19]	+11.95	B+***	W*** P***	A+***	0.33 (×1.39)
	V3-D	25→80	47.23 [41.21-53.13]	55.69 [49.34-61.69]	+8.46	B+	W P	A–***	-0.20 (×0.82)
	V4-S	10→65	47.93 [41.50-54.25]	53.04 [46.80-59.05]	+5.11	B+***	W*** P***	A+***	0.35 (×1.42)
	V4-D	30→85	48.87 [42.51-55.09]	52.43 [46.19-58.52]	+3.55	B+***	W*** P***	A+***	0.29 (×1.34)
	V5-S	10→65	37.50 [30.75-44.09]	50.85 [44.51-56.95]	+13.35	B+***	W*** P***	A+***	0.32 (×1.38)
	V5-D	35→90	42.25 [36.23-48.15]	54.32 [47.22-61.11]	+12.07	B+***	W*** P***	A–***	-0.43 (×0.65)
phi-2	V0	10→65	24.45 [19.51-29.80]	33.95 [28.25-39.68]	+9.49	B+*	W* P*	A+***	0.52 (×1.69)
	V1-S	10→65	41.76 [35.68-47.73]	53.91 [48.73-58.72]	+12.14	B+***	W*** P***	A+***	0.68 (×1.97)
	V1-D	15→70	40.06 [34.49-45.64]	51.16 [45.45-56.65]	+11.09	B+***	W*** P***	A+**	0.14 (×1.15)
	V2-S	10→65	31.48 [26.50-36.69]	32.99 [27.77-38.27]	+1.51	B+***	W*** P***	A+***	1.07 (×2.93)
	V2-D	20→75	30.79 [26.14-35.62]	32.39 [27.24-37.69]	+1.60	B+***	W*** P***	A–	-0.00 (×1.00)
	V3-S	10→65	21.34 [17.18-25.89]	23.46 [18.69-28.52]	+2.12	B+***	W*** P***	A+***	1.27 (×3.56)
	V3-D	25→80	22.60 [18.11-27.42]	25.02 [20.08-30.23]	+2.42	B+***	W*** P***	A+***	0.20 (×1.22)
	V4-S	10→65	18.13 [14.39-22.21]	27.22 [22.27-32.28]	+9.09	B+***	W*** P***	A+***	0.79 (×2.20)
	V4-D	30→85	23.42 [19.65-27.43]	28.83 [23.66-34.20]	+5.41	B+	W P	A+***	0.38 (×1.46)
	V5-S	10→65	21.64 [17.60-25.92]	27.16 [22.55-31.97]	+5.52	B+***	W*** P***	A+***	0.86 (×2.37)
	V5-D	35→90	24.68 [20.80-28.70]	29.56 [24.79-34.57]	+4.88	B+**	W* P**	A+***	0.22 (×1.25)

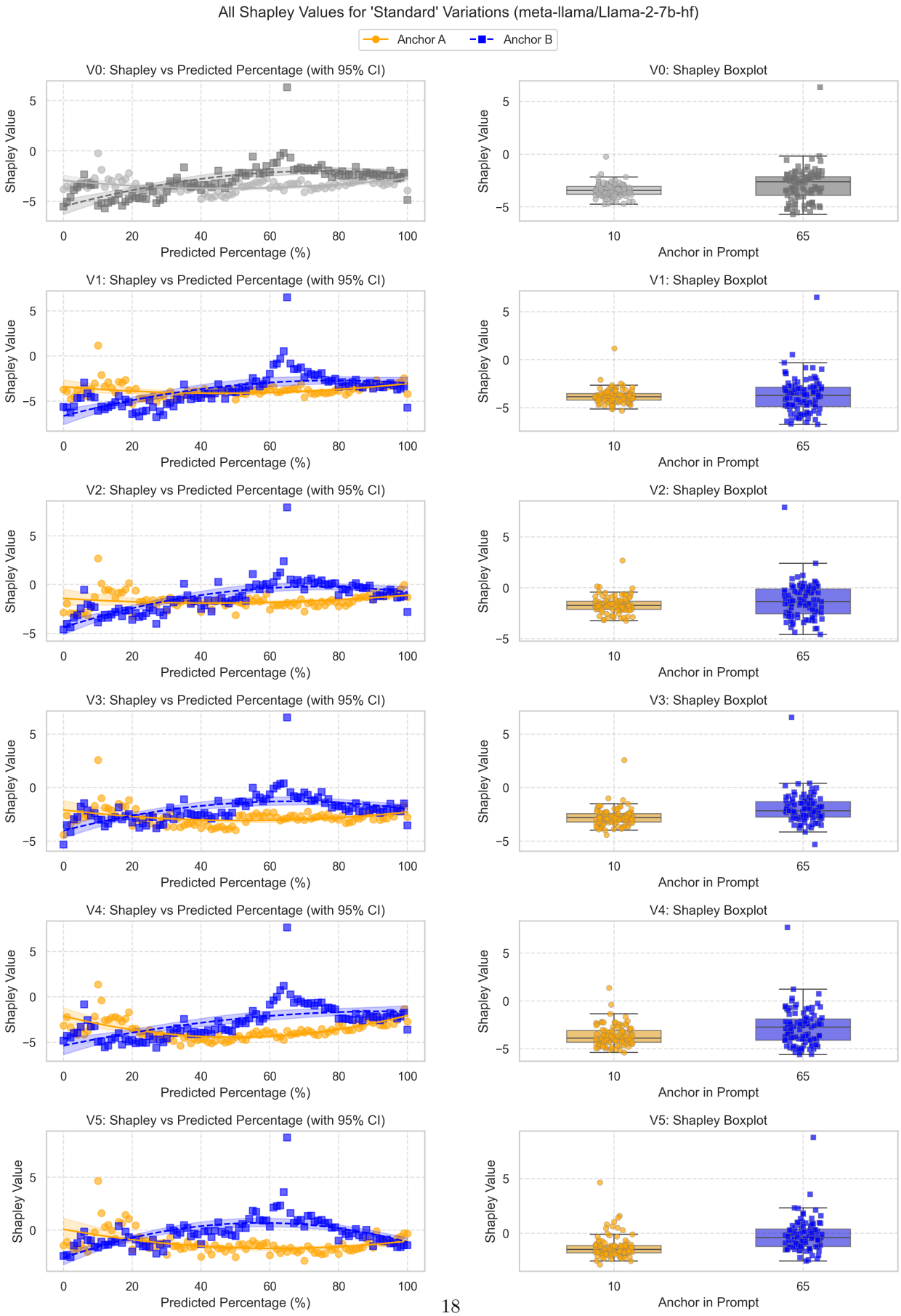

Figure 9: Attribution under standard anchors for Llama-2-7b-hf.

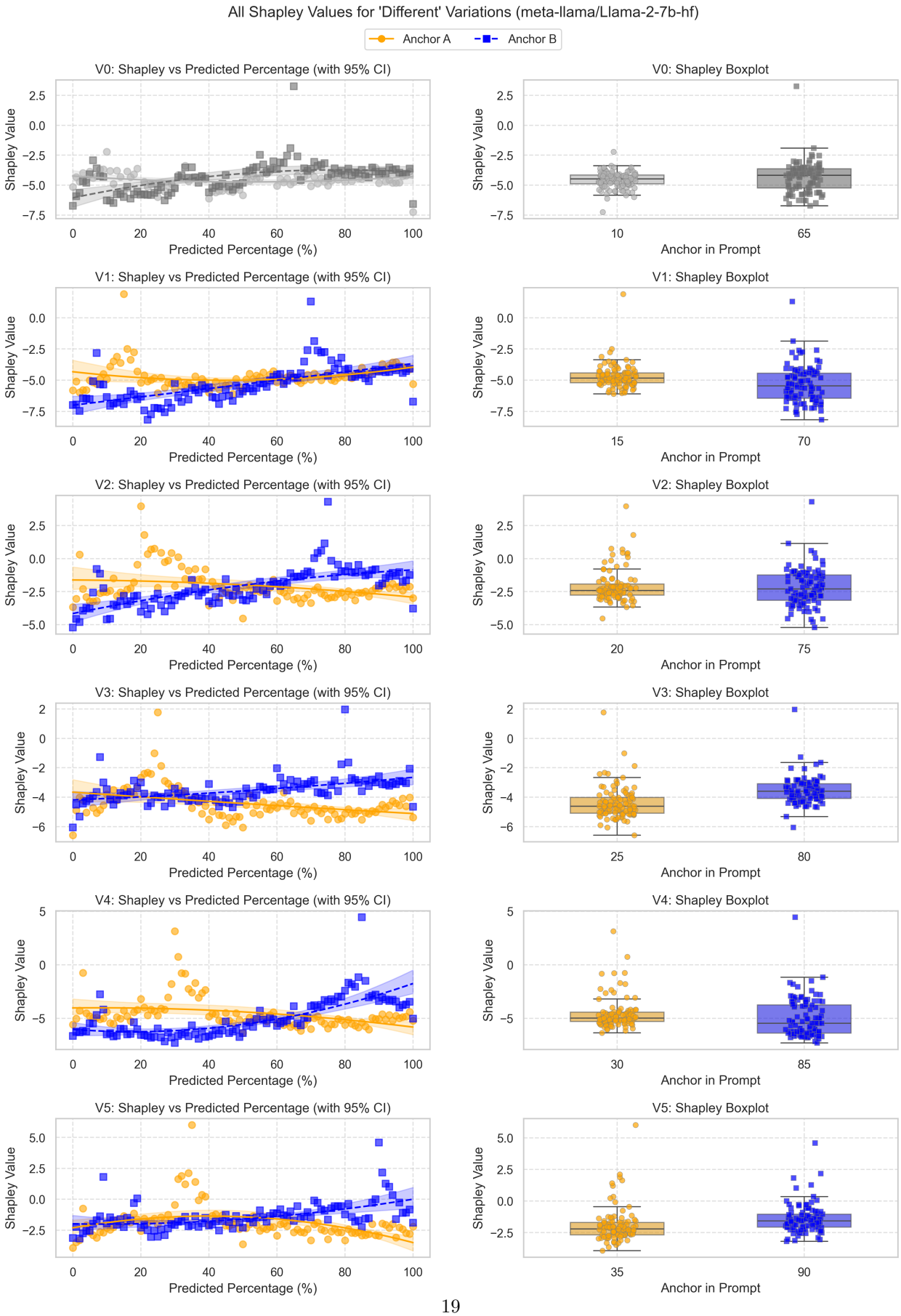

Figure 10: Attribution under different anchors for Llama-2-7b-hf.

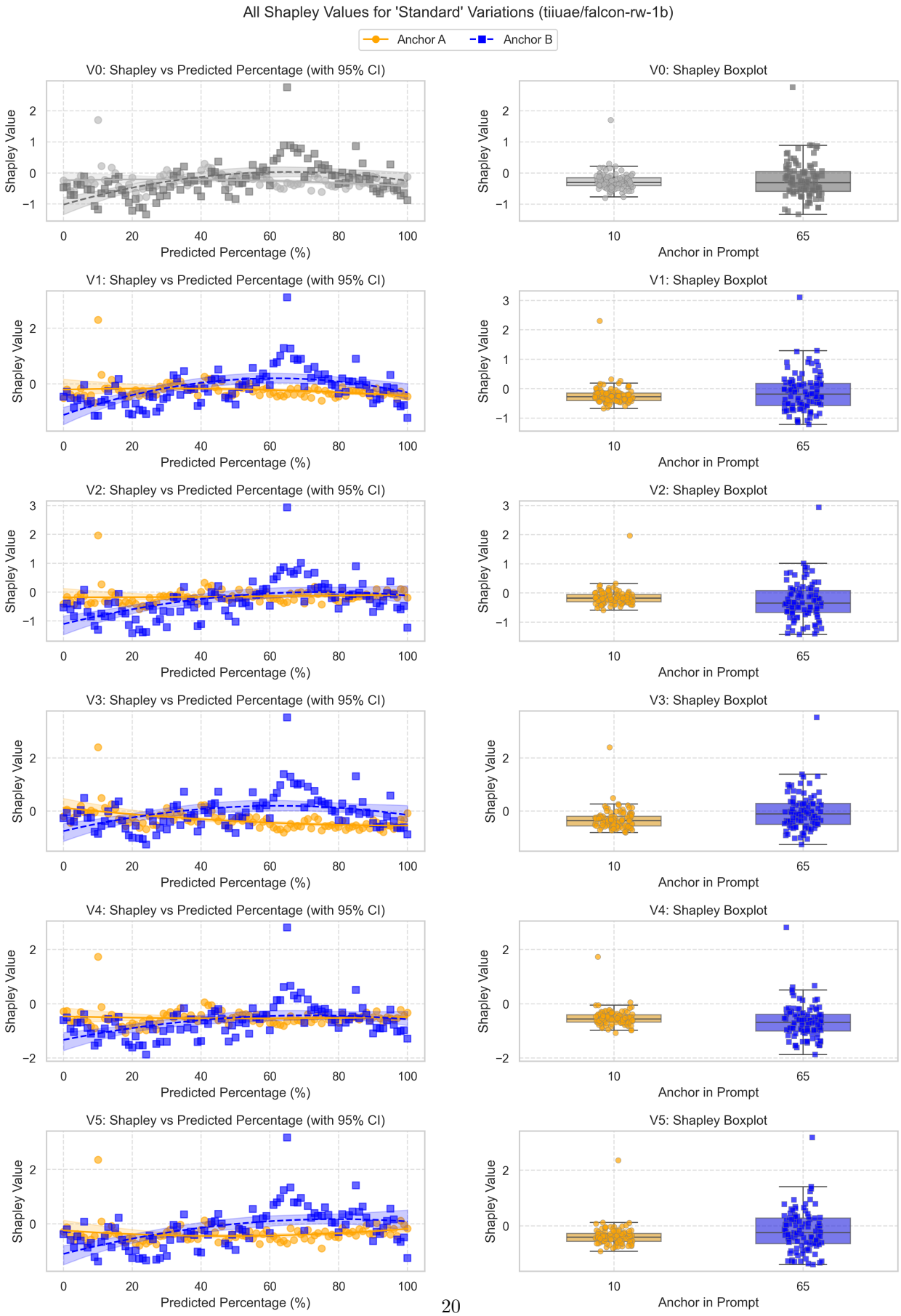

Figure 11: Attribution under standard anchors for Falcon-rw-1b.

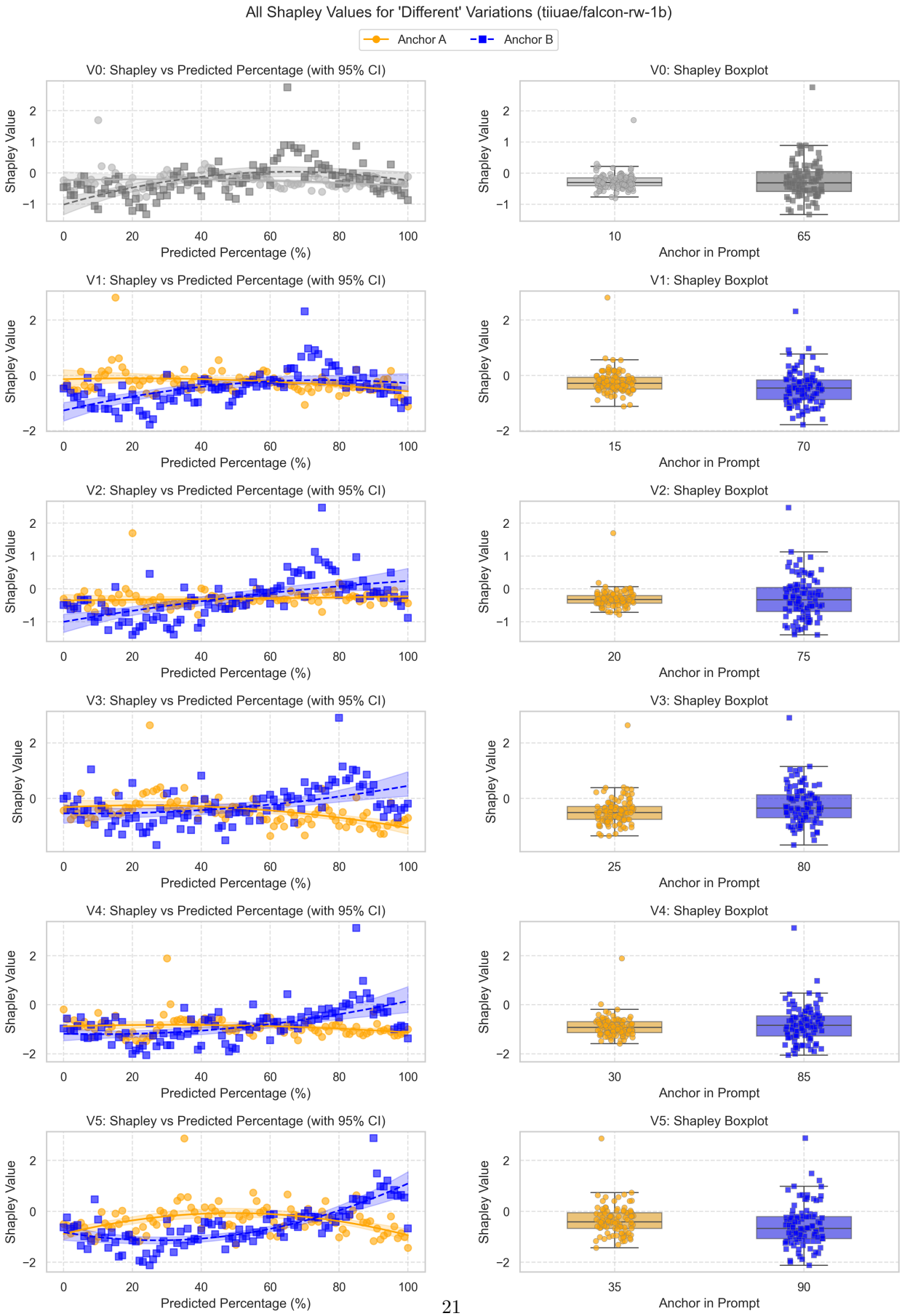

Figure 12: Attribution under different anchors for Falcon-rw-1b.

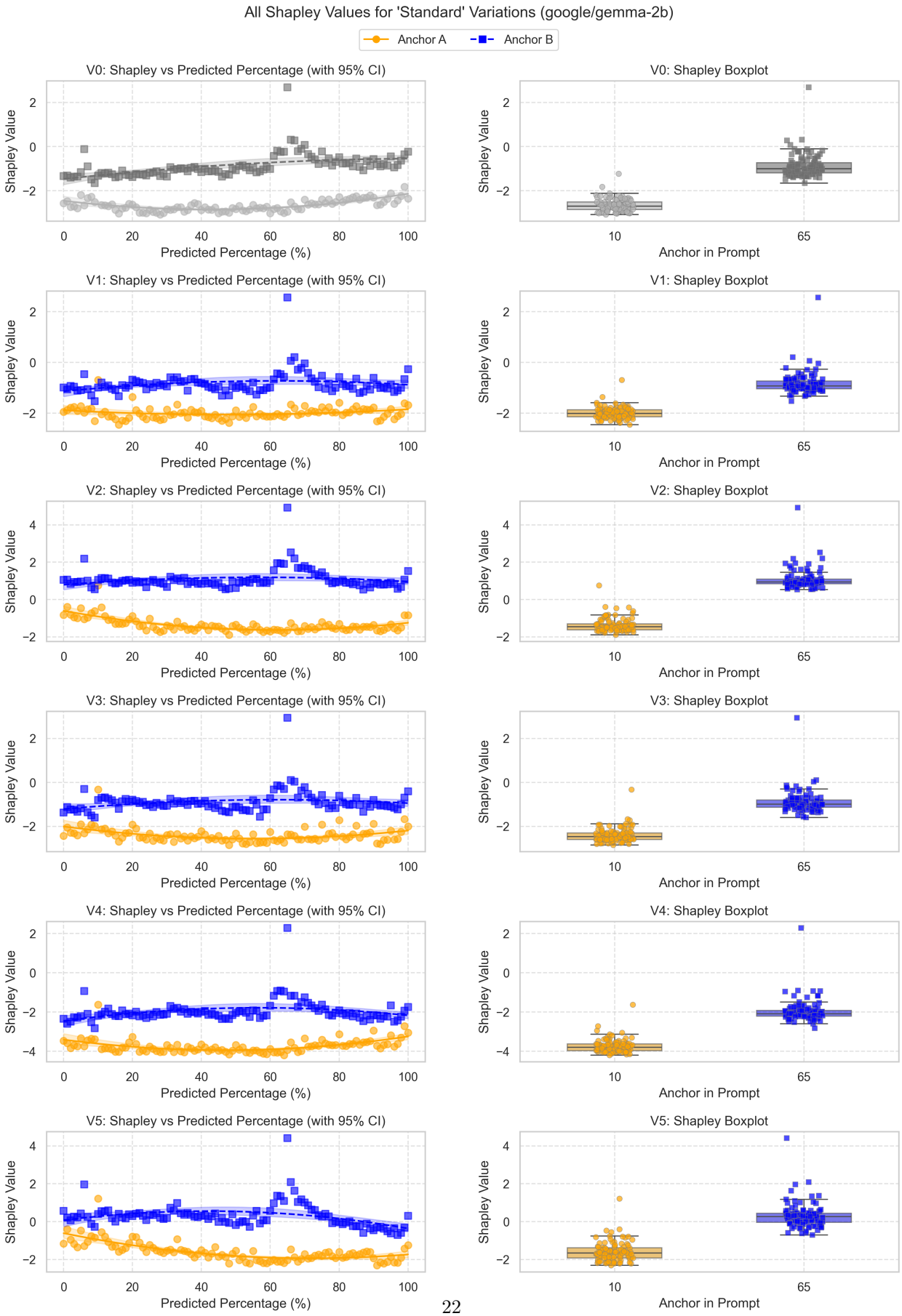

Figure 13: Attribution under standard anchors for Gemma-2b.

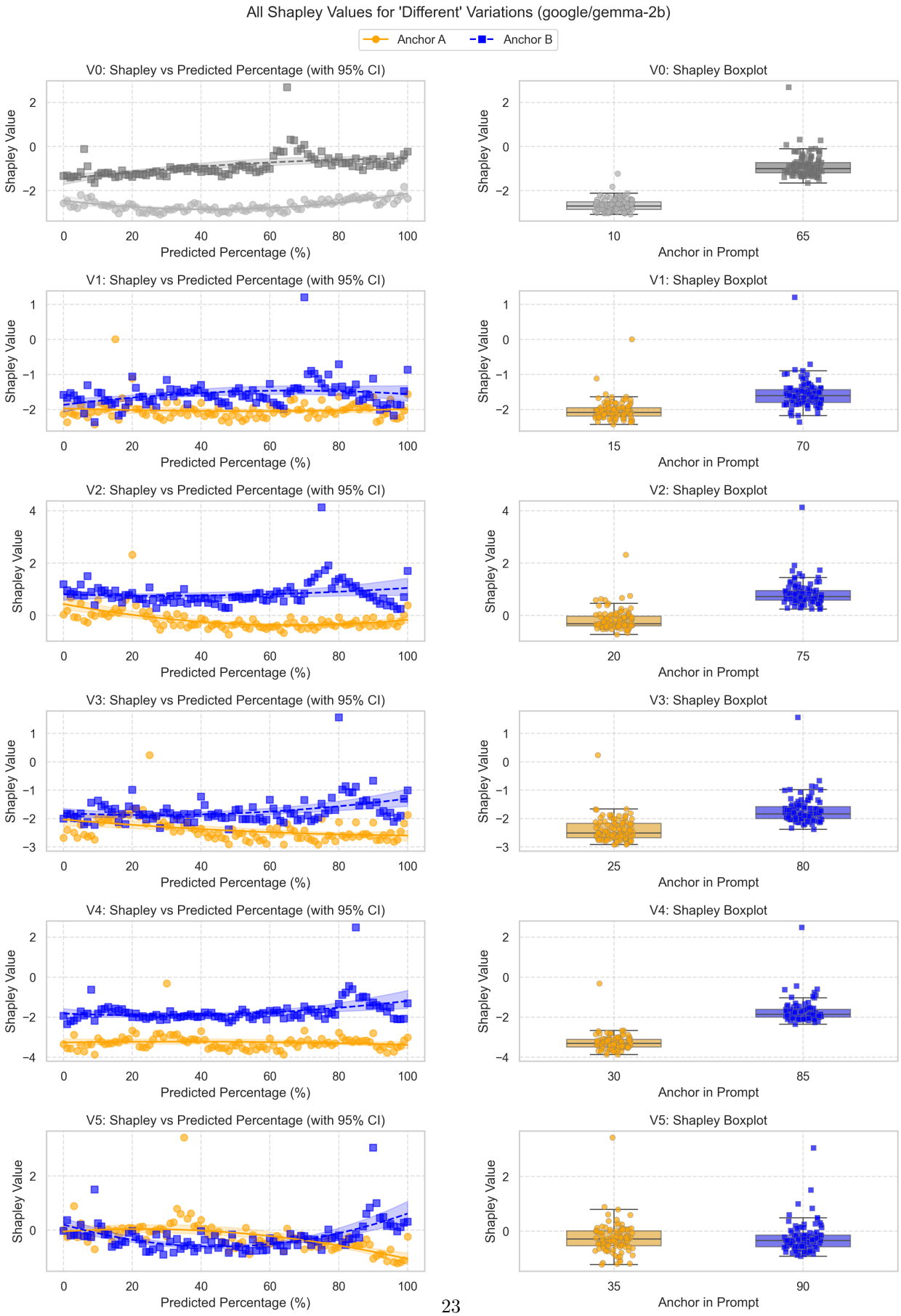

Figure 14: Attribution under different anchors for Gemma-2b.

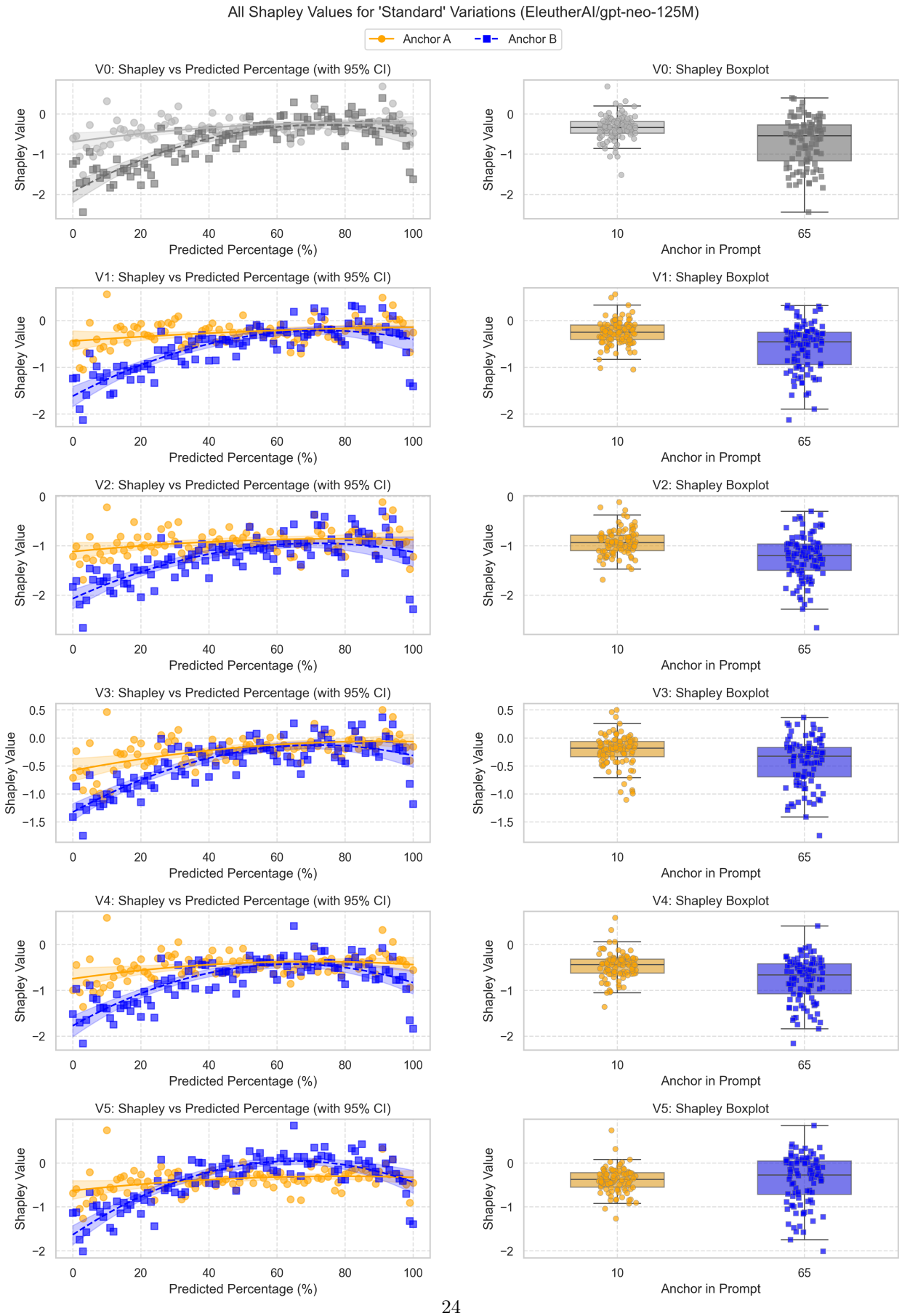

Figure 15: Attribution under standard anchors for gpt-neo-125M.

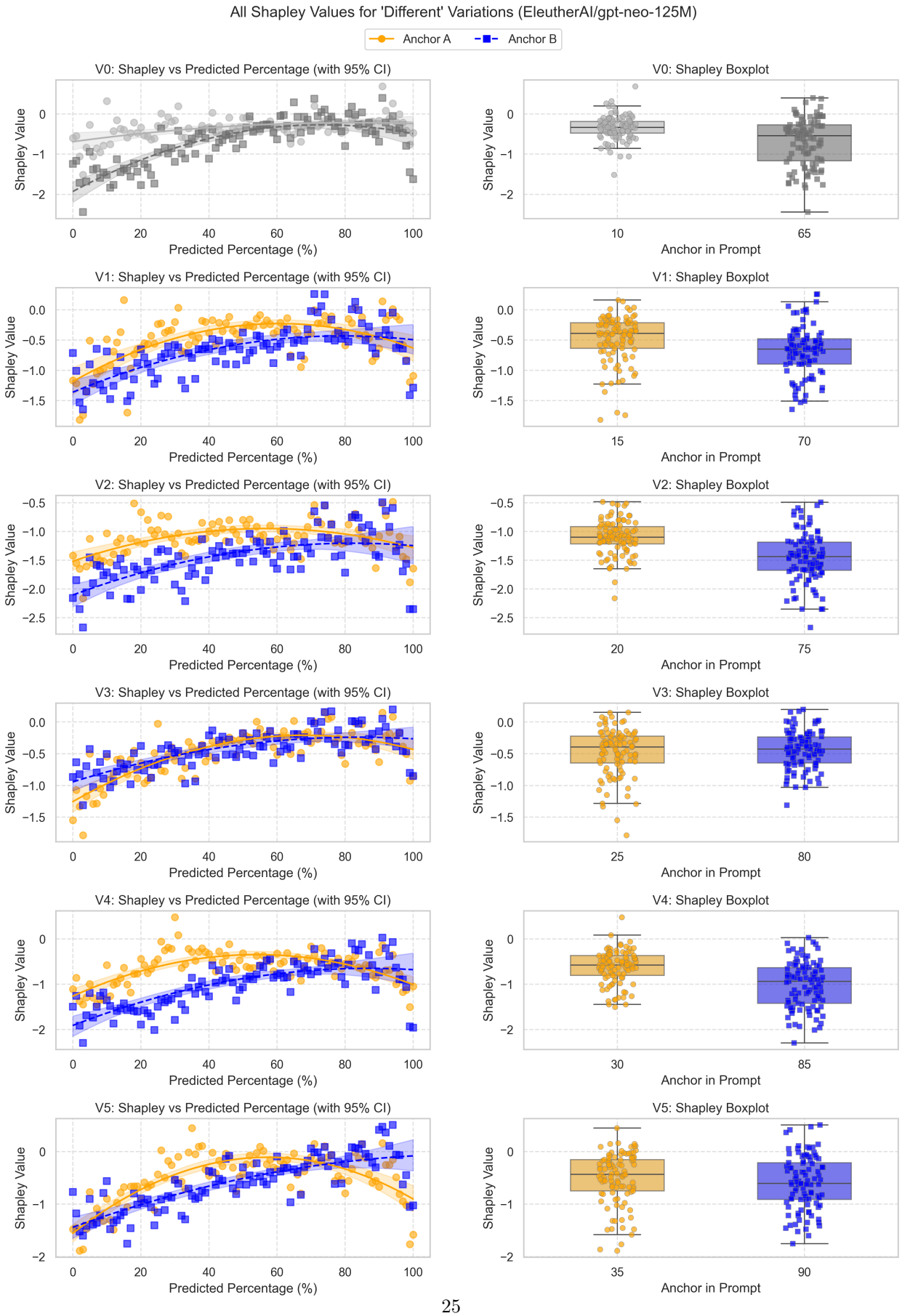

Figure 16: Attribution under different anchors for gpt-neo-125M.

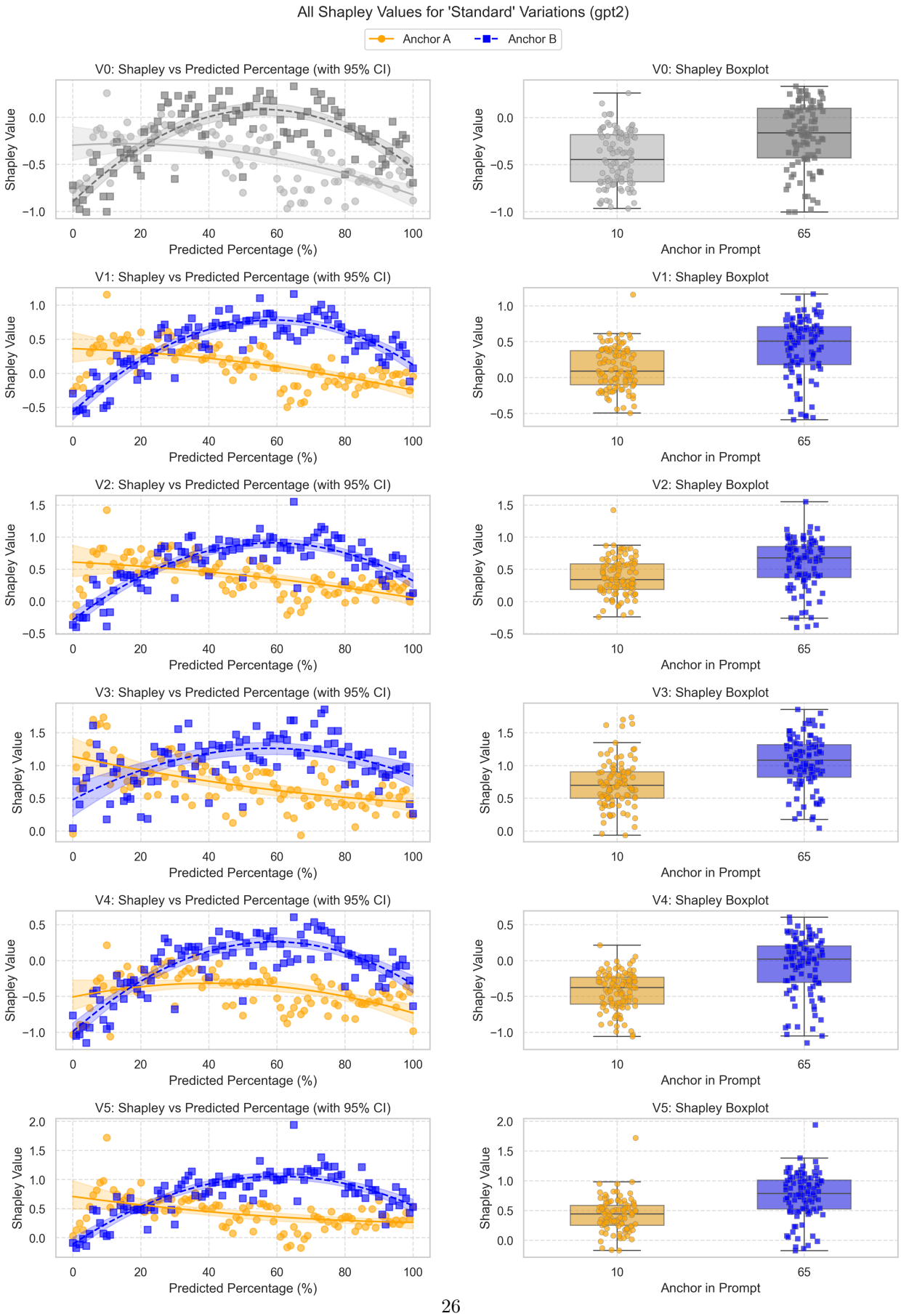

Figure 17: Attribution under standard anchors for GPT-2.

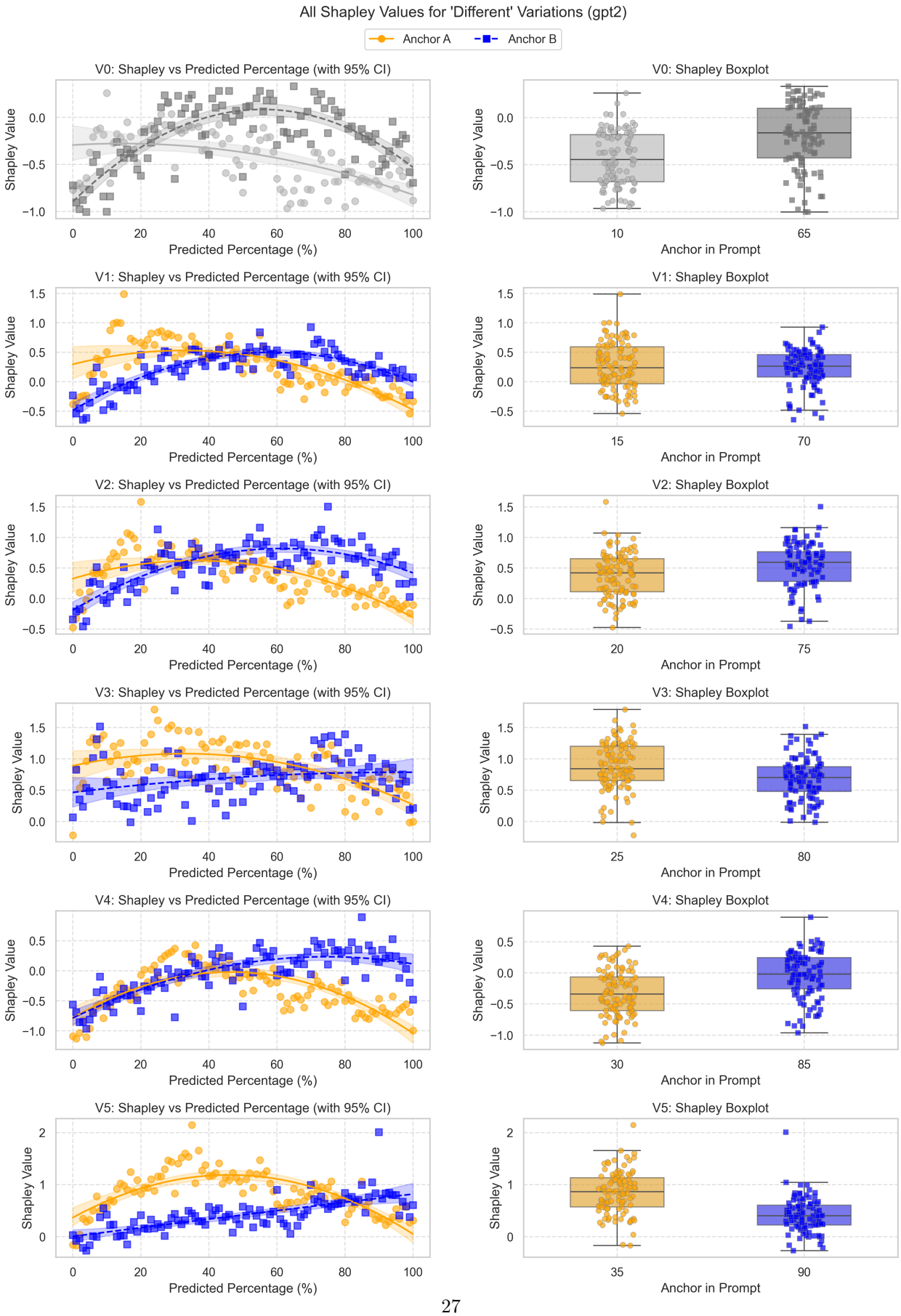

Figure 18: Attribution under different anchors for GPT-2.

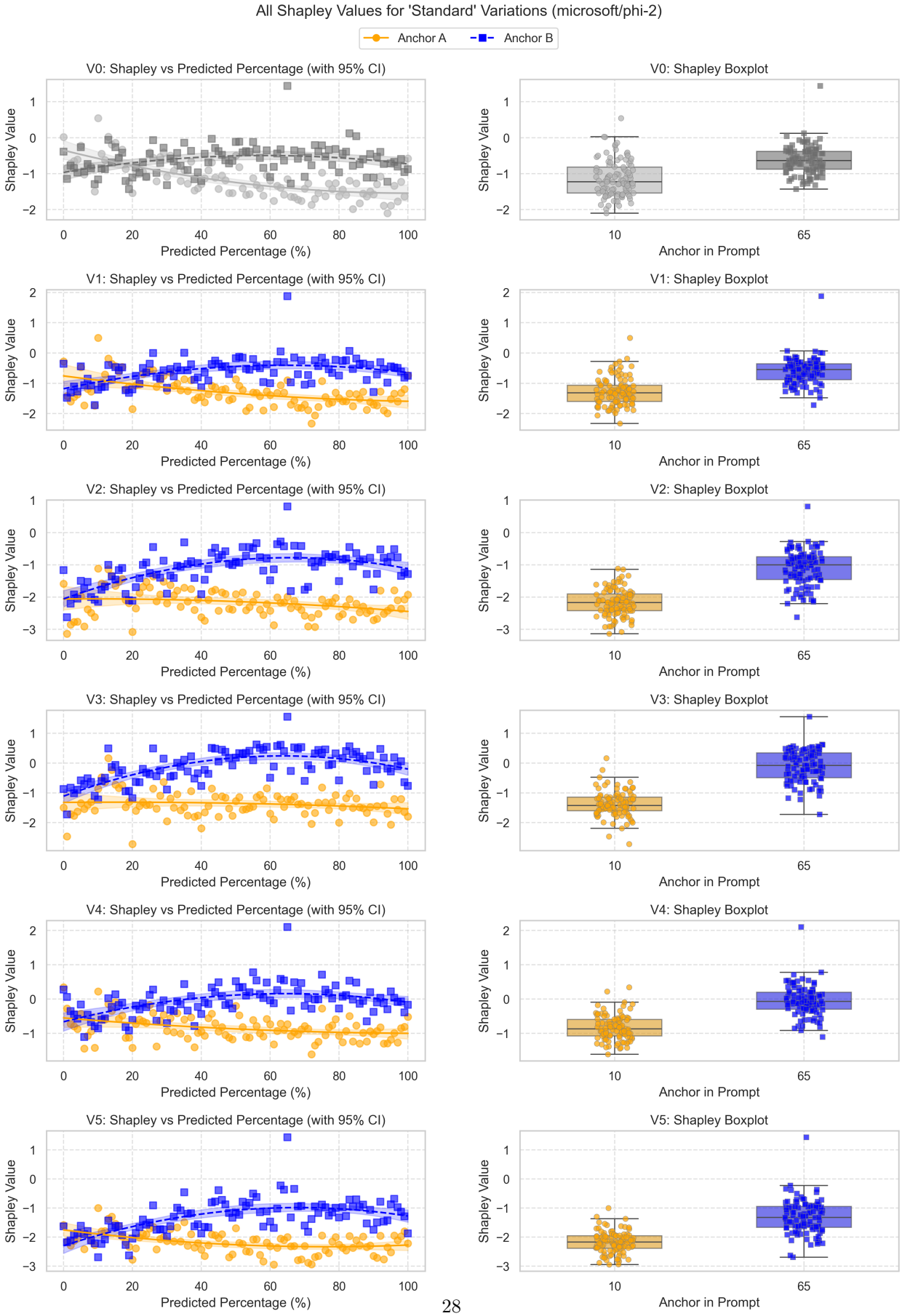

Figure 19: Attribution under standard anchors for phi-2.

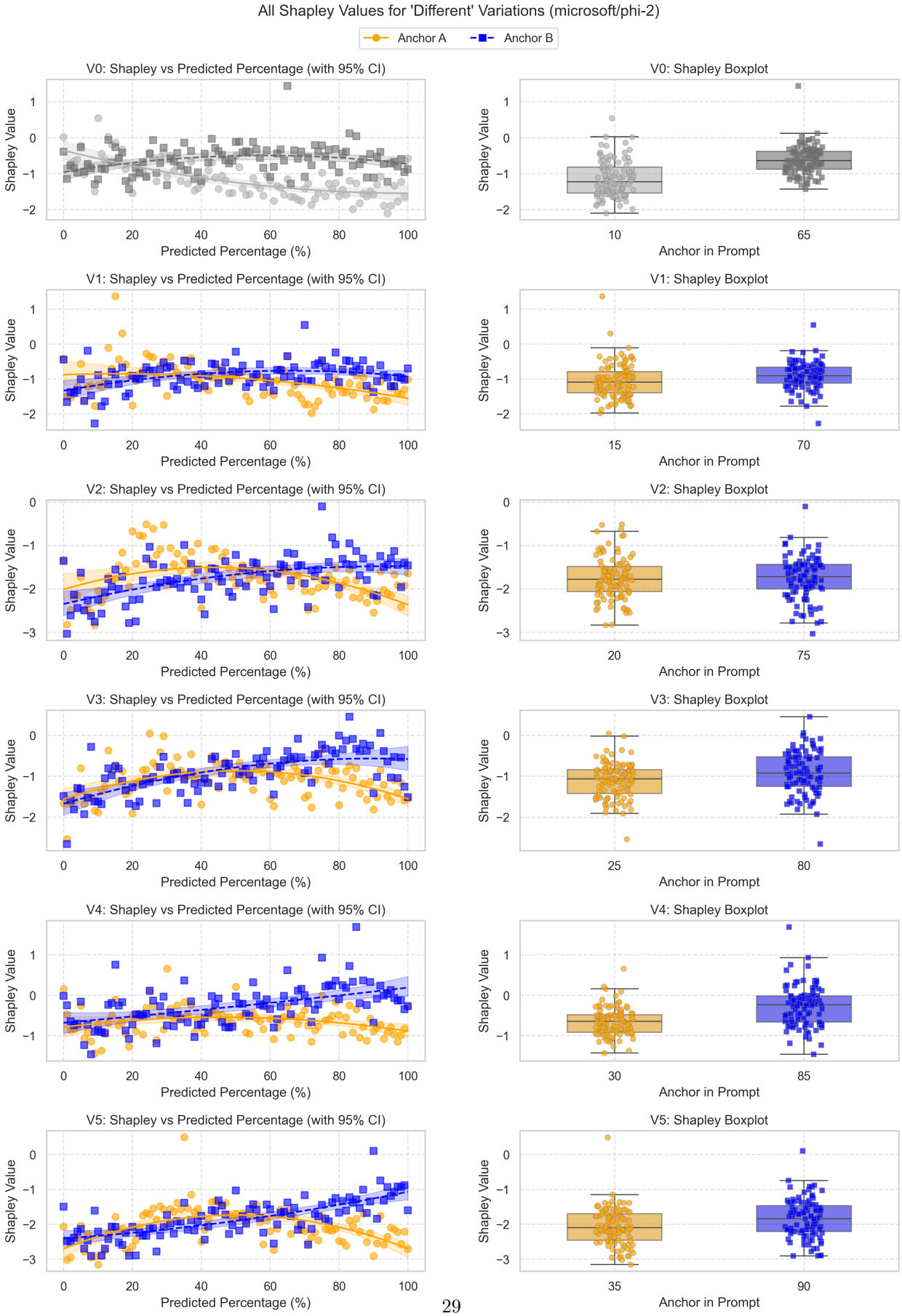

Figure 20: Attribution under moved anchors for phi-2.

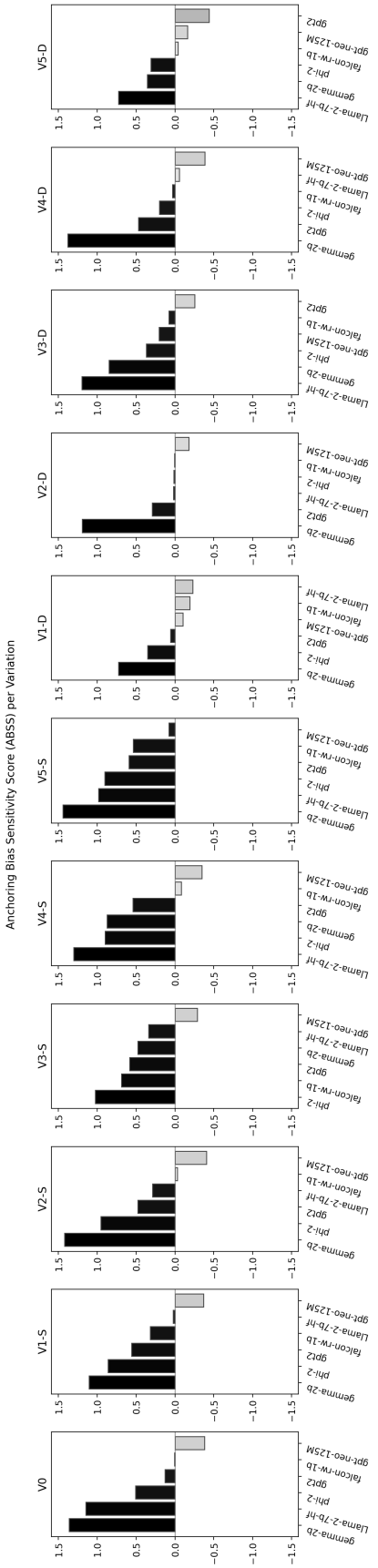

Figure 21: Anchoring Bias Sensitivity Score (ABSS) per variation across models.