AnchorBench: A Multi-Pathway Benchmark for the Anchoring Effect in LLMs

Yiderigun Borjigin[1] **Alexander Hermann**[2] **Christian Cyron**[2,3] **Roland Aydin**[1,4]
[1]Saarland University [2]Hamburg University of Technology [3]Helmholtz-Zentrum Hereon
[4]German Research Centre for Artificial Intelligence (DFKI)

Abstract

The anchoring effect is a cognitive bias in which an initial reference value shifts a later judgment toward itself. This effect is well established in human judgment and decision-making, and recent work suggests that large language models (LLMs) exhibit similar behavior. However, existing work on anchoring in LLMs typically evaluates only a narrow set of anchor pathways and rarely distinguishes irrelevant from plausible anchors. We introduce ANCHORBENCH, a benchmark for the anchoring effect in LLMs that evaluates multiple anchor pathways under an explicit anchor relevance axis. Across fourteen models, including ten open-weight models and four frontier API models, and a large set of controlled prompts, we find that (1) anchoring is strongly pathway-dependent, (2) plausible anchors usually induce larger shifts than irrelevant ones when introduced through stronger pathways, (3) anchor influence generally weakens as the anchor moves farther from the evidence-supported answer, most clearly on External and RAG, and (4) high task accuracy on the anchor-free control condition (Acc_{10}: answers within 10 points of gold) does not guarantee robustness: even frontier API models above 95% control accuracy remain susceptible to plausible anchors.

1 Introduction

Large language models (LLMs) are moving beyond text generation into decision-support tasks that require quantitative or evidence-based judgment, including medical question answering (Singhal et al., 2025) and time-series forecasting (Jin et al., 2024; Gruver et al., 2023). In these settings, average accuracy is not enough. A model may give a reasonable answer but still be influenced by a value that appears in the surrounding context, such as an earlier guess in the conversation, a value in a provided example, a retrieved document, or a tool output. This kind of error is especially concerning in high-stakes areas such as medicine. We study this failure mode through the lens of *anchoring*: the tendency for judgments to shift toward a reference value, as introduced by Tversky & Kahneman (1974).

Recent work shows that LLMs exhibit broader human-like cognitive biases across reasoning and decision-making (Jones & Steinhardt, 2022; Hagendorff et al., 2023; Echterhoff et al., 2024; Itzhak et al., 2024), and more directly that anchoring appears in LLMs (Nguyen, 2024; Lou & Sun, 2025; Huang et al., 2025; Takenami et al., 2025). However, the existing literature still leaves two important gaps. First, most prior studies test only one or two anchor pathways, usually prompt-embedded anchors. They rarely compare multiple realistic anchor pathways side by side. Second, they rarely distinguish clearly between unjustified shifts toward irrelevant anchors and potentially reasonable shifts toward plausible anchors.

We address these gaps with ANCHORBENCH,[1] a controlled diagnostic for the anchoring effect in LLMs. As shown in Figure 1, the benchmark covers five anchor pathways that correspond to standard ways context reaches a deployed model: External (the prompt), History (conversation), In-Context Learning (ICL, demonstrations), Retrieval-Augmented Generation (RAG, retrieved documents), and Tool (tool outputs). These pathways also have

[1]Code and data: https://github.com/Ydrg9989/AnchorBench.

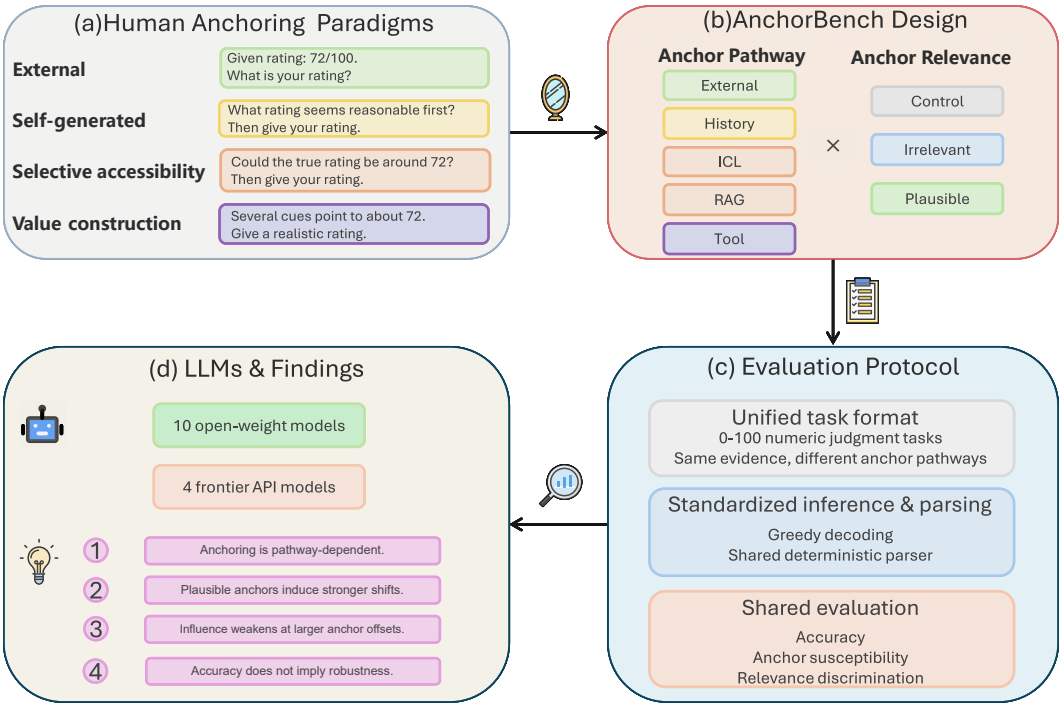

Figure 1: ANCHORBENCH is a unified benchmark for studying anchoring in LLMs, (a) inspired by four human anchoring paradigms, (b) ANCHORBENCH varies anchor pathway and anchor relevance, (c) evaluates them in a shared numeric judgment setup with standardized inference and metrics, (d) compares anchoring behavior across fourteen LLMs.

loose parallels in the human anchoring literature (Tversky & Kahneman, 1974; Epley & Gilovich, 2001; Mussweiler & Strack, 1999; Chapman & Johnson, 1999), which we use as informal motivation rather than strict one-to-one equivalences. It also introduces an explicit anchor relevance axis with control, irrelevant, and plausible conditions. Because every item contains structured numeric evidence and a deterministic gold answer, the benchmark measures not only whether outputs shift, but also whether the shift is justified.

Across the experiments, we find that (1) anchoring is strongly pathway-dependent, (2) plausible anchors usually produce larger shifts than irrelevant anchors when the pathway is strong, (3) anchor influence weakens as the anchor moves farther from the underlying evidence, and (4) accuracy and robustness come apart: even frontier API models with very high control-condition accuracy (Acc_{10}: answers within 10 points of gold) remain measurably susceptible to anchoring, though at smaller magnitudes than open-weight models.

Our contributions are as follows:

1. **A multi-pathway diagnostic benchmark for LLMs.** We design five realistic anchor delivery pathways that mirror how context reaches a deployed model, loosely informed by classic human anchoring theory, and evaluate them under a shared framework.
2. **A relevance-aware evaluation design.** We separate control, irrelevant, and plausible anchors. This lets us distinguish unambiguous bias (any shift toward an irrelevant anchor, which carries no task-relevant information) from sensitivity to plausible anchors, where a bounded shift can be consistent with rational evidence integration but a sufficiently large shift cannot.
3. **A broad empirical comparison across model families and access regimes.** We evaluate fourteen models, including ten open-weight models and four frontier API models, across five suites and 9,000 condition-controlled prompts per model.

2 Related work

Human anchoring theory. Anchoring is a classic finding in judgment and decision-making: estimates are often drawn toward an initial value, even when that value is arbitrary or only weakly informative (Tversky & Kahneman, 1974). No single mechanism explains it. For externally provided anchors, selective-accessibility accounts argue that people test anchor-consistent hypotheses and retrieve anchor-consistent knowledge (Strack & Mussweiler, 1997; Mussweiler & Strack, 1999); for self-generated anchors, anchoring-and-adjustment accounts emphasize insufficient adjustment from an internal starting point (Epley & Gilovich, 2001); and value-construction accounts tie anchoring to how values are activated and assembled during judgment (Chapman & Johnson, 1999). Anchor effectiveness also depends on plausibility and extremity (Wegener et al., 2001; Furnham & Boo, 2011). The human effect is sizable and relevance-dependent: Teovanović (2019) reports Cohen's d from 0.14 to 1.00 across 24 items, and Li et al. (2021) finds stronger anchoring for related than random anchors—a reference scale we revisit in Appendix A.17.13. These distinctions motivate our design: some pathways correspond to classic external or self-generated anchoring, while others operationalize selective-accessibility- or value-construction-style influence through demonstrations, retrieval, and tools.

Anchoring effect in LLMs. Anchoring also appears in LLMs. Frontier models shift their estimates toward previously mentioned values, and prompt-based mitigations such as chain-of-thought, reflection, or explicit instructions to ignore the anchor give only limited and inconsistent relief (Nguyen, 2024; Lou & Sun, 2025); the effect persists in multi-turn price negotiation, where reasoning models are less susceptible (Takenami et al., 2025). SynAnchors (Huang et al., 2025) is the closest prior benchmark and is complementary: it varies the *magnitude* of in-prompt anchors and uses causal tracing, whereas AnchorBench fixes the numeric value and varies its *framing* (irrelevant vs. plausible) and *delivery pathway*. Where they overlap, our findings agree. Most of this work studies one pathway at a time, which leaves open how External, History, ICL, RAG, and Tool anchoring compare within a unified setup.

Cognitive bias in LLMs. Anchoring is one of several human cognitive biases documented in LLMs, which have been used as a lens on systematic model failure (Jones & Steinhardt, 2022; Hagendorff et al., 2023), studied as an emergent effect of instruction tuning (Itzhak et al., 2024), and observed in decision-making and in survey response, where models are unreliable stand-ins for human respondents (Echterhoff et al., 2024; Tjuatja et al., 2024); broader evaluations and surveys catalogue many such biases and their mitigations (Malberg et al., 2025; Sumita et al., 2024). The pattern is especially well documented in LLM-as-a-judge pipelines, where position and related judging biases distort model-based evaluation (Koo et al., 2024; Zheng et al., 2023; Wang et al., 2024; Ye et al., 2025; Wang et al., 2025) and evaluators show anchoring in multi-attribute scoring (Stureborg et al., 2024)—so the effect reaches evaluation pipelines, not only end-user tasks.

Context influence beyond anchoring. Numerical anchoring is one instance of a broader sensitivity to non-informative context, alongside sycophancy and persuasion, where models shift toward a user's view or a confidently framed claim (Sharma et al., 2024; Ranaldi & Pucci, 2025) and toward authoritative-sounding sources (Anagnostidis & Bulian, 2024; Zhou et al., 2023; Nguyen et al., 2025). We see anchoring as narrower: it isolates a single numeric value rather than opinions or arguments, which lets us measure the shift against a deterministic gold answer.

3 Benchmark setup and evaluation design

AnchorBench is a controlled diagnostic rather than a sample of organic user queries: every item is synthetic and built so that the gold answer is fixed by the evidence alone, which lets us measure anchoring as a deviation from a known target. It combines two controlled design factors: the anchor pathway, which determines how an anchor reaches

Shared prompt (External suite)

A SaaS company is pricing a premium add-on. The following customer research scores (0–100) indicate price tolerance.

Evidence:
- Online panel score: 62
- In-store intercept: 58
- Trade-off analysis: 65
- Retention predictor: 54
- Benchmark comparison: 61

[anchor sentence inserted here]

Given these signals, what is your best estimate for the WTP index on a 0–100 scale?
Return only a single integer 0–100 on the last line.

Gold: $y^* = 60$ (mean of evidence).

(a) Shared prompt

Anchor sentence by condition

Control:
No added sentence.

Irrelevant (a=23):
Note: This is assessment case #23 in the current evaluation batch.

Plausible (a=23):
Note: A recent industry report suggested the WTP index might be around 23 .

The anchor value a is identical in both conditions; only the framing differs. Irrelevant framings use metadata (case IDs, queue numbers), whereas plausible framings cite a prior estimate.

(b) Anchor conditions

Figure 2: Example ANCHORBENCH item (External suite, pricing domain). All conditions share the same scenario, evidence, question, and answer instruction. Only the anchor sentence differs. Additional examples are in Appendix A.2.

the model, and anchor relevance, which determines how the anchor is framed relative to the task.

3.1 Item construction

Each item is a numeric judgment task on a shared 0–100 scale across six domains (pricing, operations, logistics, resource consumption, market adoption, compliance). As illustrated in Figure 2a, the model receives a set of numeric *evidence ratings* and must return a single integer estimate. The gold answer is the rounded arithmetic mean of the visible ratings:

$$y^* \;=\; \mathrm{round}\Big(\tfrac{1}{k}\textstyle\sum_{j=1}^{k} e_j\Big), \tag{1}$$

where $e_1, \ldots, e_k$ are the ratings shown to the model. This makes the task a deterministic aggregation problem: a model that correctly averages the given numbers matches gold exactly.

Item difficulty is governed by two parameters. A latent center $\theta \sim \mathrm{Unif}[30, 70]$ sets the evidence range; ratings are sampled as $e_j \sim \mathcal{N}(\theta, \sigma^2)$, clipped to $[0, 100]$. **Easy** items present all five ratings with low noise (σ=8). **Hard** items hide two of the five ratings, add a conflicting value, and raise noise (σ=15), leaving fewer and less consistent signals for aggregation. The six core domains are business-oriented, but a pilot adding medical, legal, and consumer domains shows the same $\mathrm{UAI}_{\mathrm{pls}} > \mathrm{UAI}_{\mathrm{irr}} > 0$ pattern (Appendix A.17.5).

Provenance of the items. Items come from author-designed templates: scenarios, evidence labels, questions, and anchor framings are drawn from pools generated with LLM assistance and verified by the authors, while the numeric evidence, anchor values, and gold answers are produced by seeded sampling, never by a model. Holding the numeric value and its position fixed while varying only the framing sentence is the matched control the relevance axis needs, and organic queries cannot supply it—at a cost in external validity we state rather than claim away.

3.2 Anchor conditions and relevance axis

Each item is presented under five matched conditions, illustrated in Figure 2b. The *control* condition presents the task without any anchor. The four *anchored* conditions cross two relevance levels—*irrelevant* and *plausible*—with a low and a high anchor direction each, yielding: `control`, `irrelevant_low/high`, and `plausible_low/high`.

The defining feature of the relevance axis is that an irrelevant and a plausible condition in the same direction use an *identical* numeric anchor value a; only the surrounding sentence differs. An irrelevant framing presents a as incidental metadata (e.g., "assessment case #a in the current batch"); a plausible framing presents it as a substantive prior estimate (e.g., "a recent industry report suggested around a"). Since the gold answer is determined entirely by the evidence (uninformative about a), any difference between irrelevant and plausible responses isolates the effect of framing, not the anchor value itself.

As a concrete example, on one real External item Qwen-7B answers 43 (the evidence mean) in control, still 43 under the irrelevant framing of a=85, but 75 under the plausible framing—a +32 shift that isolates the framing alone. Per-pathway worked examples and further case studies are in Appendices A.2 and A.17.12.

Anchor placement. Low and high anchors are placed symmetrically around the evidence center θ: $a_{\text{low}} = \max(0, \theta-\delta)$ and $a_{\text{high}} = \min(100, \theta+\delta)$, with offsets $\delta \in \{15, 25, 40\}$. Varying the offset lets us test how anchor influence changes as the anchor moves farther from the evidence-supported answer (§4.3). The History suite is an exception: its anchor is the model's own Stage 1 response rather than a designer-specified value (§3.3), so its offset varies by item. Full generation and validation details are in Appendix A.1.

These five conditions are each delivered through a distinct anchor pathway, implemented as a benchmark suite (§3.3).

3.3 Benchmark suites

The five suites hold the numeric judgment task fixed and vary only the pathway by which an anchor reaches the model; they are controlled operationalizations rather than end-to-end deployment systems (worked prompts in Appendix A.2).

External places an anchor sentence directly in the user prompt. The prompt format is identical across all models and conditions; only the framing sentence varies.

History uses a two-stage conversation: Stage 1 shows partial evidence to elicit a response near the target anchor, and Stage 2 reveals the full evidence, so the anchor is the model's *own* Stage-1 answer. The single-stage control creates a format asymmetry analyzed in Appendix A.10.

ICL has two variants: *ICL-metadata* (the main benchmark) places anchors only in incidental demonstration metadata (case IDs, batch numbers)—a deliberately weak manipulation; *ICL-dist* instead uses demonstration *answers* matching the anchor distribution, a stronger and more realistic few-shot signal (Appendix A.11).

RAG embeds the anchor in one document of a frozen three-document mini-corpus. The other two documents contain factual evidence consistent with the gold answer. Prompt format is identical across models.

Tool returns the anchor in a tool-call response. Models supporting structured tool messages (Qwen, Llama) receive them natively; others (Gemma, OLMo, APIs) get an equivalent plaintext rendering—a known confound compared within-model in Appendix A.12.

Cross-suite comparability. EXTERNAL and RAG have no format confounds and provide the strongest cross-model evidence. HISTORY and TOOL have format-related confounds (described above); we interpret their magnitudes qualitatively.

Each suite contains 360 items (6 domains $\times$ 60 items) in five conditions, giving 1,800 prompts per model per suite. All data generation is seed-controlled and deterministic (Appendix A.1).

3.4 Evaluation metrics

All suites share the same metrics, so *metric semantics* are comparable; recall that History and Tool magnitudes are format-confounded and should be read qualitatively (§3.3). We

measure three dimensions of behavior: *task accuracy*, *anchor susceptibility*, and *relevance discrimination*.

Notation. Let y_i^* be the gold answer for item i, $y_{\text{ctrl},i}$ the control (no-anchor) response, and $y_{\text{anchor},i,r}$ the response under relevance $r \in \{\text{irr}, \text{pls}\}$. Let $a_{i,r}$ be the anchor value, constructed in the prompt for External/ICL/RAG/Tool and equal to the model's Stage 1 answer for History.

(a) Task accuracy (control-only). We measure task accuracy only on control responses, because these reflect task performance without anchoring. We report both mean absolute error and a coarse accuracy-within-tolerance measure:

$$\text{MAE}_\text{c} = \tfrac{1}{n}\textstyle\sum_i |y_{\text{ctrl},i} - y_i^*|, \quad \text{Acc}_{10} = \tfrac{1}{n}\textstyle\sum_i \mathbf{1}\big[|y_{\text{ctrl},i} - y_i^*| \leq 10\big], \tag{2}$$

MAE_c is average prediction error on the native 0–100 scale; Acc_{10} is the fraction of control responses within 10 points (10% of the range) of gold. These capture task competence but do not by themselves measure anchoring.

(b) Anchor susceptibility. *Unified Anchor Influence* (UAI) measures what fraction of the anchor–control gap is closed by the anchored response:

$$\text{UAI}_{i,r} = \frac{y_{\text{anchor},i,r} - y_{\text{ctrl},i}}{a_{i,r} - y_{\text{ctrl},i}}. \tag{3}$$

$\text{UAI} = 0$ indicates no shift and $\text{UAI} = 1$ indicates full movement to the anchor. Items where the control response already falls within $\varepsilon = 3$ points of the anchor ($|a_{i,r} - y_{\text{ctrl},i}| < 3$) are excluded, because the near-zero denominator makes the ratio unstable. Exclusion rates are low and conclusions are stable across $\varepsilon \in \{1, 3, 5\}$ (Appendix A.6).

Because UAI is undefined for excluded items, we also report the denominator-free *Toward-Anchor Rate* (TAR):

$$\text{TAR}_{i,r} = \mathbf{1}\big[(y_{\text{anchor},i,r} - y_{\text{ctrl},i}) \cdot (a_{i,r} - y_{\text{ctrl},i}) > 0\big]. \tag{4}$$

$\text{TAR} > 0.50$ indicates a systematic shift toward the anchor. TAR covers all items and is independent of the gap magnitude, so the two metrics together guard against exclusion-rule artifacts.

(c) Relevance discrimination. The discrimination gap captures whether a model responds differently to plausible and irrelevant anchors:

$$\text{Disc}_\Delta = \text{UAI}_\text{pls} - \text{UAI}_\text{irr}. \tag{5}$$

$\text{Disc}_\Delta > 0$ indicates greater sensitivity to plausible than irrelevant anchors, and the reverse if negative. It should be read alongside absolute UAI: a model with $\text{Disc}_\Delta > 0$ but high UAI_irr is still strongly susceptible, so positive discrimination alone is not evidence of robustness.

3.5 Experimental setup

We evaluate fourteen instruction-tuned models: ten open-weight models from four families—Llama 3.1/3.2 (Grattafiori et al., 2024), Qwen2.5 (Qwen Team et al., 2025), Gemma 3 (Gemma Team, 2025), and OLMo 2 (OLMo Team et al., 2025), plus four frontier API models from OpenAI (Singh et al., 2025), Anthropic (Anthropic, 2025), Google (Comanici et al., 2025), and xAI (xAI, 2025) (full model panel: Appendix A.1). All models use greedy decoding (temperature 0); answers are extracted by deterministic hierarchical parsing (median parse rate 99.9%). A sampled-decoding check (Appendix A.15) and a prompt-based mitigation probe (Appendix A.16) confirm the main conclusions are stable.

Model	EXTERNAL			HISTORY			ICL			RAG			TOOL		
	Acc	U_{irr}	U_{pls}	Acc	U_{irr}	U_{pls}	Acc	U_{irr}	U_{pls}	Acc	U_{irr}	U_{pls}	Acc	U_{irr}	U_{pls}
2.5-1.5B	34%	0.02	0.20	32%	0.29	**0.98**	59%	0.01	0.00	22%	0.02	0.23	46%	0.26	**0.82**
2.5-3B	64%	0.01	0.09	61%	−0.08	0.49	78%	−0.04	−0.04	74%	−0.02	0.08	89%	0.08	0.18
2.5-7B	72%	0.01	0.27	79%	−0.06	0.37	88%	−0.03	−0.01	64%	−0.00	0.20	99%	0.00	0.17
3.2-1B	47%	0.13	0.15	49%	0.48	0.14	55%	0.30	**0.22**	60%	0.10	0.09	—	—	—
3.2-3B	69%	0.14	0.29	72%	0.15	0.73	81%	0.07	0.14	72%	0.14	0.22	95%	0.03	0.21
3.1-8B	83%	0.07	0.36	83%	0.13	0.23	77%	0.03	0.15	85%	−0.02	0.12	73%	0.31	0.48
3-1B-IT	37%	0.02	0.18	39%	−0.17	0.13	46%	−0.00	−0.03	37%	−0.00	0.04	26%	−0.05	−0.03
3-4B-IT	55%	0.20	**0.43**	56%	0.23	0.50	39%	0.00	−0.06	48%	0.05	0.01	63%	0.18	0.21
2-13B	89%	−0.01	0.16	86%	−0.00	0.32	57%	0.03	0.02	87%	−0.00	0.05	71%	0.09	0.07
2-32B	53%	0.05	0.40	54%	−0.30	0.75	70%	−0.02	−0.01	58%	0.01	**0.45**	72%	0.01	0.02
5.4-MINI	99%	0.00	0.14	94%	−0.02	0.04	99%	−0.00	−0.01	99%	0.01	0.09	99%	0.00	0.06
HAIKU 4.5	100%	0.00	0.12	98%	−0.02	−0.06	97%	−0.00	−0.02	98%	0.00	0.06	98%	−0.01	0.01
2.5-FLASH	96%	0.02	0.07	95%	0.03	−0.00	99%	−0.00	0.00	84%	0.08	0.17	98%	0.03	0.06
3-MINI	99%	−0.00	0.16	98%	0.10	0.10	99%	−0.00	0.00	99%	−0.01	0.02	99%	−0.01	0.02

Table 1: Task accuracy (Acc, $=\mathrm{Acc}_{10}$, control-only) and absolute anchor influence on irrelevant (U_{irr}) and plausible (U_{pls}) anchors for fourteen models across five suites. Any positive U_{irr} is an unjustified shift toward a value that carries no task information; U_{pls} is sensitivity to a plausibly framed value, and the plausible–irrelevant gap $\mathrm{Disc}_\Delta = U_{pls} - U_{irr}$ is analyzed in §4.2. Anchoring varies strongly by pathway: External and RAG show the broadest positive effects; ICL is near zero; History and Tool vary by model. Bold marks the largest U_{pls} per suite. Caveats: Llama-1B/Tool is undefined (parse rate 0.9%); History uses a single-stage control (Appendix A.10); Tool format varies by model (Appendix A.12). For space, instruction-tuned model names are abbreviated by omitting the "-Instruct" suffix; full model identifiers are listed in Table 3.

4 Empirical findings

4.1 Finding 1: Anchoring is pathway-dependent

Table 1 reports task accuracy (Acc_{10}, control-only) and absolute anchor influence on irrelevant (UAI_{irr}) and plausible (UAI_{pls}) anchors for all fourteen models across five suites. We report both UAI columns rather than only their difference, because any positive UAI_{irr} is unjustified bias while UAI_{pls} is sensitivity to a plausibly framed value. The central finding is that *susceptibility depends strongly on the anchor pathway*.

EXTERNAL and RAG show the broadest positive discrimination. ICL (metadata) is uniformly near zero. HISTORY and TOOL show substantial effects, but their magnitudes are inflated by format confounds: HISTORY compares a two-stage anchored protocol against a single-stage control (matched-format analysis reduces Disc_Δ by up to 70% for some models; Appendix A.10), and TOOL mixes structured and plaintext tool messages across model families (Appendix A.12). The Tool effect is also provenance-sensitive: a model-elicited call or a noisy JSON envelope barely moves the panel mean but splits individual models (Qwen-7B collapses, Gemma-4B strengthens; Appendix A.17.10), so we read Tool numbers as robust on average but not per model. No model is consistently best or worst across suites, and across the ten open-weight models the range of suite-mean Disc_Δ is 0.40 (bootstrap 95% CI [0.20, 0.60]; Table 10).

4.2 Finding 2: Plausible anchors induce stronger shifts

Across the 69 model–suite cells with computable Disc_Δ, 55 (80%) show positive discrimination—greater susceptibility to plausible than irrelevant anchors—rising to 48/55 (87%) once ICL, where both UAI values are near zero, is excluded. Wilcoxon signed-rank tests confirm the pattern on EXTERNAL, RAG, and TOOL ($p_{BH} < 0.01$) and HISTORY ($p_{BH} \approx 0.02$), but not ICL ($p_{BH} = 0.81$, n.s.; Table 9).

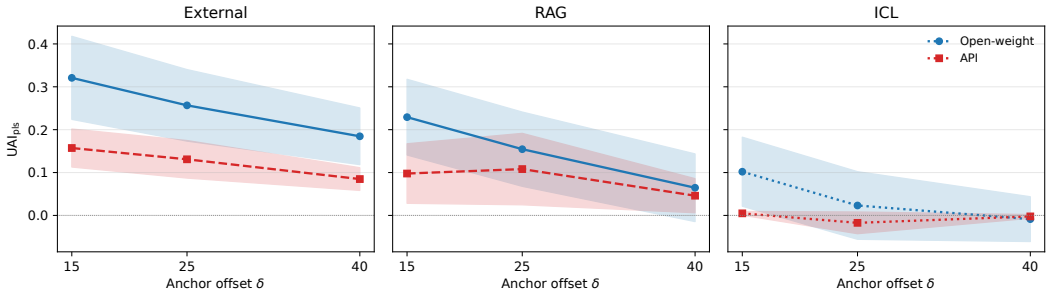

Figure 3: Anchor influence attenuates with distance (suites with designer-controlled offsets only). UAI_{pls} decreases monotonically with anchor offset δ on External and RAG; ICL (dotted) stays near zero at all distances. Lines show suite means across models; shaded bands are 95% CIs (± 1.96 SE). API models show monotonic attenuation on External; on RAG, they peak at δ=25 before decreasing. History is omitted because its offsets are not designer-controlled (§3.3); Tool is omitted due to format confounds (Appendix A.12).

The two relevance levels mean different things: a shift toward an *irrelevant* anchor is bias by construction, while a shift toward a *plausible* anchor is only a problem once it exceeds what a rational update could justify (§4.4). Either way, positive Disc_Δ is not itself evidence of robustness; low overall susceptibility remains the desirable outcome.

ICL: a weak floor, not immunity. The near-zero standard ICL result reflects a deliberately weak manipulation (anchors only in demonstration metadata); a distribution-matching ICL-dist variant raises mean UAI_{pls} from $\approx$0.03 to 0.16 ($\text{Disc}_\Delta > 0$ for 8/10 open-weight models), comparable to RAG (Appendix A.11).

Beyond the binary axis. Relevance is graded, so we check the two-level axis against a four-point credibility spectrum on EXTERNAL: panel-mean UAI is +0.09 for a *placebo* framing (the same number given as a document's age), +0.05 for irrelevant, +0.27 for plausible, and +0.49 for *authority*. Credibility modulates the shift, but the axis is not clean at the bottom—the placebo floor is above zero and above the irrelevant condition, which no purely rational account predicts (Appendix A.17.2). Applying the same manipulation at matched mild / standard / strong intensity across pathways puts RAG (0.05→0.08→0.18) one step below EXTERNAL (0.25→0.32→0.40) rather than differing in kind, with HISTORY non-monotone (Appendix A.17.4).

4.3 Finding 3: Anchor influence generally decreases with anchor offset

Figure 3 breaks down UAI_{pls} by anchor offset ($\delta \in \{15, 25, 40\}$) on the suites with designer-controlled offsets. On External and RAG, mean UAI_{pls} *decreases monotonically* with larger offsets (open-weight External: $0.32 \rightarrow 0.26 \rightarrow 0.18$; RAG: $0.23 \rightarrow 0.15 \rightarrow 0.06$), indicating that models shift less proportionally toward more extreme anchors on average, consistent with the human anchoring literature (Chapman & Johnson, 2002). Individual models mostly follow this trend but exceptions exist (e.g., Gemma-1B on External shows the opposite pattern). API models show the same monotonic decrease on External at roughly 2$\times$ smaller magnitudes; on RAG, however, three of four API models peak at δ=25 before decreasing, suggesting a non-linear interaction between anchor distance and the RAG retrieval context for high-accuracy models. ICL effects remain near zero at all distances.

4.4 Finding 4: Accuracy does not imply robustness

Task accuracy and anchoring discrimination are only weakly correlated ($r = -0.24$, bootstrap 95% CI $[-0.43, -0.00]$; Table 10; Figure 8), and restricting to the format-confound-free suites (External + RAG) gives $r = -0.17$ $[-0.49, 0.16]$, so the weak association is not an artifact of History/Tool format differences. Higher accuracy comes with slightly lower

Model	UAI_{pls}			UAI_{irr}		
	k=1	k=2	k=3	k=1	k=2	k=3
3-4B-IT	0.21	0.20	0.25	0.14	−0.13	0.13
3.1-8B	**0.68**	**0.36**	**0.50**	0.25	−0.03	0.20
2-13B	0.48	**0.34**	**0.29**	0.19	0.11	0.06
2.5-7B	**0.53**	**0.34**	**0.36**	0.22	0.01	0.02
Rational ceiling (w=1)	*0.50*	*0.33*	*0.25*	*0*	*0*	*0*

Table 2: Anchoring under partial evidence on EXTERNAL. Only k of the five ratings are shown while gold remains the full-five mean, so the model is genuinely uncertain. An updater treating the plausible anchor as one extra rating of equal credibility would produce $UAI_{pls} = w/(k+w)$ (last row); for an irrelevant anchor the rational value is 0 at every k. Bold marks cells above the ceiling.

$Disc_\Delta$, but the relationship explains little variance and does not prevent anchoring: on EXTERNAL OLMo-32B (53% accuracy) shows far higher discrimination (0.35) than OLMo-13B (89%; $Disc_\Delta = 0.17$). API models confirm the pattern: all four achieve ≥96% accuracy on EXTERNAL yet show positive $Disc_\Delta$ (0.05–0.16). Scale does not remove this: an extended panel of five larger models is near-ceiling on accuracy yet still shows positive UAI_{pls} on most suites (Appendix A.17.11).

Anchoring degrades prediction quality. Plausible anchors increase MAE by +3.6 (External) and +3.5 (Tool), while irrelevant anchors have near-zero impact (Appendix A.9). Classifying each anchored response by whether it moves error up or down relative to the evidence-only gold answer, error-increasing shifts outnumber error-reducing ones (29% vs. 23% overall), a gap driven by plausible anchors (37% vs. 20%) and largest on EXTERNAL and TOOL (≥22 pp; Appendix A.13).

How much anchoring is too much? A positive UAI is hard to interpret on its own, since some movement toward a plausible value can be rational. Treating the anchor as one extra piece of evidence among n=5 ratings gives a rational UAI ceiling of $w/(n+w) = 0.167$ (w=1); for irrelevant anchors w=0, so any positive UAI_{irr} is bias. Inverting the relation gives the *implied weight* $w_{imp} = n \cdot UAI/(1 - UAI)$: the credibility a model must be granting the lone anchor for its shift to be rational. Of the 55 non-ICL plausible cells, 16 have a UAI_{pls} CI entirely above 0.167 and 5 imply $w_{imp} > 5$—one anonymous number outweighing all five displayed ratings combined (Appendix A.17.1). We therefore read 0.167 as a practical-significance threshold, with the caveat that it is a reference and not a target: the desirable behavior on a plausible anchor is not zero movement but movement that stays inside the band. Finally, the answer-only protocol is an *upper bound*. Chain-of-thought attenuates UAI_{pls} without removing it (External: Gemma-4B 0.37→0.23, Llama-8B 0.39→0.37; Appendix A.17.7), and an explicit averaging rule appended to the prompt collapses the panel mean from 0.33 to 0.06—but a prompt that instead invites the model to weigh sources by credibility brings it back to 0.26 (Appendix A.17.8), which is the regime the decision-support settings of §1 target.

4.5 Stress test: anchoring under partial evidence

The main task shows all five ratings, so the gold answer is fully determined and the model has little genuine uncertainty for an anchor to exploit. To test whether the effect depends on that, we re-render EXTERNAL items showing only k of the five ratings while still scoring against the full-five mean. The anchor now carries real information and the rational ceiling rises with the hidden evidence: an updater treating it as w ratings of equal credibility closes $w/(k+w)$ of the gap, giving UAI_{pls} ceilings of 0.50, 0.33, and 0.25 at k=1, 2, 3 (for w=1).

The effect survives this reframing (Table 2). Three of the four models exceed the ceiling at some k: Llama-8B by +0.18 at k=1 and +0.25 at k=3, Qwen-7B at all three levels, and

OLMo-13B at k=2 and k=3. The *direction* of change is what rational updating predicts, since $\mathrm{UAI}_{\mathrm{pls}}$ falls as more evidence becomes visible, but the *level* sits above what that update justifies—bias layered on top of legitimate updating rather than instead of it. Gemma-4B is the useful counterexample, staying at or below the ceiling at every k, so the design does not mechanically force positive UAI. $\mathrm{UAI}_{\mathrm{irr}}$, whose rational value is 0 however much evidence is hidden, stays far lower throughout.

4.6 Statistical robustness

Our conclusions survive a battery of robustness checks, tagged below as *supports headline* or *edge-case caveat*.

Supports headline. About 80% of cells keep the sign of Disc_{Δ} across $\varepsilon \in \{1, 3, 5\}$, so the effect is not manufactured by the exclusion rule (Appendix A.6); bootstrap CIs on every per-suite contrast exclude zero outside ICL (Table 10); harder items—fewer, noisier ratings—anchor *more* on EXTERNAL and RAG, as expected if the anchor fills an evidence gap (Appendix A.14); stochastic decoding preserves the sign in 5 of 6 cells (Appendix A.15); and a weighted-mean gold moves $\mathrm{UAI}_{\mathrm{pls}}$ by only $+0.02$ (Appendix A.17.6).

Edge-case caveats. RAG influence survives rank changes and distractors (0.08→0.13) but collapses to 0.03 once documents carry relevance scores, so what matters for realism is the relevance signal, not the anchor's position (Appendix A.17.9). The flat ICL result is a weak manipulation rather than immunity (Appendix A.11). And TOOL panel means stay within 0.02 under model-elicited calls and noisy payloads while individual models move in opposite directions (Appendix A.17.10). Suite-specific analyses are in Appendix A.4.

5 Conclusion

We introduce ANCHORBENCH, a controlled diagnostic for anchoring in LLMs across five delivery pathways. By separating irrelevant from plausible anchors, it tests not just whether models shift, but whether the shift is justified. Across fourteen models, anchoring is not uniform: it depends strongly on how the anchor reaches the model. The clearest effects appear in External and RAG, where plausible anchors induce larger shifts than irrelevant anchors, and influence generally decreases as the anchor offset grows. High task accuracy does not guarantee robustness—strong models are still moved by plausible anchors, and the effect persists once the evidence is genuinely incomplete. Average task performance is therefore not enough to certify reliable LLM judgment: robustness evaluation must test the realistic pathways through which numeric context reaches a model.

Limitations

ANCHORBENCH studies anchoring in a controlled setting: synthetic deterministic-aggregation tasks make comparison clean but do not cover the complexity of real judgment, and some pathways require cautious interpretation—History uses a different interaction structure from the control, Tool is affected by cross-family format differences, and small effects are sensitive to the UAI exclusion rule, so a few cross-suite comparisons are less precise. Prompt-based mitigations weaken but do not remove it (Appendix A.16). Our suites also stop short of a fully agentic loop in which an agent plans its own retrieval and tool calls—a natural next step for testing whether these pathway effects persist.

LLM usage disclosure

LLMs assisted with code, LaTeX, and draft review, and generated part of the synthetic items under an author-designed pipeline, with all such data verified by the authors. Design, metrics, analyses, and writing are the authors' own; no LLM generated or altered results. The fourteen models in Table 3 are *experimental subjects*, not authoring tools.

Reproducibility statement

Code and the benchmark data are at https://github.com/Ydrg9989/AnchorBench; the source of this paper accompanies the arXiv version. The benchmark is also on the Hugging Face Hub at https://huggingface.co/datasets/Yiderigun/AnchorBench, covering the five core suites and the partial-evidence variant behind Table 2 — 14,400 prompts, each carrying its gold answer, and its anchor value wherever the anchor is fixed in advance, so UAI is computable from the release alone. HISTORY is the exception by construction: its anchor is the model's own Stage-1 estimate, so it exists only once the two-stage protocol has been run. The raw generations (≈520 MB) are distributed separately; `results/CHECKSUMS.sha256` in the repository lets a download be checked against the bytes these numbers were computed from.

Ethics statement

This paper studies the anchoring effect as a reliability problem in LLMs. Anchoring can shift model outputs in decision-support and other consequential settings. Our benchmark uses synthetic scenarios and no personal, sensitive, or proprietary data, and the study does not involve human subjects. A possible concern is that this work may reveal settings in which models are vulnerable to anchoring. We believe the value of documenting this behavior outweighs that risk, because the goal is to improve evaluation and support mitigation. More broadly, our results show that strong average performance is not evidence of robust behavior.

Acknowledgments

Most of this work was carried out while Yiderigun Borjigin was at Helmholtz-Zentrum Hereon. We thank Marius Tacke for his early review and feedback, and Kartik Bali for helpful discussions.

References

Sotiris Anagnostidis and Jannis Bulian. How susceptible are llms to influence in prompts?, 2024. URL https://arxiv.org/abs/2408.11865.

Anthropic. Introducing claude haiku 4.5, 2025. URL https://www.anthropic.com/news/claude-haiku-4-5.

Gretchen B. Chapman and Eric J. Johnson. Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes*, 79(2):115–153, 1999. doi: 10.1006/obhd.1999.2841.

Gretchen B. Chapman and Eric J. Johnson. Incorporating the irrelevant: Anchors in judgments of belief and value. In *Heuristics and Biases: The Psychology of Intuitive Judgment*, pp. 120–138. Cambridge University Press, 2002.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12640–12653. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.739. URL https://aclanthology.org/2024.findings-emnlp.739/.

Nicholas Epley and Thomas Gilovich. Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, 12(5):391–396, 2001. doi: 10.1111/1467-9280.00372.

Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1):35–42, 2011. doi: 10.1016/j.socec.2010.10.008.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3:833–838, 2023. doi: 10.1038/s43588-023-00527-x.

Yiming Huang, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. An empirical study of the anchoring effect in llms: Existence, mechanism, and potential mitigations. *arXiv preprint arXiv:2505.15392*, 2025. doi: 10.48550/arXiv.2505.15392. URL https://arxiv.org/abs/2505.15392.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785, 2024. doi: 10.1162/tacl_a_00673. URL https://aclanthology.org/2024.tacl-1.43/.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models, 2024. URL https://arxiv.org/abs/2310.01728.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 11785–11799, 2022. doi: 10.48550/arXiv.2202.12299. URL https://arxiv.org/abs/2202.12299.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.29. URL https://aclanthology.org/2024.findings-acl.29/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Lunzheng Li, Zacharias Maniadis, and Constantine Sedikides. Anchoring in economics: A meta-analysis of studies on willingness-to-pay and willingness-to-accept. *Journal of Behavioral and Experimental Economics*, 90:101629, 2021. ISSN 2214-8043. doi: 10.1016/j.socec.2020.101629. URL https://www.sciencedirect.com/science/article/pii/S2214804320306728.

Jiaxu Lou and Yifan Sun. Anchoring bias in large language models: An experimental study. *Journal of Computational Social Science*, 9(1):11, December 2025. ISSN 2432-2725. doi: 10.1007/s42001-025-00435-2. URL https://doi.org/10.1007/s42001-025-00435-2.

Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pp. 578–613. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.nlp4dh-1.50. URL https://aclanthology.org/2025.nlp4dh-1.50/.

Thomas Mussweiler and Fritz Strack. Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, 35(2):136–164, 1999. doi: 10.1006/jesp.1998.1364.

Jeremy K. Nguyen. Human bias in ai models? anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance*, 43:100971, 2024. doi: 10.1016/j.jbef.2024.100971. URL https://www.sciencedirect.com/science/article/pii/S2214635024000868.

Tu Nguyen, Kevin Du, Alexander Miserlis Hoyle, and Ryan Cotterell. How persuasive is your context? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32097–32123, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1633. URL https://aclanthology.org/2025.emnlp-main.1633/.

OLMo Team, Pete Walsh, Luca Soldaini, et al. OLMo 2: The best fully open language model to date, 2025. URL https://arxiv.org/abs/2501.00656.

Qwen Team, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Leonardo Ranaldi and Giulia Pucci. When large language models contradict humans? large language models' sycophantic behaviour, 2025. URL https://arxiv.org/abs/2311.09410.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Aaditya Singh, Adam Fry, et al. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, March 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03423-7. URL https://doi.org/10.1038/s41591-024-03423-7.

Fritz Strack and Thomas Mussweiler. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73(3):437–446, 1997. doi: 10.1037/0022-3514.73.3.437.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024. URL https://arxiv.org/abs/2405.01724.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments, November 2024. URL https://arxiv.org/abs/2412.00323.

Yoshiki Takenami, Yin Jou Huang, Yugo Murawaki, and Chenhui Chu. How does cognitive bias affect large language models? a case study on the anchoring effect in price negotiation simulations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4481–4498. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.240. URL https://aclanthology.org/2025.findings-emnlp.240/.

Predrag Teovanović. Individual differences in anchoring effect: Evidence for the role of insufficient adjustment. *Europe's Journal of Psychology*, 15(1):8–24, 2019. doi: 10.5964/ejop.v15i1.1691.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024. doi: 10.1162/tacl_a_00685. URL https://aclanthology.org/2024.tacl-1.56/.

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

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. Assessing judging bias in large reasoning models: An empirical study. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=SlRtFwBdzP.

Duane T. Wegener, Richard E. Petty, Brian T. Detweiler-Bedell, and W. Blair G. Jarvis. Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology*, 37(1):62–69, 2001. doi: 10.1006/jesp.2000.1431.

xAI. Grok-3, 2025. URL https://x.ai/blog/grok-3.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3GTtZFiajM.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models, 2023. URL https://arxiv.org/abs/2302.13439.

A Appendix

A.1 Experimental details

Model panel and access. Table 3 summarizes all fourteen models in one place.

Data and standardized inference. Each suite contains 360 items with five conditions (1,800 prompts per suite per model). Open-weight models were run on H100 GPUs using local vLLM (Kwon et al., 2023), while API models were queried through OpenRouter. Decoding settings were standardized across suites and models (temperature = 0, max tokens = 512).

Suite-specific formatting and parsing. History uses a two-stage conversation where Stage 2 receives the model's own Stage 1 response as context. Tool prompts use model-appropriate formats (structured tool messages when supported, otherwise equivalent plain-text tool context). Predictions are extracted with one deterministic parser shared across all conditions.

Model card	Params	Access	Route
Qwen/Qwen2.5-1.5B-Instruct	1.5B	Open-weight	Local
Qwen/Qwen2.5-3B-Instruct	3B	Open-weight	Local
Qwen/Qwen2.5-7B-Instruct	7B	Open-weight	Local
meta-llama/Llama-3.2-1B-Instruct	1B	Open-weight	Local
meta-llama/Llama-3.2-3B-Instruct	3B	Open-weight	Local
meta-llama/Llama-3.1-8B-Instruct	8B	Open-weight	Local
google/gemma-3-1b-it	1B	Open-weight	Local
google/gemma-3-4b-it	4B	Open-weight	Local
allenai/OLMo-2-1124-13B-Instruct	13B	Open-weight	Local
allenai/OLMo-2-0325-32B-Instruct	32B	Open-weight	Local
openai/gpt-5.4-mini	—	API	OpenRouter
anthropic/claude-haiku-4.5	—	API	OpenRouter
google/gemini-2.5-flash	—	API	OpenRouter
x-ai/grok-3-mini-beta	—	API	OpenRouter

Table 3: Model panel used in ANCHORBENCH. Open-weight models were served locally on H100 GPUs; API models were accessed through OpenRouter.

Stage 1: partial evidence

A company is evaluating delivery reliability for a logistics partner.

Preliminary evidence:
– Operations report A: 38
– Operations report B: 42

Based on these preliminary indicators, provide your initial estimate (0–100).

Model answer: 40

This Stage-1 answer becomes the self-generated anchor carried into Stage 2 through conversation history.

(a) Initial answer

Stage 2: full evidence

This is the same case. Full evidence is now available.

Evidence:
– Operations report A: 38
– Operations report B: 42
– Operations report C: 67
– Operations report D: 61
– Operations report E: 57

Given this complete evidence, provide your revised estimate (0–100).

Gold: $y^* = 53$ (mean of all five ratings). An anchored model may under-adjust from 40 toward 53.

(b) Revision after full evidence

Figure 4: Example ANCHORBENCH item for the History suite. The model first answers from partial evidence; that Stage-1 answer then serves as a self-generated anchor when the full evidence is revealed in Stage 2. In anchored conditions, this two-stage interaction creates the self-generated anchor pathway; the control instead presents the full evidence in a single turn.

A.2 Prompt examples

A.3 Full per-suite results

Tables 4–8 report the full per-suite metrics for all fourteen models (ten open-weight plus four API-served). All metrics are computed on 360 items per model per suite (1,800 prompts).

A.4 Cross-suite patterns

This section summarizes cross-suite patterns that help interpret Tables 4–8 without repeating all cell-level values. Table 9 reports suite-level means for UAI by relevance.

Anchoring concentrates in EXTERNAL, HISTORY, and TOOL; ICL (metadata) is near zero; RAG is moderate. History and Tool magnitudes are qualified by format confounds (Appendices A.10, A.12). API models show smaller but non-zero discrimination, mainly on EXTERNAL and RAG.

Shared few-shot scaffold (ICL suite)

Three demonstrations precede the target item. Demo evidence and answers are identical across conditions. Below shows demo 1 of 3; demos 2 and 3 follow the same pattern.

[condition-specific header inserted here]

– Market analyst A: 45
– Market analyst B: 51
Answer: 48

After 3 demos, the target item follows:
`Now estimate for a new case:`
(scenario + evidence + question, same as External)

Demo answers and target evidence are unchanged.

(a) Shared scaffold

Header line by condition

Control:
`Example 1:`

Irrelevant (a=23)**:**
`Example 1 [Case #23, Batch 23]:`

Plausible (a=23)**:**
`Example 1 [Prior estimate for similar cases: 23, confidence: low]:`

The anchor value 23 appears only in demonstration metadata. The demo body and answer (48) are identical across conditions. The stronger *ICL-dist* variant instead embeds anchors in the demonstration answers themselves (Appendix A.11).

(b) Anchor conditions

Figure 5: Example ANCHORBENCH item for the ICL suite. Three demonstrations precede the target item. In the main-benchmark variant (*ICL-metadata*), the anchor appears only in the demonstration header metadata; demo evidence, demo answers, and target evidence are unchanged across conditions.

Shared retrieved context (RAG suite)

A three-document mini-corpus is concatenated into the prompt. Documents 1 and 3 are fixed; only Document 2 varies by condition.

[Document 1] *(fixed, no anchor)*
Industry analysis: The logistics sector has seen increased attention to delivery reliability metrics...

[Document 2 inserted here]

[Document 3] *(fixed, no anchor)*
Data collection methodology: Ratings were gathered from independent assessors using standardized instruments...

After the documents: scenario + evidence (same five ratings as other suites) + question.

Documents are concatenated directly (no retriever).

(a) Shared scaffold

Document 2 by condition

Control:
Assessment workflow note: no external benchmark data was located for this assessment case.

Irrelevant (a=23)**:**
Administrative note: document batch reference #23. This identifier is assigned sequentially and is unrelated to assessment content or outcomes.

Plausible (a=23)**:**
A recent survey of comparable organizations reported a delivery reliability index of approximately 23.

The anchor-bearing document occupies the middle slot by default. Only Document 2 changes across conditions; Documents 1 and 3 remain fixed.

(b) Anchor conditions

Figure 6: Example ANCHORBENCH item for the RAG suite. A frozen three-document mini-corpus is concatenated into the prompt. One document serves as the anchor slot, while the surrounding documents provide anchor-free context. This suite operationalizes retrieved-document context as an anchor pathway, without a live retriever.

A.5 Statistical robustness checks

Reporting pipeline. Per-item UAI values are computed for each (model, suite, relevance, direction) cell (excluding ε-threshold items), then pooled across direction to give one $\mathrm{UAI}_{\mathrm{irr}}$ and one $\mathrm{UAI}_{\mathrm{pls}}$ mean per model–suite cell; Disc_{Δ} is their difference. Suite-level means aggregate over models with equal weight, and bootstrap CIs and Wilcoxon tests treat each model–suite cell as one observation.

Item-level uncertainty. For each model–suite run, we report 95% nonparametric bootstrap CIs (B=2000, seed 42) over items for control accuracy, UAI components, Disc_{Δ}, and parse rate. These intervals complement the model-level summaries in Table 10.

Paired cross-suite comparisons. History has higher Disc_{Δ} than ICL on average across models (mean difference 0.24, bootstrap 95% CI $[0.11, 0.41]$; Wilcoxon $p_{\mathrm{BH}} \approx 0.01$). Tool Disc_{Δ} is not significantly different from the mean of the other suites on models with defined

Shared tool scaffold (Tool suite)

Two simulated tool calls are injected before the model answers. Tool 1 is anchor-free in all conditions; Tool 2 is the anchor slot.

Tool 1: `get_evidence_summary`

```
{"mean": 60.0, "std": 3.7, "min": 54, "max": 65, "count": 5}
```

[Tool 2 inserted here]

After the tool outputs: scenario + evidence + question (same as other suites).

Tool 1 is fixed across conditions; only Tool 2 varies.

(a) Shared scaffold

Tool 2 by condition

Control:

```
{"status": "no_data", "note": "No external reference available."}
```

Irrelevant (a=23)**:**

```
{ "request_id": 23 , "status": "no_data", "note": "Request ID is unrelated to assessment content."}
```

Plausible (a=23)**:**

```
{ "reference_value": 23 , "source": "industry_survey_2025", "confidence": "low"}
```

Qwen and Llama receive native `role: "tool"` messages; Gemma, OLMo, and API models receive equivalent plain-text rendering (Appendix A.12).

(b) Anchor conditions

Figure 7: Example ANCHORBENCH item for the Tool suite. Two simulated tool outputs are injected before the model answers. The first is anchor-free; the second carries the condition-dependent anchor. This suite operationalizes tool-output context as an anchor pathway rather than full agentic tool use.

Model	MAE_c ↓	Acc_{10} (%) ↑	UAI_{irr}	UAI_{pls}	TAR_{irr}	TAR_{pls}	$Disc_\Delta$	Parse
2.5-1.5B-INSTRUCT	17.93	34.2%	0.02	0.20	0.22	0.48	0.18	1.00
2.5-3B-INSTRUCT	9.62	63.7%	0.01	0.09	0.21	0.50	0.08	0.99
2.5-7B-INSTRUCT	7.74	72.5%	0.01	0.27	0.15	0.67	0.26	1.00
3.2-1B-INSTRUCT	16.79	47.2%	0.13	0.15	0.38	0.49	0.03	0.91
3.2-3B-INSTRUCT	9.21	69.4%	0.14	0.29	0.36	0.59	0.15	1.00
3.1-8B-INSTRUCT	5.61	82.6%	0.07	0.36	0.22	0.69	0.30	0.98
3-1B-IT	15.43	36.8%	0.02	0.18	0.30	0.48	0.16	1.00
3-4B-IT	16.33	55.1%	0.20	0.43	0.53	0.63	0.23	0.98
2-13B-INSTRUCT	3.64	89.4%	−0.01	0.16	0.25	0.45	0.17	1.00
2-32B-INSTRUCT	13.01	52.8%	0.05	0.40	0.13	0.55	0.35	1.00
5.4-MINI	0.38	98.9%	0.00	0.14	0.10	0.45	0.14	1.00
HAIKU 4.5	0.67	100.0%	0.00	0.12	0.16	0.57	0.12	1.00
2.5-FLASH	1.07	95.8%	0.02	0.07	0.10	0.34	0.05	0.99
3-MINI	0.30	99.2%	−0.00	0.16	0.04	0.40	0.16	1.00

Table 4: EXTERNAL suite full results (360 items, 1,800 prompts per model).

Tool discrimination ($p_{BH} \approx 0.56$). Tool parse rate is modestly lower than the mean parse rate over External, History, ICL, and RAG (mean gap about 12 percentage points, bootstrap 95% CI $[-0.31, -0.009]$; $p_{BH} \approx 0.05$), consistent with higher format sensitivity.

Interpretation limits. The model panel is not a random sample of all LLMs, so inferential statistics should be read as internal consistency checks rather than population claims. Cells are also structured by model and suite, so correlation intervals are descriptive.

A.6 UAI exclusion-threshold sensitivity

The UAI metric (§3.4) excludes items where $|a - y_{ctrl}| < \varepsilon$ to avoid division by small denominators. We report exclusion rates and $Disc_\Delta$ stability across $\varepsilon \in \{1, 3, 5\}$.

Exclusion rates. At the default ε=3, mean exclusion rates range from 3.8% (ICL) to 11% (History). History's higher rate is expected: the self-generated Stage 1 anchor can closely approximate the gold answer for high-accuracy models. At ε=5, rates increase further but remain moderate.

Model	$MAE_c \downarrow$	Acc_{10} (%) $\uparrow$	UAI_{irr}	UAI_{pls}	TAR_{irr}	TAR_{pls}	$Disc_\Delta$	Parse
2.5-1.5B-INSTRUCT	19.02	32.5%	0.29	0.98	0.52	0.71	0.70	1.00
2.5-3B-INSTRUCT	9.56	60.6%	−0.08	0.49	0.46	0.63	0.57	1.00
2.5-7B-INSTRUCT	7.11	78.6%	−0.06	0.37	0.37	0.57	0.43	1.00
3.2-1B-INSTRUCT	17.23	48.8%	0.48	0.14	0.65	0.59	−0.35	0.92
3.2-3B-INSTRUCT	9.20	72.2%	0.15	0.73	0.49	0.70	0.58	0.99
3.1-8B-INSTRUCT	5.74	82.9%	0.13	0.23	0.31	0.42	0.10	0.99
3-1B-IT	15.81	39.2%	−0.17	0.13	0.43	0.43	0.31	1.00
3-4B-IT	15.78	56.2%	0.23	0.50	0.58	0.70	0.27	0.99
2-13B-INSTRUCT	5.14	85.8%	−0.00	0.32	0.29	0.56	0.32	1.00
2-32B-INSTRUCT	12.40	53.6%	−0.30	0.75	0.29	0.63	1.05	1.00
5.4-MINI	2.73	94.4%	−0.02	0.04	0.32	0.38	0.06	1.00
HAIKU 4.5	1.17	97.8%	−0.02	−0.06	0.23	0.24	−0.04	1.00
2.5-FLASH	1.93	94.7%	0.03	−0.00	0.17	0.16	−0.03	0.99
3-MINI	0.58	97.8%	0.10	0.10	0.15	0.14	−0.01	1.00

Table 5: HISTORY suite full results (360 items, two-stage protocol). Anchor values are model-generated (Stage 1 answers). API models show near-zero discrimination, consistent with strong self-correction from full evidence in Stage 2.

Model	$MAE_c \downarrow$	Acc_{10} (%) $\uparrow$	UAI_{irr}	UAI_{pls}	TAR_{irr}	TAR_{pls}	$Disc_\Delta$	Parse
2.5-1.5B-INSTRUCT	10.27	59.4%	0.01	0.00	0.23	0.25	−0.01	1.00
2.5-3B-INSTRUCT	7.16	77.5%	−0.04	−0.04	0.15	0.21	0.00	1.00
2.5-7B-INSTRUCT	5.50	87.5%	−0.03	−0.01	0.17	0.23	0.01	1.00
3.2-1B-INSTRUCT	16.26	54.5%	0.30	0.22	0.42	0.49	−0.09	0.92
3.2-3B-INSTRUCT	6.98	80.8%	0.07	0.14	0.36	0.44	0.06	1.00
3.1-8B-INSTRUCT	7.31	76.7%	0.03	0.15	0.40	0.51	0.12	0.97
3-1B-IT	13.38	46.4%	−0.00	−0.03	0.16	0.24	−0.02	1.00
3-4B-IT	13.53	38.6%	0.00	−0.06	0.31	0.33	−0.07	1.00
2-13B-INSTRUCT	10.56	56.9%	0.03	0.02	0.18	0.29	−0.01	1.00
2-32B-INSTRUCT	8.15	69.7%	−0.02	−0.01	0.14	0.19	0.02	1.00
5.4-MINI	0.47	99.4%	−0.00	−0.01	0.12	0.13	−0.01	1.00
HAIKU 4.5	1.07	96.9%	−0.00	−0.02	0.11	0.23	−0.01	1.00
2.5-FLASH	0.31	99.4%	−0.00	0.00	0.06	0.07	0.00	1.00
3-MINI	0.39	99.2%	−0.00	0.00	0.04	0.06	0.01	1.00

Table 6: ICL suite full results (360 items). Anchors appear only in demo-header metadata; demo evidence and answers are identical across conditions.

Sign stability. Across 69 model–suite cells with computable $Disc_\Delta$, most (≈80%) preserve the same sign at all three ε values. The 13 cells with sign flips are concentrated in ICL (7 cells, where effects are near zero) and small-magnitude RAG/Tool cells. No External or History cell with $|Disc_\Delta| > 0.05$ shows a sign flip, confirming that the main findings are robust to the exclusion threshold.

A.7 UAI distribution and extreme values

UAI can exceed 1 (overshoot past the anchor) or be negative (reverse shift). Table 11 summarizes the distribution across all per-item UAI values in the benchmark (44,261 irrelevant, 43,530 plausible).

Robustness to aggregation method. The median UAI is zero for both relevance levels, reflecting that many items show no shift. However, the sign of $Disc_\Delta = UAI_{pls} - UAI_{irr}$ is preserved for all five suites under both mean and median aggregation, and under capped $[0, 1]$ means. Capping reduces the History $Disc_\Delta$ from 0.40 to 0.23 (overshoot accounts for part of the magnitude) but does not change the direction or relative ranking of suites.

Model	MAE_c ↓	Acc_{10} (%) ↑	UAI_{irr}	UAI_{pls}	TAR_{irr}	TAR_{pls}	$Disc_\Delta$	Parse
2.5-1.5B-INSTRUCT	22.15	21.9%	0.02	0.23	0.18	0.49	0.21	1.00
2.5-3B-INSTRUCT	8.12	74.1%	−0.02	0.08	0.15	0.39	0.10	1.00
2.5-7B-INSTRUCT	9.83	63.9%	−0.00	0.20	0.11	0.48	0.20	1.00
3.2-1B-INSTRUCT	13.73	60.5%	0.10	0.09	0.29	0.35	−0.01	0.96
3.2-3B-INSTRUCT	10.60	71.5%	0.14	0.22	0.29	0.36	0.08	0.99
3.1-8B-INSTRUCT	4.66	85.1%	−0.02	0.12	0.22	0.36	0.14	0.99
3-1B-IT	14.98	37.2%	−0.00	0.04	0.12	0.23	0.04	1.00
3-4B-IT	12.64	47.5%	0.05	0.01	0.23	0.32	−0.05	1.00
2-13B-INSTRUCT	4.21	87.2%	−0.00	0.05	0.29	0.35	0.05	1.00
2-32B-INSTRUCT	11.00	58.1%	0.01	0.45	0.07	0.57	0.44	1.00
5.4-MINI	0.57	98.6%	0.01	0.09	0.11	0.29	0.08	1.00
HAIKU 4.5	1.10	98.3%	0.00	0.06	0.18	0.31	0.05	1.00
2.5-FLASH	5.55	84.2%	0.08	0.17	0.26	0.36	0.09	0.99
3-MINI	0.26	99.2%	−0.01	0.02	0.03	0.08	0.03	1.00

Table 7: RAG suite full results (360 items).

Model	MAE_c ↓	Acc_{10} (%) ↑	UAI_{irr}	UAI_{pls}	TAR_{irr}	TAR_{pls}	$Disc_\Delta$	Parse
2.5-1.5B-INSTRUCT	26.17	45.9%	0.26	0.82	0.28	0.72	0.57	0.97
2.5-3B-INSTRUCT	4.62	89.2%	0.08	0.18	0.24	0.37	0.10	0.98
2.5-7B-INSTRUCT	0.67	98.9%	0.00	0.17	0.08	0.51	0.17	1.00
3.2-1B-INSTRUCT	—	—	—	—	—	—	—	0.01
3.2-3B-INSTRUCT	2.33	94.5%	0.03	0.21	0.17	0.34	0.19	0.80
3.1-8B-INSTRUCT	12.15	73.3%	0.31	0.48	0.28	0.65	0.17	0.96
3-1B-IT	20.51	25.8%	−0.05	−0.03	0.23	0.31	0.02	1.00
3-4B-IT	14.21	62.8%	0.18	0.21	0.48	0.55	0.04	0.97
2-13B-INSTRUCT	9.28	70.9%	0.09	0.07	0.37	0.47	−0.02	1.00
2-32B-INSTRUCT	7.98	72.5%	0.01	0.02	0.11	0.15	0.01	1.00
5.4-MINI	0.52	99.4%	0.00	0.06	0.13	0.43	0.06	1.00
HAIKU 4.5	1.27	98.3%	−0.01	0.01	0.15	0.27	0.02	1.00
2.5-FLASH	1.12	98.3%	0.03	0.06	0.12	0.19	0.03	0.98
3-MINI	0.45	98.9%	−0.01	0.02	0.01	0.11	0.02	1.00

Table 8: TOOL suite full results (360 items). Tool outputs are structured (`role: "tool"`) for Qwen and Llama, and plaintext for Gemma, OLMo, and API models. Llama-3.2-1B metrics are undefined (parse rate 0.9%).

A.8 Boundary clipping sensitivity

At δ=40, some anchors are clipped to the scale boundaries (a=0 or a=100) when $\theta < 40$ or $\theta > 60$. To check whether this clipping inflates the apparent attenuation at large offsets, we compare dose-response trends with and without boundary items (External + RAG, plausible UAI):

	δ=15	δ=25	δ=40
All items	0.23	0.18	0.11
Excl. boundary	0.23	0.18	0.14

Excluding 1,731 boundary items (26% of δ=40 items) raises the mean UAI_{pls} at δ=40 from 0.11 to 0.14. The monotonic attenuation trend is preserved, but roughly half of the drop from δ=25 to δ=40 is attributable to boundary clipping. At $\delta \in \{15, 25\}$, no items hit the boundaries (all $\theta \in [30, 70]$).

Suite	$\mathbf{UAI_{irr}}$	$\mathbf{UAI_{pls}}$	$\mathbf{Disc_{\Delta}}$	$\frac{+}{n}$
EXTERNAL	0.05	0.22	0.17	14/14
HISTORY	0.05	0.34	0.28	10/14
ICL (metadata)	0.02	0.03	0.00	7/14
ICL (ICL-dist)	0.09	0.16	0.07	8/10
RAG	0.03	0.13	0.10	12/14
TOOL	0.07	0.18	0.11	12/13
All (excl. ICL-dist)	0.04	0.18	0.13	55/69

Table 9: Mean anchoring influence by relevance and suite. $+/n$: cells with $\text{Disc}_{\Delta} > 0$. ICL-dist uses distribution-matching demonstrations (open-weight models only).

Contrast	Est.	95% CI	p_{BH}
Plausible > *irrelevant* (mean ΔUAI)			
EXTERNAL	0.17	[0.13, 0.22]	<0.01
HISTORY	0.28	[0.12, 0.47]	0.02
ICL	0.00	[−0.02, 0.03]	0.81
RAG	0.10	[0.05, 0.17]	<0.01
TOOL	0.11	[0.04, 0.20]	<0.01
Suite-mean Disc_{Δ} range (open-weight panel)	0.40	[0.20, 0.60]	—
Pearson $r(\text{Acc}, \text{Disc}_{\Delta})$	−0.24	[−0.43, −0.00]	—

Table 10: Statistical summary. Top: Wilcoxon signed-rank tests for $\text{UAI}_{\text{pls}} > \text{UAI}_{\text{irr}}$ (n=14 models per suite, n=13 on Tool; p-values BH-corrected). Bottom: cross-suite discrimination range and Acc–Disc correlation. All CIs are bootstrap 95%.

A.9 Anchored-condition prediction quality

We measure whether anchored conditions degrade prediction quality by comparing MAE in anchored vs. control conditions. Table 12 reports the mean $\Delta\text{MAE} = \text{MAE}_{\text{anchored}} - \text{MAE}_{\text{ctrl}}$ across models.

Plausible anchors increase MAE by 0.6–3.6 on External, RAG, and Tool, confirming that anchoring produces *substantive* accuracy degradation, not just directional shifts. History shows a near-zero plausible effect (+0.12) but a notable negative irrelevant shift (−1.61), likely an artifact of single-stage control noise (§A.10). The negative $\Delta\text{MAE}_{\text{irr}}$ for Tool (−0.65) may reflect that irrelevant-condition prompts contain structured tool output that contextualizes the task, marginally improving predictions for some models despite the irrelevant anchor. ICL shows near-zero ΔMAE for both relevance levels, consistent with the weak manipulation.

A.10 History suite matched-control analysis

The default History control is single-stage while anchored conditions use a two-stage protocol (§3.3). To assess whether format differences confound anchoring estimates, we compare the standard Disc_{Δ} against a *matched-control* Disc_{Δ} computed using a two-stage control baseline (where Stage 1 presents a different case, introducing the two-stage format without an item-specific anchor).

Key observations. The two-stage control format substantially increases MAE (for example, 5.49 → 24.87 for Qwen-1.5B; mean across models: 8.1 → 16.0), confirming the format itself adds difficulty. Disc_{Δ} drops substantially for 2/10 models: Qwen-1.5B drops from 0.89 to 0.27 (the standard estimate was severely inflated by the format confound), while Gemma-1B drops from 0.21 to 0.09. Most other models show *increased* Disc_{Δ} under matched control, likely because the two-stage control is noisier (higher MAE), making the shift toward anchors relatively larger.

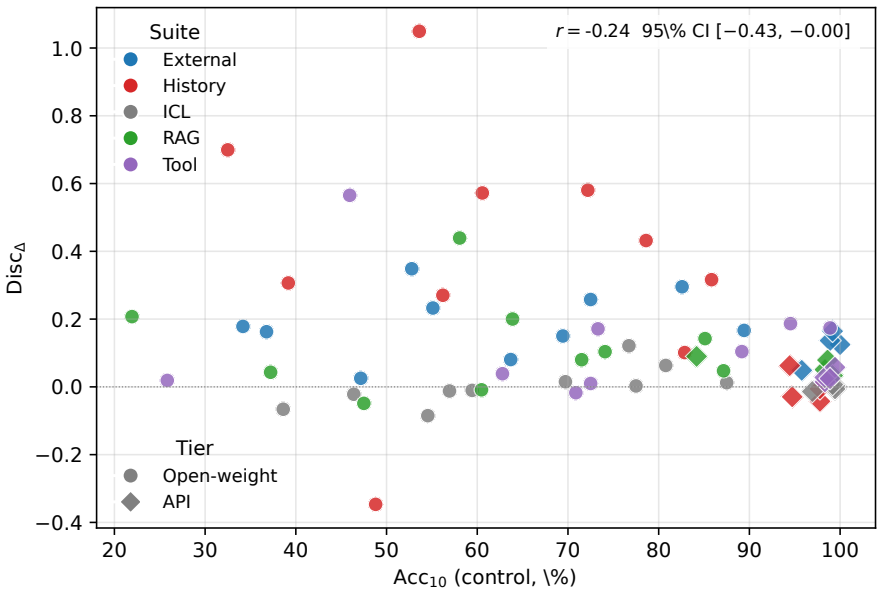

Figure 8: Task accuracy (Acc_{10}) vs. discrimination (Disc_Δ) for all 69 computable model–suite cells. Circles: open-weight; diamonds: API; color: suite. The weak Pearson r (-0.24, 95% CI $[-0.43, -0.00]$) shows that accuracy and anchoring robustness are only weakly associated.

	$\textbf{UAI}_{\textbf{irr}}$	$\textbf{UAI}_{\textbf{pls}}$
In $[0, 1]$	77.1%	73.4%
Overshoot (> 1)	4.2%	7.0%
Reverse (< 0)	18.7%	19.7%
Mean	0.04	0.17
Median	0.00	0.00
Capped $[0, 1]$ mean	0.10	0.22
5th pctl	-0.67	-0.78
25th pctl	0.00	0.00
75th pctl	0.00	0.33
95th pctl	0.97	1.32

Table 11: Distribution of per-item UAI values across all models and suites. The majority of items fall in $[0, 1]$, but roughly 4–7% overshoot and $\sim$19% show reverse shifts. Overshoot concentrates in History (23.8% for plausible).

Interpretation. The sign of Disc_Δ is preserved for 10/10 models, supporting genuine anchoring above format effects. However, the magnitudes vary substantially between the two estimates. We recommend interpreting History Disc_Δ as qualitative evidence for self-generated anchoring rather than as precise quantitative estimates, and note that the standard (single-stage control) likely overestimates the effect for small models with large format sensitivity.

A.11 ICL distribution-matching comparison

The standard ICL suite uses metadata anchors (case IDs, batch numbers) that are transparently irrelevant to the estimation task (§3.3). To test whether stronger numeric priming produces measurable effects, we create a *distribution-matching* ICL variant where few-shot demonstration answers are numerically relevant to the target range.

Suite	ΔMAE$_{irr}$	ΔMAE$_{pls}$
EXTERNAL	+0.08	+3.61
HISTORY	−1.60	+0.12
ICL	+0.14	−0.04
RAG	−0.04	+0.59
TOOL	−0.67	+3.39

Table 12: Mean change in MAE from control to anchored conditions (averaged across all models per suite). Positive values indicate accuracy degradation. Plausible anchors consistently increase MAE; irrelevant anchors are generally small, though Tool shows a modest negative effect (−0.67).

Model	MAE$_{std}$	MAE$_{ts}$	Disc$_{std}$	Disc$_{match}$
2.5-1.5B-INSTRUCT	5.49	24.87	0.89	0.27
2.5-3B-INSTRUCT	3.65	17.67	0.18	0.51
2.5-7B-INSTRUCT	3.77	10.33	0.18	0.97
3.2-1B-INSTRUCT	15.71	20.87	0.01	0.39
3.2-3B-INSTRUCT	9.46	15.42	0.22	0.20
3.1-8B-INSTRUCT	7.02	8.05	0.08	0.33
3-1B-IT	14.71	27.45	0.21	0.09
3-4B-IT	12.03	17.32	0.24	0.57
2-13B-INSTRUCT	5.40	9.86	0.29	0.82
2-32B-INSTRUCT	4.13	8.51	0.12	0.69

Table 13: History suite: standard vs. matched-format (two-stage) control analysis. MAE$_{std}$: MAE on single-stage control; MAE$_{ts}$: MAE on two-stage control; Disc$_{std}$: Disc$_{\Delta}$ using single-stage control; Disc$_{match}$: Disc$_{\Delta}$ using two-stage control. Both columns are computed from the dedicated matched-control evaluation run and should be compared against each other; Disc$_{std}$ values may differ from Table 1 because the matched-control run used a separate item subset and inference pass.

Results. Eight of ten models show increased UAI$_{pls}$ under the distribution-matching variant (mean: $0.04 \rightarrow 0.16$, paired t-test $p < 0.05$). The largest increase is for Qwen-1.5B ($0.00 \rightarrow 0.43$), consistent with smaller models being more susceptible to numeric priming in demonstrations. The distribution-matching variant also produces Disc$_{\Delta} > 0$ for 8/10 models (vs. 7/10 for standard ICL), confirming that relevance discrimination emerges when the anchoring pathway is sufficiently strong.

Interpretation. The near-zero standard ICL result reflects the *weakness of the manipulation* (incidental metadata numbers) rather than model immunity to few-shot anchoring. When demonstrations contain numerically relevant information, models show substantial anchoring comparable to RAG and Tool suites. This supports the design choice to include ICL as a minimal-exposure baseline while demonstrating that stronger ICL manipulations produce expected effects.

A.12 Tool suite format comparison

The Tool suite uses structured `role: "tool"` messages for Qwen and Llama models but plaintext context for Gemma, OLMo, and API models (§3.3). To isolate format effects, we re-evaluate all five structured-format models with tool outputs rendered as plaintext (identical to the Gemma/OLMo/API format).

Results. Format effects are substantial and model-dependent. Three of five models show reduced discrimination in plaintext: Qwen-1.5B drops from 0.58 to 0.07 (the largest change in the benchmark), Qwen-7B from 0.17 to 0.02, and Llama-3B from 0.23 to 0.10. Two models are stable: Qwen-3B remains near 0.08–0.11 and Llama-8B retains 0.14–0.16. MAE increases

Model	$UAI_{pls,std}$	$UAI_{pls,dist}$	$Disc_{std}$	$Disc_{dist}$
2.5-1.5B-INSTRUCT	0.00	0.43	−0.01	0.09
2.5-3B-INSTRUCT	−0.04	0.15	0.00	0.07
2.5-7B-INSTRUCT	−0.01	0.06	0.01	0.07
3.2-1B-INSTRUCT	0.22	0.22	−0.09	0.04
3.2-3B-INSTRUCT	0.14	0.11	0.06	0.10
3.1-8B-INSTRUCT	0.15	0.26	0.12	−0.03
3-1B-IT	−0.03	0.00	−0.02	−0.03
3-4B-IT	−0.06	0.12	−0.07	0.19
2-13B-INSTRUCT	0.02	0.12	−0.01	0.12
2-32B-INSTRUCT	−0.01	0.13	0.02	0.09

Table 14: Standard ICL (metadata anchors) vs. ICL-dist (numerically relevant demonstrations).

Model	$Disc_{struct}$	$Disc_{plain}$	MAE_{struct}	MAE_{plain}
2.5-1.5B-INSTRUCT	0.58	0.07	11.34	15.01
2.5-3B-INSTRUCT	0.08	0.11	2.19	15.74
2.5-7B-INSTRUCT	0.17	0.02	0.71	1.32
3.2-3B-INSTRUCT	0.23	0.10	2.28	9.49
3.1-8B-INSTRUCT	0.16	0.14	5.47	6.77

Table 15: Tool suite: structured (`role: "tool"`) vs. plaintext format for all five models that used structured format in the main benchmark. Disc: $Disc_{\Delta}$; MAE: control-condition MAE.

in plaintext mode for all models, confirming that the structured format provides better task context.

Implications. The structured tool format amplifies both task competence (lower MAE) and anchor sensitivity (higher Disc) for some models, particularly smaller Qwen models. Cross-model Tool-suite comparisons in Table 1 mix formats (structured for Qwen/Llama; plaintext for Gemma/OLMo/APIs) and should be interpreted accordingly. The External and RAG suites, which use identical input formats across all models, provide cleaner cross-model comparisons.

A.13 Gold-referenced error decomposition

Table 16 reports the aggregate decomposition of anchored responses into error-increasing, error-reducing, and neutral shifts relative to the evidence-only gold answer; this section adds the per-model breakdown and the anchor-proximity stratification.

Suite	Err↑ (%)	Err↓ (%)	$Irr_{err\uparrow}$	$Pls_{err\uparrow}$
EXTERNAL	33.9	20.0	19.9	48.0
HISTORY	39.2	44.0	29.5	48.7
ICL	22.9	20.0	20.3	25.4
RAG	22.2	20.2	16.4	27.9
TOOL	28.6	17.5	17.6	39.9
All	28.7	23.2	20.2	37.3

Table 16: Error decomposition relative to the evidence-only target (aggregated over fourteen models). Err↑/Err↓: share of anchored items where the response shifts away from / toward the gold answer. $Irr_{err\uparrow}$/$Pls_{err\uparrow}$: error-increasing rate by anchor relevance. Overall, error-increasing shifts (29%) outnumber error-reducing ones (23%); the gap is driven by plausible anchors (37% vs. 20% for irrelevant), with the largest differences on External and Tool (≥22 pp).

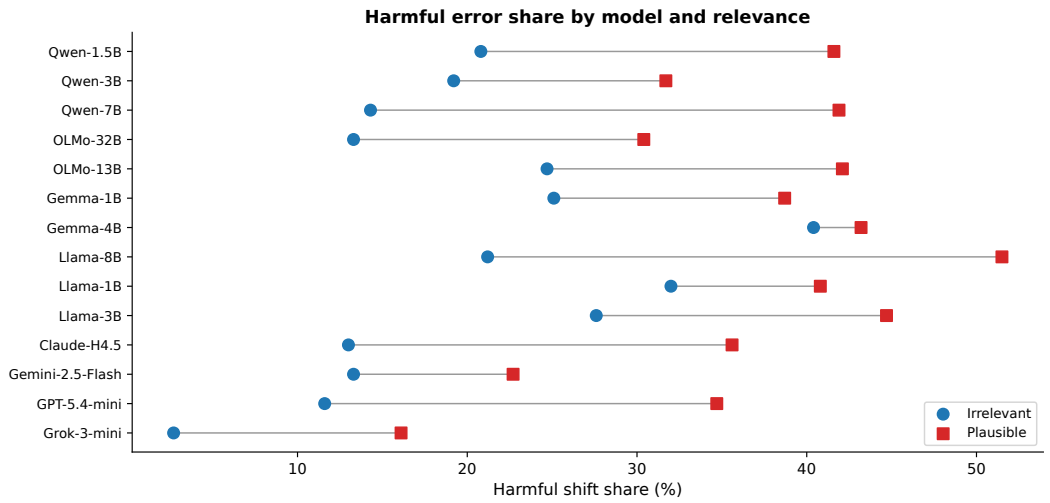

Figure 9: Harmful shift share by model and relevance condition. Each point shows the percentage of anchored items on which the model's error increased relative to the unanchored control. Plausible anchors (right) consistently produce a larger harmful share than irrelevant anchors (left), with the widest gaps for External and Tool suites.

Suite-level patterns. External shows the sharpest relevance contrast: plausible anchors produce harmful shifts 48.0% of the time versus 19.9% for irrelevant anchors ($\Delta = 28$ pp). Tool follows with a large gap (39.9% vs. 17.6%, $\Delta = 22$ pp). RAG shows a moderate gap (27.9% vs. 16.4%). ICL shows modest differentiation between relevance levels (25.4% vs. 20.3%), consistent with the weak discrimination in the standard ICL suite. History shows a 19 pp gap (48.7% vs. 29.5%).

Proximity analysis. When the anchor is farther from the gold answer than the control response, 31.0% of items show harmful shifts. When the anchor is closer to gold, only 14.9% are harmful. This directional asymmetry is expected: anchors that pull responses away from gold are more likely to increase error, while anchors that lie between the control and gold may coincidentally improve predictions. Among the "closer" items, helpful shifts are common (59.3%), confirming that the helpful category largely reflects cases where the anchor direction happens to align with gold.

Interpretation. The decomposition shows that plausible anchors increase prediction error more often than they reduce it (37.3% vs. 24.1%; the remaining 38.6% of items are unchanged), and substantially more often than irrelevant anchors. The 17 pp gap between plausible and irrelevant harmful rates (37.3% vs. 20.2%) mirrors the discrimination observed in the main UAI analysis and confirms that the anchor-induced shifts are not symmetric noise but directionally biased toward the anchor value.

Breakdown by difficulty and offset. Hard items produce more non-neutral shifts overall: harmful 31.6% vs. 25.8% for easy items, and helpful 25.4% vs. 21.1%. This is consistent with noisier evidence leaving more room for anchor-driven shifts in either direction. Across anchor offsets, plausible harmful rates are similar at δ=15 (35.5%) and δ=25 (36.1%) but decline at δ=40 (33.9%), while irrelevant harmful rates remain stable (≈18.4%). The moderate decline at δ=40 is consistent with the dose-response attenuation, partly amplified by boundary clipping (Appendix A.8).

A.14 Difficulty-conditioned analysis

Table 17 reports anchoring metrics separately for *easy* items (σ=8, all five ratings visible) and *hard* items (σ=15, two hidden, one conflicting).

Suite	Diff.	Acc_{10}	MAE_c	UAI_{irr}	UAI_{pls}	$Disc_\Delta$
EXTERNAL	easy	77.0%	7.2	0.05	0.16	0.11
	hard	65.5%	9.6	0.04	0.28	**0.23**
HISTORY	easy	77.1%	7.5	0.03	0.36	**0.33**
	hard	65.1%	10.1	0.08	0.31	0.23
ICL	easy	76.7%	6.7	0.03	0.02	−0.01
	hard	72.4%	7.8	0.02	0.03	**0.01**
RAG	easy	77.7%	6.8	0.02	0.07	0.05
	hard	63.3%	10.3	0.03	0.19	**0.16**
TOOL	easy	79.0%	8.0	0.09	0.17	0.08
	hard	79.2%	7.6	0.05	0.18	**0.13**
All	easy	—	—	0.04	0.16	0.12
	hard	—	—	0.05	0.20	**0.15**

Table 17: Anchoring metrics by item difficulty (mean across 14 models per suite, 13 on Tool). Bold highlights the higher $Disc_\Delta$ within each suite. Hard items show stronger discrimination on External and RAG; History shows the opposite pattern.

Findings. Hard items show higher overall susceptibility (mean $Disc_\Delta$: 0.15 vs. 0.12 for easy items), driven by External (0.23 vs. 0.11) and RAG (0.16 vs. 0.06). This is consistent with the expectation that models are more vulnerable to anchors when the evidence is sparser or more conflicting. History shows the opposite pattern (easy: 0.33 vs. hard: 0.23): easy items produce a stronger Stage 1 answer that serves as a clearer self-generated anchor, increasing susceptibility despite high task accuracy. ICL remains near zero for both difficulty levels.

A.15 Sampling robustness

All main results use greedy decoding (temperature 0). To assess whether conclusions depend on the decoding strategy, we re-evaluate two models (Qwen-7B, Llama-8B) on three suites (External, RAG, ICL-dist) under stochastic sampling (temperature 0.7, top-p 0.9) with three seeds.

Suite	Model	$Disc_{\Delta greedy}$	$Disc_{\Delta sampled}$	Sign
EXTERNAL	Qwen-7B	0.26	0.29 ± 0.02	✓
	Llama-8B	0.30	0.24 ± 0.07	✓
RAG	Qwen-7B	0.19	0.18 ± 0.02	✓
	Llama-8B	0.16	0.06 ± 0.04	✓
ICL-dist	Qwen-7B	0.07	0.08 ± 0.02	✓
	Llama-8B	0.03	0.00 ± 0.02	×

Table 18: Sampling robustness. Sampled values are mean ± SD over 3 seeds (temperature 0.7, top-p 0.9). Sign: whether greedy and sampled mean share the same $Disc_\Delta$ sign.

Findings. Qwen-7B produces nearly identical metrics under sampling (SD $\leq$ 0.02). Llama-8B is more variable (SD up to 0.07) but preserves the sign of $Disc_\Delta$ in 5 of 6 cells. The exception—ICL-dist, where greedy yields a small positive $Disc_\Delta$ (0.03) that averages to zero under sampling—involves a near-zero-effect cell where sign instability is expected.

Anchoring conclusions from the greedy main results are robust: plausible > irrelevant anchoring holds across decoding strategies for all suites with meaningful effect sizes.

A.16 Mitigation headroom probe

We test whether simple prompt-based strategies can attenuate anchoring effects. Three strategies are compared against the unmodified baseline on the External and RAG suites for two open-weight models:

- **Ignore:** an instruction to disregard extraneous numbers and base the answer only on the task evidence.
- **Self-check:** after producing an initial estimate, the model is asked to verify whether any extraneous numbers may have biased its answer, and correct if so.
- **CoT (chain-of-thought):** the model is asked to list the relevant evidence, compute an estimate from that evidence only, then state its final answer.

Suite	Strategy	Disc_Δ		MAE_{ctrl}	
		Qwen-7B	Llama-8B	Qwen-7B	Llama-8B
EXTERNAL	baseline	0.26	0.30	7.7	5.5
	ignore	0.16	0.10	5.0	3.4
	self-check	0.45	0.28	8.8	7.5
	CoT	0.14	0.31	1.1	6.1
RAG	baseline	0.20	0.16	9.8	4.8
	ignore	0.14	0.04	6.8	4.4
	self-check	0.30	0.11	12.2	6.0
	CoT	0.03	0.08	1.3	4.5

Table 19: Mitigation headroom probe. Disc_Δ and control MAE for three prompt-based strategies vs. unmodified baseline. *Ignore* consistently reduces discrimination across both models and suites. *CoT* strongly helps Qwen-7B but is inconsistent for Llama-8B. *Self-check* increases discrimination for Qwen-7B on both suites but slightly reduces it for Llama-8B.

Results. The *ignore* strategy consistently reduces Disc_Δ across all four model–suite cells (mean reduction: -0.12), with concurrent MAE improvement on External for both models. *CoT* strongly reduces discrimination for Qwen-7B (-0.12 on External, -0.17 on RAG) and moderately for Llama-8B on RAG (-0.08), but slightly increases it for Llama-8B on External ($+0.01$). The *self-check* strategy is model-dependent: it increases Disc_Δ for Qwen-7B on both suites (External: $+0.19$; RAG: $+0.10$) while slightly reducing it for Llama-8B (External: -0.01; RAG: -0.05). Differences of this size should not be read directionally. Re-running the Llama-8B$\times$External cell under an identical configuration moves Disc_Δ by up to 0.06 and UAI_{pls} by up to 0.08, because batched GPU inference is not bitwise reproducible even at temperature 0 (Appendix A.17.7, footnote 2). Only the larger effects ($|\Delta| \gtrsim 0.10$) in this table are resolved by a single run.

Interpretation. Simple prompt-based mitigations can partially attenuate anchoring, but no single strategy eliminates it and effectiveness is model-dependent. The *ignore* strategy is the most reliable, reducing discrimination by 30–75% on External and RAG. *CoT* is highly effective for Qwen-7B (which also sees large MAE improvements, down to 1.1–1.3) but inconsistent for Llama-8B. *Self-check* backfires for Qwen-7B, likely because revisiting the context gives the anchor a second opportunity to bias the response; for Llama-8B, the effect is small. These results suggest headroom for targeted interventions but indicate that robust debiasing will require more than instruction-level changes.

A.17 Additional analyses

This section collects supplementary analyses that probe the robustness, interpretation, and generality of the main findings. All use the standardized inference pipeline of §3.5 and, unless noted, the locked four-model open-weight panel (Gemma-4B, Llama-8B, OLMo-13B, Qwen-7B), chosen to span families and accuracy levels at manageable cost.

A.17.1 A rational-updating reference for UAI

A bare UAI value does not say whether a shift is excessive. We give it a reference point with a simple Gaussian-conjugate Bayesian model: treat the visible evidence as n ratings of unit credibility and the anchor as worth w ratings. A rational updater then closes a fraction

Model	External	History	Rag	Tool	Icl
Claude-H4.5	0.71	0.00	0.29	0.03	0.00
GPT-5.4-mini	0.79	0.23	0.47	0.33	0.00
Gemini-2.5-Flash	0.38	0.00	1.03	0.29	0.00
Gemma-1B	1.10	0.76	0.22	0.00	0.00
Gemma-4B	3.74*	5.00*	0.03	1.37	0.00
Grok-3-mini	0.97	0.53	0.11	0.08	0.02
Llama-1B	0.91	0.79	0.51	—	1.38
Llama-3B	2.09*	13.32*	1.43	1.35	0.78
Llama-8B	2.82*	1.52	0.69	4.58*	0.88
OLMo-13B	0.94	2.30*	0.25	0.40	0.08
OLMo-32B	3.27*	15.17*	4.04*	0.10	0.00
Qwen-1.5B	1.25	319.68*	1.46*	23.51*	0.02
Qwen-3B	0.50	4.88*	0.43	1.13	0.00
Qwen-7B	1.86*	2.99*	1.22	1.05	0.00

Table 20: Implied anchor weight w_{imp} on *plausible* anchors, defined as the smallest weight under a Gaussian-conjugate Bayesian model that would make the observed mean UAI rational (with $n = 5$ evidence ratings). Under the reference assumption that an anonymous anchor is worth one evidence item ($w = 1$), the rational UAI ceiling is 0.167; $w_{\text{imp}} > 1$ means the model treats the anchor as more credible than a single evidence item. $w_{\text{imp}} > n$ means the model treats the anchor as more credible than all visible evidence combined. * marks cells where the 95% CI for UAI_{pls} lies entirely above the rational ceiling.

Model	External	History	Rag	Tool	Icl
Claude-H4.5	-0.042	-0.226	-0.112	-0.161	-0.184
GPT-5.4-mini	-0.030	-0.123	-0.080	-0.105	-0.173
Gemini-2.5-Flash	-0.096	-0.171	+0.005	-0.111	-0.166
Gemma-1B	+0.013	-0.034	-0.125	-0.195	-0.193
Gemma-4B	+0.261*	+0.333*	-0.161	+0.048	-0.228
Grok-3-mini	-0.004	-0.071	-0.145	-0.151	-0.163
Llama-1B	-0.013	-0.031	-0.074	—	+0.050
Llama-3B	+0.128*	+0.560*	+0.055	+0.046	-0.031
Llama-8B	+0.194*	+0.067	-0.045	+0.311*	-0.018
OLMo-13B	-0.008	+0.149*	-0.120	-0.093	-0.150
OLMo-32B	+0.228*	+0.585*	+0.280*	-0.148	-0.174
Qwen-1.5B	+0.033	+0.818*	+0.060*	+0.658*	-0.162
Qwen-3B	-0.076	+0.327*	-0.088	+0.018	-0.204
Qwen-7B	+0.104*	+0.207*	+0.030	+0.007	-0.180

Table 21: Excess UAI on plausible anchors above the rational Bayesian ceiling 0.167 (assuming the anchor is worth one evidence item, $w = 1$, $n = 5$ evidence ratings). Positive values indicate the model shifted further toward the anchor than a rational Bayesian update could justify. * marks cells where the 95% CI for UAI_{pls} lies entirely above the ceiling.

$w/(n+w)$ of the control–anchor gap, so with n=5 and a reference weight w=1 the rational UAI ceiling is 0.167. For an irrelevant anchor the rational weight is 0, so the ceiling is 0 and any positive UAI_{irr} is bias. For a plausible anchor, we invert the relation to read off the *implied weight* $w_{\text{imp}} = \text{UAI}/(1 - \text{UAI}) \cdot n$, the smallest weight that would make the observed shift rational (Table 20). Table 21 reports the matching excess over the ceiling.

Across the 55 non-ICL model–suite cells, 16 have a UAI_{pls} whose 95% CI lies entirely above 0.167, and 5 imply $w_{\text{imp}} > n$=5—treating one anonymous anchor as more credible than all five ratings combined. These are the cells we flag as practically significant in §4.4; the rest sit inside the band that a rational update at w=1 would produce.

Model	$UAI_{placebo}$	UAI_{irr}	UAI_{pls}	UAI_{auth}
GPT-5.4-mini	-0.01	+0.01	+0.13	+0.40
Gemma-4B	+0.38	+0.20	+0.43	+0.58
Llama-8B	+0.08	+0.07	+0.36	+0.42
OLMo-13B	+0.02	-0.01	+0.16	+0.49
Qwen-7B	-0.00	+0.01	+0.27	+0.56
Mean	**+0.09**	**+0.05**	**+0.27**	**+0.49**

Table 22: Plausibility spectrum on External: mean UAI for each of four anchor types — placebo (anchor refers to an extraneous numeric property such as document length), irrelevant (case-number anchor), plausible ("a recent report suggested ∼23"), and authority ("a panel of senior analysts estimated ∼23"). A monotone increase along the plausibility axis would be consistent with rational use of source credibility; any positive UAI on the placebo column is unambiguously anchoring bias.

Model	Acc_{10}			UAI_{pls}			UAI_{irr}		
	k=1	k=2	k=3	k=1	k=2	k=3	k=1	k=2	k=3
Gemma-4B	54%	60%	64%	0.21	0.20	0.25	0.14	−0.13	0.13
Llama-8B	47%	61%	69%	0.68	0.36	0.50	0.25	−0.03	0.20
OLMo-13B	54%	76%	81%	0.48	0.34	0.29	0.19	0.11	0.06
Qwen-7B	42%	64%	75%	0.53	0.34	0.36	0.22	0.01	0.02
Rational ceiling (w=1)	—	—	—	0.50	0.33	0.25	—	—	—

Table 23: Uncertain-judgment results. Only k of 5 ratings are shown to the model; the gold answer remains the full-5 mean. An ideal Bayesian updater treating the plausible anchor as one additional rating of equal credibility (w=1) would produce $UAI_{pls} = w/(k+w)$ (Rational ceiling row). Measured UAI_{pls} above the ceiling indicates classical (super-rational) anchoring; below indicates the model weights the anchor less than one full additional evidence point. UAI_{irr} should stay near zero for capable models at every k.

A.17.2 Plausibility spectrum and a placebo floor

To separate anchoring on the numeric value from rational use of source credibility, we extend the binary relevance axis to a four-point spectrum on External: *placebo* (the number is an extraneous property such as a document's age in days), *irrelevant* (a case number), *plausible* ("a recent report suggested ∼23"), and *authority* ("a panel of senior analysts estimated ∼23"). A purely rational model would ignore placebo and irrelevant values equally and move only for plausible/authority framings.

Two observations argue against a purely rational reading. The placebo column is not zero (mean UAI 0.09) and sits close to irrelevant (0.05)—a shift toward a number that is literally a document length. And the increase along the axis is not clean: for some models the placebo shift rivals the plausible one. This is the behavior we expect if the numeric value itself is salient, not only its stated source.

A.17.3 Uncertain judgment under partial evidence

The main task shows all five ratings, so the gold answer is a fully determined mean. To introduce genuine epistemic uncertainty, we re-render External items showing only k of the five ratings while still scoring against the full-five mean. Now the anchor could rationally carry information, and the rational ceiling rises to $w/(k+w)$: 0.50, 0.33, 0.25 at k=1, 2, 3.

Measured UAI_{pls} frequently meets or exceeds the rational ceiling at each k (Table 23), and decreases as more evidence becomes visible—models do treat the anchor as one piece of evidence among several, but several cells still sit above what rational updating allows. UAI_{irr} stays comparatively low at every k.

Model	External			RAG			History		
	Mild	Std	Strong	Mild	Std	Strong	Mild	Std	Strong
Gemma-4B	0.45	0.42	0.40	0.00	−0.06	0.04	0.61	0.58	0.60
Llama-8B	0.26	0.39	0.42	0.10	0.09	0.29	0.44	0.20	1.57
OLMo-13B	0.07	0.16	0.29	0.02	0.09	0.10	0.45	0.24	0.91
Qwen-7B	0.23	0.31	0.49	0.07	0.19	0.29	0.37	0.24	0.52
Mean	**0.25**	**0.32**	**0.40**	**0.05**	**0.08**	**0.18**	**0.47**	**0.31**	**0.90**

Table 24: Cross-pathway plausibility-intensity dose-response. The same mild/standard/strong source-credibility manipulation is applied across three pathways. Positive Strong > Mild gaps indicate models weight source credibility; near-flat curves at high UAI indicate numeric anchoring dominates regardless of source. UAI values are mean across the panel; details in supplementary CSV.

		UAI_{irr}			UAI_{pls}			$Disc_{\Delta}$		
Suite	Model	bus	med	oth	bus	med	oth	bus	med	oth
external	Gemma-4B	0.20	0.28	0.20	0.43	0.21	0.67	0.23	-0.07	0.46
	Llama-8B	0.07	0.03	0.20	0.36	0.25	0.46	0.30	0.21	0.26
	OLMo-13B	-0.01	0.08	-0.03	0.16	0.09	0.22	0.17	0.01	0.24
	Qwen-7B	0.01	0.07	0.01	0.27	0.38	0.45	0.26	0.32	0.44
history	Gemma-4B	0.23	0.17	0.03	0.50	0.41	0.40	0.27	0.24	0.37
	Llama-8B	0.13	-0.23	0.42	0.23	0.25	0.36	0.10	0.48	-0.06
	OLMo-13B	-0.00	0.43	0.46	0.32	0.53	0.51	0.32	0.10	0.05
	Qwen-7B	-0.06	-0.37	-0.02	0.37	0.65	0.17	0.43	1.02	0.19

Table 25: Eleven-domain robustness panel: 6 business domains (published benchmark), 3 medical pilot domains, and 2 further pilot domains (legal contract compliance, consumer purchase decision). The pattern $UAI_{pls} > UAI_{irr} > 0$ replicates across all three domain families on both External and History, demonstrating that anchoring is not a business-domain artifact.

A.17.4 Cross-pathway credibility intensity

We apply a matched three-level source-credibility manipulation (mild / standard / strong) to three pathways (External, RAG, History). If models weight source credibility, the curve should rise from mild to strong; a flat-but-high curve indicates numeric anchoring that ignores the source.

External and RAG show a clean monotone increase, consistent with credibility-sensitive updating. History is non-monotone and much larger, because the credibility cue compounds with the model's own self-generated Stage-1 anchor; we therefore read the History intensity numbers qualitatively.

A.17.5 Domain extension beyond business scenarios

The six core domains are business-oriented. To check that anchoring is not a business-domain artifact, we run an 11-domain panel that adds three medical domains and two further extension domains (legal contract compliance, consumer purchase decisions), keeping the same item structure.

The ordering $UAI_{pls} > UAI_{irr} > 0$ holds in all three domain families on both External and History (Table 25). Magnitudes vary by family and model, but the qualitative pattern is stable, so the effect generalizes beyond the original domains.

	UAI_{irr}		UAI_{pls}		$Disc_{\Delta}$	
Model	mean	wmean	mean	wmean	mean	wmean
Gemma-4B	0.20	0.20	0.43	0.47	0.23	0.28
Llama-8B	0.07	0.06	0.36	0.34	0.30	0.28
OLMo-13B	-0.01	0.03	0.16	0.20	0.17	0.17
Qwen-7B	0.01	0.01	0.27	0.27	0.26	0.26

Table 26: Anchor influence on External under the published mean gold standard ("mean") vs. a weighted-mean gold standard with weights $[1, 1, 1.5, 1.5, 2]$ ("wmean"). The bias structure ($UAI_{pls} \gg UAI_{irr}$) is preserved under either scoring choice; the size of the anchoring effect is not an artifact of the simple-mean aggregation.

		UAI_{irr}			UAI_{pls}			$Disc_{\Delta}$		
Suite	Model	base	CoT	Δ	base	CoT	Δ	base	CoT	Δ
external	GPT-5.4-mini	0.01	-0.01	-0.01	0.13	0.14	+0.01	0.12	0.14	+0.02
	Gemma-4B	0.09	0.01	-0.08	0.37	0.23	-0.14	0.28	0.22	-0.06
	Llama-8B	0.08	0.06	-0.03	0.39	0.37	-0.03	0.31	0.31	-0.00
	OLMo-13B	0.06	0.05	-0.01	0.15	0.18	+0.03	0.08	0.13	+0.04
	Qwen-7B	0.01	0.00	-0.01	0.27	0.16	-0.12	0.26	0.15	-0.11
history	GPT-5.4-mini	-0.01	0.01	+0.01	0.02	-0.03	-0.05	0.03	-0.03	-0.06
	Gemma-4B	0.24	0.31	+0.07	0.57	0.64	+0.07	0.32	0.33	+0.00
	Llama-8B	0.08	0.27	+0.19	0.29	0.28	-0.01	0.21	0.01	-0.20
	OLMo-13B	-0.09	0.09	+0.18	0.51	0.17	-0.34	0.60	0.08	-0.52
	Qwen-7B	-0.10	-0.02	+0.08	0.42	0.13	-0.29	0.52	0.15	-0.37
rag	GPT-5.4-mini	0.00	-0.00	-0.01	0.09	0.08	-0.01	0.09	0.08	-0.00
	Gemma-4B	-0.03	0.07	+0.09	-0.05	0.25	+0.31	-0.03	0.19	+0.22
	Llama-8B	0.08	0.07	-0.02	0.20	0.12	-0.08	0.12	0.06	-0.06
	OLMo-13B	0.08	0.04	-0.04	0.05	-0.01	-0.06	-0.03	-0.05	-0.02
	Qwen-7B	-0.00	-0.00	+0.00	0.19	0.03	-0.16	0.19	0.03	-0.16

Table 27: Reasoning-allowed (Chain-of-Thought) prompting vs. "final answer only" baseline across the locked 5-model panel and all three suites (External, RAG, History). CoT lowers plausible-anchor influence in 11 of the 15 model–suite cells, but the effect is model- and pathway-dependent rather than uniform: it lowers both UAI_{irr} and UAI_{pls} in only 6 cells, and raises UAI_{pls} in 4. UAI_{pls} remains positive in 13 of 15 cells, so anchoring is attenuated but not an artifact of the answer-only protocol.

A.17.6 Robustness to the gold-standard choice

The gold answer is the simple mean of the ratings. One might worry that the anchoring magnitude is an artifact of this particular aggregation. We recompute UAI against a weighted-mean gold (weights $[1, 1, 1.5, 1.5, 2]$ over the five ratings).

The bias structure ($UAI_{pls} \gg UAI_{irr}$) is preserved, and the mean change in UAI_{pls} is only $+0.02$ (Table 26), so the effect is not an artifact of the simple-mean scoring.

A.17.7 Reasoning-allowed (chain-of-thought) evaluation

The main protocol asks for a final integer only. Here we let the model reason first (chain-of-thought) and compare to the answer-only baseline across five models and three suites.

CoT lowers anchor influence in 11 of 15 (model, suite) cells (mean $\Delta UAI_{pls} = -0.06$) but rarely removes it: $UAI_{pls} > 0$ under CoT in 13/15 cells.[2] The answer-only numbers in

[2]The Llama-8B×External×CoT cell reported here (UAI_{pls} $0.39 \rightarrow 0.37$, $\Delta = -0.03$) differs from the separate mitigation-headroom run in Appendix A.16 ($0.336 \rightarrow 0.338$, $\Delta = +0.00$). We traced this: the two runs render byte-identical prompts, gold answers and anchor values, and both use greedy

Model	UAI_{pls}			UAI_{irr}		
	Baseline	+Rule	+Judgment	Baseline	+Rule	+Judgment
Gemma-4B	0.51	0.08	0.38	0.27	0.02	0.25
Llama-8B	0.32	0.01	0.28	0.05	0.00	0.14
OLMo-13B	0.20	0.01	0.12	0.04	0.00	0.08
Qwen-7B	0.27	0.13	0.27	0.01	0.02	−0.00
Mean	**0.33**	**0.06**	**0.26**	**0.09**	**0.01**	**0.12**

Table 28: Task-specification ablation on External. Baseline = published prompt (return integer only); +Rule appends 'Your estimate should be the unweighted arithmetic mean of the visible ratings'; +Judgment appends 'Your estimate should be a weighted average ... weighting each piece according to which sources you find more or less credible'. A drop from Baseline to +Rule indicates the published anchoring effect is partly driven by task underspecification.

the main paper are therefore an upper bound on a reasoning-allowed deployment, with Llama-8B and the History pathway as the residual stress test.

A.17.8 Task specification: rule vs. judgment

To test how much of the anchoring effect is task underspecification rather than a deep numeric bias, we append two prompt suffixes on External: *+Rule* ("report the unweighted arithmetic mean of the ratings") and *+Judgment* ("report a weighted average, weighting sources by credibility").

The explicit rule cuts mean UAI_{pls} from 0.33 to 0.06, so a substantial share of the published effect is attributable to underspecification. But it does not vanish, and the residual under +Rule is the portion that confusion alone cannot explain. The +Judgment prompt, which invites credibility weighting, keeps UAI_{pls} below the published baseline, consistent with the rational ceiling of Appendix A.17.1. This addresses the concern that weak-model behavior reflects task confusion: even when told exactly how to aggregate, models retain measurable anchoring.

A.17.9 RAG realism: rank, distractors, and scores

The published RAG suite places the anchor document in a fixed slot of a three-document corpus. We vary the realism: move the anchor to the top (R1) or bottom of a five-document list with two plausible distractors (R5+D), and optionally expose synthetic per-document relevance scores (R5+D+Sc).

Plausible-anchor influence persists across rank and distractor changes, and is slightly higher at the top of the ranking—models privilege top-ranked retrieved documents, as a real RAG system would. It falls toward zero only when explicit relevance scores let the model down-weight a mid-ranked anchor, suggesting that production pipelines with good relevance signals would see weaker anchoring than the controlled suite.

decoding, so the difference is not a prompt or protocol difference. It is run-to-run variation of batched GPU inference, which is not bitwise reproducible even at temperature 0 because batch composition changes the reduction order inside the attention and GEMM kernels. An independent third run of this cell under the configuration used here gives $0.32 \rightarrow 0.35$ ($\Delta = +0.03$): across the three runs the baseline UAI_{pls} spans 0.32–0.39 and ≈8% of parsed answers change. The *sign* of the CoT effect for this single cell is therefore not resolved by one run. The aggregate conclusions above are unaffected: they are driven by cells whose effects (Δ up to -0.34) are an order of magnitude larger than this variation, and UAI_{pls} stays positive under CoT in all three runs. We report each run with its own numbers rather than re-normalizing across them.

Model	Plausible (UAI)				Irrelevant (UAI)			
	Base	R1	R5+D	R5+D+Sc	Base	R1	R5+D	R5+D+Sc
Gemma-4B	−0.06	0.00	0.01	0.01	0.03	−0.04	0.01	0.02
Llama-8B	0.09	0.12	0.14	−0.04	−0.02	0.01	0.01	0.01
OLMo-13B	0.09	0.02	0.21	0.04	0.04	0.09	0.13	0.05
Qwen-7B	0.19	0.23	0.14	0.11	0.01	−0.02	−0.02	0.00
Mean	**0.08**	**0.10**	**0.13**	**0.03**	**0.02**	**0.01**	**0.03**	**0.02**

Table 29: RAG realism ablation. Base = anchored doc at position 2 (published RAG layout); R1 = anchor at top of retrieval (rank 1); R5+D = anchor at bottom of a 5-doc list with 2 plausible distractor documents adjacent to the anchor; R5+D+Sc adds synthetic per-doc relevance scores. A drop from R1 to R5+D indicates models down-weight low-ranked retrieved anchors.

Model	Plausible (UAI)			Irrelevant (UAI)		
	Base	Elicited	Noisy	Base	Elicited	Noisy
Gemma-4B	0.13	0.28	0.29	0.31	0.35	0.39
Llama-8B	0.62	0.54	0.58	0.45	0.40	0.45
OLMo-13B	0.06	0.06	0.15	0.06	0.02	−0.01
Qwen-7B	0.17	0.02	−0.01	−0.00	0.01	0.00
Mean	**0.25**	**0.23**	**0.25**	**0.20**	**0.19**	**0.21**

Table 30: Tool realism ablation. Base reproduces the published Tool suite (externally injected tool output); Elicited prepends a synthetic model-planned tool-call turn (provenance signal); Noisy wraps the anchor value in a realistic JSON envelope containing additional metadata fields. Closing the gap between Base and Elicited indicates the anchor effect is not driven by the externally-injected framing; reductions under Noisy indicate the anchor's salience competes with surrounding metadata.

A.17.10 Tool realism: provenance and noise

We vary two tool-realism axes: *Elicited* prepends a synthetic model-planned tool-call turn (a provenance signal), and *Noisy* wraps the anchor value in a realistic JSON envelope with extra metadata fields.

On average the panel barely moves (mean $\mathrm{UAI}_{\mathrm{pls}}$ $0.25 \rightarrow 0.23 \rightarrow 0.25$), so the effect is not an artifact of the externally-injected framing. But the per-model picture splits: Qwen-7B collapses to near zero under both variants while Gemma-4B strengthens, which is why we treat Tool numbers as robust on average but not reliable per model (§4.1).

A.17.11 Extended large-model panel

To test whether scale removes anchoring, we evaluate five additional larger models on all five suites: two 70B-class open-weight models (Llama-3.3-70B, Qwen2.5-72B) and three frontier API models (GPT-5.4, Claude-Sonnet-4.6, Grok-4.3).

The main-panel pattern persists at scale (Table 31): irrelevant anchors stay near zero, plausible anchors are positive on 21/25 cells, and even GPT-5.4 (near-ceiling accuracy on every suite) shows a small but strictly positive $\mathrm{UAI}_{\mathrm{pls}}$ on 4/5 suites. Llama-3.3-70B on External ($\mathrm{UAI}_{\mathrm{pls}} = 0.23$) is the worst-case in this panel, comparable to Llama-3.1-8B in the main results.

A.17.12 Per-item case studies

To show that the aggregate UAI pattern reflects interpretable per-item behavior rather than an averaging artifact, we surface, for each panel model, the External items where

Model	EXTERNAL			HISTORY			ICL			RAG			TOOL		
	Acc_{10}	UAI_{irr}	UAI_{pls}	Acc_{10}	UAI_{irr}	UAI_{pls}	Acc_{10}	UAI_{irr}	UAI_{pls}	Acc_{10}	UAI_{irr}	UAI_{pls}	Acc_{10}	UAI_{irr}	UAI_{pls}
LLAMA-3.3-70B	95%	−0.02	0.23	91%	−0.02	0.01	87%	0.01	0.01	95%	−0.02	0.04	98%	0.01	0.05
QWEN2.5-72B	91%	0.01	0.16	93%	0.01	0.12	96%	−0.01	−0.03	88%	0.00	0.04	99%	0.01	−0.00
GPT-5.4	100%	0.01	0.19	99%	0.03	0.02	100%	0.00	0.00	100%	0.00	0.09	100%	0.00	0.03
CLAUDE-SONNET-4.6	99%	−0.00	0.19	99%	0.02	−0.02	72%	−0.02	−0.01	95%	0.03	0.07	95%	−0.00	0.06
GROK-4.3	97%	0.05	0.25	99%	0.09	0.03	97%	0.04	0.05	96%	0.02	0.18	91%	0.10	0.25

Table 31: Extended large-model panel. Task accuracy (Acc_{10}, control-only) and absolute anchor influence UAI_{irr}/UAI_{pls} across all five suites for two 70B-class open-weight models (top) and three frontier API models (bottom). The qualitative pattern of the main panel persists at scale: irrelevant anchors are near-zero, plausible anchors are positive.

Model	Item	a^{hi}	y_c	y_p^{hi}	y_i^{hi}	$\Delta_p - \Delta_i$
Qwen-7B	EXT-legal_policy-e-off40-005	85	43	75	43	+32
Qwen-7B	EXT-resource_consumption-e-off15-002	50	75	44	75	+31
Qwen-7B	EXT-pricing_wtp-e-off40-006	83	44	74	44	+30
Llama-8B	EXT-market_demographics-h-off40-005	70	23	100	26	+74
Llama-8B	EXT-resource_consumption-h-off25-010	65	100	14	56	+42
Llama-8B	EXT-transportation_logistics-h-off40-002	97	38	100	45	+55
Gemma-4B	EXT-transportation_logistics-e-off40-010	71	0	90	0	+90
Gemma-4B	EXT-resource_consumption-e-off15-010	66	0	100	57	+43
Gemma-4B	EXT-operations_time-h-off40-002	98	100	9	58	+49
OLMo-13B	EXT-market_demographics-e-off40-006	75	27	100	27	+73
OLMo-13B	EXT-resource_consumption-e-off15-002	50	35	100	35	+65
OLMo-13B	EXT-transportation_logistics-e-off40-009	82	43	100	43	+57

Table 32: Per-item case studies (External): for each model we surface the top discriminative items where the plausible-anchor shift cleanly exceeds the irrelevant- anchor shift, illustrating that the aggregate UAI pattern is driven by interpretable per-item behaviour rather than averaging artifacts. Full scenarios and per-condition answers are tabulated in the supplementary material.

the plausible-anchor shift most clearly exceeds the irrelevant-anchor shift (Table 32). In each case the control answer tracks the evidence mean, the irrelevant-anchor answer barely moves, and the plausible-anchor answer jumps toward the anchor value—the same value in both conditions.

A.17.13 LLM anchoring on the human effect-size scale

To position LLM anchoring against the human literature, we express the External-suite effect in |Cohen's d|, averaging the absolute per-direction effect (pooling high and low anchors, whose signed shifts cancel), following the convention of Furnham & Boo (2011).

Suite	$\lvert d_{\mathbf{pls}}\rvert$	$\lvert d_{\mathbf{irr}}\rvert$	$\lvert d_{\mathbf{hi\text{-}lo}}\rvert$
EXTERNAL	0.35	0.11	0.59
HISTORY	0.24	0.19	0.46
RAG	0.16	0.09	0.28
TOOL	0.23	0.12	0.36

Table 33: LLM anchoring magnitude in |Cohen's d| (open-weight panel mean). Read against the conventional small / medium / large bands of $d = 0.2/0.5/0.8$.

On External, the panel reaches $|d_{\text{pls}}|$ in the 0.2–0.5 range and $|d_{\text{hi-lo}}|$ in the 0.3–0.9 range: small to medium by the conventional bands, and reaching large for the high-versus-low contrast on the strongest models. The pattern $|d_{\text{pls}}| > |d_{\text{irr}}|$ replicates across RAG, Tool, and History at smaller magnitudes.

We deliberately stop short of a numerical head-to-head with human effect sizes. Human anchoring is normally measured between subjects on a single judgment, whereas UAI and the d values above are within-item contrasts on matched prompts, so the two are not on a

common scale, and the classical studies most often quoted for anchoring magnitude report medians, percentages, or F-statistics rather than a pooled d that can be reused here. What does transfer is the qualitative structure, and our data reproduce it: numbers that carry no task information still move judgments, plausibility modulates the size of the shift, and presentation format modulates it further—the same regularities reported by Tversky & Kahneman (1974), Strack & Mussweiler (1997), and Furnham & Boo (2011). For studies that do report effect sizes on comparable scales, Teovanović (2019) spans $d = 0.14$–1.00 across 24 items and Li et al. (2021) finds stronger anchoring for related than random anchors ($r = 0.42$ vs. 0.21), which is the same relevance ordering we measure. A paired study running humans and models on identical ANCHORBENCH items is the clean way to close this gap and is our primary follow-up.