

An Empirical Study of the Anchoring Effect in LLMs: Existence, Mechanism, and Potential Mitigations

Yiming Huang^{1*}, Biquan Bie^{2*}, Zuqiu Na¹, Weilin Ruan¹, Songxin Lei¹, Yutao Yue^{1,†}, Xinlei He^{1,†}

¹The Hong Kong University of Science and Technology, Guangzhou

²Independent Researcher

†Correspondence: {yutaoyue, xinleih}@hkust-gz.edu.cn

Abstract

The rise of Large Language Models (LLMs) like ChatGPT has advanced natural language processing, yet concerns about cognitive biases are growing. In this paper, we investigate the anchoring effect, a cognitive bias where the mind relies heavily on the first information as anchors to make affected judgments. We explore whether LLMs are affected by anchoring, the underlying mechanisms, and potential mitigation strategies. To facilitate studies at scale on the anchoring effect, we introduce a new dataset, *SynAnchors*. Combining refined evaluation metrics, we benchmark current widely used LLMs. Our findings show that LLMs' anchoring bias exists commonly with shallow-layer acting and is not eliminated by conventional strategies, while reasoning can offer some mitigation. This recontextualization via cognitive psychology urges that LLM evaluations focus not on standard benchmarks or over-optimized robustness tests, but on cognitive-bias-aware trustworthy evaluation.

1 Introduction

Since the appearance of milestone work ChatGPT (OpenAI, 2023), Large Language Models (LLMs) have profoundly shifted human-computer interaction and its applications to a brand new level. The growing demand for trustworthy LLM assistants makes cognitive biases inherited from biased human-mimicking features in training corpora a critical concern. Although LLMs surpass humans in standard benchmarks, their psychological traits remain understudied despite their growing importance. The anchoring effect (Tversky and Kahneman, 1974a), a prominent example of such biases, lacks comprehensive research in LLM. The anchoring effect refers to the phenomenon that humans unconsciously attach too much importance to the first piece of information (also called anchors) in

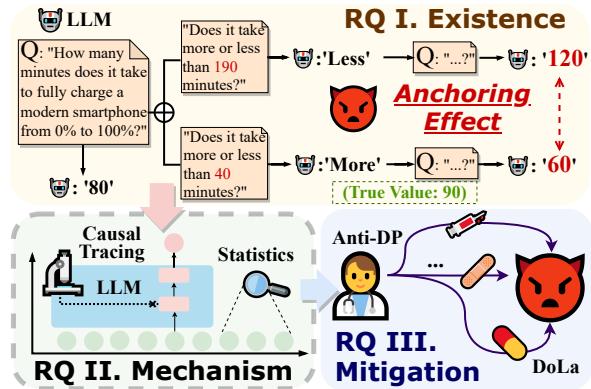


Figure 1: Detailed illustration of our research on LLM's anchoring effect from three key aspects: (1) Existence: showing significant biases toward different anchor values for identical questions. (2) Mechanism: using causal tracing and statistics to explore underlying patterns. (3) Mitigation: evaluating across varied mitigation strategies. ‘Q: “...?”’ refers to asking the same question again.

the decision-making process, causing subsequent judgments biasedly affected by these anchors. For example, if people were asked to estimate today's stock price of a certain company without any reference, their guesses may vary widely. However, if they were first told the stock prices of different companies in the same industry (e.g., \$10, \$50, or \$300), their estimates would be unconsciously biased toward these values, even if those prices are impractical to common sense. Current LLM research excessively prioritizes over-optimized evolution metrics, while common trustworthiness works (e.g., Adversarial Attack, LLM Red Teaming) remain disconnected from human cognition. This disconnection needs to be bridged through quantified measurement of cognitive biases, where the anchoring effect with our refined metrics matches this need. To be specific, these anchors are usually concrete values of indirectly related or even irrelevant objects in context, and we measure the discrepancy with/without different anchors. Based

* Equal contribution.

on that, we mainly focus on three research questions (RQs) shown in Figure 1 of the anchoring effect:

1. **RQ1: Are current modern LLMs easily affected by anchoring hints?**
2. **RQ2: What is the underlying mechanism of the anchoring effect in LLMs?**
3. **RQ3: What are the possible mitigations to this anchoring effect?**

We study the two most common paradigms of anchoring effects: semantic priming (Jacowitz and Kahneman, 1995) and numerical priming (Wilson et al., 1996). The former uses a “standard two-step procedure” that first lets the testee LLM give qualitative estimates about high/low anchors in question then asks again. The latter compares the answers with and without the interfering anchors. Different from the previous studies (Lou and Sun, 2024; Nguyen, 2024; E. O’Leary, 2025), we are the first to use an advanced LLM (DeepSeek-R1) to synthesize the anchoring question dataset named *SynAnchors* for evaluation. The curation of questions and anchors is carefully conducted through the human-LLM-loop, yielding questions that effectively elicit anchoring effects. Combining refined evaluation metrics proposed in RQ1, it is suitable for benchmarking the anchoring effect in LLMs. In RQ1, we quantitatively finalized the evaluation metrics of the presence and intensity of the anchoring effect in LLMs. The empirical results in RQ1 show that modern LLMs are anchored at different levels across a wide range of questions, with 22% to 61% questions anchored. While the reasoning models are relatively mild, achieving the lowest ratio. In RQ2, we are the first work to explore internal patterns of LLM cognitive biases through the internal causal tracing technology. We achieve it by practicing the commonly acknowledged activation patching method (Meng et al., 2022a; Zhang and Nanda, 2023) in mechanistic interpretability studies. The causal tracing outcomes show that the input tokens related to this effect are not as significant as we initially infer. It showcases the shallow salience of the anchoring effect within LLMs’ inner workings. In addition, the statistical results of the reasoning models’ outputs indicate that reasoning has a certain correlation to alleviate these effects. Based on these findings, we propose our mitigation strategy **Anti-DP** (Anti-Dual-Process,

detailed in Section 6.1) and compare it with various conventional approaches, achieving up to 10% decrease. From the results of the mitigation experiment, although our method achieves better alleviation than conventional approaches, all mitigation strategies fail to eradicate cognitive bias at its root. These result shows a valid direction for debiasing LLMs, rather than optimizing on standard benchmarks Overall, the contribution of our paper can be summarized as follows: (1) We provide the dataset *SynAnchors* with relative evaluation metrics for benchmarking and studying the anchoring effect of LLM. (2) We comprehensively measure the severity of this cognitive bias in modern LLMs. (3) We mechanistically explore the inner pattern of the anchoring effects and highlight that reasoning is a certain cure to break this shallow anchoring. (4) We test possible mitigation strategies, including our proposal, stressing the ineffectiveness of the conventional methods, and identifying promising directions starting from our limited proposal.

2 Background

2.1 Anchoring Effect & LLM

The anchoring effect is a ubiquitous cognitive bias (Furnham and Boo, 2011) and influences decisions in many fields (Jacowitz and Kahneman, 1995). Under uncertainty, people’s decisions tend to be influenced by initial information, or “anchor”, which causes their subsequent judgment to drift closer to it (Tversky and Kahneman, 1974b). The anchoring effect can be categorized into semantic and numerical priming paradigms. Subjects are asked a relative qualitative judgment question and are subsequently asked the specific value. It is known as the semantic priming paradigm (Jacowitz and Kahneman, 1995; Mussweiler and Strack, 2001) and the two-step procedure where both questions are semantically related (Wong and Kwong, 2000). Wilson et al. (1996) introduced the numerical priming paradigm. Exposure to an irrelevant number can bias numerical estimates. This effect occurs in a one-step procedure, hence the term numerical priming paradigm. Recent studies have explored whether LLMs exhibit human-like anchoring effects. CBEval (Shaikh et al., 2024) examines several biases, including anchoring, but lacks clear paradigms and diverges from established methods like the two-step procedure, limiting meaningful comparison. Nguyen (2024) and Lou and Sun (2024) adopt semantic priming paradigms, though

Nguyen focused only on financial data or limited domain questions, making the scope too narrow and limiting robustness. They test three types of anchoring hints, showing their impact on model output. However, these works focus mainly on surface-level evaluation and mitigation, without probing the mechanisms behind LLM anchoring. Our study draws directly from cognitive psychology and clearly distinguishes between semantic and numerical priming paradigms. We design dedicated experiments with clear operational definitions and tighter control, enabling more precise evaluation and mechanical insight into LLM anchoring. As for a deeper theoretical level, the anchoring effect is often explained by the Dual Process Model (Kahneman, 2011): **Automatic Adjustment (System 1)**: Humans initially rely heavily on the first piece of information, producing a rapid and unconscious judgment. **Effortful Correction (System 2)**: Subsequently, humans may attempt to adjust their judgment away from the anchor through a more deliberate and effortful process. More theoretical basis is demonstrated in Appendix E.

2.2 Trustworthy Artificial Intelligence: a Psychology-Combined Perspective

Research shows LLM exhibit human-like biases, underscoring the need for psychological insights in AI to understand and mitigate these effects (Liu et al., 2022; Brown et al., 2020). Existing psychological-combining research on LLMs has evolved into three main directions: (1) comprehensive reviews and benchmark construction; (2) empirical comparisons of cognitive behaviors between LLMs and humans; and (3) conventional safety-oriented trustworthy AI. The first of them focuses on reviews and develops benchmarks for evaluating bias in LLMs. Survey studies document bias across diverse architectures and applications (Gallegos et al., 2024; Li et al., 2023; Dong et al., 2025). Koo et al. (2023) benchmark translates cognitive bias categories into tasks to evaluate LLMs' susceptibility. However, current works remain high-level, lacking detailed conclusions regarding the mechanisms behind specific biases. The second direction applies psychological theories and cognitive frameworks to evaluate LLM behavior. For instance, Wang et al. (2023) explores the primacy effect in ChatGPT, and tools like CB-Eval (Shaikh et al., 2024) and CognitiveLLM (Wu et al., 2024) have benchmarked various biases. Yet these methods often yield gener-

alized, surface-level correlations—tagging lexical patterns without differentiating bias types or revealing the mechanisms by which biases arise in language models. Consequently, they offer only a “thin description” of LLM cognition. The third range across a wide trustworthy AI domain, like knowledge conflicts and various humanized poisoning attacks (Luo et al., 2025; He et al., 2025; Zhang et al., 2024). But they finally locates their contribution to general LLM robustness issues rather than granular cognitive bias. This study confirms the anchoring effect in LLMs and goes beyond remaining at the surface level of lexical correlations by tracing how models process cues of varying semantic depth (Meng et al., 2022b; Zhang and Nanda, 2023) across model layers. Our analysis reveals internal processing dynamics across layers and informs targeted mitigation strategies that address the underlying causes.

3 Psychological Test Formalization

3.1 Targeted Anchoring Effect Paradigms

The first paradigm we aim to study is semantic priming anchors. As demonstrated in Section 2.1, it is triggered by a “standard two-step procedure”. The first step is asking the testee (LLM) to give a qualitative estimate “higher” or “lower” compared to asking values (anchors), which implicitly introduces anchors into the decision-making process. The second step asks the testee (LLM) again by same question, but requests its direct answer with the previous context. The collection of conversation results is suitable for quantitative research, and we also make our dataset in this format as below:

Conversation Format of Semantic Test

System Prompt: *You are a helpful assistant. Answer user questions concisely, providing only the necessary information. Avoid full sentences. You cannot refuse to answer, at least answer your estimation. For numerical answers, please provide a number without any spaces, and keep the same units as the question asked.*

Question 1: <question> + <anchor_text> + *Your response should be only the word ‘Higher’ or ‘Lower’.*

Expected Answer 1: <‘Higher’ / ‘Lower’>

Question 2: <question>

Expected Answer 2: <value> .

Here, the question <question> and the an-

choring hints `<anchor_text>` are prepared in our dataset, e.g., ‘How many fjords are there in Norway?’ and ‘Is it higher or lower than 2500?’. The expected answer in the first step `<‘Higher’/‘Lower’>` and the expected answer in the second step `<value>` are extracted by the assistance model Qwen2.5-1.5B-Instruct. Another paradigm is the numerical priming anchors. As Section 2.1 mentioned, it judges the discrepancy of answers with and without irrelevant numerical statements as anchors. We settle this paradigm in the following conversation format:

Conversation Format of Numerical Test

System Prompt: *You are a helpful assistant. Answer user questions concisely, providing only the necessary information. Avoid full sentences. You cannot refuse to answer, at least answer your estimation. For numerical answers, please provide a number without any spaces, and keep the same units as the question asked.*

Question: `<anchor_text> + <question>`

Expected Answer: `<value>`.

Similar to semantic ones, `<question>` and `<anchor_text>` are prepared in our dataset, e.g., ‘What is the weight of a pelican (kg)?’ and ‘The slot machine stopped on 114.’ Expected answer `<value>` is extracted by the same assistance model.

3.2 Dataset Construction

For semantic anchoring questions, we constructed 60 questions over 10 diverse topics, covering various aspects of factual knowledge domains and subjective real-world decision scenarios. For numerical anchoring questions, we constructed 40 questions across 10 topics, primarily focusing on quantifiable factual measurements. The questions were intentionally kept concise to emphasize the impact of numerical priming effects. The dataset is constructed by using a human-in-the-loop methodology to ensure data quality, diversity, and relevance. The iterative process involved the following steps: 1. **Initial Generation:** Based on a small set of seed questions provided by human annotators, we employ DeepSeek-R1 to generate a larger pool of candidate questions, mimicking the structural and topical patterns of the seeds. 2. **Human Curation and Filtering:** Human annotators meticulously review the generated questions against a set of principles: verifiable with true value, unfamiliarity to common LLMs, balance of topics, di-

versity of anchoring items, and linguistic diversity.

3. **True Value Determination:** Human annotators determine the true value of filtered questions through rigorous web searching and verification. 4. **Iterative Refinement:** Human annotators iteratively refine the LLM’s output through structured feedback, until the target volume of high-quality questions is achieved. This rigorous construction process mitigates the risk of LLM memorization and promotes significant diversity and richness in both content and linguistic style. The resulting dataset serves as a well-curated and comprehensive benchmark, enhancing the generalizability and validity of the findings regarding anchoring effects in LLMs. More details about the dataset making are in Appendix B.

4 Testing Anchoring Effect Level in Current LLMs (RQ1)

4.1 Evaluation Standards and Setups

For testing the semantic priming anchoring effect, we query LLM 100 times (sampling decoding) for both high anchor and low anchor, each question, in the conversation format illustrated in Section 3.1. We collect pure value answers in `<value>` and perform the independent t-test over answers of the high anchor group and low anchor group. The statistical significance p indicates the presence of the anchoring effect. Besides the presence, the intensity is also important. In line with typical psychology research mentioned in Section 2.1, we use the Anchor Index (Jacowitz and Kahneman, 1995) (**A-Index**) for evaluation:

$$\text{A-Index} = \left| \frac{\text{Median}_{high} - \text{Median}_{low}}{\text{Anchor}_{high} - \text{Anchor}_{low}} \right|.$$

Here, Median_{high} and Median_{low} are the median value in each group, Anchor_{high} and Anchor_{low} are the specific high anchor and low anchor value of each group. Because some LLM has completely opposite cognition on a few questions, leading to the minus value **A-Index**, we take the absolute value for computing undistorted overall intensity across datasets. This widely used index in the psychology domain is around 0.4~0.6 across human tests (Jacowitz and Kahneman, 1995; Zong and Guo, 2022; Yasseri and Reher, 2022). Therefore, comprehensively considering the presence and intensity of the anchoring effect, we count the question yields $p < 0.05$ and $\text{A-Index} > 0.4$ as an obvious occurrence of the anchoring effect. As for numerical ones, we query the testee LLM 100 times for

Model / Metrics	Semantic		Numerical		Total Ratio% ↓
	A-Index ↓	Ratio% ↓	R-Error ↓	Ratio% ↓	
Qwen2.5-0.5B-Instruct	0.500 [‡]	50.0% [‡]	0.540 [†]	61.1% [†]	60.5%
Llama-3.2-1B-Instruct	0.618 [‡]	47.7% [‡]	0.623	67.5%	57.1%
Phi-3.5-mini-instruct	0.590 [‡]	58.6% [‡]	0.299 [†]	36.8% [†]	46.3%
Qwen2.5-7B-Instruct	0.463	43.3%	0.233	35.0%	40.0%
Mistral-7B-Instruct-v0.3	0.606 [†]	63.0% [†]	0.230 [†]	34.2% [†]	51.1%
Falcon3-7B-Instruct	0.389 [‡]	38.5% [‡]	0.332	45.0%	42.4%
Llama-3.1-8B-Instruct	0.394	38.3%	0.270 [†]	29.0% [†]	34.7%
GPT-4o-mini	0.475	48.3%	0.164	20.0%	37.0%
GPT-4o	0.340	36.7%	0.114 [†]	12.8% [†]	27.3%
Qwen3-235B-A22B (<i>Thinking Mode</i>)	0.321 [#]	33.3% [#]	0.080	5.0%	22.0%
DeepSeek-R1 (<i>DeepThink Mode</i>)	0.278	31.7%	0.112	15.0%	25.0%

Table 1: Evaluation of semantic and numerical performance. Total ratio represents the overall occurrence of the anchoring effect across both semantic and numerical questions. Superscript indicates the percentage of invalid results (if exists): ‘[†]’ is < 10%, ‘[‡]’ is ≥ 10%. ‘[#]’ indicates results are based on 30 samples per question due to budget limits. A deeper red background of the row means a stronger anchoring effect.

each question with and without the anchoring hints `<anchor_text>`, which contains random irrelevant numerical anchors. Thus, we get 100 pairs of answers, and the presence anchoring effect is judged by the statistical significance p of the paired t-test. To measure the intensity, we use the relative error (**R-Error**) as a metric:

$$\textbf{R-Error} = \text{Mean}\left(\left|\frac{v_{\text{anchor}} - v_{\text{orig}}}{v_{\text{orig}}}\right|\right).$$

Here, v_{anchor} is the answer `<value>` with anchoring hints and v_{orig} is the without ones. We regard $p < 0.05$ and **R-Error** > 0.2 as sufficient proof of the occurrence of the anchoring effect. During the process of statistical computation, we also implemented the following data preprocessing procedure: (1) Same as previous psychology studies (Yasseri and Reher, 2022), we take out the largest 15% and smallest 15% for eliminating the extreme cases. (2) Once LLM’s answer is unextractable for our assistance model, we remove the null answer from the high/low anchor group, and we pairily remove the pair in numerical answers (with/without anchor hints groups). If there are fewer than 30 answers in each group or 30 pairs in total, this question will not be counted for the measurement of the anchoring effect. Thus, we present anchoring ratios in Table 1, and the severity of failed-to-follow-instruction is marked in these ratios. (3) To better present overall effects across all questions, we apply the maximum truncation (1.0) to copy the corner case of several extremely large anchoring indices or relative er-

rors. Equipped with the above evaluation metrics and corresponding procedures, our dataset *SynAnchors* is suitable for benchmarking the anchoring bias in LLM.

4.2 Empirical Results

Target Models. To ensure the diversity of the model’s output, we use the default hyperparameters for sampling of the LLM. We choose 4 representative sorts of LLM for our evaluation:

- **Tiny Models:** Qwen2.5-0.5B-Instruct (Yang et al., 2024) and Llama-3.2-1B-Instruct (Grattafiori et al., 2024).
- **Standard Light Models:** Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), Phi-3.5-mini-instruct (Abdin et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Falcon3-7B-Instruct (Team, 2024).
- **Advanced Large Models:** GPT-4o-mini (Hurst et al., 2024) and GPT-4o (Hurst et al., 2024) (through their API).
- **Advanced Reasoning Models:** DeepSeek-R1 (Guo et al., 2025) and Qwen3 (Yang et al., 2025) (through their API).

Results. We perform the anchoring effect evaluation as Table 1 presents. Above all, it is intuitive that the anchoring effect is widely occurring in current LLMs, even powerful reasoning models exhibit a non-negligible extent of this biased pattern.

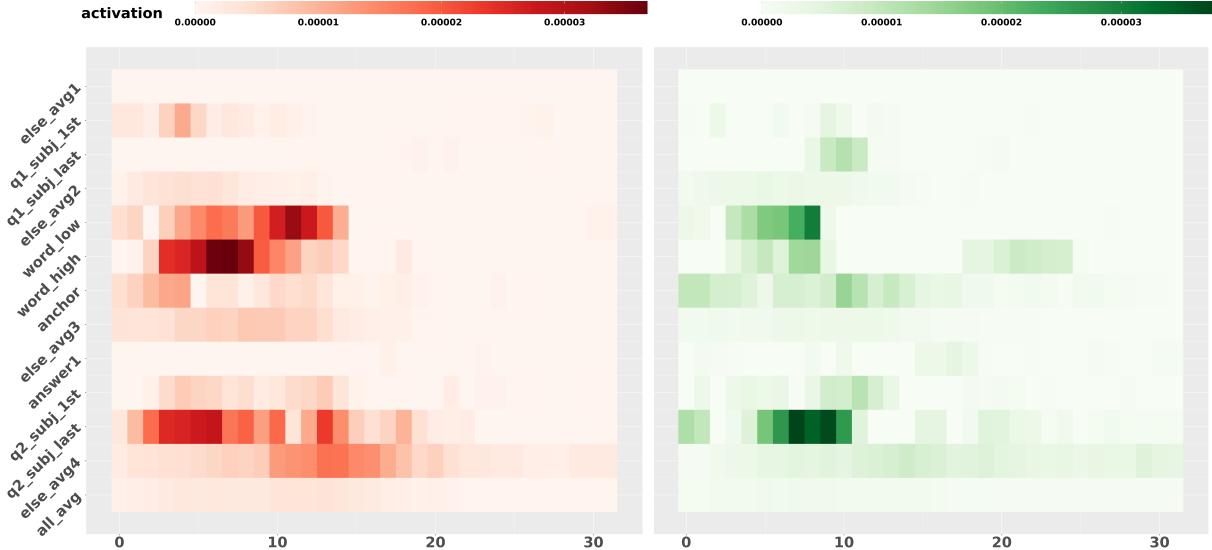


Figure 2: Causal tracing on attention (red) and FFN (green) modules of Llama-3.1-8B-Instruct about semantic anchoring questions. The X-axis represents the layer index of the model (32 layers). The Y-axis is the ROI tokens.

Naturally, advanced models show a mild anchoring effect compared to less-advanced ones, as the bigger models are less biased than small models, and reasoning models achieve the best performance in our experiments. As expected, numerical questions trigger fewer anchoring effects due to the irrelevant anchors are weaker in constructing cognitive connections for LLMs. Compared to the general 0.4~0.6 **A-Index** value of human, most LLM yields a more serious or the same level of anchoring effect in our result. To conclude, the anchoring effect is prevalent in the majority of LLMs, which is more challenging to LLMs’ trustworthiness compared to normal benchmarks like MMLU (Hendrycks et al.) or over-optimized safety test TruthfulQA (Lin et al., 2022) (where these LLMs achieve human-comparable performance).

5 Mechanistically Explorating LLMs’ Anchoring Effect (RQ2)

5.1 Causal Tracing Analysis

Methodology. Following previous work (Meng et al., 2022b; Zhang and Nanda, 2023), we adopt the same activation patching measures for causal tracing. It executes three types of runs to trace the causal effects of the anchoring hints: the clean run, the corrupted run, and the restoration run. These runs allow us to quantify how the hidden states at specific layers of key tokens influence the model’s final response through recovering patching (activation patching), which is detailed in Appendix A.

Setup. We excutes the experiment on **Llama-3.1-8B-Instruct**. We randomly select 6 semantic questions and 4 numerical questions that are judged as anchored in the last section. We mainly focus on the internal pattern of the two modules. The one is the attention modules, we perform the activation patching at hidden states before the attention output weight matrix multiplication. The other is the feed-forward networks (FFN), where we perform the activation patching at hidden states before the second down-project weights multiplication.

Target Tokens. Since the research goal is to find out how tokens related to anchoring hints work internally within the LLM, these special tokens need to be marked as ROI tokens for causal tracing. Besides tokens highly related to anchoring hints, like tokens of anchors, tokens about “higher” or “lower”, etc., these Region of Interest (ROI) tokens also include subject/object tokens for comparison, like **q1_subj_1st**, **q1_subj_last**, **anchor**, and **word_high**. In addition, all remaining tokens’ internal restore results are recorded for comparison as well. Concrete correspondence of the target tokens’ marks is listed in Appendix A.

Results. The results of the causal tracing analysis are shown in Figure 2 and Figure 4 (in Appendix A), which generally indicate the shallowness of the anchoring effect. To be specific, it is obvious in the attention module of Figure 2 that the words ‘higher’ and ‘lower’ and the anchor value in the anchor hint text <anchor_text> affect the

LLM’s prediction on a crucial level. Compared to subject tokens in **Question 2** that we regard as LLMs’ nature to pay attention to subject tokens for answering questions, the significance of these tokens does not dominantly surpass common subject tokens in the question, and even the answer of estimation barely has any significance. FFN modules in Figure 2 are similar, the tokens related to anchoring hints show smaller significance than the subject tokens of **Question 2**. Meanwhile, all of the hidden states corresponding to anchoring hints only exhibit significance before the middle layers (layer 16) and do not produce a high-level semantic shift in the high layers. We can conclude that the semantic priming anchoring effect mainly happens in the early stage of LLM, especially the detokenization stage (Kaplan et al., 2025; Kamoda et al., 2025), when deeply related tokens are close together as the basic ingredients of high-level semantics. As for numerical ones, tokens of irrelevant numerical anchoring hints contribute to some extent to both attention and FFN modules. These contributions are still lighter than the subject tokens in question and rarely demonstrate salience in higher layers.

5.2 Statistics of Horizontal Reasoning

In Section 4.2, the reasoning models exhibit relatively mild symptoms of anchoring, and the anchoring effect’s influence tends to be shallow in the LLM’s inner workings as discussed above. To better understand whether “horizontal reasoning” can alleviate this shallow vertical anchoring effect, we perform further analysis. Figure 3 shows statistics

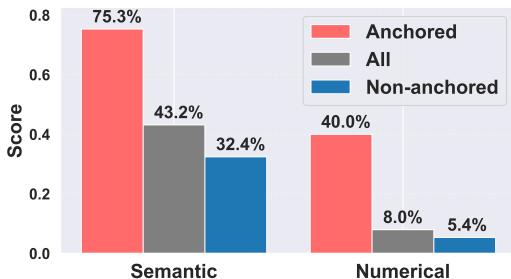


Figure 3: Percentages of sufficient anchor information mentions in DeepSeek-R1 reasoning contents. **Legend:** “Anchored” refers to the percentages of questions judged as anchored based on the metrics introduced in Section 4.1; “All” and “Non-anchored” indicate the percentages over all questions and those judged non-anchored, respectively. We employ an LLM-as-a-Judge approach to automatically detect explicit mentions of anchor-influenced features in reasoning contents, guided by detailed criteria defining what extent can be counted as sufficient mention (see more in Appendix C).

on the ratio of anchor mentions within reasoning contents. It suggests that questions exhibiting mild anchoring effects tend to include fewer explicit mentions of anchor information during reasoning. This implies that other information gradually dilutes the shallow anchoring effect in early conversation. This finding is the basis of our proposed mitigation strategy.

6 Possible Strategies for Mitigation (RQ3)

6.1 Mitigation Setup

We evaluate several mitigation strategies on two currently widely used instruction-tuned LLMs: Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct. For all experiments. For prompt-free strategies, the conversation format remains identical to that used in Section 4.1. Prompt-based strategies involve additional texts and/or conversational turns. We compare conventional strategies (See Appendix D) widely adopted in research on trustworthy AI works, such as those aimed at LLM unbiasing and hallucination elimination. Furthermore, combining insights from RQ1 (Section 4.2) and RQ2 (Section 5.2), we propose a novel strategy, Anti-DP (Anti-Dual-Process), by integrating reasoning. The considered strategies are (detailed in Appendix D):

- **Question-Aware Prompt:** Ask the LLM to be aware of the question, such as “Please think carefully and cautiously about the question before providing your answer.”
- **Knowledge Enhancement:** Provide the LLM a piece of helpful background knowledge without revealing the answer.
- **Self-Improving:** Give the LLM an additional turn to refine its previous answer.
- **Adversarial Finetuning:** Finetune the LLM with its unbiased conversations using the LoRA (Hu et al.) (Low-Rank Adaptation).
- **DoLa Decoding:** (Chuang et al.) Modify LLM decoding strategy by contrasting the prediction between an early (high or low) layer and the final layer.
- **Anti-DP (Ours):** implement a two-phase reasoning intervention. Upon receiving the question, the LLM is first instructed to establish its standard to guide the rethinking of the question. Then, the LLM produces the final answer based on prior thoughts. This approach explicitly opposes the automatic dual-process by encouraging integrated, iterative reasoning.

Mitigation Strategy	Semantic		Numerical		Total Ratio% ↓
	A-Index ↓	Ratio% ↓	R-Error ↓	Ratio% ↓	
Llama-3.1-8B-Instruct	0.394	38.3%	0.270	29.0%	34.7%
+ Question-Aware Prompt	0.368 ↓	38.3%	0.286*	34.3%*	36.8%
+ Knowledge Enhancement	0.410	43.3%	0.447*	62.2%*	50.5%
+ Self-Improving	0.368 ↓	36.7% ↓	0.298*	41.0%*	38.4%
+ Adversarial Finetuning [◇]	0.394	38.3%	0.257* ↓	25.0%* ↓	33.3% ↓
+ DoLa Decoding (<i>low</i>)	0.369 ↓	36.7% ↓	0.308*	38.9%*	37.5%
+ DoLa Decoding (<i>high</i>)	0.377 ↓	38.3%	0.308*	38.9%*	38.5%
+ Anti-DP (Ours)	0.305* ↓	19.0%* ↓	0.250* ↓	33.3%*	24.7% ↓
Qwen2.5-7B-Instruct	0.463	43.3%	0.233	35.0%	40.0%
+ Question-Aware Prompt	0.470	46.7%	0.312	37.5%	43.0%
+ Knowledge Enhancement	0.418 ↓	38.3% ↓	0.318	42.5%	40.0%
+ Self-Improving	0.557	55.0%	0.291	37.5%	48.0%
+ Adversarial Finetuning [◇]	0.464	43.3%	0.252	27.5% ↓	37.0% ↓
+ DoLa Decoding (<i>low</i>)	0.420 ↓	43.3%	0.283	35.0%	40.0%
+ DoLa Decoding (<i>high</i>)	0.393 ↓	40.0% ↓	0.283	35.0%	38.0% ↓
+ Anti-DP (Ours)	0.344* ↓	34.5%* ↓	0.315	50.0%	40.8%

Table 2: Evaluation of mitigation strategies on semantic and numerical tasks. Green arrows (↓) indicate the degree of mitigation, with a deeper color representing better mitigation. ‘*’ denotes cases with $\leq 10\%$ invalid results (if exist). ‘ \diamond ’ indicates results are derived on test splits, which exclude train splits of LoRA.

6.2 Mitigation Result

Table 2 presents the performance of the evaluated mitigation strategies on Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct under the specified sampling setting. Mitigation strategies achieved up to a 10% alleviation in anchoring effects, confirming the failure of elimination, while some reductions are possible. Performances are inconsistent, with DoLa Decoding notably demonstrating a more pronounced decrease on semantic tasks for both models. Nevertheless, our proposed **Anti-DP** strategy reduces the anchoring effect (except numerical questions on the Qwen model), reinforcing our finding that incorporating an explicit reasoning process during inference helps to erode shallow anchoring.

7 Discussion

Our results show that anchoring effects are prominent in LLMs (RQ1), Enhancing reasoning may offer a path to reduce such shallow biases (RQ2), and current mitigation methods fail to fully address them (RQ3). The **Anti-DP** intervention introduces an alternative activation path that may interfere with or override the anchor-primed semantic activation, akin to inhibitory control mechanisms (Rueda et al., 2005; Munakata et al., 2011) observed in human cognition, allowing the model to suppress automatic responses in favor of goal-directed be-

havior. This effect parallels human cognition: in dual-process theories (Kahneman, 2011), intuitive judgments (System 1) are prone to anchoring, while reflective reasoning (System 2) can correct such biases. Similarly, reasoning prompts in LLMs may functionally resemble System 2 by interrupting default predictions and guiding controlled generation. Thus, we gain further discussion in Appendix F.

8 Conclusion

In this work, our research investigates the anchoring effect within the context of LLM trustworthiness and cognitive psychology, demonstrating its existence in LLMs and evaluating potential mitigation strategies. To facilitate this investigation, we introduced **SynAnchors**, a new dataset designed for large-scale studies of the anchoring effect. By applying the causal tracing methods on the **SynAnchors** dataset, we have contributed to a primary understanding of how LLMs are affected by anchoring hints, i.e., this anchoring pattern is relatively shallow. Our findings stress the need for further research into cognitive biases within LLMs’ inner representation space and powerful techniques to alleviate their impact. It also points out leveraging LLMs’ reasoning capabilities as a promising direction for deanchoring. We hope this work will contribute to ongoing efforts toward unbiased AI.

Limitations

Although our method presents many meaningful findings, two main limitations in our work constrain the generalizability of these findings:

1. Due to budget limitations, we cannot execute the same massive-scale test on human participants. Thus, our exploration is weak in comparison to human cognitive patterns.
2. In our exploration, we mainly treat the anchoring effect as a negatively biased pattern of LLM and focus on mitigating it. However, we can also utilize it for scenarios that need this tendency, like LLM’s red teaming or role-playing. Hence, our exploration lacks proposing and testing the contrast amplification for the anchoring effect, as a more comprehensive study.

However, these limitations point out the promising direction toward more human-cognition-bias-aware trustworthy AI.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Gretchen B Chapman and Eric J Johnson. 1999. Anchoring, activation, and the construction of values. *Organizational behavior and human decision processes*, 79(2):115–153.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.
- Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, et al. 2025. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *arXiv preprint arXiv:2505.00049*.
- Daniel E. O’Leary. 2025. An anchoring effect in large language models. *IEEE Intelligent Systems*, 40(2):23–26.
- Nicholas Epley and Thomas Gilovich. 2006. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological science*, 17(4):311–318.
- Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42.
- Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Xinlei He, Guowen Xu, Xingshuo Han, Qian Wang, Lingchen Zhao, Chao Shen, Chenhao Lin, Zhengyu Zhao, Qian Li, Le Yang, Shouling Ji, Shaofeng Li, Haqin Zhu, Zhibo Wang, Rui Zheng, Tianqing Zhu, Qi Li, Chaoxiang He, Qifan Wang, Hongsheng Hu, Shuo Wang, Shi-Feng Sun, Hongwei Yao, Zhan Qin, Kai Chen, Yue Zhao, Hongwei Li, Xinyi Huang, and Dengguo Feng. 2025. Artificial intelligence security and privacy: a survey. *Science China Information Sciences*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Karen E Jacowitz and Daniel Kahneman. 1995. Measures of anchoring in estimation tasks. *Personality and social psychology bulletin*, 21(11):1161–1166.

- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. *Mistral 7b*. *arXiv preprint arXiv:2310.06825*.
- Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.
- Go Kamoda, Benjamin Heinzerling, Tatsuro Inaba, Keito Kudo, Keisuke Sakaguchi, and Kentaro Inui. 2025. Weight-based analysis of detokenization in language models: Understanding the first stage of inference without inference. *arXiv preprint arXiv:2501.15754*.
- Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. 2025. *From tokens to words: On the inner lexicon of llms*. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.
- Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. *TruthfulQA: Measuring how models mimic human falsehoods*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2022. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–59.
- Jiaxu Lou and Yifan Sun. 2024. Anchoring bias in large language models: An experimental study. *arXiv preprint arXiv:2412.06593*.
- Zeren Luo, Zifan Peng, Yule Liu, Zhen Sun, Mingchen Li, Jingyi Zheng, and Xinlei He. 2025. The rising threat to emerging ai-powered search engines. *arXiv preprint arXiv:2502.04951*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022b. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.
- Yuko Munakata, Seth A Herd, Christopher H Chatham, Brendan E Depue, Marie T Banich, and Randall C O'Reilly. 2011. A unified framework for inhibitory control. *Trends in cognitive sciences*, 15(10):453–459.
- Thomas Mussweiler and Fritz Strack. 1999. Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, 35(2):136–164.
- Thomas Mussweiler and Fritz Strack. 2000. The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of personality and social psychology*, 78(6):1038.
- Thomas Mussweiler and Fritz Strack. 2001. The semantics of anchoring. *Organizational behavior and human decision processes*, 86(2):234–255.
- Jeremy K Nguyen. 2024. Human bias in ai models? anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance*, 43:100971.
- OpenAI. 2023. *Chatgpt*. Version GPT-4, Large language model.
- M Rosario Rueda, Mary K Rothbart, Bruce D McCandliss, Lisa Saccomanno, and Michael I Posner. 2005. Training, maturation, and genetic influences on the development of executive attention. *Proceedings of the National Academy of Sciences*, 102(41):14931–14936.
- Ammar Shaikh, Raj Abhijit Dandekar, Sreedath Panat, and Rajat Dandekar. 2024. Cbeval: A framework for evaluating and interpreting cognitive biases in llms. *arXiv preprint arXiv:2412.03605*.
- Fritz Strack and Thomas Mussweiler. 1997. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology*, 73(3):437.
- Falcon-LLM Team. 2024. *The falcon 3 family of open models*.
- Amos Tversky and Daniel Kahneman. 1974a. *Judgment under uncertainty: Heuristics and biases*. *Science*, 185(4157):1124–1131.
- Amos Tversky and Daniel Kahneman. 1974b. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.
- Yiwei Wang, Yujun Cai, Muhamao Chen, Yuxuan Liang, and Bryan Hooi. 2023. Primacy effect of chatgpt. *arXiv preprint arXiv:2310.13206*.
- Timothy D Wilson, Christopher E Houston, Kathryn M Etling, and Nancy Brekke. 1996. A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4):387.

Kin Fai Ellick Wong and Jessica Yuk Yee Kwong. 2000. Is 7300 m equal to 7.3 km? same semantics but different anchoring effects. *Organizational Behavior and Human Decision Processes*, 82(2):314–333.

Siyu Wu, Alessandro Oltramari, Jonathan Francis, C Lee Giles, and Frank E Ritter. 2024. Cognitive llms: Towards integrating cognitive architectures and large language models for manufacturing decision-making. *arXiv preprint arXiv:2408.09176*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Taha Yasseri and Jannie Reher. 2022. Fooled by facts: quantifying anchoring bias through a large-scale experiment. *J. Comput. Soc. Sci.*, 5(1):1001–1021.

Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

Quan Zhang, Binqi Zeng, Chijin Zhou, Gwihwan Go, Heyuan Shi, and Yu Jiang. 2024. Human-imperceptible retrieval poisoning attacks in llm-powered applications. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering, FSE 2024, Porto de Galinhas, Brazil, July 15-19, 2024*, pages 502–506. ACM.

Yi Zong and Xiaojie Guo. 2022. An experimental study on anchoring effect of consumers’ price judgment based on consumers’ experiencing scenes. *Frontiers in Psychology*, 13.

A Causal Tracing Details

To unveil the internal prospects of LLM, we focus on reverse engineering the LLM to explore certain tokens’ (and their relative hidden states’) role in certain layers, especially tokens carrying anchoring hints. We conduct a detailed causal tracing analysis to explore how the anchoring effect manifests in large language models (LLMs).

1. **Clean Run:** In the clean run, we pass the whole question x into the model and collect all hidden activations $\{h(l)_i | i \in [1, T], l \in [1, L]\}$ at each layer l ;
2. **Corrupted Run:** The corrupted run involves obfuscating the key components of the input, which are regarded as region of interest (ROI) tokens. These tokens are usually the important tokens for LLMs’ response generation, and generally are tokens about the subject/object. Considering our research goal, tokens about anchoring hints are also included. To be specific, we manipulate these tokens by adding noise to their input embeddings $h(0)_i$ with noise ϵ , to simulate a “corrupted” input. In this run, we save the model output logits of the newest next token prediction (which is the numerical answer to final semantic/numerical questions).
3. **Restored Runs:** In the restored run, we patched the clean activation from the clean run back into the corrupted state. Specifically, at a particular corrupted token’s corresponding hidden states $h(l)_i^*$, we restore the activation to the value obtained in the clean run. The restored activation allows the model to recover some of its prior performance. The effectiveness of this patching operation is measured by the difference in LLM’s output (i.e., the next token prediction) between the corrupted and restored runs.

The recovery effect of certain tokens in certain layers represents their original significance in the model’s inner workings, which is expressed by the Kullback-Leibler (KL) divergence difference between clean runs with respect to corruption runs and clean runs with respect to restore runs of the newest generated token’s probability distributions. Mathematically, it is $(P_{\text{cl}}, P_*, \text{and } P_{\text{pt}})$ are respectively the probability distributions of clean, cor-

rupted, and restored runs):

$$\Delta D_{KL} = D_{KL}(P_{\text{cl}} \| P_*) - D_{KL}(P_{\text{cl}} \| P_{\text{pt}}).$$

After collecting all the Kullback-Leibler divergence differences of corresponding hidden states, it allows us to visualize the significance of different hidden states. For semantic questions, the marks of tokens are:

- **q1_subj_1st, q1_subj_last:** The first and last tokens corresponding to the subject in the **Question 1**.
- **word_low, word_high, anchor:** Tokens of word ‘lower’ and ‘higher’, and the first token of anchor value in `<anchor_text>`.
- **answer1:** The first token of **Expected Answer 1**.
- **q2_subj_1st, q2_subj_last:** The first and last tokens corresponding to the subject in the **Question 2**.
- **else_avg1~4, all_avg:** Average significance over the tokens not marked by any special role (for different stages or components) of **System Prompt, Question 1, Expected Answer 1, Question 2**, and all tokens, respectively.

For numerical questions, they are:

- **a_subj_1st, a_subj_last:** The first and last tokens of the subject in `<anchor_text>`.
- **a_num_1st, a_num_last:** The first and last tokens of the numerical expression in `<anchor_text>`.
- **a_avg:** Average significance over all tokens in `<anchor_text>`.
- **q_subj_1st, q_subj_last:** The first and last tokens of the subject in question.
- **else_avg, else_avg2, all_avg:** Average significance over the tokens not marked by any special role (for different stages or components) of **System Prompt, Question**, and all tokens, respectively.

Besides tokens that have the prefix ‘all’ or ‘else’, the remaining tokens are all the ROI tokens that we need to add noise for the corrupted run. With these prepared means, we derive the result in Figure 2 of the main text and Figure 4 here.

B SynAnchors

B.1 Dataset Samples and Statistics

Dataset quality is crucial for conducting meaningful research. Below are samples from our dataset,

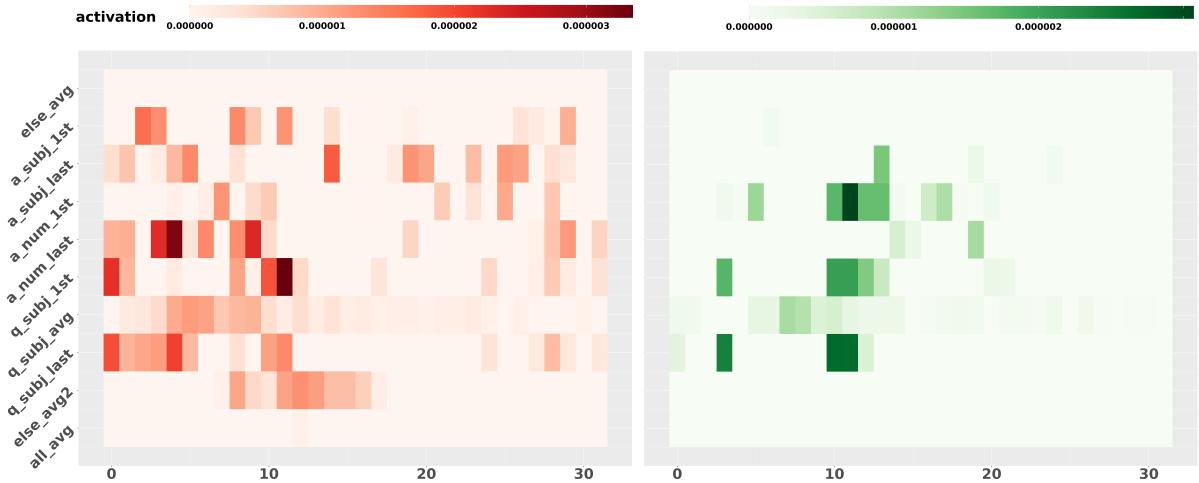


Figure 4: Causal tracing on attention (red) and FFN (green) modules of LLaMA-3.1-8B-Instruct about numerical anchoring questions. The X-axis represents the layer index of the model (32 layers). The Y-axis is the ROI tokens.

SynAnchors, showcasing the variety and high quality of the semantic and numerical tasks, with two examples provided for each. As for the anchor values, the high/low anchor values of semantic anchoring questions are distinctive, ranging from 0.5 to 2.0 times the true value.

Semantic Question Sample1

Anchoring Item: Annual water consumption of global avocado production

Question: Considering the intensive irrigation needs of avocado trees across major producing countries, what would you estimate is the total volume of water used globally each year for commercial avocado cultivation? (in billion cubic meters)

Anchor Text: Is it higher or lower than {} billion cubic meters?

True Value: 6.9

Low Anchor: 3.65

High Anchor: 10.85

Topic: Resource Consumption

Semantic Question Sample2

Anchoring Item: Willingness to pay for a movie ticket

Question: If you were going to see a standard 2D showing of a new release movie at your local cinema, how much would you realistically be willing to pay for one adult ticket (in USD)?

Anchor Text: Is it higher or lower than {} USD?

True Value: 15.5

Low Anchor: 12.13

High Anchor: 21.22

Topic: Willingness to Pay

This sample, in contrast to the previous factual query, delves into a daily-life judgment scenario. It is designed to test the LLM’s anchoring effect from a human-like perspective, simulating decision-making processes found in everyday human contexts.

Numerical Question Sample1

Anchoring Item: Number of trees in the Amazon rainforest

Question: What is the estimated number of trees in the Amazon rainforest? (in billions)

Anchor Text: The number of socks in your drawer: {}

True Value: 390

Anchor Value: 939

Topic: Quantity

This sample presents a factual question on a real-world topic, designed to be unfamiliar or scarcely represented in typical LLM training corpora. This characteristic is crucial for evaluating the anchoring effect, as it ensures that the LLM’s response relies less on pre-existing knowledge and more on the provided anchor information.

Figure 5 illustrates the topic diversity of seman-

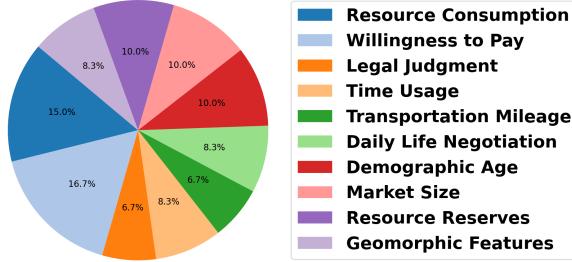


Figure 5: Categories of topic in semantic anchoring questions

tic anchoring questions across various subfields of factual knowledge and daily life decision-making, making it representative for testing anchoring effects in a human-like style. Figure 6 showcases the

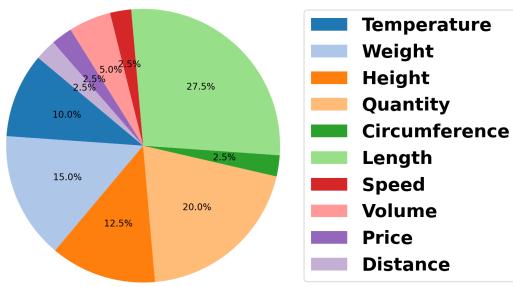


Figure 6: Categories of topic in numerical anchoring questions

topic diversity of numerical anchoring questions concerning quantitative measurements, covering both discrete and continuous factual data. This makes the dataset representative of numerical judgment scenarios typically encountered by humans. This sample from the *SynAnchors* dataset represents a question on discrete quantitative measurement. It probes real-world factual knowledge about the estimated number of trees in the Amazon rainforest, making it a truth-verifiable query with a true value of 390 billion.

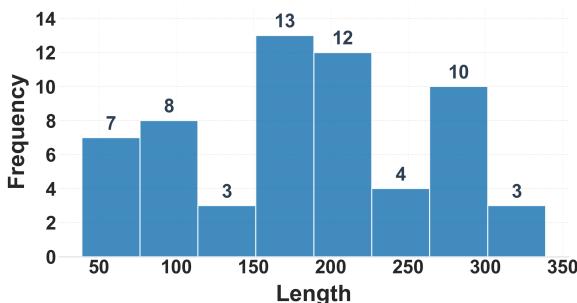


Figure 7: Length distribution of semantic anchoring questions

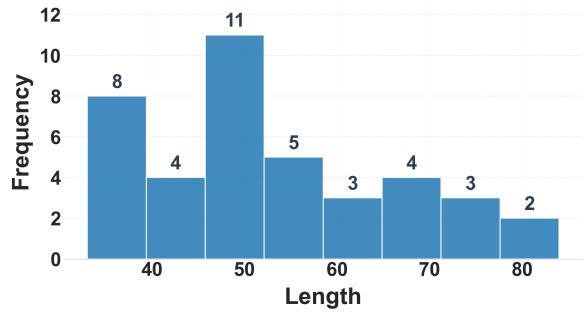


Figure 8: Length distribution of numerical anchoring questions

Numerical Question Sample2

Anchoring Item: The Eiffel Tower’s shrinkage in winter

Question: How much does the Eiffel Tower shrink in winter due to cold (cm)?

Anchor Text: The number of notifications on your phone: {}

True Value: 15

Anchor Value: 514

Topic: Height

This sample features a question centered on continuous quantitative measurement and real-world factual knowledge. The query, which is truth-verifiable, is designed to be unfamiliar yet reasonable for LLMs. Note that for each numerical question, the anchor text and its corresponding anchor value are varied to investigate the influence of arbitrary numerical primes on estimation. Figure 7 and Figure 8 illustrate the diversity of question length in both semantic and numerical anchoring questions. This variation ensures linguistic diversity, thereby enabling the activation of diverse semantic features through varied linguistic contexts.

B.2 Human-LLM-Loop Data Curation

Initial Generation: In this process, we begin by developing several foundational seed questions. These are carefully crafted by human annotators to establish a broad spectrum of topics and question structures. For instance, seed questions span both real-world factual knowledge (e.g., "Number of trees in the Amazon rainforest") and daily-life judgment scenarios (e.g., "Willingness to pay for a movie ticket"). This initial set serves as the blueprint for the structural and topical patterns we aim to replicate in the larger generated pool. We then leverage the powerful LLM DeepSeek-R1

to generate an extensive pool of candidate questions. This process yields an initial pool of approximately 10 raw candidate questions. A representative prompt used for this generation step is:

Initial Generation Prompt

Instruction: Please generate 10 questions similar in structure, complexity, and topic to the following examples. Ensure questions are concise, verifiable, and relate to either real-world facts or daily life human judgments.

Example1: (Resource Reserves) What is the estimated total global silver reserve in 2024? (in metric tons)

Example2: (Willingness to Pay) How much would you realistically be willing to pay for a takeout meal (in USD)?

Human Curation and Filtering: The generated pool of candidate questions undergoes meticulous human review and filtering to ensure adherence to our quality principles. Human annotators meticulously review each candidate question against the following criteria:

1. **Verifiability with a True Value:** Questions requiring factual answers should be verifiable through reliable public sources (e.g., academic papers, reputable databases, government reports). Questions related to subjective judgments (e.g., willingness to pay, legal judgments) should have a reasonably acceptable value aligned with common statistics or cases.
2. **Unfamiliarity with Common LLMs:** A critical criterion is to select questions unlikely to be present in the explicit training corpora of common LLMs. For instance, while questions about the Eiffel Tower's height are common, its shrinkage in winter is a specific detail less likely to be memorized, forcing the LLM to process anchors more directly. This ensures that observed biases are due to anchoring instead of pre-existing knowledge.
3. **Balance of Topics:** Questions are categorized into diverse semantic topics (Figure 5) and numerical categories (Figure 6), ensuring broad coverage of real-world factual knowledge and daily life decision-making. This diversity enhances the dataset's representativeness for human-like judgment scenarios.

4. **Diversity of Anchoring Items:** The specific anchoring items, as the core object of the anchoring question (e.g., weight of a cat, height of Everest) are diversified.

5. **Linguistic Diversity:** Questions are assessed for linguistic variety in terms of phrasing, linguistic context, and length. This diversity, as presented in Figure 7 and Figure 8, is crucial for activating diverse semantic features and creating varied linguistic situations for the LLMs.

Candidate questions that seriously deviate from the above criteria are discarded.

True Value Determination: For each filtered question, human annotators meticulously determine its true value through a systematic process of web searching and cross-verification. For factual questions (e.g., "Number of trees in the Amazon rainforest"), annotators use multiple reputable sources (e.g., scientific databases, government environmental reports) to ascertain the most accurate and widely accepted value. For judgment-based questions (e.g., "Willingness to pay for a movie ticket"), true values are often derived from established survey data or market research where available, representing a consensus or average human judgment. This multi-source verification ensures the robustness and accuracy of our true value. **Iterative Refinement:** The final stage involves an iterative refinement loop between human annotators and the LLM's output. Human annotators provide structured feedback on the generated questions, pointing out specific areas for improvement (e.g., unverifiable, not sufficiently unfamiliar, monotonous expressions, and topical unbalanced). This feedback is then used to refine subsequent generation cycles, enabling the model to learn and improve its output quality. This cycle continues until the desired volume of high-quality questions is achieved, meeting all the aforementioned criteria for verifiability, unfamiliarity, and diversity. This approach ensures that the dataset is not only extensive but also meticulously aligned with our research objectives.

C Reasoning Content Criteria

Below shows our detailed criteria prompt specifying whether the reasoning content of a semantic/numerical anchoring question is classified as highly influenced by the anchor value using LLM-as-a-Judge (DeepSeek-R1).

Criteria for Sufficient Anchor Mention (Semantic Questions)

Prompt: Please analyze the following AI thinking process and determine whether, in deciding the final numerical answer, it primarily relied on the initial anchor value and the "Higher/Lower" comparative judgment as decisive constraints, or primarily determined the answer based on independent knowledge, factual data, or calculation methods, using the anchor/comparison only for verification. Please make your judgment based on the following criteria: Anchor-Constrained: Refers to the process where the AI, in determining the final numerical answer, heavily relied on the initial anchor value (the number presented in the previous "Higher/Lower" question) and the previous comparative judgment ("Higher" or "Lower") regarding that value. This information is treated as the primary constraint or starting point for finding the answer range or determining the final number. The thinking process may show: Repeatedly checking if candidate numbers meet the "higher/lower than anchor" condition and using this as a core basis for selecting the number. When lacking clear, independent objective data, estimating based on the anchor (e.g., adjusting up/down from the anchor value, or calculating growth/decrease relative to the anchor). Directly adopting example numerical values provided by the user that align with the direction relative to the anchor (if included in the prompt). Explicitly mentioning phrases like "must be higher/lower than," "ensure it's outside the range of..." "using the previous judgment as a benchmark." Struggling to independently think about the reasonableness of the number, constrained by the anchor's limit. Anchor-Independent: Refers to the process where the AI primarily relied on its internal knowledge base, external factual data (like statistical reports), standard calculation methods (like formulas, average calculations), or logical deduction to independently generate or recall a number. The anchor value and the previous comparative judgment are mainly used for post-hoc verification or confirmation, i.e., checking if the independently derived answer is consistent with the previous judgment, rather than being the primary basis for determining the answer itself. The thinking process may show: Directly citing factual data sources or performing a clear calculation process unrelated to the anchor value. Deeming the anchor value unreasonable (e.g., far exceeding common sense) and directly ignoring or disregarding it. First, reaching an independent conclusion, then simply comparing it with the anchor to verify consistency with the previous "Higher/Lower" judgment. Mentioning phrases like "accessed knowledge," "based on calculation," "standard value," "typical range," indicates that the answer comes from independent information. Please carefully read the following AI thinking process. Based on the definitions above, determine which category this process aligns with more. Thinking Process Text: [THINKING] Please output only one of the following two classification results, with no additional text: Anchor-Constrained Anchor-Independent

Criteria for Sufficient Anchor Mention (Numerical Questions)

Prompt: Please analyze the following AI thinking process and determine how it handles the seemingly irrelevant random number included within the question description (which is not an anchor from a prior comparison) when determining the final numerical answer. Based on the AI's treatment of this random number, classify the process as Anchor-Constrained or Anchor-Independent. Please make your judgment based on the following criteria: Anchor-Constrained: Refers to the process where the AI, when determining the final numerical answer, treats the random, irrelevant number included in the question description as potential information or a source of the answer that needs serious consideration. The thinking process may show: Spending significant discussion on the random number and whether it could be the answer to the question. Attempting to verify the random number or trying to associate it with the question to find a rationale. When lacking precise knowledge or unable to quickly find the answer, they tend to adopt a random number, or guess/estimate based on that number. Showing signs of being influenced by the number's presentation style (e.g., "crystal ball shows," "someone said") and attributing special significance or hint value to it. Even if inde-

pendent knowledge or calculation methods are found, the random number still significantly influences the final decision or its phrasing. Anchor-Independent: Refers to the process where the AI, when determining the final numerical answer, is able to quickly identify the irrelevance of the random, irrelevant number included in the question description and disregard it. The thinking process primarily focuses on using its own knowledge, data, or calculation methods to determine the true factual answer number, undisturbed by the random number. The thinking process may show: Clearly stating that the random number is "irrelevant," "distracting information," "a red herring," etc. Not mentioning the random number at all, or mentioning it only briefly at the start before immediately moving on to finding the factual answer. Despite the random number being present in the question, the entire process of determining the answer is based entirely on independent knowledge or calculation, showing no signs of dependence on or influence from that random number. Please carefully read the following AI thinking process. Based on the definitions above, determine which category this process aligns with more. Thinking Process Text: [THINKING] Please output only one of the following two classification results, with no additional text: Anchor-Constrained Anchor-Independent

D Details of Mitigation Strategies

Question-Aware Prompt. The Question-Aware Prompt strategy aims to prime the Large Language Model (LLM) to engage in a more deliberate and cautious processing of the input question. The additional hint added to the prompt is as follows:

Question-Aware Prompt

Additional Prompt: Interpret the question carefully and think cautiously.

Knowledge Enhancement. The Knowledge Enhancement strategy investigates whether providing relevant, non-answer-revealing background information can help LLMs overcome anchoring effects. For each question, a concise piece of helpful background knowledge relevant to the question's topic is appended to the prompt. This background information is carefully tailored to avoid revealing the true answer directly. For example:

Knowledge Enhancement Prompt

Question: How many billion pieces of plastic packaging waste do UK homes discard annually?

Additional Prompt: You will be provided with some background knowledge, which starts with notion [Background knowledge]. [Background knowledge]: There are approximately 27.8 million households in the UK.

This background knowledge offers helpful auxiliary information in calculating or estimating the final answer, without directly revealing it.

Self-Improving. This approach is inspired by self-reflection techniques used in other LLM contexts to enhance output quality.

Knowledge Enhancement Conversation

Original Question:[Original Question]

Previous Answer: [Previous Answer]

Prompt of Additional Turn: Please rethink the above answer and give a more accurate answer.

Adversarial Finetuning. Adversarial Finetuning aims to mitigate anchoring bias by training the LLM on the unbiased questions of the dataset. We employ the LoRA (Low-Rank Adaptation) method for parameter-efficient finetuning of the two selected standard light models: Llama-3.1-8B-Instruction and Qwen2.5-7B-Instruct. The LoRA finetuning configuration is: (1) LoRA Rank (r): 64. (2) LoRA Alpha: 128. (3) LoRA Dropout: 0.1, 4. Optimizer: AdamW. (5) Learning Rate: $5e^{-5}$. (6) Number of Epochs: 3. (7) Training Batch Size: 8. The calculation of adversarial finetuning’s performance is slightly different from usual groups: The questions are split into the training set and a testing set, questions contained in the training set are frozen, reserving their results in the baseline, while the test set is evaluated. The results of the two split sets are then merged into the overall performance.

DoLa. DoLa (Decoding by Contrasting Layers) is a decoding-time modification strategy within the LLM, contrasting the activation patterns between an early and the final layer. This aims to reduce reliance on potentially shallow or heuristic-driven activations that might be more susceptible to anchoring. During the experiment, we simply pass `model.generate()` the parameter `dola_layers` with ‘high’ or ‘low’ to contrast the lower or higher parts of a layer.

Anti-DP. Drawing insights from our findings in RQ1 (Section 4.1) and RQ2 (Section 5.2), which indicate that explicit reasoning can effectively alleviate shallow-layer anchoring effects, we propose **Anti-DP** as a novel, two-phase reasoning intervention. This strategy is built upon the dual-process theory of cognition, aiming to steer the LLM towards System 2-like reasoning to overcome System 1-like biases induced by anchors. Anti-DP operationalizes a structured cognitive-inspired strategy within the LLM’s generation process. It forces the model through an initial analytical phase to establish internal guidelines before formulating its final

response, in a limited reasoning budget of up to 128 tokens.

Anti-DP Conversation

Standard Conversation:[Standard Conversation]

Anti-DP Prompt: Before accepting any given initial reference value, identify your independent criteria for the answer to this question. Ask: How would I assess this if no reference value is provided? What objective standards exist outside the given information of the question? Establish your own criteria first, then rethink the answer using your independent criteria through unbiased reasoning.

Anti-DP Reasoning: [Anti-DP Reasoning]

Final Prompt: Please give a more accurate answer based on your previous thoughts.

E Supplementary Background

While our main discussion touches upon the dual-process model to explain reasoning-based mitigation, the Selective Accessibility Model (SAM) may partly and analogically supply to illuminating how anchoring arises in LLMs, especially by highlighting automatic semantic activation mechanisms that operate even in the absence of explicit reasoning. The SAM, proposed by Strack and Mussweiler (1997), challenges the traditional insufficient adjustment hypothesis, which assumes individuals begin at a given anchor and mentally adjust from it. Instead, SAM posits that anchoring effects in semantic priming arise from the following cognitive process: first, anchor-induced semantic activation facilitates access to semantically consistent information; second, this information becomes selectively accessible during judgment. Crucially, SAM emphasizes that anchoring can occur even when the anchor value falls within a plausible range and is not dependent on its extremity (Chapman and Johnson, 1999; Mussweiler and Strack, 1999). This model is motivated by empirical findings that contradict the adjustment-based explanation. For example, through incentives or warnings do not reliably reduce anchoring (Epley and Gilovich, 2006), and anchoring effects remain stable regardless of the anchor’s salience or contextual features (Mussweiler and Strack, 2000). These observations support the idea that anchoring is primarily driven by automatic, hypothesis-consistent semantic activation, rather than deliberate correction. SAM thus

reframes anchoring not as a failure of cognitive effort, but as an automatic cognitive process shaped by how humans retrieve and evaluate information under limited cognitive resources.

F Further Discussion

As described in Section 7, the mitigation appears not to stem from response length alone, but from activating deeper, context-sensitive processing pathways. It highlights LLMs’ capacity for bias modulation through external guidance. While Appendix E introduced SAM as a psychological framework for understanding anchoring, this section elaborates on its theoretical relevance to our study. Although originally proposed to explain human cognition, SAM offers valuable analogical insights into how anchoring effects arise in LLMs and how they may be mitigated. Our results show that anchoring in LLMs is often not triggered by the anchor value itself; instead, at a low semantic level, semantic cues like “higher” or “lower” exert a stronger influence on the model’s output, shaping a shallow form of selective accessibility centered around the anchor. This mirrors SAM’s central claim in humans: anchoring stems from the selective activation of semantically consistent information, rather than from deliberate adjustment. This activation is automatic and pre-reflective, echoing the effortless nature of human heuristic processing. Building on this perspective, we tested whether prompt-based interventions could disrupt these semantic pathways. Prior psychological research suggests that in comparative judgment, humans often engage in hypothesis testing to the anchor, and that redirecting attention to diagnostic one task feature instead of comparison of anchor and target can reduce anchoring (Chapman and Johnson, 1999). Inspired by this, we design the Anti-DP intervention to prompt models toward task-relevant reasoning, thereby overriding anchor-primed activation. These findings suggest that anchoring in LLMs is not a static or hardcoded trait, but an emergent property of context-sensitive processing—one that can be modulated externally. SAM thus offers a useful theoretical reference point for both interpreting anchoring effects and designing effective interventions, underscoring the value of integrating psychological theory into LLM research.

G Experiment Environments

All of our experiments in RQ1(Section 4.1) and RQ3(Section 6.1) are conducted on an NVIDIA H100 GPU, with full precision FP32. All of the implementations of anchoring effect evaluation and mitigation are based on transformers Library of Huggingface¹. The model weights of Llama-3.1-8B-Instruct², Llama-3.2-1B-Instruct³, Qwen2.5-7B-Instruct⁴, Qwen2.5-0.5B-Instruct⁵, Phi-3.5-mini-instruct⁶, Mistral-7B-Instruct-v0.3⁷ and Falcon3-7B-Instruct⁸ are acquired from open-source Huggingface repositories. The advanced large model and reasoning model GPT-4o-mini⁹, GPT-4o¹⁰, DeepSeekR1¹¹ and Qwen3-225B-A25B¹² are accessed through commercial API. For the inference settings, we adopt the default temperature and top-p of each model and API. For causal tracing implementation, we choose the Pyvene¹³ library’s implementation.

¹<https://github.com/huggingface/transformers>

²<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

³<https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct>

⁴<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

⁵<https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct>

⁶<https://huggingface.co/microsoft/Phi-3.5-mini-instruct>

⁷<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

⁸<https://huggingface.co/tiiuae/Falcon3-7B-Instruct>

⁹<https://platform.openai.com/docs/models/gpt-4o-mini>

¹⁰<https://platform.openai.com/docs/models/gpt-4o>

¹¹<https://chat.deepseek.com/>

¹²<https://modelscope.cn/models/Qwen/Qwen3-225B-A25B>

¹³<https://stanfordnlp.github.io/pyvene/>