UNDERSTANDING THE ANCHORING EFFECT OF LLM WITH SYNTHETIC DATA: EXISTENCE, MECHANISM, AND POTENTIAL MITIGATIONS

Yiming Huang[1,*] **Biquan Bie**[2,*] **Zuqiu Na**[1] **Weilin Ruan**[1] **Songxin Lei**[1]
Yutao Yue[1,†] **Xinlei He**[1,†]
[1]The Hong Kong University of Science and Technology, Guangzhou
[2]Independent Researcher
[*]Equal Contribution [†]Corresponding Author
yhuang033@connect.hkust-gz.edu.cn BiquanBie@outlook.com
{yutaoyue,xinleihe}@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*** (https://huggingface.co/datasets/TimTargaryen/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 can not be eliminated by conventional strategies, while reasoning can offer some mitigation.

1 INTRODUCTION

Since the appearance of ChatGPT (OpenAI, 2023), Large Language Models (LLMs) have profoundly shifted human-computer interaction and its applications. 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. The anchoring effect (Tversky & Kahneman, 1974a), a prominent example of such biases, lacks comprehensive research in LLMs.

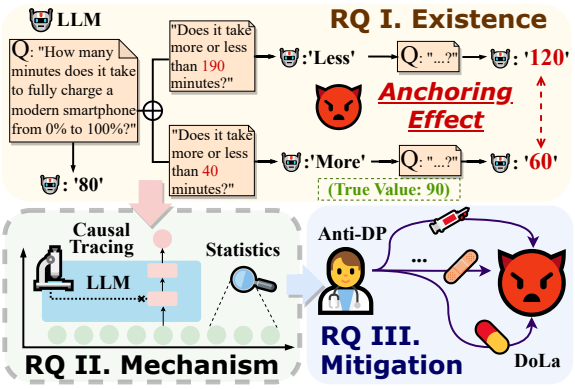

Figure 1: Detailed illustration of LLM's anchoring effect from three key aspects: (1) Existence: showing significant biases toward different anchor values for identical questions. (2) Mechanism: using causal tracing and statistics to explore underlying patterns. (3) Mitigation: evaluating varied mitigation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

The anchoring effect refers to the phenomenon that humans unconsciously attach too much importance to the first piece of information (anchors) in the decision-making process, causing subsequent judgments to be 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 performance metrics (Laskar et al., 2024), and common trustworthiness works (e.g., Adversarial Attack, Red Teaming) remain disconnected from human cognition (Wang et al., 2025). This disconnection needs to be bridged through quantified measurement of cognitive biases, where our refined metrics for anchoring effect match 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 on that, we mainly focus on three research questions (RQ) shown in Figure 1 of the anchoring effect:

RQ1: Are current modern LLMs easily affected by anchoring hints?

RQ2: What is the underlying mechanism of the anchoring effect in LLMs?

RQ3: What are the possible mitigations to this anchoring effect?

We study the two typical paradigms of anchoring effects: semantic priming (Jacowitz & 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 the question then asks again. The latter compares the answers with and without the interfering anchors. Different from the previous studies (Lou & Sun, 2026; 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% of questions anchored. While the reasoning models are relatively mild, they achieve the lowest ratio. In RQ2, we are the first 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., 2022b; Zhang & 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 inferred. It showcases the shallow salience of the anchoring effect within LLMs' early layers. In addition, the statistical proportion of the reasoning traces that mention the anchor information indicates that reasoning has a certain correlation to alleviate these effects. As for RQ3, all mitigation strategies fail to eradicate cognitive bias at its root, but reasoning shows the best alleviation. These results suggest a promising direction for debiasing LLMs: diluting the shallow anchoring influence by providing more balanced contextual signals that shift the model's focus away from the initial anchor.

Overall, the contribution of our paper can be summarized as follows: (1) We provide the dataset ***SynAnchors*** with relative evaluation metrics for 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 cure to break this shallow anchoring. (4) We test possible mitigation strategies, stressing the ineffectiveness of the conventional methods, and identifying promising directions for future anchoring-proof AI.

2 BACKGROUND

2.1 ANCHORING EFFECT & LLM

The anchoring effect is a ubiquitous cognitive bias (Furnham & Boo, 2011) and influences decisions in many fields (Jacowitz & 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 & 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 & Kahneman, 1995; Mussweiler & Strack, 2001) and the two-step procedure where both questions are semantically related (Wong & 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), Takenami et al. (2025) and Lou & Sun (2026) adopt semantic priming paradigms, though Nguyen (2024) and Takenami et al. (2025) focused only on financial prediction or price negotiation, making the scope too narrow and limiting robustness. While Lou & Sun (2026) demonstrates the prevalence of anchoring bias in LLMs using fact and expert-opinion hints, but lack the further exploration towards mechanism of anchoring effect. In short, these works focus mainly on surface-level evaluation and mitigation, without probing the mechanisms behind 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. To extend Tversky & Kahneman (1974b)'s experimental methodology, Jones & Steinhardt (2022) introduce a powerful framework for eliciting qualitative failure modes in large language models. However, while their work provides an excellent framework for broad failure analysis by transforming existing prompts to elicit task-level failures, our approach fundamentally differs. Our work contributes ***SynAnchors***, a new, standalone benchmark specifically structured to test the two established psychological paradigms of anchoring—Semantic Priming tasks (using a two-step conversational format) and Numerical Priming tasks (introducing a demonstrably irrelevant number). This targeted design allows for a more controlled and psychologically-grounded investigation of the anchoring effect itself. Furthermore, while Jones & Steinhardt (2022) primarily evaluate the effect by measuring the drop in functional accuracy on a task, our work complements this by adopting evaluation methods directly from cognitive science to quantify the bias itself, independent of task accuracy. Our methods include statistical significance testing (t-tests) and quantitative metrics like the Anchor Index (A-Index) and Relative Error (R-Error) to measure the magnitude of the bias, providing a quantitative basis for comparing model behavior to human behavior without presuming an identical underlying cognitive process.

As for a deeper theoretical level, the anchoring effect is often explained by the Dual Process (DP) Model (Kahneman, 2011): (1) **Automatic Adjustment (System 1):** Humans initially rely heavily on the first piece of information, producing a rapid and unconscious judgment. (2) **Effortful Correction (System 2):** Subsequently, humans may attempt to adjust their judgment away from the anchor through a more deliberate and effortful process. While our main discussion touches upon the dual-process model to explain reasoning-based mitigation, the Selective Accessibility Model (SAM, detailed in Section A) may offer a partial analogy to illuminate how anchoring arises in LLMs, especially by highlighting automatic semantic activation mechanisms that operate even in the absence of explicit reasoning.

2.2 Trustworthy Artificial Intelligence: a Psychology-Combined Perspective

Research shows LLMs 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 literature reviews and benchmark construction; (2) empirical comparisons of cognitive behaviors between LLMs and humans; and (3) conventional safety-oriented trustworthy AI. Firstly, a part of these works focuses on literature 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). The benchmark of Koo et al. (2023) 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 generalized, 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 locate their contribution to general LLM robustness issues rather than granular cognitive bias. This study confirms the anchoring effect in LLMs and goes beyond surface-level lexical correlations by tracing how models process cues of varying semantic (Meng et al., 2022a; Zhang & Nanda, 2023) across layers.

3 PSYCHOLOGICAL TEST FORMALIZATION

3.1 TARGETED ANCHORING EFFECT PARADIGMS

The first paradigm we aim to study is semantic priming anchors. As discussed 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, which are implicitly introduced as anchors into the decision-making process. The second step asks the testee the same question again, but requests its direct answer with the previous context. The collection of conversations is suitable for quantitative research, and we make our dataset in the following format:

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 with 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 anchoring 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 with 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 construct 60 questions over 10 diverse topics, covering various aspects of factual knowledge domains and subjective real-world decision scenarios. For numerical anchoring questions, we construct 40 questions across 10 topics, primarily focusing on quantifiable factual measurements. The questions are 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 involves 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, diversity of anchoring items, and linguistic diversity. This deliberate mix of factual and subjective questions enables a comprehensive evaluation across different task types, with tests conducted to ensure that the models possessed no pre-existing knowledge of the selected questions.
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 Section B.

Table 1: Evaluation of semantic and numerical performance. The 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 $\geq 10\%$. '$^{\#}$' indicates results are based on 30 samples per question due to budget limits. A deeper red background color of the row means a stronger anchoring effect.

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 (*Thinking Mode*)	0.321$^{\#}$	33.3%$^{\#}$	0.080	5.0%	22.0%
DeepSeek-R1 (*DeepThink Mode*)	0.278	31.7%	0.112	15.0%	25.0%

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 & Kahneman, 1995) (**A-Index**) for evaluating intensity of anchoring effect:

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

Here, Median_{high} and Median_{low} are the median values in each group, Anchor_{high} and Anchor_{low} are the specific high anchor and low anchor values 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 & Kahneman, 1995; Zong & Guo, 2022; Yasseri & Reher, 2022). Therefore, comprehensively considering the presence and intensity of the anchoring effect, we count the question yields $\mathrm{p} < 0.05$ and **A-Index** > 0.4 as an obvious occurrence of the anchoring effect.

As for numerical ones, we query the testee LLM 100 times for each question with and without the anchoring hints `<anchor_text>`, which contain 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:

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

Here, v_{anchor} is the answer `<value>` with anchoring hints and v_{orig} is the value without them. We regard $\mathrm{p} < 0.05$ 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 & 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 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 errors.

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. 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 humans, most LLMs yield 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).

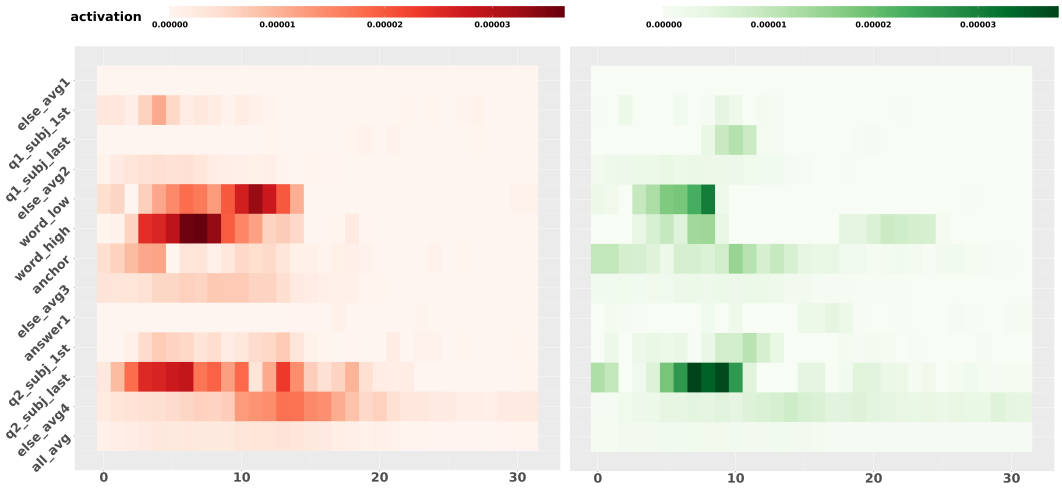

Figure 2: Causal tracing on attention (red) and FFN (green) modules of LLama-3.1-8B-Instruct about semantic anchoring questions. The X-axis represents the layer index of the model (32 layers). The Y-axis is the ROI tokens.

5 Mechanistically Exploring LLMs' Anchoring Effect (RQ2)

5.1 Causal Tracing Analysis

Methodology. Following previous work (Meng et al., 2022a; Zhang & 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 Section C.

Setup. We execute 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, where 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 restoring results are recorded for comparison as well. Concrete correspondence of the target tokens' marks is listed below. For semantic questions, the notions are (numerical questions are detailed in Section D):

- **`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.

Results. The results of the causal tracing analysis are shown in Figure 2 and Figure 8 (in Section D), 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.

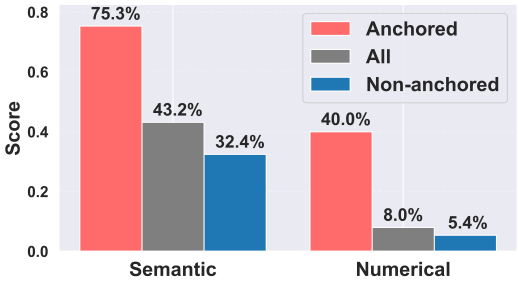

Figure 3: Percentages of sufficient anchor information mentions in DeepSeek-R1 reasoning contents. We employ an LLM-as-a-Judge approach to automatically detect explicit mentions of anchor-influenced features in the reasoning trace. "Anchored" refers to the percentages of questions judged as anchored 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.

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, and 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 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.

5.2 Statistics of Reasoning Trace

In Section 4.2, the reasoning models exhibit 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 reasoning can alleviate this shallow anchoring effect, we perform further analysis.

Figure 3 shows statistics on the proportion 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.

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 may involve additional text, conversation turns, or both.

We compare conventional strategies (detailed in Section E) 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). The considered strategies are:

- **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 with a piece of helpful background knowledge without a direct answer.
- **Self-Improving:** Give the LLM an additional turn to refine its answer with given prompt.
- **Adversarial Finetuning:** Finetune the LLM with its unbiased conversations using the LoRA (Hu et al.).

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). '◇' indicates results are derived on test splits, which exclude train splits of LoRA.

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 (*low*)	0.369 ↓	36.7% ↓	0.308*	38.9%*	37.5%
+ DoLa Decoding (*high*)	0.377 ↓	38.3%	0.308*	38.9%*	38.5%
+ Anti-DP	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 (*low*)	0.420 ↓	43.3%	0.283	35.0%	40.0%
+ DoLa Decoding (*high*)	0.393 ↓	40.0% ↓	0.283	35.0%	38.0% ↓
+ Anti-DP	0.344* ↓	34.5%* ↓	0.315	50.0%	40.8%

- **DoLa Decoding:** (Chuang et al.) Modify the LLM decoding strategy by contrasting the prediction between an early (high or low) layer and the final layer.
- **Anti-DP (Anti-Dual-Process):** Implement a two-phase reasoning intervention. Upon receiving the question, the LLM is first instructed to establish its standard to guide the rethinking. Then, the LLM produces the final answer based on prior thoughts. This approach clearly opposes the automatic dual-process by encouraging integrated, iterative reasoning.

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, **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 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. Our study (as discussed in Section F) also profoundly points out leveraging LLMs' reasoning capabilities as a promising direction for deanchoring. Through quantifying these biases, we reveal how such cognitive vulnerabilities can undermine LLM reliability in decision-making tasks, where contextual hints may lead to deviations from the objective truth. Future work will focus on scaling the dataset to a larger collection across comprehensive dimensions to ensure broader generalizability. We hope this work will make contributions toward unbiased 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. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 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. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *arXiv preprint arXiv:2505.00049*, 2025.

Daniel E. O'Leary. An anchoring effect in large language models. *IEEE Intelligent Systems*, 40(2): 23–26, 2025. doi: 10.1109/MIS.2025.3544939.

Nicholas Epley and Thomas Gilovich. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological science*, 17(4):311–318, 2006.

Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42, 2011.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xinlei He, Guowen Xu, Xingshuo Han, Qian Wang, Lingchen Zhao, Chao Shen, Chenhao Lin, Zhengyu Zhao, Qian Li, Le Yang, Shouling Ji, Shaofeng Li, Haojin 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. Artificial intelligence security and privacy: a survey. *Science China Information Sciences*, 2025.

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. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://doi.org/10.48550/arXiv.2310.06825.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799, 2022.

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

Go Kamoda, Benjamin Heinzerling, Tatsuro Inaba, Keito Kudo, Keisuke Sakaguchi, and Kentaro Inui. Weight-based analysis of detokenization in language models: Understanding the first stage of inference without inference. *arXiv preprint arXiv:2501.15754*, 2025.

Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. 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, 2025. URL https://openreview.net/forum?id=328vch6tRs.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*, 2023.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, et al. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13785–13816, 2024.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–59, 2022.

Jiaxu Lou and Yifan Sun. Anchoring bias in large language models: An experimental study. *Journal of Computational Social Science*, 9(1):11, 2026.

Zeren Luo, Zifan Peng, Yule Liu, Zhen Sun, Mingchen Li, Jingyi Zheng, and Xinlei He. The rising threat to emerging ai-powered search engines. *arXiv preprint arXiv:2502.04951*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022a.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022b.

Yuko Munakata, Seth A Herd, Christopher H Chatham, Brendan E Depue, Marie T Banich, and Randall C O'Reilly. A unified framework for inhibitory control. *Trends in cognitive sciences*, 15 (10):453–459, 2011.

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.

Thomas Mussweiler and Fritz Strack. The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of personality and social psychology*, 78(6):1038, 2000.

Thomas Mussweiler and Fritz Strack. The semantics of anchoring. *Organizational behavior and human decision processes*, 86(2):234–255, 2001.

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.

OpenAI. Chatgpt, 2023. URL https://openai.com/chatgpt. Version GPT-4, Large language model.

M Rosario Rueda, Mary K Rothbart, Bruce D McCandliss, Lisa Saccomanno, and Michael I Posner. Training, maturation, and genetic influences on the development of executive attention. *Proceedings of the National Academy of Sciences*, 102(41):14931–14936, 2005.

Ammar Shaikh, Raj Abhijit Dandekar, Sreedath Panat, and Rajat Dandekar. Cbeval: A framework for evaluating and interpreting cognitive biases in llms. *arXiv preprint arXiv:2412.03605*, 2024.

Fritz Strack and Thomas Mussweiler. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology*, 73(3):437, 1997.

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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4481–4498, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.240. URL https://aclanthology.org/2025.findings-emnlp.240/.

Falcon-LLM Team. The falcon 3 family of open models, December 2024. URL https://huggingface.co/blog/falcon3.

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

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974b.

Jun Wang, Ninglun Gu, Kailai Zhang, Zijiao Zhang, Yelun Bao, Jin Yang, Xu Yin, Liwei Liu, Yihuan Liu, Pengyong Li, et al. Beyond benchmark: Llms evaluation with an anthropomorphic and value-oriented roadmap. *arXiv preprint arXiv:2508.18646*, 2025.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. Primacy effect of chatgpt. *arXiv preprint arXiv:2310.13206*, 2023.

Timothy D Wilson, Christopher E Houston, Kathryn M Etling, and Nancy Brekke. A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4):387, 1996.

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

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

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

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

Taha Yasseri and Jannie Reher. Fooled by facts: quantifying anchoring bias through a large-scale experiment. *J. Comput. Soc. Sci.*, 5(1):1001–1021, 2022. doi: 10.1007/S42001-021-00158-0. URL https://doi.org/10.1007/s42001-021-00158-0.

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

Quan Zhang, Binqi Zeng, Chijin Zhou, Gwihwan Go, Heyuan Shi, and Yu Jiang. Human-imperceptible retrieval poisoning attacks in llm-powered applications. In Marcelo d'Amorim (ed.), *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering, FSE 2024, Porto de Galinhas, Brazil, July 15-19, 2024*, pp. 502–506. ACM, 2024. doi: 10.1145/3663529.3663786. URL https://doi.org/10.1145/3663529.3663786.

Yi Zong and Xiaojie Guo. An experimental study on anchoring effect of consumers' price judgment based on consumers' experiencing scenes. *Frontiers in Psychology*, 13, 2022. URL https://api.semanticscholar.org/CorpusID:246635839.

A Supplementary Background

The SAM, proposed by Strack & 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 & Johnson, 1999; Mussweiler & Strack, 1999).

This model is motivated by empirical findings that contradict the adjustment-based explanation. Through incentives or warnings, anchoring is not reliably reduced (Epley & Gilovich, 2006), and anchoring effects remain stable regardless of the anchor's salience or contextual features (Mussweiler & 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 with limited cognitive resources.

B *SynAnchors*

B.1 Dataset Samples and Statistics

Dataset quality is crucial for conducting meaningful research. Below are samples from our dataset, ***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 Sample 1

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 each year globally 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

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.

Semantic Question Sample 2

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.

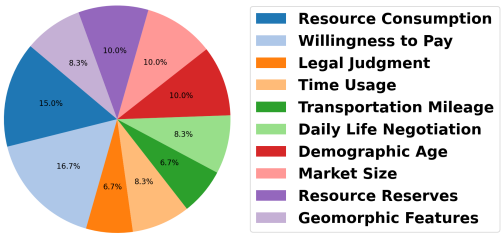

Figure 4: Categories of topic in semantic questions.

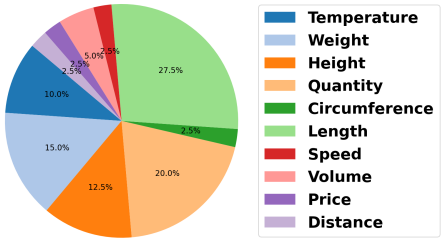

Figure 5: Categories of topic in numerical questions.

Numerical Question Sample 1

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

Figure 4 illustrates the topic diversity of semantic 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 5 showcases the 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 a 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.

Numerical Question Sample 2

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 knowledge. The query is truth-verifiable and 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 6 and Figure 7 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

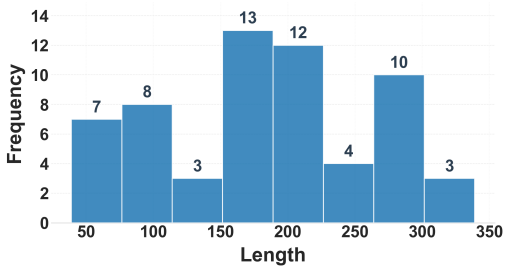

Figure 6: Length distribution of semantic questions.

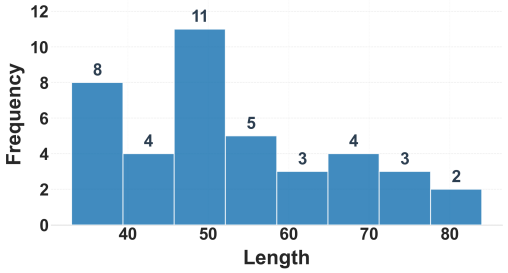

Figure 7: Length distribution of numerical questions.

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.*
Example 1: (Resource Reserves) What is the estimated total global silver reserve in 2024? (in metric tons)
Example 2: (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. Annotators meticulously review each candidate question with 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 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 4) and numerical categories (Figure 5), ensuring coverage of real-world 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 6 and Figure 7, 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 on 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., unverifiability, insufficient unfamiliarity, monotonous expressions, and topical imbalance). 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 Causal Tracing Details

To be specific, 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 the final questions).
3. **Restored Run**: 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. In our case, the clean run provides a reference, the corrupted run introduces the anchoring manipulation of all tokens we are interested in, and the restoration run helps us gauge how much the model relies on specific tokens (such as those corresponding to the anchor values) to generate the correct answer. By comparing the model's performance across these different runs, we can quantify the total effect (TE) and indirect effect (IE) of the anchoring hint on the model's answer. Mathematically, it is (P_{cl}, P_*, and P_{pt} are respectively the distributions of clean, corrupted, and restored runs):

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

After collecting all the Kullback-Leibler divergence differences of corresponding hidden states, it allows us to visualize the significance of different hidden states.

D Causal Tracing Results on Numerical Questions

As for numerical questions, target tokens 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.

With the prepared means in the main text and above, we derive the result in Figure 8 here. As

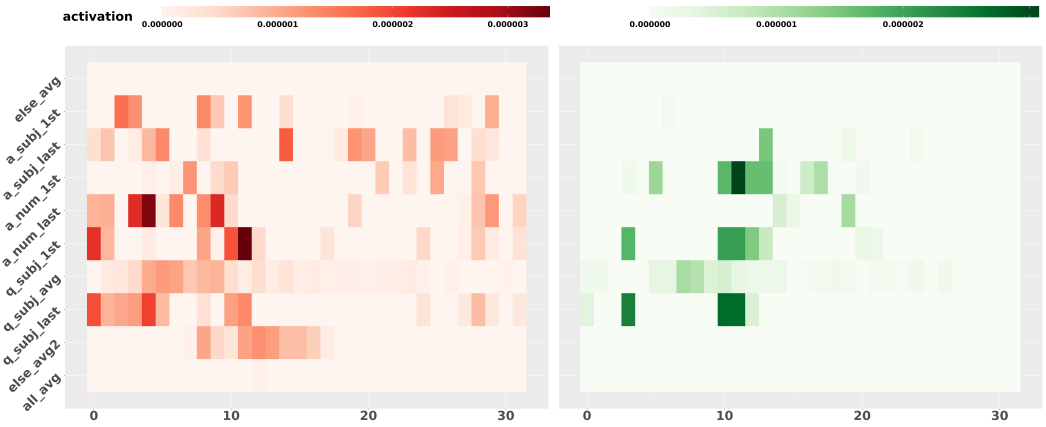

Figure 8: Causal tracing on attention (red) and FFN (green) modules of LLama-3.1-8B-Instruct about numerical anchoring questions. The X-axis represents the layer index of the model (32 layers). The Y-axis is the ROI tokens.

shown in Figure 8, 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.

E Details of Mitigation Strategies

Their concrete implementations are demonstrated below.

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.

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

Self-Improving 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-Instruct 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 layer and the final layer. This aims to reduce reliance on potentially shallow or heuristic-driven activations that might be more susceptible to anchoring.

Anti-DP From a cognitive-aware mitigation perspective, we implement **Anti-DP** as a two-phase reasoning intervention. This strategy is built upon the dual-process (DP) 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.

F Discussion

Our results show that anchoring effects are prominent in LLMs (RQ1), that enhancing reasoning may offer a path to reduce such shallow biases (RQ2), and that 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 behavior. 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.

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 feature of the task instead of comparison of anchor and target can reduce anchoring (Chapman & Johnson, 1999). The Anti-DP intervention prompts models toward task-relevant reasoning, thereby overriding anchor-primed activation to some extent.

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.