# Modeling and Predicting Multi-Turn Answer Instability in Large Language Models

**Jiahang He*** Algoverse AI Research

**Rishi Ramachandran*** Algoverse AI Research

**Neel Ramachandran** Algoverse AI Research

**Aryan Katakam** Algoverse AI Research

**Kevin Zhu** Algoverse AI Research

**Sunishchal Dev** Algoverse AI Research

**Ashwinee Panda** Algoverse AI Research

**Aryan Shrivastava** University of Chicago

## Abstract

As large language models (LLMs) are adopted in an increasingly wide range of applications, user–model interactions have grown in both frequency and scale. Consequently, research has focused on evaluating the robustness of LLMs, an essential quality for real-world tasks. In this paper, we employ simple multi-turn follow-up prompts to evaluate models' answer changes, model accuracy dynamics across turns with Markov chains, and examine whether linear probes can predict these changes. Our results show significant vulnerabilities in LLM robustness: a simple "Think again" prompt led to an approximate 10% accuracy drop for Gemini 1.5 Flash over nine turns, while combining this prompt with a semantically equivalent reworded question caused a 7.5% drop for Claude 3.5 Haiku. Additionally, we find that model accuracy across turns can be effectively modeled using Markov chains, enabling the prediction of accuracy probabilities over time. This allows for estimation of the model's stationary (long-run) accuracy, which we find to be on average approximately 8% lower than its first-turn accuracy for Gemini 1.5 Flash. Our results from a model's hidden states also reveal evidence that linear probes can help predict future answer changes. Together, these results establish stationary accuracy as a principled robustness metric for interactive settings and expose the fragility of models under repeated questioning. Addressing this instability will be essential for deploying LLMs in high-stakes and interactive settings where consistent reasoning is as important as initial accuracy.

## 1 Introduction

The use of large language models (LLMs) in interactive applications has greatly expanded in recent years (Kumar, 2024). As a result, research has increasingly focused on evaluating model robustness, a quality essential for real-world tasks such as decision making and classification (Li et al., 2025a). Prior research has shown that a simple "rethink" prompt could reduce model performance on question-answering tasks (Pawitan & Holmes, 2024). Further research on single-turn accuracy has also found that models are highly sensitive to even small variations in prompts (Salinas & Morstatter, 2024).

This work investigates the following research question: Given repeated prompts without new evidence, how does a model's accuracy evolve? Addressing this question provides insight into LLM stability and

*denotes equal contribution. Correspondence to aashrivastava@uchicago.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

the prevalence of sycophantic behavior while also enabling the prediction of accuracy dynamics for more reliable and interpretable human-AI interactions. This contribution is key to real-world settings where users repeatedly query AI systems—such as education, coding, or research assistants—without introducing new information. To explore this, we used simple multi-turn follow-up prompts to evaluate models' answer changes, model accuracy dynamics across turns with Markov chains, and examine whether linear probes can predict these changes. Our research reveals significant vulnerabilities in LLM robustness: models frequently revise originally correct answers when re-questioned or slightly challenged, even without being presented new evidence. Additionally, we find that a model's accuracy across multiple turns—when subjected to both simple and adversarial prompts—can be successfully modeled using Markov chains. Upon examining the model's hidden states, we also find evidence that future answer changes can be predicted using linear probes. Overall, we quantitatively characterize LLMs' multi-turn answer stability and reveal internal state patterns linked to robustness. We hope our results can guide future research and model design to enhance reliability in practical, interactive settings.

In summary, the main contributions of this paper are as follows:

- We provide insights into how simple follow-up prompts and semantically rephrased prompts influence a model's likelihood of changing its answer, evaluating performance across datasets with diverse question types and difficulty levels.
- We demonstrate that model accuracy across multiple turns can be effectively modeled using a Markov process, which often converges to a stationary accuracy that is below the initial first-turn performance, as illustrated in Figure 1.
- We find that probing the models' hidden states yields a notable layer-wise improvement in predicting whether the model will change its answer or not, providing evidence that probes are predictive of forthcoming answer shifts.

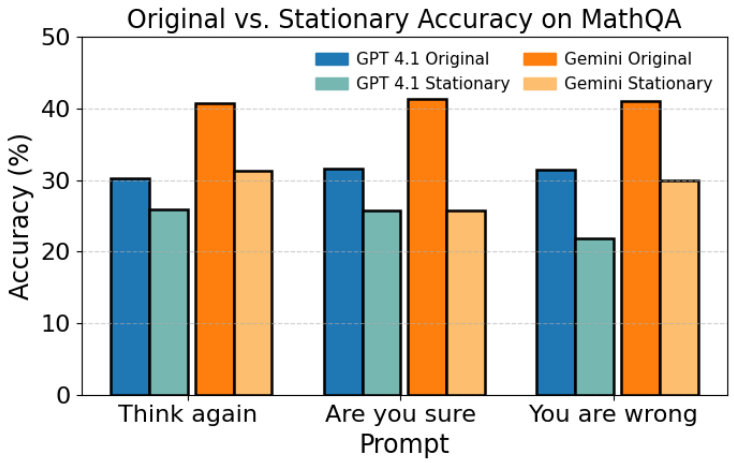


Figure 1: MathQA first-turn vs. stationary accuracies for GPT-4.1-nano and Gemini 1.5 Flash. GPT shows declines of $4.4$–$9.7\%$ across prompts (e.g., "You are wrong" drops from $31.5\% \rightarrow 21.8\%$), while Gemini declines $9.4$–$15.6\%$ (largest under "Are you sure," $41.4\% \rightarrow 25.8\%$). On average, GPT degrades by about $6.6\%$ from first-turn to stationary accuracy, while Gemini degrades by $12\%$.

## 2 Related works

**Robustness in LLMs** The robustness of LLMs has become an important field to study as these systems are increasingly being deployed in critical and complex applications (Ma et al., 2025; Wu et al., 2024). A popular line of investigation involves examining how LLMs respond to prompts that subtly alter the semantics of the original question (Salinas & Morstatter, 2024; Seleznyov et al., 2025). These prompts range from adding an extra space to purposely misspelling a word. Prior research has shown that even the smallest of changes can lead to a significant decrease in model performance across tasks (Zhu et al., 2023). Our work advances this line of research by extending

prompt variations into a multi-turn setting and modeling the resulting interactions with a Markov chain transition framework.

**Multi-turn conversations with LLMs** Recent research on multi-turn interactions with LLMs has highlighted challenges in maintaining accuracy and confidence over multiple reasoning steps (Zhang et al., 2025; Li et al., 2025b; Sirdeshmukh et al., 2025; Laban et al., 2025). One line of work investigates the models' confidence by measuring whether they adhere to initial answers when given adversarial follow-up prompts (Xie et al., 2024). These studies show that models often fail to maintain their original answers, leading to degraded performance. Our paper builds on these results by testing models with not only adversarial prompts, but also simple prompts and rephrased questions that preserve the original semantic meaning.

**Sycophancy** Research on the willingness of large language models (LLMs) to conform to user beliefs—known as sycophancy—has shown that state-of-the-art models frequently exhibit untruthful behavior across a range of tasks (Sharma et al., 2025; Malmqvist, 2024; Liu et al., 2025). Frameworks such as SycEval assess sycophancy by presenting LLMs with user rebuttals following their initial responses (Fanous et al., 2025). Their results reveal that sycophantic behavior persists across multi-turn interactions, with 58.19% of all samples showing signs of answer changes in response to user pressure. Building on prior work, we probe the model's internal hidden states to assess whether such answer changes can be predicted.

## 3 Experimental setup

### 3.1 Datasets

We select four datasets for our experiments, covering a range of difficulty levels and domains.

- **MMLU**: A dataset with approximately 16,000 questions spanning 57 subjects (Hendrycks et al., 2020). From it, we sample 3,000 questions to evaluate the robustness of LLMs across a broad range of domains.
- **MathQA**: A large-scale dataset of math word problems extending AQuA (Amini et al., 2019; Ling et al., 2017). We use 2,985 of its MCQ questions to evaluate how robust LLMs are in the specific field of mathematical reasoning and quantitative problem solving.
- **Humanity's Last Exam**: A dataset of 2,500 challenging questions across 100+ subjects, with state-of-the-art performance at only 25% (Phan et al., 2025). We use the dataset's multiple-choice questions to evaluate how models perform in a multi-turn interaction when initial accuracy is low.
- **GlobalOpinionsQA**: A subjective dataset built with the goal of developing AI to be more inclusive and serve all people worldwide (Durmus et al., 2023). The dataset is composed of 2,556 multiple-choice questions, and we utilize it to evaluate a model's tendency to change its answer on subjective questions.

### 3.2 Multi-turn prompting protocol

For our initial experiments, we begin by prompting the model with a question from the dataset. After the original question, one of three simple follow-up prompts—"Think again," "Are you sure?" or "You are wrong"—is applied repeatedly across nine subsequent turns. These prompts gradually increase the pressure on the model, with "You are wrong" being the most adversarial. In selecting these prompts, we prioritize simplicity to evaluate whether straightforward, uncomplicated prompts influence the model's answer. From now on, we will refer to these prompts as "TA," "RUS," and "URW" respectively. Between turns, no additional information about the initial problem is provided.

### 3.3 Models and hyperparameters

Our experiments testing model robustness through simple follow-up prompts were conducted primarily on Gemini 1.5 Flash (Reid et al., 2024) and GPT-4.1-nano (Achiam et al., 2023). However, we performed a smaller-scale study using Claude 3.5 Haiku (Anthropic, 2024) and GPT-4o (Achiam

et al., 2023) to validate that our findings generalize across models of different capabilities. The temperature of each model was set at 0 for deterministic answers.

### 3.4 Rephrased prompt variant

We developed a complementary experiment to evaluate whether models change their answer when a question is rephrased. Prompt rephrasings are more ecologically valid than our prior three simple prompts, and we aim to see whether such rephrasings also influence the model's answers. These experiments were only conducted on Claude 3.5 Haiku and GPT-4o using the MathQA and MMLU datasets due to budget constraints. To avoid confusion and redundancy from excessive rephrasings, we only use five subsequent prompts, each featuring a distinct question variant generated by GPT-4o (example in Appendix A). We repeat this experiment with all three follow-up prompts outlined below:

- `"Think again. Think about it this way: "` + *variation*
- `"Are you sure? Think about it this way: "` + *variation*
- `"You are wrong. Think about it this way: "` + *variation*

The goal of this procedure is to test whether LLMs would remain consistent in their answers across multiple semantically identical reworded prompts.

## 4 Models frequently change their minds

### 4.1 Results for simple follow-up prompts

Across GPT-4.1-nano and Gemini 1.5 Flash evaluated on the MathQA dataset, we observed a consistent decline in accuracy over the course of multi-turn prompting (see Figure 2). The RUS prompt caused the smallest accuracy degradation, approximately 5% for both models. In contrast, the adversarial URW prompt produced the largest drop, with accuracies decreasing by 12.4% for GPT and 11.9% for Gemini. As illustrated in Figure 8 and Figure 9, these trends were also observed in other models such as GPT-4o, and on subjective datasets, such as GlobalOpinionsQA (GOQA).[1] The accuracy decrease was smaller for GPT-4o, suggesting that it exhibits greater robustness.

Overall, Gemini 1.5 Flash demonstrated higher accuracy levels, but also a steeper accuracy decline. Another notable observation is the fluctuation in accuracy across most prompts. We hypothesize that this instability arises from the model's uncertainty on certain problems, causing it to oscillate between correct and incorrect answers over successive turns.

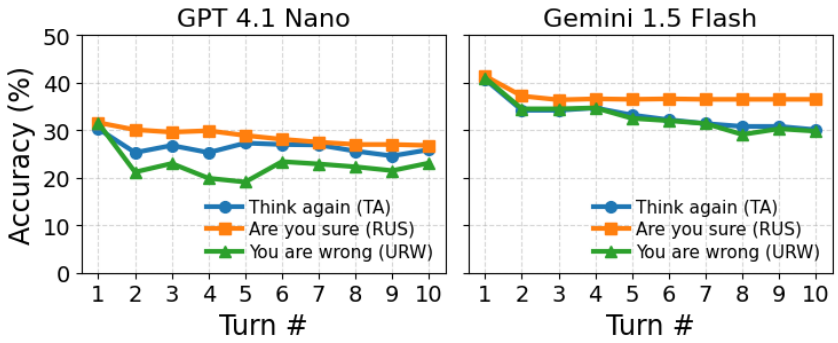


Figure 2: Accuracy drift across 10 turns for GPT-4.1-nano and Gemini 1.5 Flash on MathQA questions. For GPT-4.1-nano, the maximum accuracy decline (from the first turn to the lowest-performing turn) for each prompt was $30.3\% \rightarrow 24.6\%$ on TA, $31.6\% \rightarrow 26.8\%$ on RUS, and $31.5\% \rightarrow 19.1\%$ on URW. For Gemini 1.5 Flash, the maximum accuracy decline for each prompt was $40.7\% \rightarrow 30.1\%$ on TA, $41.4\% \rightarrow 36.4\%$ on RUS, and $41.0\% \rightarrow 29.1\%$ on URW.

[1] Since GlobalOpinionsQA is subjective, we set the model's initial response as the "correct" answer.

To address concerns that accuracy degradation may be due to other factors such as model fatigue, we conducted a control experiment using Gemini 1.5 Flash on 500 MathQA questions where each question was repeated nine times without a simple follow-up prompt. Accuracies deviated much less, by only 0.2% to 2.8% across turns, indicating that accuracy loss is primarily caused by prompt pressure (see Appendix E).

Furthermore, we applied our three simple follow-up prompts to the multiple choice questions of the Humanities Last Exam (HLE) dataset. The purpose of this was to analyze results on a dataset where the initial accuracy is low. As shown in Figure 10, GPT-4.1-nano begins with an accuracy of approximately 10%, which rises by roughly 2% over successive turns for all prompts. We explore why this increase occurs in Section 5.3, where we utilize Markov chains to assess the models' stationary accuracies.

### 4.2 Results for rephrased prompts

We conducted experiments on GPT-4.1-nano and Gemini 1.5 Flash (see Figure 11), and then further tested on Claude 3.5 Haiku and GPT-4o. These first two models overall showed slight decreases in accuracy, with Gemini 1.5 Flash showing higher accuracy degradation, most notably 2.5% for prompt URW. In contrast to the simple follow-up prompt setting used for the first two models, the latter two models employed a Chain-of-Thought prompting approach, as detailed in Appendix B (Wei et al., 2022). Our experiments with rephrased prompts show that slightly reworded questions combined with multi-turn prompting produce effects similar to those of simple follow-up prompts, with all three prompts resulting in an average accuracy drop of 15.7% for Claude 3.5 Haiku on MathQA and approximately 3% for GPT-4o on MMLU (see Figure 3). The URW prompt again induced the highest rate of answer changes. These findings suggest that even slight prompt rewording can induce multi-turn accuracy degradation, underscoring the current limitations in model robustness.

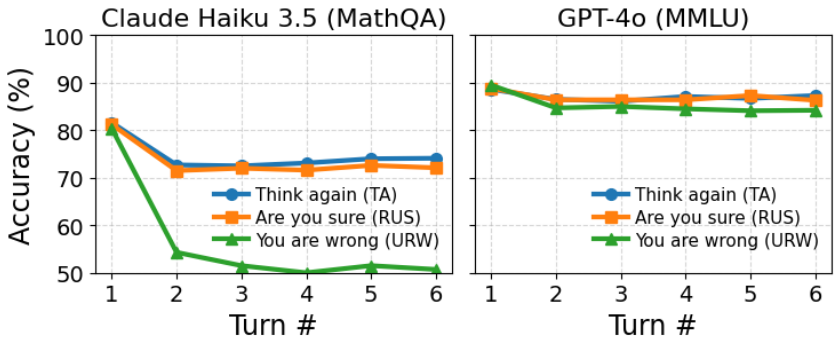


Figure 3: Accuracy drift across 6 turns for Claude 3.5 Haiku (MathQA) and GPT-4o (MMLU). For Claude 3.5 Haiku, the maximum accuracy decline for each prompt was $81.6\% \rightarrow 72.5\%$ on TA, $81.2\% \rightarrow 71.5\%$ on RUS, and $80.4\% \rightarrow 50.0\%$ on URW. For GPT-4o, the maximum accuracy decline for each prompt was from $88.6\% \rightarrow 86.1\%$ on TA, from $88.8\% \rightarrow 86.3\%$ on RUS, and from $89.5\% \rightarrow 84.1\%$ on URW.

## 5 Modeling multi-turn interactions with Markov chains

### 5.1 Markov chain introduction

Markov chains are probabilistic models that describe the likelihood of transitions between a finite set of discrete states (Pasanisi et al., 2012). They provide a simple yet powerful framework to capture how a model's answers evolve across multiple turns using probabilities. This approach allows us to analyze and predict the probability of answer changes over time, rather than considering each response independently. This is useful for uncovering systematic patterns in the fluctuations of model predictions over multiple turns.

### 5.2 Methodology

We model accuracy changes over turns using a two-state Markov chain, where states represent correct (1) or incorrect (0) answers. At each turn, the model has some probability of being in the correct state and the complementary probability of being in the incorrect state. To estimate the transition dynamics, we split the dataset into 80% for training and 20% for validation. From the training data, we count how often the model stays correct, flips from correct to incorrect, flips from incorrect to correct, or stays incorrect. These counts are then used to estimate the probabilities of switching between states: specifically, the chance of going from correct to incorrect ($p_{TF}$), and the chance of going from incorrect to correct ($p_{FT}$).

Using these probabilities, we construct a transition matrix that tells us how likely the model is to move between states from one turn to the next. Starting from the validation set's initial accuracy, we simulate how the probability of correctness evolves across turns by repeatedly applying the transition matrix, as seen in Equation 1:

$$\begin{bmatrix} a_{i+1} \\ 1 - a_{i+1} \end{bmatrix} = \begin{bmatrix} 1 - p_{TF} & p_{FT} \\ p_{TF} & 1 - p_{FT} \end{bmatrix} \begin{bmatrix} a_i \\ 1 - a_i \end{bmatrix} \tag{1}$$

where $a_i$ represents the simulated accuracy of the model at turn $i$. This allows us to see how accuracy changes across multiple reconsiderations, up to ten turns in our experiments.

Over many turns, the system converges to a stationary accuracy: the long-run probability that the model will be correct if the process were repeated indefinitely (Equation 2). If this stationary accuracy is lower than the starting accuracy, it means the model's answers tend to destabilize with more reconsiderations. If it is higher, it suggests the model has a tendency to self-correct and improve over time.

$$\text{Acc}_\infty = \frac{p_{FT}}{p_{TF} + p_{FT}} \tag{2}$$

We use log loss and mean squared error (MSE) to assess how well the model's predicted probabilities align with actual outcomes. A log loss of 0 indicates that the predicted probabilities exactly match the observed outcomes, with higher values reflecting poorer probabilistic calibration. MSE, on the other hand, quantifies the average squared deviation between predicted probabilities and actual model outcomes.

### 5.3 Results for simple follow-up prompts

Figure 4 displays the true and Markov simulated accuracies for GPT-4.1-nano and Gemini 1.5 Flash on MathQA, both with the TA prompt. We found that the simulated accuracy accurately approximates the true multi-turn dynamics of both models.[2] These results align with the RUS and URW prompts, seen in Figure 13 and Figure 14. In the subjective GlobalOpinionsQA dataset, the Markov model closely simulated the observed trends as well, especially as the number of turns increased (see Figures 15-17).

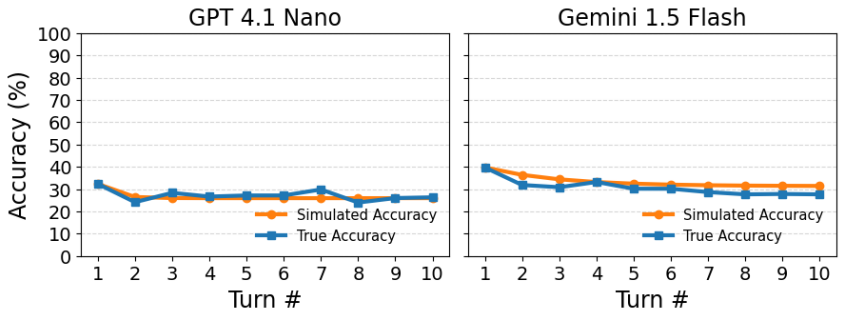


Figure 4: True vs. simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash. Both models were prompted using TA on MathQA questions. For GPT-4.1-nano, accuracies on turn 10 deviated by 0.38%. For Gemini 1.5 Flash, accuracies on turn 10 deviated by 3.76%. The close match between simulated and true accuracy shows that the Markov simulation accurately captures the model's multi-turn dynamics.

[2]Tables for log loss and MSE are shown in Appendix G

After noting the accuracy increase in HLE, we attempt to explain this using Markov chains. This increase in accuracy is in contrast with other datasets, possibly due to random answer switching from initially incorrect guesses to correct ones. To provide some intuition for this conjecture, we modeled the expected random-guess accuracy as a two-state Markov chain, providing a baseline that shows that even random guessing can lead to an increase in stationary accuracy (see Figures 23-25). Our results also indicate that answer dynamics on the HLE dataset can be effectively modeled using Markov chains, even when initial precision is extremely low and when stationary accuracy increases. That said, it is important to note that the decline in stationary accuracy for other datasets was much more substantial than the increase for HLE.

### 5.4 Results for rephrased prompts

Figures 28-30 plots true and simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash, and Figure 5 plots true and simulated accuracy for GPT-4o and Claude 3.5 Haiku on the RUS prompt. Again, our Markov model is able to well-approximate the multi-turn dynamics of GPT-4o and Claude 3.5 Haiku for rephrased prompts. These results align with the TA and URW prompts as well, seen in Figures 26 and 27.

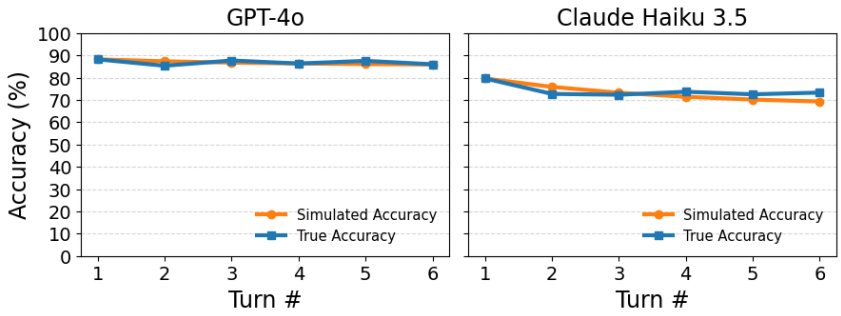


Figure 5: True vs. simulated accuracy for GPT-4o (MMLU) and Claude 3.5 Haiku (MathQA) with the rephrased RUS prompt. For GPT-4o, accuracies on turn 6 deviated by 0.11%. For Claude 3.5 Haiku, accuracies on turn 6 deviated by 3.99%. The close match between simulated and true accuracy shows that the Markov simulation accurately captures the model's multi-turn dynamics.

GPT-4o on MMLU and Claude 3.5 Haiku on MathQA exhibit patterns consistent with those seen under simple follow-up prompts, with the URW prompt producing the largest discrepancies between stationary and original accuracy (Figure 31). Specifically, GPT-4o on MMLU exhibits drops of 2.47% for prompt TA, 3.31% for prompt RUS, and 6.33% for prompt URW. In contrast, Claude 3.5 Haiku on MathQA shows substantially larger decreases of 12.73%, 13.9%, and 34.82% for the same prompts.

### 5.5 Comparing simple follow-up and rephrased prompts

By comparing stationary accuracy degradation gathered from Markov Chains we can assess whether vulnerabilities in model robustness are more pronounced under simple follow-up prompts or semantically rephrased prompts. A complete stationary accuracy degradation table can be viewed in Appendix H, with Figure 6 showing a comparison for Gemini 1.5 Flash.

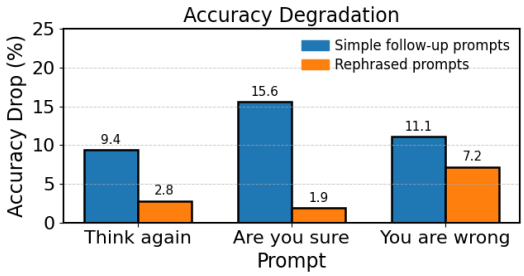


Figure 6: Accuracy degradation of Gemini 1.5 Flash on MathQA. Overall, the model's accuracy decreases by an average of 12.03% for simple follow-up prompts and 3.97% for rephrased prompts. This suggests that a model is more robust to reworded questions than to simple follow-up prompts.

# 6 Can probes predict when a model will change its mind?

## 6.1 Linear probing

Linear probes are a commonly used technique to analyze representations learned by neural networks (Alain & Bengio, 2016). Applying linear probes to an LLM's hidden states provides insight into internal model dynamics by revealing whether specific information is implicitly represented in intermediate layers (Skean et al., 2025). In this work, we assess whether linear probes can be used to effectively predict future answer changes and identify the layers in which these predictive signals first emerge.

## 6.2 Methodology

To investigate the model's internal representations, we conducted our probing experiments using the open-source Gemma 3 4B model (Team et al., 2025). We start by extracting the hidden state vectors for the last token in every layer using a simplified user prompt, seen in Appendix B. This hidden vector encodes the model's internal contextual representations at each step of the processing, reflecting what the model has integrated so far. In order to analyze the relationship between these internal representations and the model's answer stability, we pair each hidden vector with a binary label indicating whether the model changed its answer on that subsequent turn. Both the hidden vectors and labels are then used to train a linear probe using ridge regression to predict, from the internal state of the model at each turn, whether the model changes its answer on the next reconsideration. For brevity, we omit the low-level implementation details of linear probing and refer readers to Gurnee & Tegmark (2023) and Marks & Tegmark (2023) for reference.

After training the classifier on 80% of the labeled data, we assess its generalization performance by comparing the predicted outputs on the held-out test set to its true labels. Because the model retained its original answer in more turns than it revised, we applied stratified sampling to select an equal number of questions from turns with unchanged answers and turns with changed answers. This balanced sampling approach ensures that our analysis fairly compares model behavior across these conditions. We then evaluate probe performance using accuracy, reporting the proportion of correct predictions made by the trained linear probes on the test set. This experiment is repeated for all three reconsideration prompts on Gemma 3 4B on MathQA.

## 6.3 Results

Figure 7 illustrates how the probe's predicted probability of an answer change increases in the early layers under the TA prompt, then stabilizes after layer 3. This pattern suggests that signals indicative of potential answer changes are present in the early layers, and that our linear probes can detect them effectively. Under the adversarial URW prompt,[3] we observe a weaker trend: probabilities rise slightly in the initial layers before fluctuating, making its results hard to interpret. This suggests that adversarial prompts make it harder to use probes to predict when a model is going to change its answer. Our approach could be further enhanced by employing more capable models, by evaluating a larger set of questions, or by training non-linear probes, which may reveal stronger and more robust evidence. However, these preliminary results show that probing for answer changes could be valuable for tasks such as early intervention during inference. By detecting signals that a model is likely to change its answer, the system could alert users in advance. This allows for users to decide whether to modify the input, request additional clarification, or re-run the model—potentially saving compute resources.

[3]The RUS prompt was excluded from analysis due to the model producing too few answer changes, which provided inadequate training data.

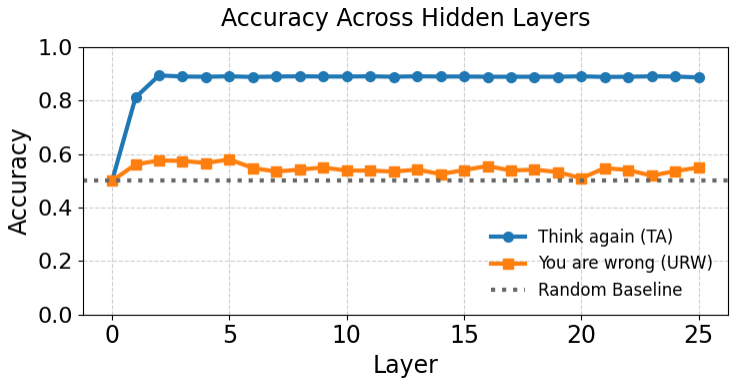


Figure 7: Probed hidden layer predictions across 26 layers for Gemma 3 4B on dataset MathQA. Under the TA prompt, the linear probe's predicted probabilities of answer changes rise sharply in the early layers from 0.50 at layer 0 to 0.89 at layer 3, then stabilize around 0.88–0.89 through layer 3 to 25. Under the adversarial URW prompt, the linear probe's predicted probabilities increase more modestly from 0.50 at layer 0 to 0.58 by layer 5 and fluctuate between 0.51 and 0.55 in higher layers.

## 7 Conclusion

While our findings provide insights into LLM stability and answer dynamics, there are several limitations to consider. Although we tested on a broad range of models, we could not include other potentially more capable models due to budget constraints. Another limitation of this study is that our prompts do not fully reflect how users naturally interact with language models. Phrases such as "Think again" or systematically rephrased questions were deliberately constructed to probe robustness, but they differ from the informal and indirect ways users typically express uncertainty or disagreement. As a result, the model behaviors observed here may not entirely generalize to real-world interactions. Additionally, the scope of our probing experiments was limited, as they were conducted only on one set of models and datasets. Consequently, we cannot yet determine the extent to which these preliminary findings generalize to broader use cases. Finally, we did not compute error bars or significance testing for our evaluation across runs, again due to cost constraints.

That being said, our findings demonstrate a consistent decrease in model accuracy over multiple turns, without new evidence, highlighting the limited robustness of current models. Especially in high-stakes domains such as healthcare or law, ensuring such robustness is key to reliable deployment. Additionally, the successful modeling of accuracy dynamics across multiple turns using Markov chains enables for the prediction of future accuracies. Combined with preliminary evidence that linear probes can anticipate future answer changes, these results allow for more interpretable and reliable human-AI interactions by revealing when a model's confidence and correctness begins to diverge.

Ultimately, our study highlights that multi-turn prompting often degrades model performance across different follow-up prompts. This accuracy degradation was successfully modeled using Markov chains, which allowed us to compare stationary accuracy with original accuracy, while hidden-state probing provided preliminary evidence that future answer changes may be predictable. These insights underscore the need for future work focused on enhancing LLM robustness, particularly in high-stakes applications where consistent reasoning is critical.

## References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Made laine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen,

Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, An drey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *ArXiv*, abs/1610.01644, 2016. URL https://api.semanticscholar.org/CorpusID:9794990.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:173188048.

Anthropic. Introducing claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2025-08-19.

Esin Durmus, Karina Nyugen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. *ArXiv*, abs/2306.16388, 2023. URL https://api.semanticscholar.org/CorpusID:259275051.

Aaron Fanous, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. Syceval: Evaluating llm sycophancy, 2025. URL https://arxiv.org/abs/2502.08177.

Wes Gurnee and Max Tegmark. Language models represent space and time. *ArXiv*, abs/2310.02207, 2023. URL https://api.semanticscholar.org/CorpusID:263608756.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020. URL https://api.semanticscholar.org/CorpusID:221516475.

Pranjal Kumar. Large language models (llms): survey, technical frameworks, and future challenges. *Artif. Intell. Rev.*, 57:260, 2024. URL https://api.semanticscholar.org/CorpusID:271961846.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *ArXiv*, abs/2505.06120, 2025. URL https://api.semanticscholar.org/CorpusID:278481320.

Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Firm or fickle? evaluating large language models consistency in sequential interactions. In *Annual Meeting of the Association for Computational Linguistics*, 2025a. URL https://api.semanticscholar.org/CorpusID:277435669.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. Beyond single-turn: A survey on multi-turn interactions with large language models. *ArXiv*, abs/2504.04717, 2025b. URL https://api.semanticscholar.org/CorpusID:277621374.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Annual Meeting of the Association for Computational Linguistics*, 2017. URL https://api.semanticscholar.org/CorpusID:12777818.

Joshua Liu, Aarav Jain, Soham Takuri, Srihan Vege, Aslihan Akalin, Kevin Zhu, Sean O'brien, and Vasu Sharma. Truth decay: Quantifying multi-turn sycophancy in language models. *ArXiv*, abs/2503.11656, 2025. URL https://api.semanticscholar.org/CorpusID:277065947.

Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, Hanxun Huang, Yige Li, Jiaming Zhang, Xiang Zheng, Yang Bai, Henghui Ding, Zuxuan Wu, Xipeng Qiu, Jingfeng Zhang, Yiming Li, Jun Sun, Cong Wang, Jindong Gu, Baoyuan Wu, Siheng Chen, Tianwei Zhang, Yang Liu, Min Gong, Tongliang Liu, Shirui Pan, Cihang Xie, Tianyu Pang, Yinpeng Dong, Ruoxi Jia, Yang Zhang, Shi jie Ma, Xiangyu Zhang, Neil Gong, Chaowei Xiao, Sarah Erfani, Bo Li, Masashi Sugiyama, Dacheng Tao, James Bailey, and Yu-Gang Jiang. Safety at scale: A comprehensive survey of large model safety. *ArXiv*, abs/2502.05206, 2025. URL https://api.semanticscholar.org/CorpusID:276250478.

Lars Malmqvist. Sycophancy in large language models: Causes and mitigations. *ArXiv*, abs/2411.15287, 2024. URL https://api.semanticscholar.org/CorpusID:274234383.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *ArXiv*, abs/2310.06824, 2023. URL https://api.semanticscholar.org/CorpusID:263831277.

Alberto Pasanisi, Shuai Fu, and Nicolas Bousquet. Estimating discrete markov models from various incomplete data schemes. *Computational Statistics & Data Analysis*, 56(9):2609–2625, September 2012. ISSN 0167-9473. doi: 10.1016/j.csda.2012.02.027. URL http://dx.doi.org/10.1016/j.csda.2012.02.027.

Yudi Pawitan and Chris Holmes. Confidence in the reasoning of large language models. *ArXiv*, abs/2412.15296, 2024. URL https://api.semanticscholar.org/CorpusID:274816808.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Richard Ren, Ryan Kim, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Daron Anderson, Tung Nguyen, Imad Ali Shah, Mobeen Mahmood, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Prashant Gangal, Chelsea Zou, Zihan Wang, Jessica P. Wang, Pawan Kumar, Olek sandr Pokutnyi, Robert Gerbicz, Serguei Popov, John-Clark Levin, Mstyslav Kazakov, Johannes Schmitt, Geoff Galgon, Alvaro Sanchez, Yongki Lee, Will Yeadon, Scott Sauers, Marc Roth, Chidozie Victor Agu, Søren Riis, Fabian Giska, Saiteja Utpala, Zachary Giboney, Gashaw M Goshu, Joan of Arc Xavier, Sarah-Jane Crowson, Mohinder Maheshbhai Naiya, Noah Burns, Lennart Finke, Zerui Cheng, Hyunwoo Park, Francesco Fournier-Facio, John B. Wydallis, Mark J. Nandor, Ankit Singh, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Darling Duclosel, Jungbae Nam, Jennifer A. Zampese, Ryan G. Hoerr, Aras Bacho, Gautier Abou Loume, Abdallah Galal, Hangrui Cao, Alexis C Garretson, Damien Sileo, Qiuyu Ren, Doru Cojoc, Pavel Arkhipov, Usman Qazi, Lianghui Li, Sumeet Ramesh Motwani, Christian Schröder de Witt, Edwin Taylor, Johannes Veith, Eric Singer, Taylor D. Hartman, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G. Willcocks, Joshua Robinson, Aleksandar Mikov, Ameya Prabhu, Longke Tang, Xavier Alapont, Justine Leon Uro, Kevin Zhou, Emily de Oliveira Santos, Andrey Pu pasov Maksimov, Edward Vendrow, Kengo Zenitani, Julien Guillod, Yuqi Li, Joshua Vendrow, Vladyslav M. Kuchkin, Ze-An Ng, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Andrew Gritsevskiy, Dakotah Martinez, Benjamin Thomas Pageler, Nicholas Crispino, Dimitri Zvonkine, Natanael Wildner Fraga, Saeed Soori, Ori Press, Henry Tang, Julian Salazar, Sean R. Green, Lina Brussel, Moon Twayana, Aymeric Dieuleveut, T. Ryan Rogers, Wenjin Zhang, Bikun Li, Jinzhou Yang, Arun Rao, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Subrata Mishra, Ariel Ghislain Kemogne Kamdoum, Tobias Kreiman, Tad Hogg, Alvin Jin, Carlo Bosio, Gongbo Sun, Brian P. Coppola, Tim Tarver, Haline Heidinger, Rafael Sayous, Stefan Ivanov, Joe Cavanagh, Jiawei Shen, Joseph Marvin Imperial, Philippe Schwaller, Shaipranesh Senthilkuma, Andrés M Bran, Ali Dehghan, Andres Algaba, Brecht Verbeken, David A. Noever, V RagavendranP, Lisa Schut, Ilia Sucholutsky, Evgenii Zheltonozhskii, Derek Lim, Richard Stanley, Shankar N. Sivarajan, Tong Yang, John Maar, Julian Wykowski, Mart'i Oller, Jennifer Sandlin, Anmol Sahu, Yuzheng Hu, Sara Fish, Nasser Heydari, Archimedes T. Apronti, Kaivalya Rawal, Tobías García Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Jeremy Nguyen, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Alan Goldfarb, Sergey Ivanov, Rafal Poswiata, Chenguang Wang, Daofeng Li, Donato Crisostomi, Donato Crisostomi, Benjamin Myklebust, Archan Sen, David Perrella, Nurdin Kaparov, Mark H Inlow, Allen Zang, Elliott Thornley, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Dan Bar Hava, Aleksey Kuchkin, Robert Lauff, David Holmes, Frank Sommerhage, Keith Schneider, Zakayo Kazibwe, Nate Stambaugh, Mukhwinder Singh, Ilias Magoulas, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Dae Hyun Kim, Kanu Priya Agarwal, Victor Efren Guadarrama Vilchis, Immo Klose, Christoph Demian, Ujjwala Anantheswaran, Adam Zweiger, Guglielmo Albani, Jeffery Li, Nicolas Daans, Maksim Radionov, V'aclav Rozhovn, Ziqiao Ma, Christian Stump, Mohammed Berkani, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Marco Piccardo, Ferenc Jeanplong, Niv Cohen, Josef Tkadlec, Paul Rosu, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Aline Menezes, Arkil Patel, Zixuan Wang, Jamie Tucker-Foltz, Jack Stade, Tom Goertzen, Fereshteh Kazemi, Jeremiah Milbauer, John Arnold Ambay, Abhishek Shukla, Yan Carlos Leyva Labrador, Alan Givr'e, Hew Wolff, Vivien Rossbach, Muhammad Fayez Aziz, Younesse Kaddar, Yanxu Chen, Robin Zhang, Jiayi Pan, Antonio Terpin, Jiayi Pan, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Adam Jones, Jainam Shah, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Andrew Ho, Shaul Barkan, Jiaqi Wang, Martin Stehberger, Egor I. Kretov, Kaustubh Sridhar, Kaustubh Sridhar, Anji Zhang, Daniel Pyda, Joanna Tam, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, Daniel Bugas, David Aldous, Jesyin Lai, Shannon Coleman, Mohsen Bahaloo, Jiangnan Xu, Sangwon Lee, Sandy Zhao, Ning Tang, Michael K. Cohen, Micah Carroll, Micah Carroll, Jan Hendrik Kirchner, Stefan Steinerberger, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Benedito Alves de Oliveira Junior, Michael Wang, Yuzhou Nie, Paolo Giordano, Philipp Petersen, Anna Sztyber-Betley, Priti Shukla, Jonathan Crozier, Antonella Pinto, Shreyas Verma, Prashant Joshi, Zheng-Xin Yong, Allison Tee, Zheng-Xin Yong, Orion Weller, Raghav Singhal, Gang Zhang, Alexander Ivanov, Seri Khoury, Hamid Mostaghimi, Kunvar Thaman, Qijia Chen, Tran Quoc Kh'anh, Jacob Loader, Stefano Cavalleri, Hannah Szlyk, Zachary Brown, Jonathan Roberts, William Alley, Kunyang Sun, Ryan T. Stendall, Max Lamparth, Anka Reuel, Ting Wang, Hanmeng Xu, Sreenivas Goud Raparthi, Pablo

Hern'andez-C'amara, Freddie Martin, Dmitry Malishev, Thomas Preu, Tomasz Korbak, Marcus Abramovitch, Dominic J. Williamson, Ziye Chen, Bir'o B'alint, M Saiful Bari, Peyman Hosseinzajeh Kassani, Zihao Wang, Behzad Ansarinejad, Laxman Prasad Goswami, Yewen Sun, Hossam Elgnainy, Daniel Tordera, George Balabanian, Earth Anderson, Lynna Kvistad, Alejandro Jos'e Moyano, Rajat Maheshwari, Ahmad Sakor, Murat Eron, Isaac C. McAlister, Javier Gimenez, Innocent Enyekwe, Andrew Favre D.O., Shailesh Shah, Xiaox iang Zhou, Firuz Kamalov, Ronald Clark, Sherwin Abdoli, Tim Santens, Khalida Meer, Harrison K Wang, K. K. Ramakrishnan, Evan Chen, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Niels Mundler, Avi Semler, Emma Rodman, Jacob Drori, Carl J Fossum, Milind Jagota, Ronak Pradeep, Honglu Fan, Tej Shah, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Carter Harris, Jason Gross, Ilya Gusev, Asankhaya Sharma, Shashank Agnihotri, Pavel Zhelnov, Siranut Usawasutsakorn, Mohammadreza Mofayezi, Sergei Bogdanov, Alexander Piperski, Marc Carauleanu, David K. Zhang, Dylan Ler, Roman Leventov, Ignat Soroko, Thorben Jansen, Pascal Lauer, Joshua Duersch, Vage Taamazyan, Wiktor Morak, Wenjie Ma, William Held, Tran DJuc Huy, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Hossein Shahrtash, Edson Oliveira, Joseph W. Jackson, Daniel Espinosa Gonzalez, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Al Dasouqi, Alexander Shen, Emilien Duc, Bita Golshani, David Stap, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Lukas Lewark, M'aty'as Vincze, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Muzhen Jiang, Fredrik Ekstrom, Angela Hammon, Oam Patel, Nicolas Remy, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Penaflor, Haile Kassahun, Alena Friedrich, Claire Sparrow, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Mike Battaglia, Mohammad Maghsoudimehrabani, Hieu Hoang, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Stephen Mensah, Nathan Andre, Anton Peristyy, Chris Harjadi, Himanshu Gupta, Stephen Malina, Samuel Albanie, Will Cai, Mustafa Mehkary, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Jasdeep Sidhu, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Brian Weber, Harsh Kumar, Tong Jiang, Arunim Agarwal, Chiara Ceconello, Warren S. Vaz, Chao Zhuang, Haon Park, Andrew R. Tawfeek, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Michael Kirchhof, Johan Ferret, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Shreen Gul, Gunjan Chhablani, Zhehang Du, Adrian Cosma, Colin White, Robin Riblet, Prajvi Saxena, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Shiv Halasyamani, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Renas Bacho, Vincent Ginis, Aleksandr Maksapetyan, Florencia de la Rosa, Xiuyu Li, Xiuyu Li, Leon Lang, Julien Laurendeau, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Yiugit Yalin, Gbenga D. Obikoya, Luca Arnaboldi, Rai, Filippo Bigi, Kaniuar Bacho, Pierre Clavier, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C.H. Lux, Ben Rank, Colin Ni, Alesia Yakimchyk, Huanxu Liu, Olle Haggstrom, Emil Verkama, Hi manshu Narayan, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Yiyang Fan, Gabriel Poesia Reis e Silva, Linwei Xin, Yosi Kratish, Jakub Lucki, Wen-Ding Li, Justin Xu, Kevin Scaria, Justin Xu, Farzad Habibi, Long Lian, Emanuele Rodolà, Jules Robins, Vincent Cheng, Declan Grabb, Ida Bosio, Tony Fruhauff, Ido Akov, Eve J. Y. Lo, Hao Qi, Xi Jiang, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Yibo Jiang, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Muhammad Rehan Siddiqi, Alon Ragoler, Justin Tan, Deepakkumar Patil, Rebeka Plecnik, Aaron Kirtland, Roselynn Grace G. Montecillo, Stephane Durand, Omer Faruk Bodur, Zahra Adoul, Mohamed Zekry, Guillaume Douville, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Sarah Hoback, Rodrigo De Oliveira Pena, Glen Sherman, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Gozdenur Demir, Sandra Mendoza, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Hsiaoyun Milliron, Mohammad Safdari, Liangti Dai, Yunze Xiao, Siriphan Arthornthurasuk, Alexey Pronin, Jingxuan Fan, Ángel Ramírez-Trinidad, Ashley Cartwright, Daphiny Pottmaier, Omid Taheri, David Outevsky, Stanley Stepanic, Mike Zhang, Samuel Perry, Luke Askew, Ra'ul Adri'an Huerta Rodr'iguez, Abdelkader Dendane, Sam Ali, Ricardo Lorena, Krishnamurthy Iyer, Sk Md Salauddin, Murat Islam, Juan Carlos Gonzalez, Josh Ducey, Russell Campbell, Maja Somrak, Vasilios Mavroudis, Eric Vergo, Juehang Qin, Benj'amin Borb'as, Eric Chu, Jack Lindsey, Anil Radhakrishnan, Antoine Jallon, I.M.J. McInnis, Alex Hoover, Soren Moller, Song Bian, John Lai, Tejal Patwardhan, David Anugraha, Xing Han Lù, Xuandong Zhao, Summer Yue, Alexandr Wang, Dan Hendrycks, Anjiang Wei, Francisco-Javier Rodrigo-Ginés, Gabriele Sarti, A. H. Elneklawy,

Bruno Hebling Vieira, and Wei Hao. Humanity's last exam. *ArXiv*, abs/2501.14249, 2025. URL https://api.semanticscholar.org/CorpusID:275906652.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Ben jamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross Mcilroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem W. Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Os car Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomás Kociský, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, J Christopher Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Ying-Qi Miao, Lukás Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontan'on, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, A.E. Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Venkatesh Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matt Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara N. Sainath, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela de Castro Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Au rko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo-Yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Joshua Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost R. van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya B Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xi ance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, S'ebastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Michael B. Chang, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Roșca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravichandra Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Luvci'c, Rajku mar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjosund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos L. Araya, Nikhil

Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Zhufeng Pan, Zachary Nado, Stephanie Winkler, Dian Yu, Mohammad Saleh, Lorenzo Maggiore, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxi aoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabriel Barth-Maron, Craig Swanson, Dominika Rogozi'nska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Ren shen Wang, Dave Lacey, Anastasija Ili'c, Yao Zhao, Woohyun Han, Lora Aroyo, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Raphael Lopez Kaufman, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, T. Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anais White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Kiran Vodrahalli, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mah moud Alnahlawi, Christo pher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, cCauglar Unlu, David Reid, Zora Tung, Daniel F. Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Rui Zhu, Ricardo Aguilar, Mai Gim'enez, Jiawei Xia, Olivier Dousse, Willi Gierke, Soheil Hassas Yeganeh, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Livio Baldini Soares, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Daniel Niels Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Daniel Toyama, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nicholas Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, Donghyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alexey Yakubovich, Nilesh Tripuraneni, James Manyika, Ha roon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Anna Bulanova, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clément Farabet, Pedro Valenzuela, Quan Yuan, Christoper A. Welty, Ananth Agarwal, Mianna Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Elnaz Davoodi, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Re beca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, A. Ya. Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Alejandro Lince, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiří Šimša, Anna Koop, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Kalpesh Krishna, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Lucas Dixon, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas Fitzgerald, Aviral Kumar, Nicholas FitzGerald, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Ilia Shumailov, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Kather ine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel El Kaed, Jing Li, Jakub Sygnowski, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Junwen Bai, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, Oriol Vinyals, and Alexandra Chronopoulou. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024. URL `https://api.semanticscholar.org/CorpusID:268297180`.

Abel Salinas and Fred Morstatter. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:266844185.

Mikhail Seleznyov, Mikhail Chaichuk, Gleb Yu. Ershov, Alexander Panchenko, Elena Tutubalina, and Oleg Somov. When punctuation matters: A large-scale comparison of prompt robustness methods for llms. 2025. URL https://api.semanticscholar.org/CorpusID:280671474.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025. URL https://arxiv.org/abs/2310.13548.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:275954328.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *ArXiv*, abs/2502.02013, 2025. URL https://api.semanticscholar.org/CorpusID:276107264.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin,

Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. URL https://api.semanticscholar.org/CorpusID:246411621.

Xiyang Wu, Ruiqi Xian, Tianrui Guan, Jing Liang, Souradip Chakraborty, Fuxiao Liu, Brian M. Sadler, Dinesh Manocha, and A. S. Bedi. On the vulnerability of llm/vlm-controlled robotics. 2024. URL https://api.semanticscholar.org/CorpusID:267740494.

Qiming Xie, Zengzhi Wang, Yihao Feng, Rui Xia, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Juraf-sky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Pang Omar Khattab, Wei Koh, Mark S. Krass, Ranjay Krishna, Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph Gonzalez, Aakanksha Chowdh-ery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Hyung Paul Barham, Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vin-odkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghe-mawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fe-dus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Oleksandr Polozov, Kather ine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Hee woo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo, P. Privitera, Al-berto Eugenio Ferrag-ina, Tozzi Caterina, Rizzo, Chatgpt, Deep Ganguli, Liane Lovitt, John Kernion, Yuntao Bai, Saurav Kadavath, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Con-erly, Nova Dassarma, Dawn Drain, Sheer Nelson El-hage, El Showk, Stanislav Fort, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Nicholas Joseph, Chris Olah, Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, Jonathan Berant. 2021, Did Aristo-tle, Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Mohammad Hosseini, Catherine A Gao, David M. Liebovitz, Faraz Alexandre M Car-valho, S Ahmad, Yuan Luo, Ngan MacDonald, Kristi L. Holmes, Abel Kho. 2023, An, Edward J. Hu, Yelong Shen, Zeyuan Phillip Wallis, Kevin B. Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E. Frisse, Karl E. Misulis, Kyu Rhee, Juan Zhao, Tom Conerly, Nelson Elhage, Tristan Hume, Kamal Ndousse, Stephanie Lin, Owain Evans. 2022, Yao Lu, Max Bartolo, Alastair Moore, Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Ouyang Long, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Katarina Slama, Alex Ray, John Schulman, Fraser Kelton, Luke Miller, Maddie Simens, Peter Welinder, Paul F. Christiano, Jan Leike, Ryan Lowe. 2022, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Sandipan Kundu, Saurav Kada-vath, Brian Israel, Bryan Seethor, Cameron McKinnon, Da Yan, Daniela Amodei, Dustin Li, Guro Khundadze, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Lan-dau, Landon Goldberg, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Noem'i Mercado, Oliver Rausch, Robin Larson, Tamera Lanham, Timothy Telleen-Lawton, Roger Grosse, Evan Hubinger, Ansh Radhakrishnan, Carol Chen, Carson E. Denison, Esin Durmus, Newton Cheng, Sheer Sam McCan-dlish, Tamera Lanham, Tim Maxwell, and Venkatesa Chandrasekaran. Ask again, then fail: Large language models' vacillations in judgment. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL https://api.semanticscholar.org/CorpusID:271894946.

Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. A survey on multi-turn interaction capabilities of large language models. *ArXiv*, abs/2501.09959, 2025. URL https://api.semanticscholar.org/CorpusID:275606308.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Weirong Ye, Neil Zhenqiang Gong, Yue Zhang, and Xingxu Xie. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, 2023. URL https://api.semanticscholar.org/CorpusID:259095572.

# A Rephrased prompts variation generation process

Our variations were generated with explicit instructions to preserve semantic meaning, answer choices, and difficulty level. Here is a sample question and its variation:

- Original: "Mark bought a set of 6 flower pots of different sizes at a total cost of 8.25. Each pot cost 0.1 more than the next one below it in size. What was the cost, in dollars, of the largest pot?"
- Variation: "Samantha purchased 6 uniquely sized picture frames for a combined price of 8.25. Each frame was priced 0.10 more than the smaller one before it. What was the price of the largest frame in dollars?"

# B Prompts

## System prompt non-COT

```
"You are an assistant skilled in solving multiple choice questions.
Only respond with the final answer letter as the only character in your
message. Do not include any other text or explanations. This letter
should correspond with the correct answer to the question provided."
```

## User prompt non-COT

```
"Always respond with just the letter that corresponds to the correct
answer. Remember: Only respond with the final answer letter (A-E) as
the only character in your message. Do not include any other text or
explanations."
```

## System prompt COT:

```
"You are an assistant skilled in solving multiple choice questions.
Always provide the final letter as the last character in your message
in the format 'Final Answer: X' where X is the correct letter choice."
```

## User prompt COT:

```
 "Please think step by step, but always end your response with just the
final letter on the last line." + question + "Remember: You must ALWAYS
give your final answer as a capital letter as the LAST character in your
message, in the format 'Final Answer: X' where X is the correct letter
choice."
```

**Simplified user prompt for hidden-state experiments:**

```
"Answer with only the letter A, B, C, D, or E." + question
```

## C Additional simple follow-up prompt results

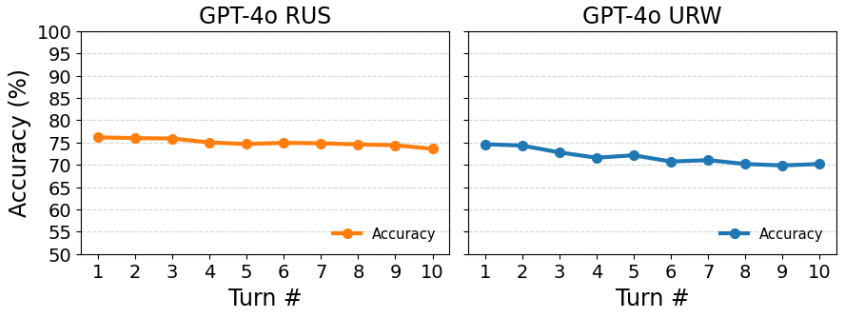


Figure 8: Accuracy drift across ten turns for GPT-4o on MathQA. Only two prompts were ran due to budget restraints.

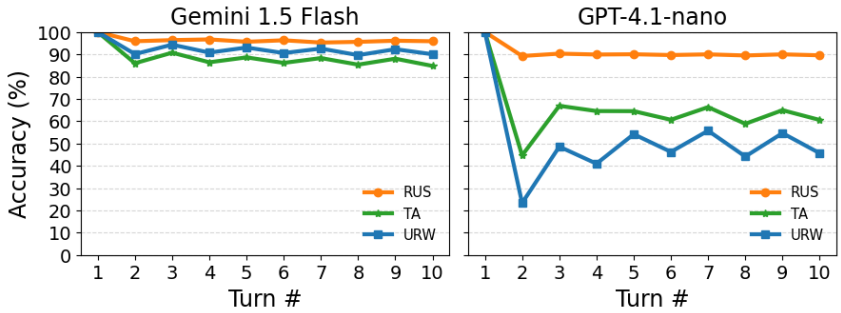


Figure 9: Accuracy drift across ten turns for Gemini 1.5 Flash and GPT-4.1-nano on GOQA.

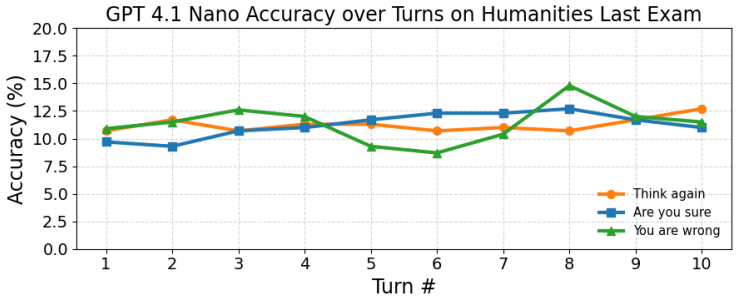


Figure 10: GPT-4.1-nano accuracy increases over turns on Humanities Last Exam.

## D Additional rephrased prompts

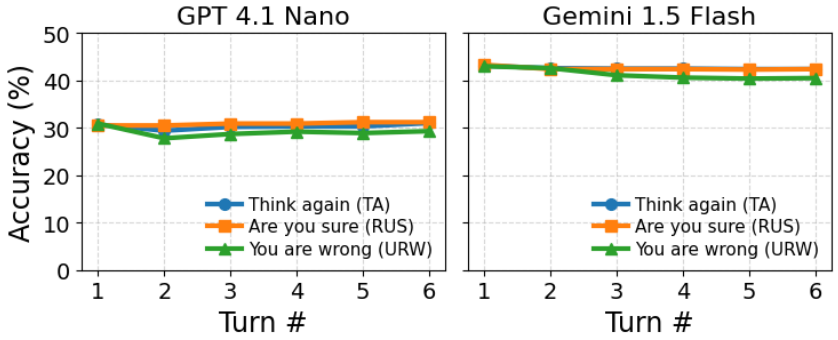


Figure 11: Accuracy drift across six turns for GPT-4.1-nano and Gemini 1.5 Flash on MathQA.

## E Control experiment

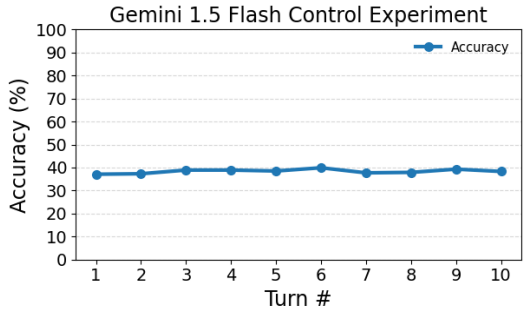


Figure 12: Gemini 1.5 Flash on 500 MathQA questions that are repeated nine times without a simple follow-up prompt.

## F Additional Markov modeling results

### Simple follow-up prompts:

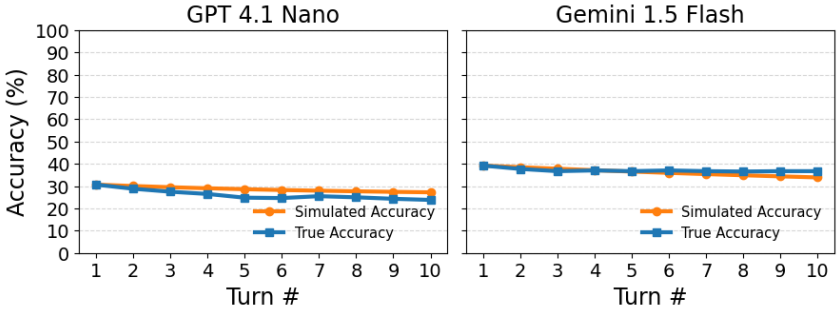


Figure 13: True vs simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash on dataset MathQA for the prompt RUS.

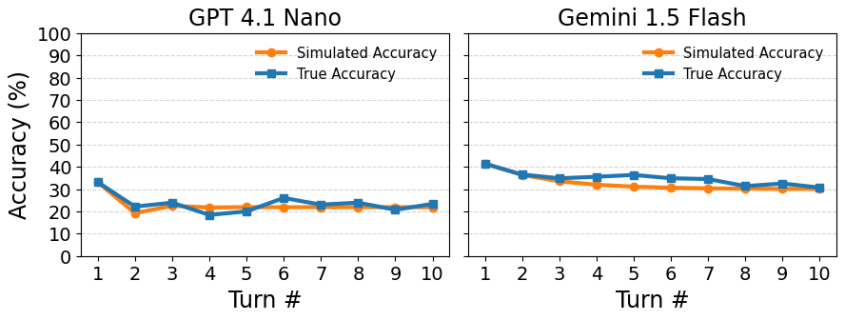


Figure 14: True vs simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash on dataset MathQA for the prompt URW.

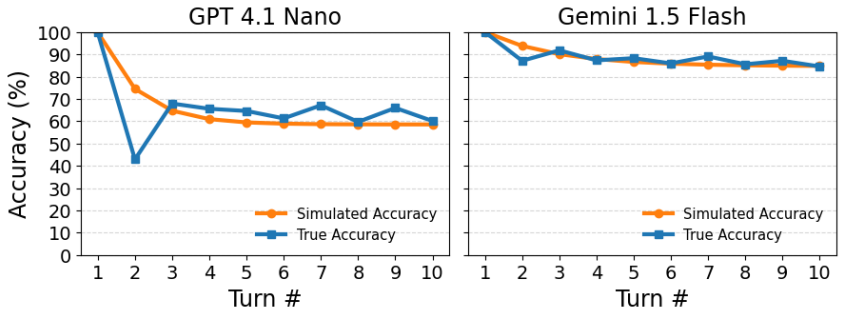


Figure 15: True vs simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash on dataset GOQA for the prompt TA.

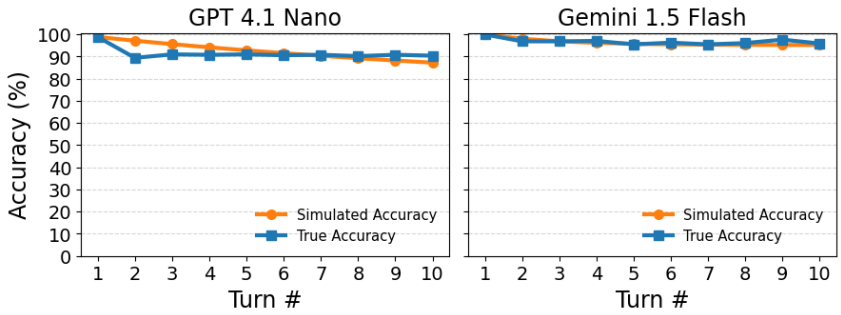


Figure 16: True vs simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash on dataset GOQA for the prompt RUS.

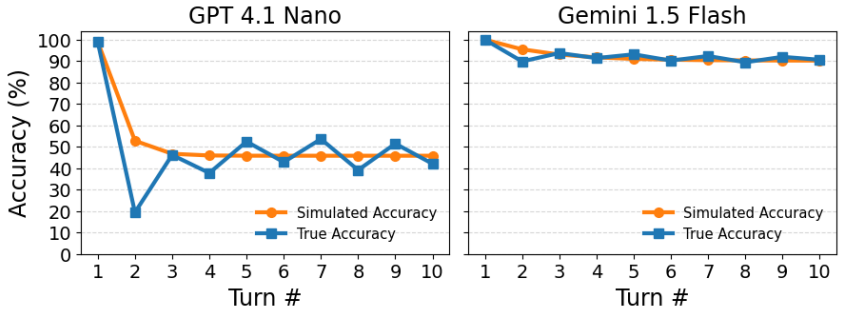


Figure 17: True vs simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash on dataset GOQA for the prompt URW.

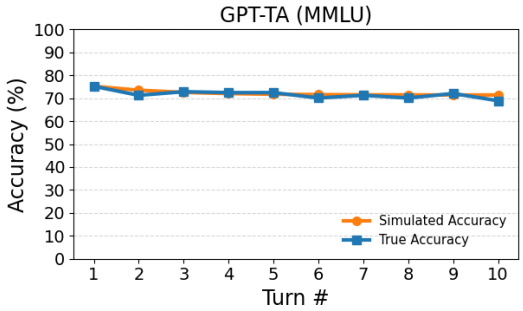


Figure 18: True vs simulated accuracy for GPT-4.1-nano on MMLU for the prompt TA.

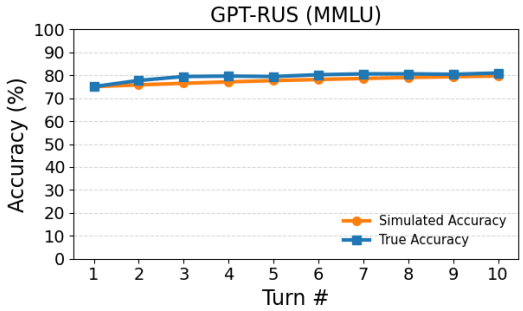


Figure 19: True vs simulated accuracy for GPT-4.1-nano on MMLU for the prompt RUS.

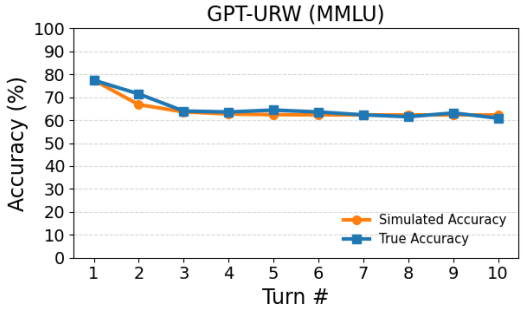


Figure 20: True vs simulated accuracy for GPT-4.1-nano on MMLU for the prompt URW.

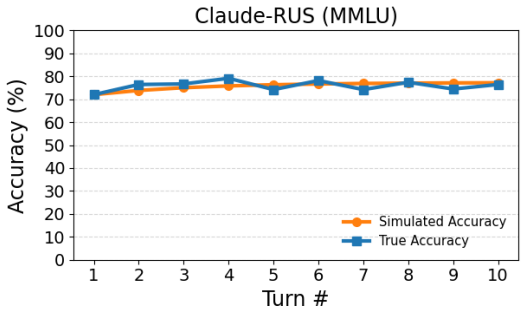


Figure 21: True vs simulated accuracy for Claude 3.5 Haiku on MMLU for the prompt RUS.

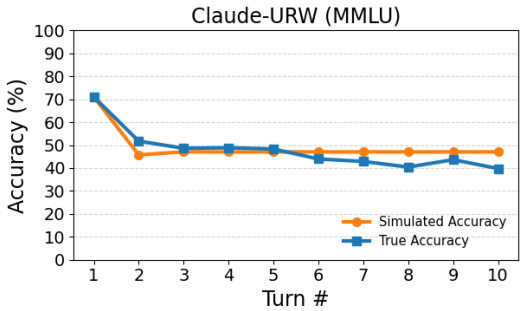


Figure 22: True vs simulated accuracy for Claude 3.5 Haiku on MMLU for the prompt URW.

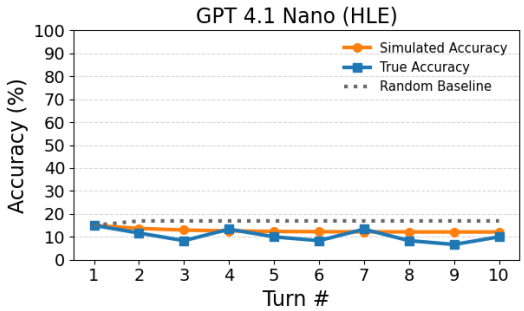


Figure 23: True vs simulated accuracy for GPT-4.1-nano on HLE for prompt TA.

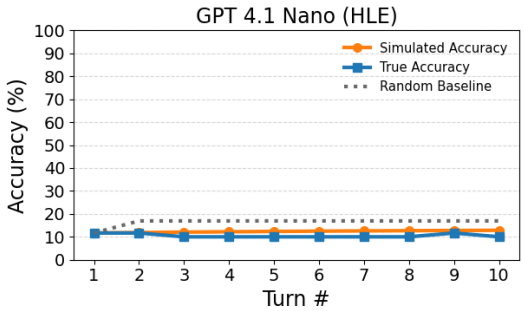


Figure 24: True vs simulated accuracy for GPT-4.1-nano on HLE for prompt RUS.

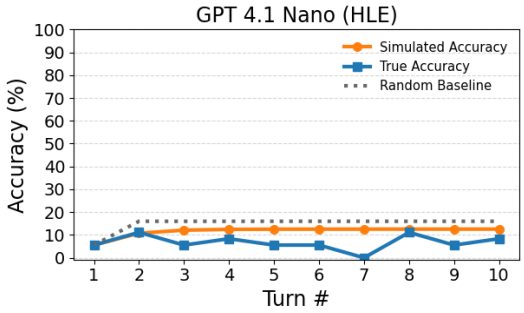

Figure 25: True vs simulated accuracy for GPT-4.1-nano on HLE for prompt URW.

## Rephrased follow-up prompts:

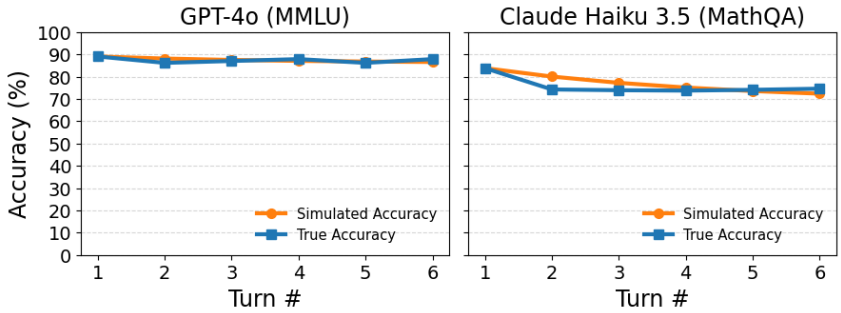

Figure 26: True vs simulated accuracy for GPT-4o on MMLU and Claude 3.5 Haiku on MathQA for the prompt TA.

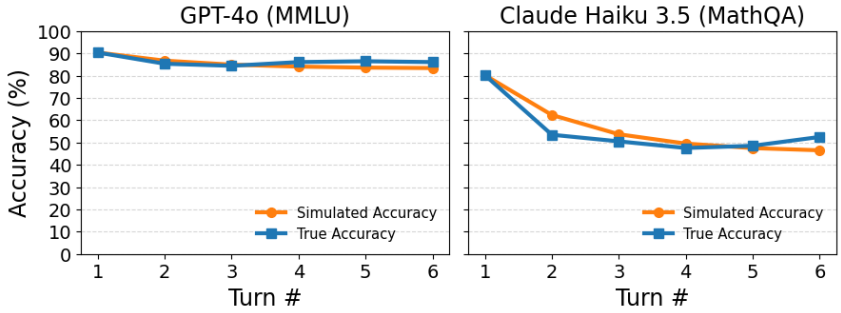

Figure 27: True vs simulated accuracy for GPT-4o on MMLU and Claude 3.5 Haiku on MathQA for the prompt URW.

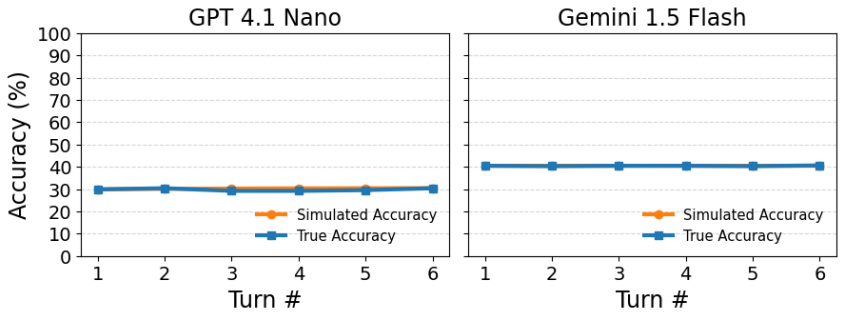


Figure 28: True vs simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash on MathQA for the prompt TA.

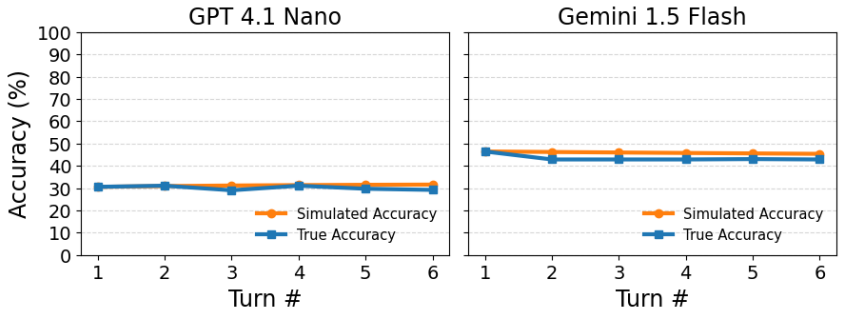


Figure 29: True vs simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash on MathQA for the prompt RUS.

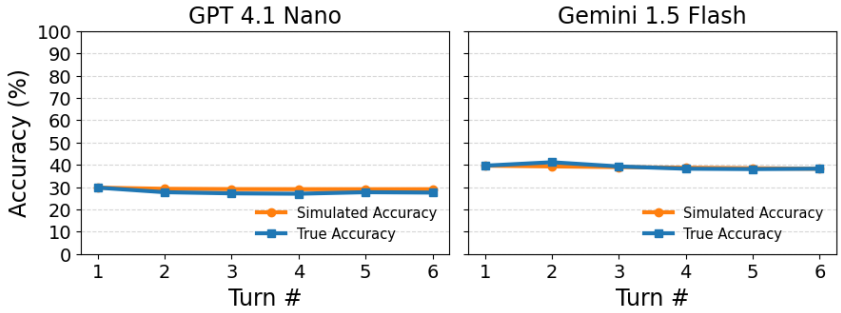


Figure 30: True vs simulated accuracy for GPT-4.1-nano and Gemini 1.5 Flash on MathQA for the prompt URW.

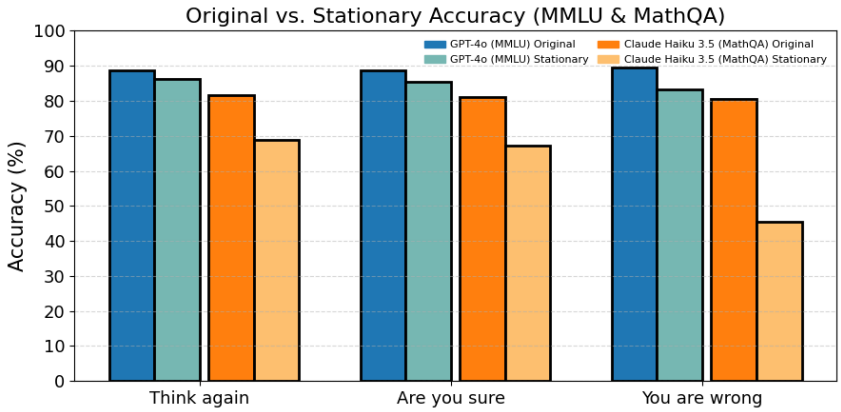

Figure 31: Comparison of original and stationary accuracies across three prompt types (TA, RUS, and URW) for GPT-4o on the MMLU dataset and Claude Haiku 3.5 on the MathQA dataset.

## G Error metrics

Table 1: Average log loss and MSE for Gemini 1.5 Flash on MathQA and GOQA.

(a) MathQA

| Prompt | Log Loss | MSE |
|---|---|---|
| RUS | 0.1118 | 0.0234 |
| TA | 0.4444 | 0.1388 |
| URW | 0.4743 | 0.1505 |

(b) GlobalOpinionsQA

| Prompt | Log Loss | MSE |
|---|---|---|
| RUS | 0.1094 | 0.0249 |
| TA | 0.2790 | 0.0771 |
| URW | 0.2294 | 0.0608 |

Table 2: Average log loss and MSE for GPT-4.1-nano on MathQA and GOQA.

(a) MathQA

| Prompt | Log Loss | MSE |
|---|---|---|
| RUS | 0.1930 | 0.0480 |
| TA | 0.5746 | 0.1934 |
| URW | 0.4924 | 0.1632 |

(b) GlobalOpinionsQA

| Prompt | Log Loss | MSE |
|---|---|---|
| RUS | 0.0915 | 0.0184 |
| TA | 0.6143 | 0.2119 |
| URW | 0.6736 | 0.2403 |

Table 3: Average log loss and MSE for Claude 3.5 Haiku and GPT-4o across MMLU and MathQA datasets.

(a) Claude 3.5 Haiku

| Prompt | Log Loss | MSE |
|---|---|---|
| RUS | 0.3349 | 0.0948 |
| TA | 0.2935 | 0.0796 |
| URW | 0.5436 | 0.1791 |

(b) GPT-4o

| Prompt | Log Loss | MSE |
|---|---|---|
| RUS | 0.2917 | 0.0821 |
| TA | 0.2080 | 0.0540 |
| URW | 0.3452 | 0.1012 |

## H Stationary accuracy change table

Table 4: Stationary accuracy change (%) across models and prompt types.

| Model | Type | Prompt | Dataset | Stationary Accuracy Change |
|---|---|---|---|---|
| Gemini 1.5 Flash | Rephrased | TA | MathQA | $-2.8$ |
| | Rephrased | RUS | MathQA | $-1.9$ |
| | Rephrased | URW | MathQA | $-7.2$ |
| | Simple Follow-Up | TA | MathQA | $-9.4$ |
| | Simple Follow-Up | RUS | MathQA | $-15.6$ |
| | Simple Follow-Up | URW | MathQA | $-11.1$ |
| GPT-4.1 Nano | Rephrased | TA | MathQA | $-0.3$ |
| | Rephrased | RUS | MathQA | $+1.3$ |
| | Rephrased | URW | MathQA | $-1.9$ |
| | Simple Follow-Up | TA | MathQA | $-4.4$ |
| | Simple Follow-Up | RUS | MathQA | $-5.8$ |
| | Simple Follow-Up | URW | MathQA | $-9.7$ |
| Claude 3.5 Haiku | Rephrased | TA | MathQA | $-12.73$ |
| | Rephrased | RUS | MathQA | $-13.90$ |
| | Rephrased | URW | MathQA | $-34.82$ |
| GPT-4o | Rephrased | TA | MMLU | $-2.47$ |
| | Rephrased | RUS | MMLU | $-3.31$ |
| | Rephrased | URW | MMLU | $-6.33$ |
| Gemini 1.5 Flash | Simple Follow-Up | TA | Global Opinions QA | $-5.0$ |
| | Simple Follow-Up | RUS | Global Opinions QA | $-15.4$ |
| | Simple Follow-Up | URW | Global Opinions QA | $-9.8$ |
| GPT-4.1 | Simple Follow-Up | TA | Global Opinions QA | $-26.6$ |
| | Simple Follow-Up | RUS | Global Opinions QA | $-41.5$ |
| | Simple Follow-Up | URW | Global Opinions QA | $-54.2$ |