A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners

Bowen Jiang [1 *] **Yangxinyu Xie** [1 2 *] **Zhuoqun Hao** [1] **Xiaomeng Wang** [1]
Tanwi Mallick [2] **Weijie J. Su** [1] **Camillo J. Taylor** [1] **Dan Roth** [1]

Abstract

This study proposes a hypothesis-testing framework to determine whether large language models (LLMs) possess genuine reasoning abilities or rely on token bias. Carefully-controlled synthetic datasets are generated, and null hypotheses assuming LLMs' reasoning capabilities are tested with statistical guarantees. Inconsistent behavior during experiments leads to the rejection of null hypotheses. Our findings, using the conjunction fallacy as a quintessential example, suggest that current LLMs still struggle with probabilistic reasoning, with apparent performance improvements largely attributable to token bias.

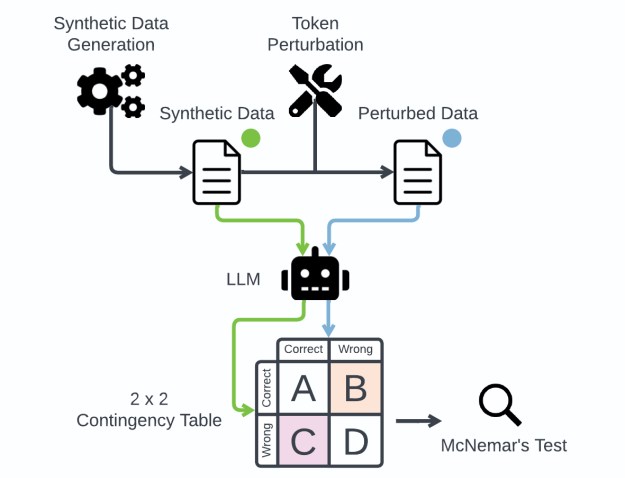

Figure 1: An illustration of the overall framework. We generate synthetic data, perform systematic token perturbations, and evaluate an LLM for comparative studies. The resulting contingency table, where A to D are integer values of counts, allows for statistical tests.

1. Introduction

Large language models (LLMs) have achieved remarkable progress in understanding and generating human-like text, triggering growing interest in the LLMs' theory of mind (Kosinski, 2023; Jamali et al., 2023; Bubeck et al., 2023) and decision-making abilities (Merrill & Sabharwal, 2023; Lyu et al., 2023; Prasad et al., 2023). However, there is ongoing debate about whether LLMs possess genuine reasoning capabilities, as evidence suggests that performance of LLMs on reasoning tasks is correlated with how much the input's semantic content supports a correct logical inference (Dasgupta et al., 2022; Li et al., 2023). If true reasoning has been applied, such a correlation won't exist, since a genuine reasoner should be able to derive the correct inference regardless of the semantic content.

In this paper, we formalize this observation and say that an LLM is subject to **token bias** in a reasoning task if, for a given reasoning task prompt, systematic changes to some or all tokens — while keeping the underlying logic intact — allow us to predict the direction of the shift in the model's output. As an example, consider the following reasoning task: *Suppose Taylor Swift embarks on another tour in 2027. Which outcome do you think is more likely?*

> *(a) Her first show is a flop.*
> *(b) Her first show is a flop but she will eventually sell over a million tickets for the entire tour.*

LLMs tend to prefer option (b), reasoning that despite a potential initial setback, Taylor Swift's consistent and immense popularity and success suggest a strong likelihood of overall tour success.[1] This choice, however, exemplifies the well-known **conjunction fallacy**, also known as the Linda Problem (Tversky & Kahneman, 1983; Kahneman, 2011): the probability of a conjunction of two events (e.g., Taylor Swift's first show is a flop *and* she will eventually sell over a million tickets for the entire tour) is never higher than the probability of either event alone. However, when we change the name "Taylor Swift" to "Nancy", disentangling the semantic narrative that concerts Taylor Swift's

[1]University of Pennsylvania, Philadelphia, PA, 19104, USA [2]Argonne National Laboratory, Lemont, IL, 60439, USA. Correspondence to: Bowen Jiang <bwjiang@seas.upenn.edu>, Yangxinyu Xie <xinyux@wharton.upenn.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]https://chatgpt.com/share/2ed7b17f-322a-4c6d-b22f-268a96463560

current success, LLMs recognize that from a probabilistic standpoint, outcome (a) is more likely than outcome (b).[2]

Building on this definition of token bias, we reconceptualize the evaluation of reasoning capabilities into a general, statistically rigorous framework with three critical components: **synthetic data generation, token perturbation, and statistical hypothesis testing**. This framework allows us to bypass the complications of evaluation set contamination (Zhou et al., 2023; Ravaut et al., 2024), leverage insights and tools from controlled experiments, and draw statistically valid conclusions.

Unlike concurrent works (Mukherjee & Chang, 2024; Wang et al., 2024; Suri et al., 2024), our objective is not to engineer prompts to yield nearly perfect benchmark results; rather, we aim to systematically examine and validate our hypotheses of reasoning behavior through carefully controlled experiments. We only leverage common prompting techniques that are sufficient to provide robust statistical evidence of whether LLMs tend to exploit biased tokens as shortcuts or consistently apply genuine logical reasoning.

2. The General Framework

Our framework is summarized in Figure 1. This general framework is grounded on the premise that for a given reasoning task, a capable reasoning agent will consistently reach the same conclusion regardless of how the task is framed, as long as the underlying logic remains the same (Hastie & Dawes, 2009). This assumption lays the foundation of our null hypothesis, $\boldsymbol{H_0}$. In our setup, a rational agent should consistently apply reasoning in its decision-making process. Under this paradigm, the only source of failure should be the procedural mistakes during the agent's abstract reasoning steps, which we assume to come up in an i.i.d. fashion. Our general framework contains three major parts as follows.

Synthetic Data Generation Once the underlying logic of a reasoning task is defined, we create an algorithm to generate a synthetic dataset with n samples. While it is helpful to leverage LLMs for linguistic coherence in the process, the data generation should be carefully controlled, utilizing information from real-world data or established datasets to mitigate potential biases from purely AI-generated texts. The process begins with the creation of a curated list of entities and a textual template that dictates the structure of the task description. By sampling from this list, we generate task descriptions that maintain the integrity and novelty of the dataset. This method ensures that while the LLM of interest might be familiar with the individual entities, it has never seen the specific combinations of these entities and narratives, thus bypassing the risk of data contamination.

The synthetic dataset can be dynamically generated, precluding its prior existence in any training datasets. It also allows the algorithm designers to control the dataset size, efficiently scaling their data based on the sample size required for achieving statistical validity.

Token Perturbation We hypothesize that if the LLM is not a capable reasoning agent, its performance on reasoning tasks will consistently improve (or degrade) as we alter some tokens in a systematic manner. This process of token perturbation generates n matched pairs of samples, enabling us to evaluate the LLM on both the original and perturbed datasets and create a 2×2 contingency table below, where $n = n_{11} + n_{12} + n_{21} + n_{22}$.

		Perturbed	
		Correct	Wrong
Original	Correct	n_{11}	n_{12}
	Wrong	n_{21}	n_{22}

Table 1: A template for the contingency table.

Statistical Hypothesis Testing for Matched Pairs For each of n matched pairs, let π_{ab} denote the underlying probability of outcome a for the original dataset and b for the perturbed dataset. As n_{ab} count the number of such pairs, n_{ab}/n is the sample proportion, a consistent estimate of π_{ab}. The null hypothesis assumes the marginal homogeneity for binary matched pairs, i.e. $\pi_{12} = \pi_{21}$. For small samples, we apply an exact test conditioned on $n^* = n_{21} + n_{12}$ (Mosteller, 1952; Agresti, 2012). Under $\boldsymbol{H_0}, n_{21}$ follows a $\text{binomial}(n^*, 1/2)$ distribution, and the corresponding p-value is the binomial tail probability. As a rule of thumb, when $n^* > 10$, the reference binomial distribution is approximately normal, and we can compute the standardized normal test statistics $z_0 = (n_{21} - n_{12})/\sqrt{n_{21} + n_{12}}$, which is identical to the McNemar statistic (McNemar, 1947). To test the same hypotheses for a group of models, we apply the Benjamini-Hochberg Procedure (Benjamini & Hochberg, 1995) to control the false discovery rate at a predetermined significance level α.

3. Peek into Token Bias via the Linda Problem

In this section, we use the task of reasoning against conjunction fallacy as an example of our general framework and introduce several variants of the token perturbation mechanism to probe whether LLMs are susceptible to token biases.

3.1. Synthetic Data Generation

Given that LLMs likely encounter the original Linda Problem (see Appendix A) in their training datasets, we con-

[2] https://chatgpt.com/share/46d025c0-f205-4f4a-972a-55d01a04dad1

struct conjunction reasoning problems in a narrative similar to the one about Taylor Swift in Section 1. Specifically, we curate a set of celebrity names from the Times Person of the Year (Rosenberg, 2021) and Forbes Celebrity 100 (Wikipedia contributors, 2024), and harness the in-context learning abilities of GPT-4 (Achiam et al., 2023) to generate new problems of the following form:

> Suppose [celebrity is going to do something]. Which is more likely:
> *(a) [Something unlikely for this person].*
> *(b) [Something unlikely for this person], but [something extremely likely for this person].*

3.2. Token Perturbation and Alternative Hypotheses

Token Perturbation in the Task Description The presence of a celebrity's name might trigger irrelevant associations, such as misleading the model's attention into the celebrity's background. We posit that if we switch the name of the celebrity to a generic one, an LLM that relies on semantic shortcuts will observe an increase in its success rate of identifying the logic of conjunction:

Hypothesis 1 Genuine Reasoning LLMs withstand irrelevant token changes.

Sub-hypothesis 1.1 (Token Bias on Celebrities): The performance of a reasoning LLM should remain consistent if we change the name of a celebrity, if any, in the problem to a generic name.

Assume P is a conjunction fallacy problem that involves a celebrity name, which may mislead the LLM, while P' changes the name to a generic one.

$\boldsymbol{H_0}$: $\pi_{12} = \pi_{21}$.

$\boldsymbol{H_a}$: $\pi_{12} < \pi_{21}$. ($\pi_{12} > \pi_{21}$ is invalid.)

Token Perturbation in the Task Instruction In this setup, we study the model behavior under the in-context learning (Brown et al., 2020) setting, where we present a single instance of the Linda problem, either in its original form or a rephrased one, as the one-shot exemplar.

Our token perturbation stems from the intuition that a model might associate the occurrence of "conjunction fallacy" with the specific name "Linda" learned from training data. Hence, if we change the Linda problem into an equivalent one, but this time about a made-up persona called "Bob" (see Appendix A), the LLM's one-shot performance may degrade.

Sub-hypothesis 1.2 (Token Bias on Linda): The performance of a reasoning LLM should remain consistent if we replace the persona "Linda" in the one-shot examplar.

Assume one-shot ICL scenarios. P has the original Linda Problem as the one-shot exemplar, while P' rephrases the exemplar to a persona called "Bob".

$\boldsymbol{H_0}$: $\pi_{12} = \pi_{21}$.

$\boldsymbol{H_a}$: $\pi_{12} > \pi_{21}$. ($\pi_{12} < \pi_{21}$ is invalid.)

3.3. Token Perturbation via Additional Hints

Just as a proficient student doesn't need hints to excel in a math exam, a truly rational LLM should solve logical problems effectively without explicit cues. Besides, even if a student answers all problems correctly but the examlet provides all the reasoning steps, we may still question whether the student really understands the reasoning.

We evaluate the LLM's dependence on hint tokens that go beyond a single exemplar, such as the phrase "conjunction fallacy" or manually crafted chain-of-thought instructions (Wei et al., 2022) that demonstrate correct reasoning, as shown in Appendix A. If injecting these hints into prompts results in additional performance gains, it implies that that the model's reasoning may be superficial, relying familiar tokens or adhering to language patterns from the instructions.

Hypothesis 2 Genuine Reasoning LLMs do not rely on hint tokens to derive correct inferences.

Sub-hypothesis 2.1 (Leaking Hint Tokens): A reasoning LLM does not rely on the prompt to tell them that the reasoning task involves "conjunction fallacy."

Assume one-shot ICL scenarios. P' explicitly points out the "conjunction fallacy" or includes detailed guidance on how to reason in its given prompts, while P does not.

$\boldsymbol{H_0}$: $\pi_{12} = \pi_{21}$.

$\boldsymbol{H_a}$: $\pi_{12} < \pi_{21}$. ($\pi_{12} > \pi_{21}$ is invalid.)

---

4. Experiments

We experiment with a variety of the state-of-the-art LLMs, including OpenAI gpt-4-turbo (Achiam et al., 2023), Meta llama-3-70b-instruct (Touvron et al., 2023) and Anthropic claude-3-opus-20240229 (Anthropic, 2024).

We evaluate the performance of each LLM using appropriate prompting methods with synthetic data of sample size $n = 100$. For each sub-hypothesis, we conduct a McNemar test for every (model, prompting-method) pair and apply the Benjamini-Hochberg procedure with a fixed α of 0.05 to correct for multiple testing. For Sub-Hypothesis 1.1, we consistently reject the null hypothesis with predictable performance shifts under irrelevant token perturbations. Under Sub-Hypothesis 1.2, all tests lead to the rejection of the null, except for GPT-4 using the "os_cot" prompting method, suggesting models still have a strong tendency to rely on specific tokens like "linda" rather than the genuine reasoning. For Sub-Hypothesis 2.1, we compare the performance of one-shot ICL with and without additional hint tokens to evaluate the marginal benefits hints can offer. We reject the null for GPT-4 and Claude, showing that although an LLM can achieve almost perfect scores, its performance will be

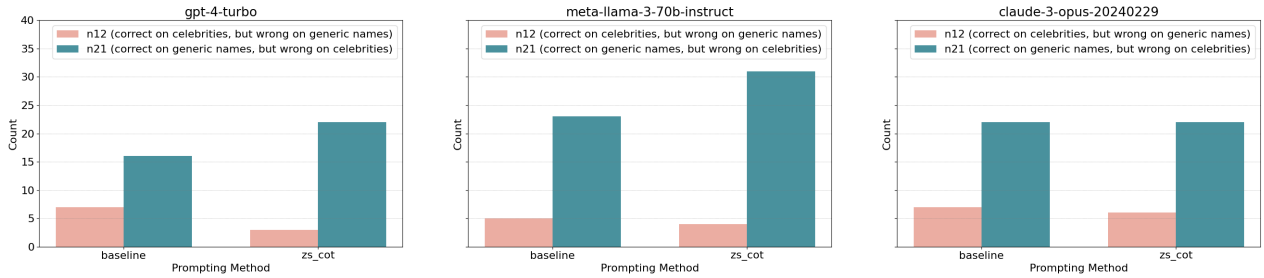

Figure 2: Experimental results for Sub-Hypothesis 1.1.

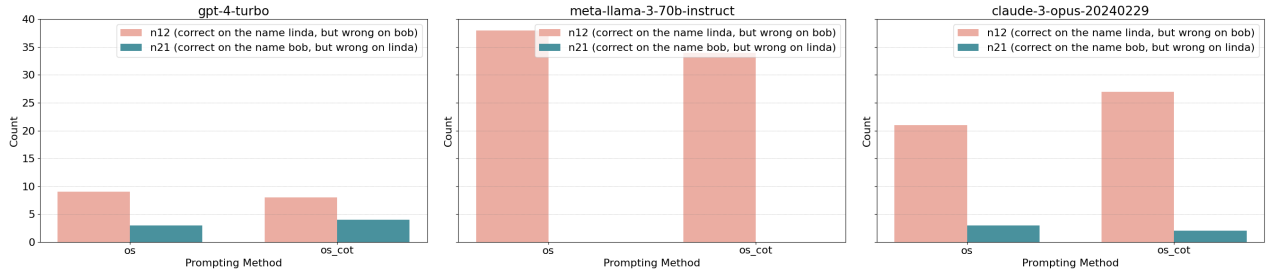

Figure 3: Experimental results for Sub-Hypothesis 1.2.

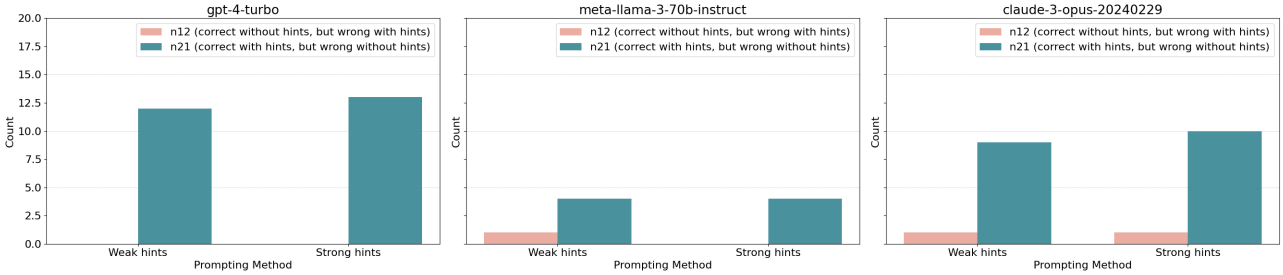

Figure 4: Experimental results for Sub-Hypothesis 2.1.

Figure 5: Experimental results. Our controlled experiments cast doubt on the capability of LLMs to function as rational thinkers, as we reject most of the null hypotheses. We implement different prompting techniques: "baseline" asks the model to answer directly, "zs" and "os" refer to zero-shot and one-shot prompting, "cot" means chain-of-thought ("*think step-by-step*"), "weak " and "strong" hints are detailed in Appendix B.

significantly impacted if hints become unavailable. Detailed testing results are included in Appendix C.

5. Discussion and Future Work

The statistical evidence presented in this paper contributes to the larger discussion that LLMs do not apply reasoning consistently in their decision-making processes. Instead, they primarily rely on token bias for response generation. This suggests that chain-of-thought prompting (CoT) (Wei et al., 2022; Wang et al., 2022) or in-context learning (ICL) (Brown et al., 2020; Min et al., 2022) may not elicit actual reasoning but instead result in semantic shortcuts for LLMs to imitate desired behavior. In fact, earlier work on CoT prompting and ICL found that even with invalid demonstration exemplars, these prompts can improve LLMs' performance on some tasks (Lyu et al., 2022; Wang et al., 2022). These findings raise questions about the extent to which LLMs truly engage in reasoning when responding to prompts, and further investigations are needed to uncover the underlying mechanisms and limitations of LLMs' reasoning capabilities.

This preliminary work reconceptualizes the evaluation of the reasoning behavior of LLMs. It combines controlled experiments with statistical hypothesis testing to complement traditional benchmarking methods. The proposed framework is general and can be adapted to many logical reasoning tasks beyond the scope of this study. In future work, we aim to expand this study by increasing the diversity of the synthetic data and LLMs being tested. Additionally, we intend to explore a wider range of logical fallacies, mathematical problems, and set-based reasoning tasks. By broadening the scope, we aim to uncover and characterize other interesting token biases that may exist in language models.

6. Limitations

This hypothesis testing framework is specifically designed for multiple choice questions and is not applicable to open-ended responses. It relies on LLMs with strong instruction-following capabilities to consistently produce responses that include either (a) or (b), but we find that LLMs can generally follow these instructions in most cases. Moreover, we acknowledge that there are likely other hypotheses and assumptions that a genuine reasoner should satisfy. Our current study focuses solely on the conjunction fallacy, i.e., the Linda problem and its variants, using three commercial LLMs to demonstrate our framework. This is merely a quintessential example, and we plan to expand our scope in the near future to include a broader range of hypotheses, LLMs, and reasoning tasks

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agresti, A. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.

Anthropic. Claude-3-opus-20240229. Software available from Anthropic, 2024. URL https://docs.anthropic.com/en/docs/models-overview. Accessed: 2024-05-20.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

Hastie, R. and Dawes, R. M. *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications, 2009.

Jamali, M., Williams, Z. M., and Cai, J. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain. *arXiv preprint arXiv:2309.01660*, 2023.

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

Kosinski, M. Evaluating large language models in theory of mind tasks. *arXiv e-prints*, pp. arXiv–2302, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, B., Zhou, B., Wang, F., Fu, X., Roth, D., and Chen, M. Deceiving semantic shortcuts on reasoning chains: How far can models go without hallucination? *arXiv preprint arXiv:2311.09702*, 2023.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*, 2023.

Lyu, X., Min, S., Beltagy, I., Zettlemoyer, L., and Hajishirzi, H. Z-icl: zero-shot in-context learning with pseudo-demonstrations. *arXiv preprint arXiv:2212.09865*, 2022.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Merrill, W. and Sabharwal, A. The expresssive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*, 2023.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Mosteller, F. Some statistical problems in measuring the subjective response to drugs. *Biometrics*, 8(3):220–226, 1952.

Mukherjee, A. and Chang, H. H. Heuristic reasoning in ai: Instrumental use and mimetic absorption. *arXiv preprint arXiv:2403.09404*, 2024.

Prasad, A., Koller, A., Hartmann, M., Clark, P., Sabharwal, A., Bansal, M., and Khot, T. Adapt: As-needed decomposition and planning with language models. *arXiv preprint arXiv:2311.05772*, 2023.

Ravaut, M., Ding, B., Jiao, F., Chen, H., Li, X., Zhao, R., Qin, C., Xiong, C., and Joty, S. How much are llms contaminated? a comprehensive survey and the llmsanitize library. *arXiv preprint arXiv:2404.00699*, 2024.

Rosenberg, J. Times man of the year list, 2021. URL https://www.thoughtco.com/times-man-of-the-year-list-1779824. Accessed: 05-05-2024.

Suri, G., Slater, L. R., Ziaee, A., and Nguyen, M. Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *Journal of Experimental Psychology: General*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tversky, A. and Kahneman, D. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293, 1983.

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., and Sun, H. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022.

Wang, P., Xiao, Z., Chen, H., and Oswald, F. L. Will the real linda please stand up... to large language models? examining the representativeness heuristic in llms. *arXiv preprint arXiv:2404.01461*, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wikipedia contributors. Forbes celebrity 100, 2024. URL https://en.wikipedia.org/wiki/Forbes_Celebrity_100. Accessed: 2024-05-20.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

A. The Original Linda Problem (Tversky & Kahneman, 1983)

The original Linda problem is framed as follows (Tversky & Kahneman, 1983):

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations. Which is more probable?
>
> 1. Linda is a bank teller.
> 2. Linda is a bank teller and is active in the feminist movement.

Here is an example of GPT-4o explaining the Linda Problem: https://chatgpt.com/share/eff10b9d-d219-4806-9cb9-d2d9104c0e83.

Our "Bob" version of this problem is as follows:

> Bob is 29 years old, deeply passionate about environmental conservation, and volunteers his weekends at local park clean-ups. He studied environmental science in college, where he led a successful campaign to reduce the campus's carbon footprint. Bob is also an avid cyclist and promotes sustainable living practices whenever possible. Based on this information, which is more possible?
>
> 1. Bob works for a renewable energy company and is an active member of a local environmental advocacy group.
> 2. Bob works for a renewable energy company.

The original form of the problem about Taylor Swift discussed in Section 1 was also introduced in Tversky & Kahneman (1983):

> Suppose Bjorn Borg reaches the Wimbledon finals in 1981. Please rank order the following outcomes from most to least likely.
>
> 1. Borg will win the match
> 2. Borg will lose the first set
> 3. Borg will lose the first set but win the match
> 4. Borg will win the first set but lose the match

B. Prompts in Hypothesis 2

This section includes the detailed prompts we use to evaluate the influences from weak and strong hints. We also include the original Linda Problem as the one-shot in-context learning exemplar before the following prompts, which is not shown here.

Weak Hint Your task is to answer the following question by explicitly selecting either option (a), (b), etc. Please aware that this is a Linda Problem designed to explore the concept of the conjunction fallacy. Here is the question and let's think step by step.

Strong Hint Your task is to answer the following question by explicitly selecting either option (a), (b), etc. Please aware that this is a Linda Problem designed to explore the concept of the conjunction fallacy. The conjunction fallacy occurs when individuals incorrectly judge the conjunction of two events as more probable than one of the events alone. For instance, many might believe that Linda, who is described as a bright, single woman deeply concerned with discrimination and social justice, is more likely to be both a bank teller and active in the feminist movement than just a bank teller. This judgment violates the basic probability rule: the probability of a conjunction, P(A and B), is always less than or equal to the probabilities of its constituents, P(A) or P(B). This error often stems from the representativeness heuristic, where people estimate the likelihood of an event by how closely it matches their mental prototype. To correctly solve problems like this, you must adopt probabilistic thinking: abstract the problem from its narrative context and focus solely on the probabilistic models. Ignore all extraneous background information and consistently choose the option involving a single event as it statistically holds a higher likelihood than the conjunction of multiple events. Here is the question and let's think step by step.

C. Hypothesis Testing Results

Table 2: Hypothesis Testing Outputs for Sub-Hypothesis 1.1

model	prompting method	n_{12}	n_{21}	n	raw p-value	adjusted p-value	reject
gpt-4-turbo	baseline	7	16	23	0.030284	0.030284	True
gpt-4-turbo	zs-cot	3	22	25	0.000072	0.000217	True
meta-llama-3-70b-instruct	baseline	5	23	28	0.000335	0.000670	True
meta-llama-3-70b-instruct	zs-cot	4	31	35	0.000003	0.000015	True
claude-3-opus-20240229	baseline	7	22	29	0.002673	0.003207	True
claude-3-opus-20240229	zs-cot	6	22	28	0.001248	0.001873	True

Table 3: Hypothesis Testing Outputs for Sub-Hypothesis 1.2

model	prompting method	n_{12}	n_{21}	n	raw p-value	adjusted p-value	reject
gpt-4-turbo	os	9	3	12	0.041632	0.049959	True
gpt-4-turbo	os-cot	8	4	12	0.124107	0.124107	False
meta-llama-3-70b-instruct	os	38	0	38	0.000000	0.000000	True
meta-llama-3-70b-instruct	os-cot	34	0	34	0.000000	0.000000	True
claude-3-opus-20240229	os	21	3	24	0.000119	0.000179	True
claude-3-opus-20240229	os-cot	27	2	29	0.000002	0.000003	True

Table 4: Hypothesis Testing Outputs for Sub-Hypothesis 2.1

model	prompting method	n_{12}	n_{21}	n	raw p-value	adjusted p-value	reject
gpt-4-turbo	weak-hint-os-cot	0	12	12	0.000266	0.000798	True
gpt-4-turbo	strong-hint-os-cot	0	13	13	0.000156	0.000798	True
meta-llama-3-70b-instruct	weak-hint-os-cot	1	4	5	0.371100	0.371100	False
meta-llama-3-70b-instruct	strong-hint-os-cot	0	4	4	0.133600	0.160320	False
claude-3-opus-20240229	weak-hint-os-cot	1	9	10	0.005706	0.008559	True
claude-3-opus-20240229	strong-hint-os-cot	1	10	11	0.003328	0.006656	True