Deduction under Perturbed Evidence: Probing Student Simulation Capabilities of Large Language Models[*]

Shashank Sonkar[1,*], Richard G. Baraniuk[1]

[1]*Rice University, Houston, Texas, USA*

Abstract

We explore whether Large Language Models (LLMs) are capable of logical reasoning with distorted facts, which we call Deduction under Perturbed Evidence (DUPE). DUPE presents a unique challenge to LLMs since they typically rely on their parameters, which encode mostly accurate information, to reason and make inferences. However, in DUPE, LLMs must reason over manipulated or falsified evidence present in their prompts, which can result in false conclusions that are valid only under the manipulated evidence. Our goal with DUPE is to determine whether LLMs can arrive at these false conclusions and identify whether the dominant factor influencing the deduction process is the encoded data in the parameters or the manipulated evidence in the prompts. To evaluate the DUPE capabilities of LLMs, we create a DUPEd version of the StrategyQA dataset, where facts are manipulated to reverse the answer to the question. Our findings show that even the most advanced GPT models struggle to reason on manipulated facts – showcasing poor DUPE skills – with accuracy dropping by 45% compared to the original dataset. We also investigate prompt settings inspired from student simulation models, which mitigate the accuracy drop to some extent. Our findings have practical implications for understanding the performance of LLMs in real-world applications such as student simulation models that involve reasoning over inaccurate information.

Keywords
Large Language Models, Reasoning, GPT, Student Simulation Models, Knowledge Tracing

1. Introduction

Over the last several years, Transformer models have played a significant role in shaping the field of Natural Language Processing (NLP) [1, 2, 3, 4, 5, 6]. Their exceptional ability to reason across a broad range of NLP tasks [7, 8, 9] has been a key factor contributing to their success. The success of LLMs on challenging datasets like HellaSwag [10], AI2 Reasoning Challenge (ARC) [11], WinoGrande [12], and GSM-8K [13] is a testament to their advanced reasoning skills and their potential to address challenging NLP tasks.

In this paper, we investigate the reasoning abilities of LLMs models under a novel paradigm we dub Deduction under Perturbed Evidence (DUPE for short). By testing LLMs' capacity to

---

[*]You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.
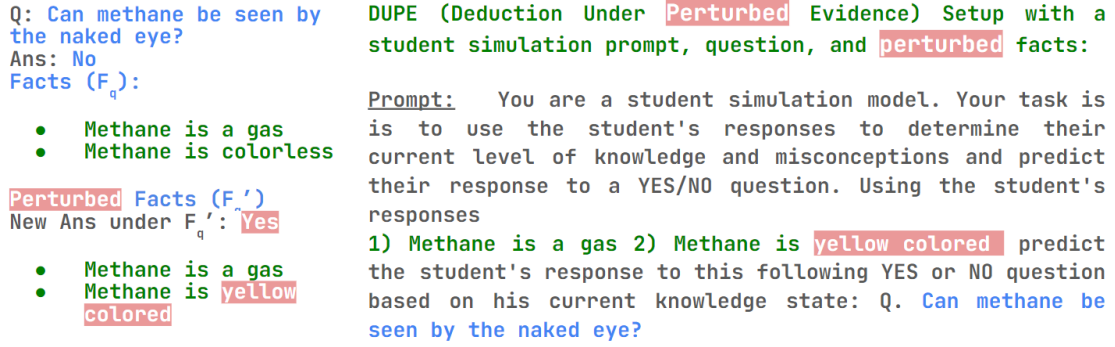
[*]Corresponding author.

✉ ss164@rice.edu (S. Sonkar); richb@rice.edu (R. G. Baraniuk)

CEUR Workshop Proceedings (CEUR-WS.org)

```
Q: Can methane be seen by          DUPE (Deduction Under  Perturbed  Evidence) Setup with a
the naked eye?                     student simulation prompt, question, and  perturbed  facts:
Ans: No
Facts (F_q):                       Prompt:   You are a student simulation model. Your task is
                                   is to use the student's responses to determine their
    • Methane is a gas             current level of knowledge and misconceptions and predict
    • Methane is colorless         their response to a YES/NO question. Using the student's
                                   responses
 Perturbed  Facts (F_q')            1) Methane is a gas 2) Methane is  yellow colored  predict
New Ans under F_q':  Yes            the student's response to this following YES or NO question
                                   based on his current knowledge state: Q. Can methane be
    • Methane is a gas             seen by the naked eye?
    • Methane is  yellow
      colored
```

Figure 1: Setup of the Deduction under Perturbed Evidence (DUPE) reasoning framework. On the left is a question-fact pair in StrategyQA dataset. To test DUPE skills of a model, we change facts provided with each question such that the response to the question flips. On the right is a prompting setup to probe DUPE skills of LLMs. We use a custom prompt tailored to student simulation setting that takes in the input question, perturbed (DUPEd) facts, and requests a *yes/no* response from LLMs. Perturbed facts represent a realistic student simulation setting since they mirror the inaccurate nature/ misconceptions of students' responses.

reason with flawed or perturbed evidence, we aim to determine whether LLMs can generate logically sound yet erroneous conclusions when presented with misleading information. Strong DUPE skills are critical in NLP applications like student simulations [14, 15], where models simulate student responses to understand how they may respond in certain scenarios. As student responses often contain inaccuracies and misconceptions, it is important for a model to analyze and utilize these inaccuracies and misconceptions as evidence to arrive at the same conclusion as the student. For instance, a student may have the misconception that the heavier an object is, the faster it falls, leading them to conclude that a bowling ball will fall faster than a ball bearing. If we provide LLMs with evidence that a heavier object falls faster, would LLMs also arrive at the conclusion that a bowling ball will fall faster than a ball bearing? We introduce DUPE as our approach to investigate this question.

Contributions: This paper develops a novel reasoning paradigm – Deduction under Perturbed Evidence (DUPE) – to examine whether LLMs arrive at different conclusions when presented with distorted initial facts. To test the DUPE capabilities of LLMs, we create a DUPEd version of StrategyQA dataset (Figures 1, 2). StrategyQA [16] is an open-domain QA dataset that is characterized by its explicit provision of the necessary facts required to answer each *yes-no* question. In the DUPEd version of the dataset, we manipulate the facts provided in a way that results in a different answer to the original question.

Our findings reveal that state-of-the-art LLMs, , including GPT3.5 and GPT4, struggle significantly on the newly introduced DUPEd-StrategyQA dataset. The accuracy of these models dropped drastically by approximately 45%, falling from an impressive 91.9% on the original dataset to only 46.7% on the DUPEd-StrategyQA dataset. In addition, we conduct an ablation study on the DUPEd-StrategyQA dataset by categorizing it into two distinct parts based on the type of manipulation used – one involving language perturbations and the other involving mathematical manipulations. Furthermore, our results demonstrate that the accuracy drop

```
Q1 (NLP - Bio): Is            Q2 (NLP - History): Did Karl Marx    Q3 (NLP - Bio): Can you cure
Cholera alive?                influence the communist party of     hepatitis with a tonsillectomy?
                              China?
Facts                                                              Facts
a) Cholera are a type of      Facts                                a) A tonsillectomy removes the
bacteria                      a) Communist ideology is foundation  tonsils, glands found in the back
b) Bacteria are               of communist party of China          of the throat
considered living             b) Marx produced a political pamphlet b) Hepatitis is a disease that
creatures                     known as communist manifesto         targets liver

Perturbed Facts               Perturbed Facts                      Perturbed Facts
a) Cholera are a type         a) Communist ideology is foundation   a) A tonsillectomy removes the
of bacteria                   of communist party of China          tonsils, glands found in the back
b) Bacteria are               b) Marx produced a political pamphlet of the throat
considered dead               known as capitalist manifesto        b) Hepatitis is a disease that
creatures                                                          targets tonsils


Q4 (Math - Chem) : Is         Q5 (Math - Physics): Would a pear     Q6 (Math+NLP - History): Was Hundred
nickel a better payout        sink in water?                       Years' War a misnomer?
than mercury if given a
dollar per atomic number?     Facts                                Facts
                              a) Density of pear is 0.59 g/cm^3     a) A misnomer is an inaccurate name
Facts                         b) Density of water is 1 g/cm^3       b) Hundred Years' War lasted for
a) Nickel has atomic          c) Objects sink if they are          longer than one hundred years
number of 28                  denser than the surrounding          c) Hundred Years' War lasted from
b) Mercury has atomic         fluid.                               1337-1453
number of 80
                              Perturbed Facts                      Perturbed Facts
Perturbed Facts               a) Density of pear is 2  g/cm^3       a) A misnomer is an inaccurate name
a) Nickel has atomic          b) Density of water is 1 g/cm^3       b) Hundred Years' War lasted for
number of 92                  c) Objects sink if they are          longer than one hundred years
b) Mercury has a atomic       denser than the surrounding          c) Hundred Years' War lasted from
number of 80                  fluid.                               1337-1437
```

Figure 2: Six examples from our DUPEd-StrategyQA dataset. We flip the answer to a *yes-no* question by altering facts provided with each question. First three questions on the top are examples of natural language perturbations, while the bottom three questions involves manipulating numerical digits. The DUPEd version was designed with minimal modifications to the facts, usually involving only one to two word changes in the original facts. Additionally, we refrained from using explicit negation words like *not*.

can be mitigated by using prompt settings inspired by student simulation models. This approach reduced the accuracy drop to 29%, with the models achieving an accuracy of 62.7% on the DUPEd-StrategyQA dataset. Our findings carry crucial implications for practical LLMs applications, particularly in the realm of student simulation models that demand reasoning over erroneous information.

2. Methodology, Dataset, and Prompting

In this section, we overview the DUPE reasoning framework, provide details on the DUPEd version of AllenAI's StrategyQA dataset, and then explore customized prompt settings designed to assess the DUPE skills of LLMs.

2.1. DUPE

Given a *true-false* question q, the correct response $r_q \in \{true, false\}$ and facts F_q that determine the truth or falsehood of Q (r_q), we change F_q to F_q' s.t. the correct response to q flips to $\neg r_q$

Table 1

We evaluate the DUPE capabilities of the two largest GPT models under two different prompt settings using the DUPEd-StrategyQA dataset. Prompt P1 asks GPT models to answer a question based on provided evidence. Under Prompt P1 setting, both GPT3.5 and GPT4 perform poorly on DUPEd version of the dataset with around 45% accuracy drop. We also find that both models are more robust to mathematical perturbation compared to natural language perturbations. Prompt P2 is inspired from student simulation settings. P2 primes the models that evidence provided may be incorrect. We find that prompt P2 achieves better accuracy than Prompt P1 by 16.0 points for GPT4, but we still see a substantial 29.2% drop in accuracy compared to GPT4's accuracy on original dataset.

Dataset	Model	Prompt	Accuracy (Overall)	Accuracy (NLP)	Accuracy (Math)
StrategyQA	GPT3.5	P1	84.6	94.1	74.4
DUPEd-StrategyQA	GPT3.5	P1	38.6 (46.0↓)	35.4 (58.7↓)	42.0 (32.4↓)
StrategyQA	GPT4	P1	91.9	94.1	89.4
DUPEd-StrategyQA	GPT4	P1	46.7 (45.2↓)	43.8 (50.3↓)	50.0 (39.4↓)
DUPEd-StrategyQA	GPT4	P2	62.7 (29.2↓)	63.1 (31.0↓)	62.2 (27.2↓)

under altered facts F_q',

$$\mathrm{DUPE}\big((q, F_q, r)\big) = (q, F_q', r')$$
$$\text{s.t. } r' = \neg r, \ \mathrm{edit}_{\mathrm{dist}}(F_q, F_q') < \tau, \tag{1}$$

where $\mathrm{edit}_{\mathrm{dist}}$ ensures that the edit distance between the fact strings F_q and F_q' is less than a threshold τ. The threshold τ is generally set to two to three words to ensure minimal changes to underlying facts (examples in figure 2). The new DUPEd-tuple (q, F_q', r') can be used to probe the DUPE capabilities of LLMs as shown in Figure 1.

2.2. DUPEd-StrategyQA

We use AllenAI's StrategyQA dataset [16] to assess the DUPE skills of LLMs. StrategyQA dataset provides explicit facts for answering open-domain questions. We create a DUPEd version of StrategyQA dataset composed of a total of 325 examples, of which 173 introduce natural language perturbations, while the remainder introduce mathematical errors (refer to examples in figure 2).

While designing the DUPEd version, we were careful to modify the facts in the most minimal way possible As a result, we made a conscious effort to only alter one or two words in the original facts whenever possible, in order to preserve the overall meaning and context of the original question. Additionally, we refrained from using explicit negation, such as the word *not*, to modify the facts, since our intent is not to evaluate the reasoning proficiency of LLMs in handling negation.

2.3. Student Simulation and Prompt Design

DUPE is highly relevant to *student simulation models* [14, 17, 15], which are widely used in education and cognitive psychology research. These models help in predicting and understanding student responses to various tasks, and thus their ability to reason over false information is critical to their success. Given this strong connection between simulation models and DUPE,

these models can inspire innovative approaches to prompt design, which can be used to probe DUPE skills of LLMs [8, 18]. An example of such a prompt is illustrated in figure 1 and section 3.

DUPE and Counterfactual Reasoning: Counterfactual reasoning and student simulation models require different types of reasoning. In counterfactual reasoning, the focus is on exploring hypothetical scenarios that may or may not correspond to actual reality. The fact that the information being considered is hypothetical or counterfactual is usually known beforehand.

In contrast, a student simulation model needs to reason about both true and false information, and may not know beforehand whether the information being considered is true or false. For example, in figure 2, the model lacks prior knowledge about which facts are true and which ones are perturbed. The model must identify incorrect answers from the student to make inferences about future questions, which requires robust and nuanced reasoning capabilities beyond those needed for counterfactual reasoning.

3. Experiments

We evaluate the DUPE capabilities of the two largest GPT models – GPT3.5 (version gpt-3.5-turbo-0301) and the latest GPT4 model (version gpt-4-0314) – via experiments under two different prompt settings, P1) "You are a question answering model. Your task is reason on provided evidence to answer a YES or NO question", and P2) "You are a student simulation model. Your task is reason on student's responses to accurately measure the student's current knowledge state and predict the student's response to a YES or NO question based on the student's current knowledge state" from section 2.3. An example is illustrated in Figure 1.

3.1. Main Results

In the prompt setting P1, both GPT3.5 and GPT4 performed poorly on the DUPEd version of the dataset, with a decrease in accuracy by 46.0%. and 45.2% respectively. As expected, the latest GPT4 model demonstrates superior performance to GPT3.5 on both the original and the DUPEd StrategyQA dataset.

3.1.1. Student Simulation Prompt

Prompt P2 inspired by student simulation setting informs/ primes the models that the provided evidence may be incorrect since the evidence reflects the erroneous nature of students' responses. We found that prompt setting P2 performs significantly better than P1 by a margin of 16.0% for the GPT4 model. However, there was still a significant 29.2% drop in accuracy compared to GPT4's performance on the original dataset.

3.1.2. Language vs. Math Perturbations

While curating the DUPEd-StrategyQA dataset, we divided the perturbations introduced into two distinct categories - one that involved language perturbations, while the other manipulated mathematical information (see figure 2). Our finding suggest that both GPT models are more resilient to math perturbations compared to language perturbations. E.g. for GPT3.5 there was

accuracy drop of 58.7% and 32.4 for language and math Perturbations respectively, while for GPT4 the accuracy drops were 50.3% and 39.4.

3.2. Root Cause of Poor DUPE Skills

To explain the GPT models' poor performance on the DUPEd dataset, we need to identify the main factor influencing their reasoning process, i.e., whether it is the encoded information in parameters or the manipulated evidence in prompts. Recent studies have shed light on this issue, suggesting that factual information encoded in the parameters of LLMs plays a dominant role in governing the generated output. For instance, the feed-forward layers in transformer models function as key-value memories, which implies that they encode factual information, as noted by Geva et al. [19]. Moreover, Meng et al. [20] demonstrated that localized computations, such as Rank-One Model Editing (ROME), can modify these factual associations, leading to alternative conclusions. These findings suggest that the encoded information in parameters has a significant impact on LLMs' reasoning process; further investigation is left for future work.

4. Conclusions

In this paper, we have introduced a new reasoning paradigm we call Deduction under Perturbed Evidence (DUPE for short). Through DUPE , we have assessed the ability of LLMs models to arrive at logically sound yet erroneous conclusions when faced with distorted initial facts. Our study, which used a carefully curated dataset to evaluate DUPE abilities, has revealed that even the most advanced GPT models struggle with logical reasoning in the presence of falsified information. Moving forward, we plan to investigate into the performance of different LLMs with our dataset in varied prompt settings.

5. Limitations

Due to limitations in both financial and computational resources, we had to limit our testing to only the most advanced LLMs – the GPT models. Consequently, we directed our attention towards developing a dataset for evaluating proposed reasoning scenarios. As a result of these limitations, we chose to focus specifically on the evaluation of the two largest models offered by OpenAI. While we recognize that other LLMs may produce different outcomes, we believe that our dataset could serve as a valuable resource for further research into the capabilities and limitations of LLMs .

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017).

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances In Neural Information Processing Systems 33 (2020) 1877–1901.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[6] OpenAI, GPT-4 technical report, 2023. arXiv:2303.08774.

[7] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, et al., Language models are multilingual chain-of-thought reasoners, arXiv preprint arXiv:2210.03057 (2022).

[8] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, O. Bousquet, Q. Le, E. Chi, Least-to-most prompting enables complex reasoning in Large Language Models, arXiv preprint arXiv:2205.10625 (2022).

[9] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al., Sparks of Artificial General Intelligence: Early experiments with GPT-4, arXiv preprint arXiv:2303.12712 (2023).

[10] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, Hellaswag: Can a machine really finish your sentence?, arXiv preprint arXiv:1905.07830 (2019).

[11] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved Question Answering? Try ARC, the AI2 reasoning challenge, arXiv preprint arXiv:1803.05457 (2018).

[12] K. Sakaguchi, R. L. Bras, C. Bhagavatula, Y. Choi, Winogrande: An adversarial winograd schema challenge at scale, Communications of the ACM 64 (2021) 99–106.

[13] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al., Training verifiers to solve math word problems, arXiv preprint arXiv:2110.14168 (2021).

[14] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, J. Sohl-Dickstein, Deep Knowledge Tracing, Advances in Neural Information Processing Systems 28 (2015).

[15] N. Liu, Z. Wang, R. Baraniuk, A. Lan, Open-ended knowledge tracing for computer science education, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 3849–3862.

[16] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, J. Berant, Did Aristotle use a laptop? A Question Answering benchmark with implicit reasoning strategies, Transactions of the Association for Computational Linguistics 9 (2021) 346–361.

[17] S. Sonkar, A. E. Waters, A. S. Lan, P. J. Grimaldi, R. G. Baraniuk, qdkt: Question-centric Deep Knowledge Tracing, arXiv preprint arXiv:2005.12442 (2020).

[18] M. Bommarito II, D. M. Katz, GPT takes the Bar Exam, arXiv preprint arXiv:2212.14402 (2022).

[19] M. Geva, R. Schuster, J. Berant, O. Levy, Transformer feed-forward layers are key-value memories, arXiv preprint arXiv:2012.14913 (2020).

[20] K. Meng, D. Bau, A. Andonian, Y. Belinkov, Locating and editing factual associations in

gpt, Advances in Neural Information Processing Systems 35 (2022) 17359–17372.