Evaluating the Robustness of rStar: A Novel Framework for Enhanced Reasoning in Small Language Models

Technical University of Denmark
Department of Applied and Computer Science, Mathematics
Kgs. Lyngby, Denmark

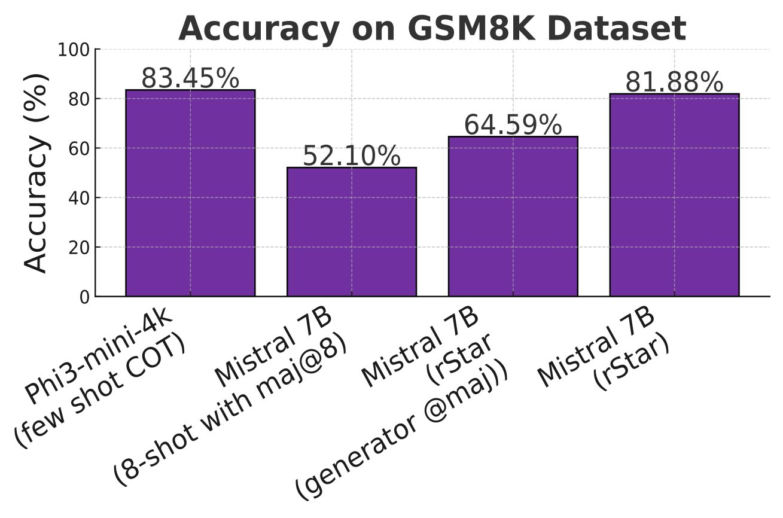
Jone Egon Steinhoff (s243867), Lukas Rasocha (s233498), Panagiota Emmanouilidi (s223531), Petr Boska Nylander (s240466) & Robert Spralja (s243658)

Introduction

- **Small language models** (SLMs) show strong reasoning abilities, but benchmarks like **Grade School Math 8K** (GSM8K) may overestimate their true reasoning capabilities [1]
- Recent methods **enhance reasoning** but **may neglect robustness** to diverse input variations [2]
- This study evaluates the **robustness of rStar** by evaluating its performance across diverse variations of inputs to identify its **strengths** and **limitations**, offering a more accurate assessment of its reasoning abilities in **mathematical problem-solving**.

GSM8K

Q: John buys purple pairs of shoes for each of his 3 children. They cost 60 each. How much did he pay?
A: "He bought $3 * 2 =<< 3 * 2 = 6 >>$ 6 pairs of shoes So he spent $60 * 6 = \$ << 60 * 6 = 360 >>$ 360
360"

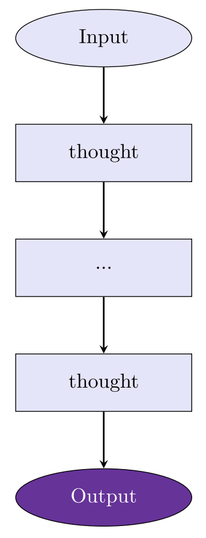


rStar Background

Prompting Language Models to Reason
- Chain of Thought (CoT)

Sampling Reasoning Paths
- Self-consistency

Answer Verification
- **@maj**
- Self verification
- **Mutual consistency**



Chain of Thought-Prompting (CoT)

Related Work

- GSM1k (newly crafter analogous to GSM8k – up to 8% drop in accuracy) [3]
- **Token bias** with logical problems [4]
    - **Token bias** is the idea that the models are biased to the specific tokens that appear in the GSM8k dataset (the names and numbers)
- PMC benchmark – Problems with missing and contradictory conditions [5]

References

[1] Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. arXiv. https://arxiv.org/abs/2410.05229

[2] Zhenting Qi and Mingyuan Ma and Jiahang Xu and Li Lyna Zhang and Fan Yang and Mao Yang (2024), Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solvers

[3] H. Zhang, J. Da, D. Lee, V. Robinson, C. Wu, W. Song, T. Zhao, P. Raja, C. Zhuang, D. Slack, Q. Lyu, S. Hendryx, R. Kaplan, M. Lunati, and S. Yue, "A Careful Examination of Large Language Model Performance on Grade School Arithmetic," *arXiv preprint arXiv:2405.00332*, 2024. [Online]. Available: https://arxiv.org/abs/2405.00332

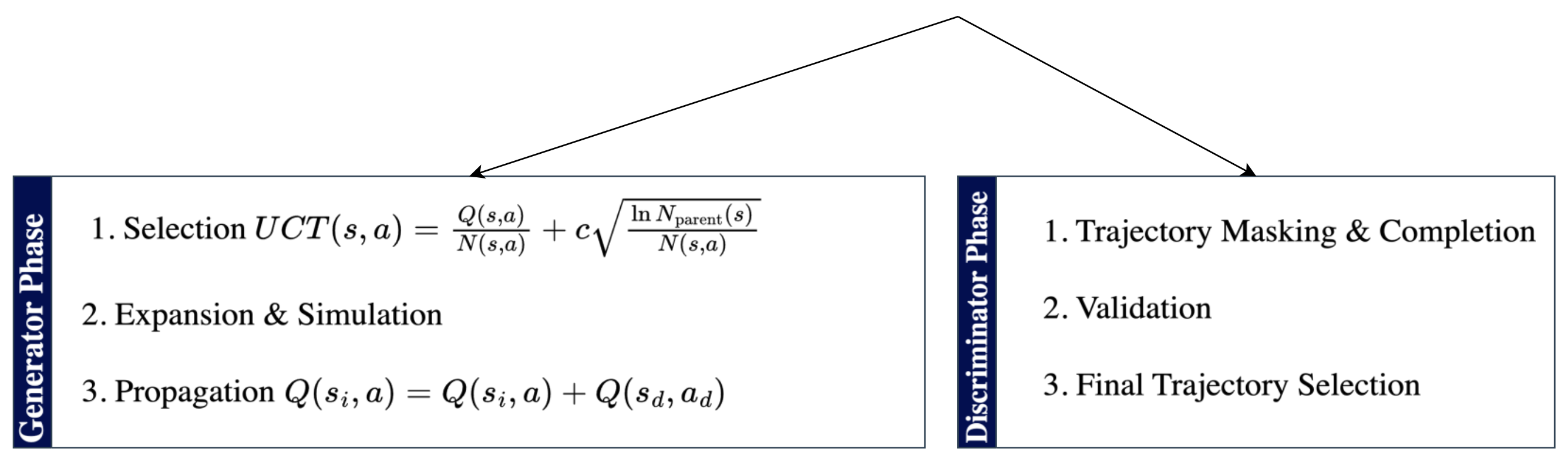
[4] B. Jiang, Y. Xie, Z. Hao, X. Wang, T. Mallick, W. J. Su, C. J. Taylor, and D. Roth, "A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA, Nov. 2024, pp. 4722–4756. Association for Computational Linguistics. [Online]. Available: https://aclanthology.org/2024.emnlp-main.272. doi: 10.18653/v1/2024.emnlp-main.272

[5] S.-Y. Tian, Z. Zhou, L.-H. Jia, L.-Z. Guo, and Y.-F. Li, "Robustness Assessment of Mathematical Reasoning in the Presence of Missing and Contradictory Conditions," *arXiv preprint arXiv:2406.05055*, 2024. [Online]. Available: https://arxiv.org/abs/2406.05055

ChatGPT was used to explain complicated topics and help with text refinement

Methodology

rStar



Generator Phase

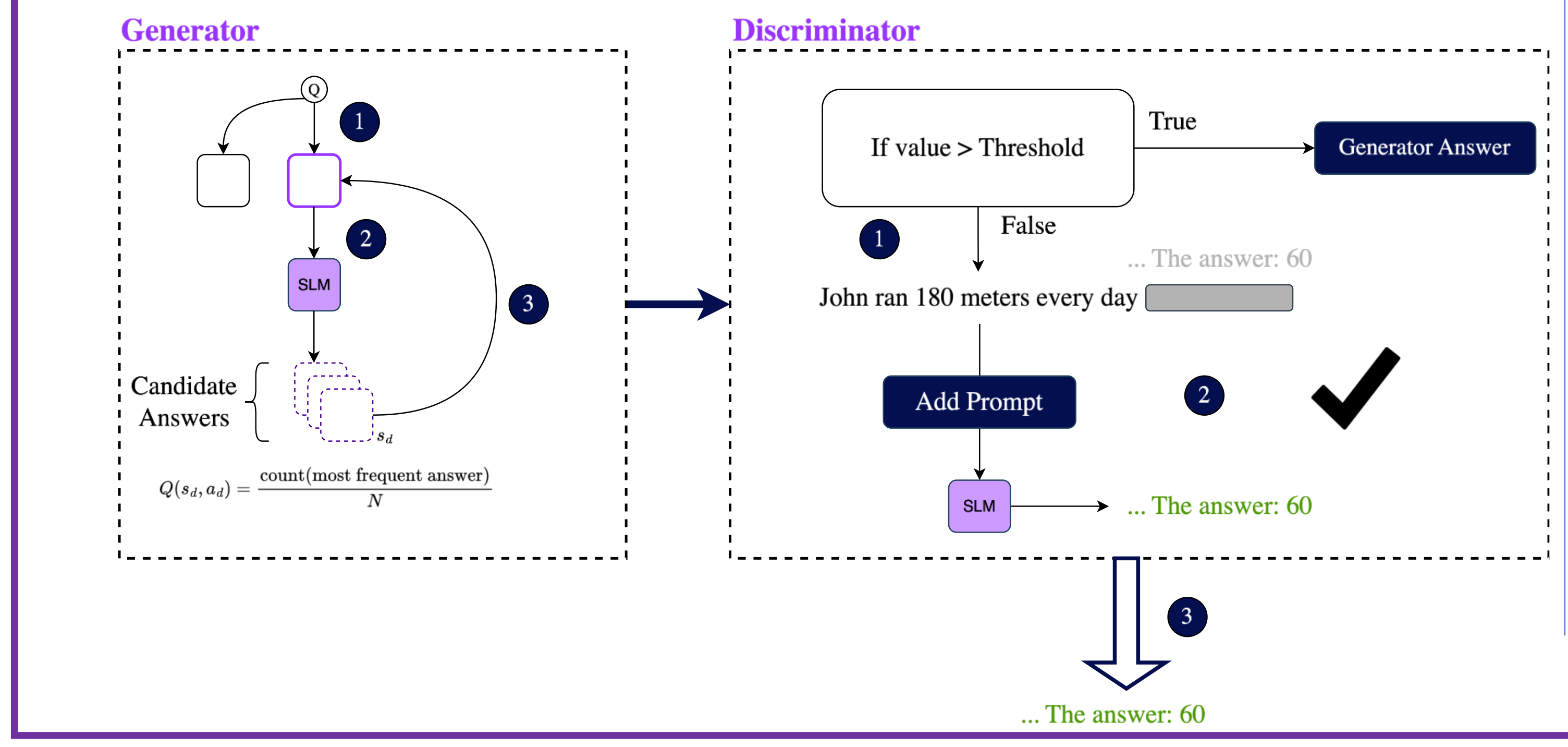
1. Selection $UCT(s,a) = \frac{Q(s,a)}{N(s,a)} + c\sqrt{\frac{\ln N_{\text{parent}}(s)}{N(s,a)}}$
2. Expansion & Simulation
3. Propagation $Q(s_i,a) = Q(s_i,a) + Q(s_d,a_d)$

Discriminator Phase

1. Trajectory Masking & Completion
2. Validation
3. Final Trajectory Selection

Actions:
- Propose a one-step thought (A1)
- Propose the remaining thought steps (A2)
- Propose next sub-question along with its answer (A3)
- Answer the sub-question again (A4)

Generator SLM: Mistral 7B-v0.1
Discriminator SLM: Phi-3-mini-4k-instruct



Variations

- Enhance GSM8K dataset by introducing 5 new variation types
- Each variation type is represented by generic templates
- Generate novel math questions derived from the original GSM8K dataset

Template (Names + Numbers)

A company sold {total_gallons} gallons of milk in jars to Mr. {name}'s store at the cost of $ {price_per_gallon} per gallon. However, Mr. {name} later realized {exp_frac_num} / {exp_frac_den} of the amount of milk he purchased had passed the expiry date and could not be sold. He returned the sour milk to the company and ordered a refund. Calculate how much he got in refunds.

#Variables:
- name = sample(names)
- total_gallons = range(1000, 5000)
- price_per_gallon = range(2, 5)
- exp_frac_num = range(1, 4)
- exp_frac_den = range(5, 8)

#Conditions:
- expired_gallons = range(1, 1000)
- refund_amount = range(1, 5000)

#Calculations:
expired_fraction = exp_frac_num / exp_frac_den
expired_gallons = expired_fraction × total_gallons
refund_amount = price_per_gallon × expired_gallons

#Solution: If Mr. {name} purchased {total_gallons} gallons, the number of expired gallons is: {exp_frac_num} / {exp_frac_den} × {total_gallons} =

<< {exp_frac_num} / {exp_frac_den} × {total_gallons} = {expired_gallons

{expired_gallons} gallons.
If each gallon costs $ {price_per_gallon}, and he returned {expired_gallons} gallons for a refund, the total refund amount is:

<< {price_per_gallon} × {expired_gallons} = {refund_amount} >>

Final Answer: $refund_amount

. . . .

- **Experiment 1: Only Names**
- **Experiment 2: Only Numbers**
- **Experiment 3: Mixed**

Template (NoOps + Irrelevant Informations)

A company sold 4000 gallons of milk in jars to Mr. Marcellus' store at the cost of $3.5 per gallon. However, Mr. Marcellus later realized 2/5 of the amount of milk he purchased had passed the expiry date and could not be sold. He returned the sour milk to the company and ordered a refund. {sentence} Calculate how much he got in refunds.

. . . .

- **Experiment 4: NoOps**

{sentence} = {"The milk was delivered in eco-friendly jars.", Mr. Marcellus' store is known for its strict quality control.", ..., }

- **Experiment 5: Irrelevant informations**

{sentence} = {"A marathon is 42.195 kilometers long.", "The Great Wall of China is over 21,000 kilometers long.", ..., }

Conclusion

- **Lower Performance**: All experiments, apart from changing names, resulted in a drop in rStar's performance beyond one standard deviation.
- **Lack of Robustness**: Experiments reveal that rStar is not robust to various input variations.
- **Largest Accuracy Drop**: Numeric variations caused the largest performance decline.
- **Combined Variations**: Combining names and numbers led to a much smaller drop in accuracy compared to numbers alone.
- **NoOps vs. Irrelevant Info**: NoOps caused a slightly larger accuracy drop than irrelevant information, suggesting rStar handles ignoring irrelevant information better than handling seemingly relevant information.

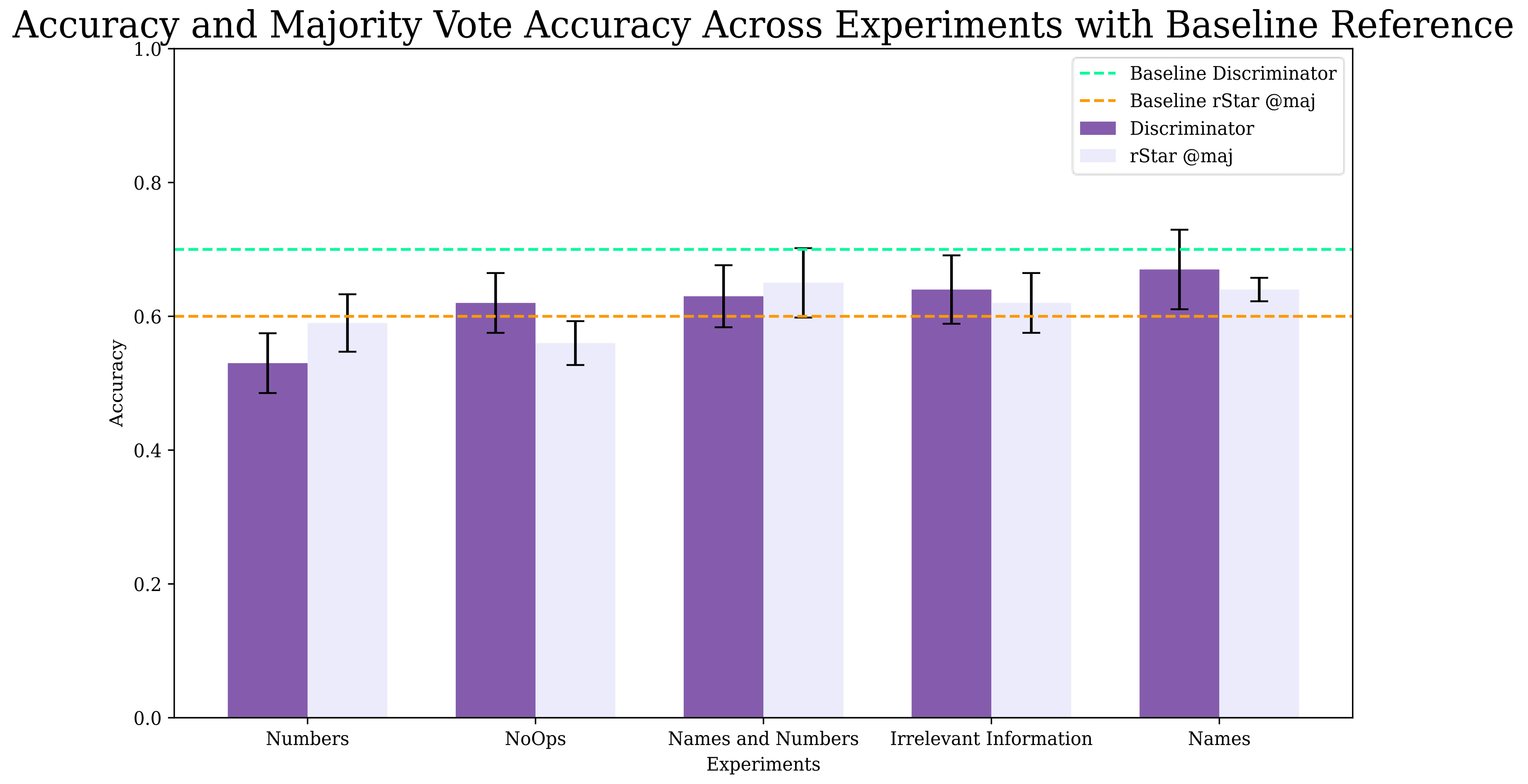
Future Work

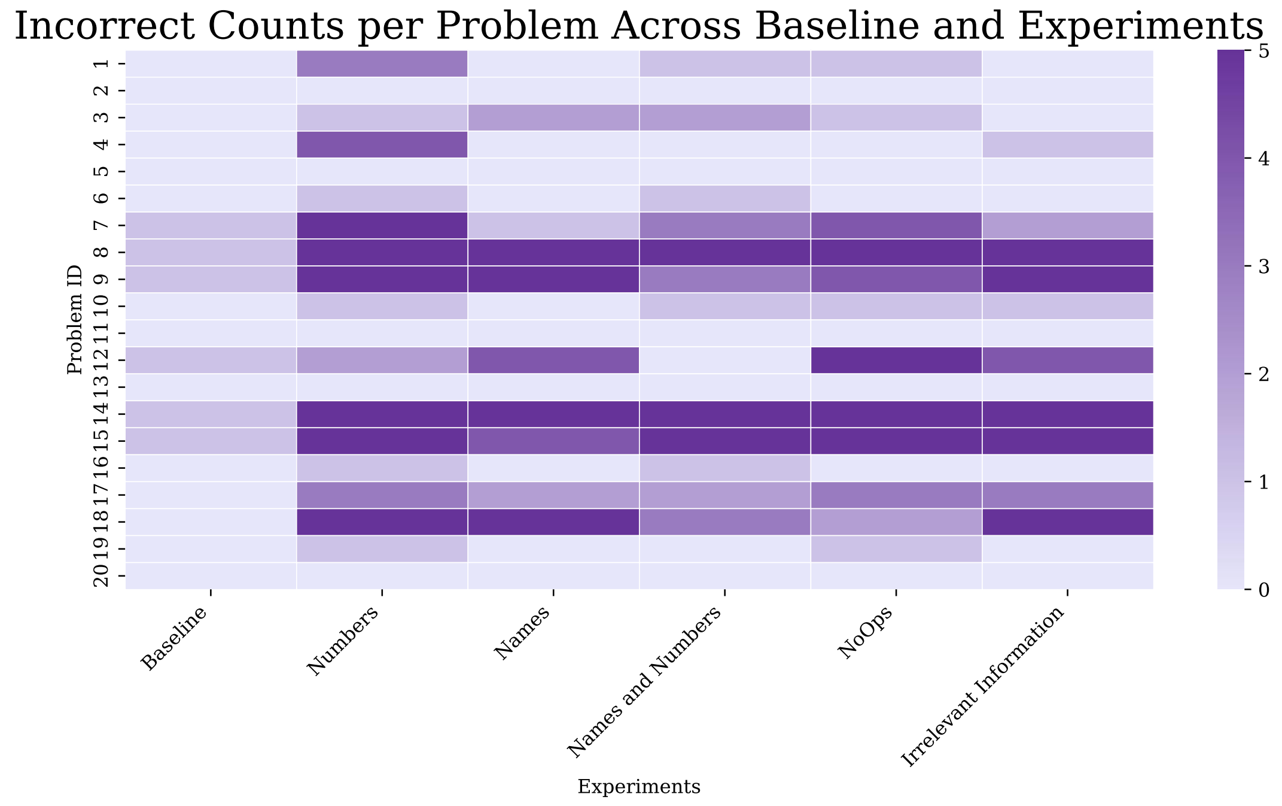
- More variations & bigger subset of questions
- Develop another action that is adapted to the newly introduced variances
- Run experiment more times to account for the inherent variance of language models

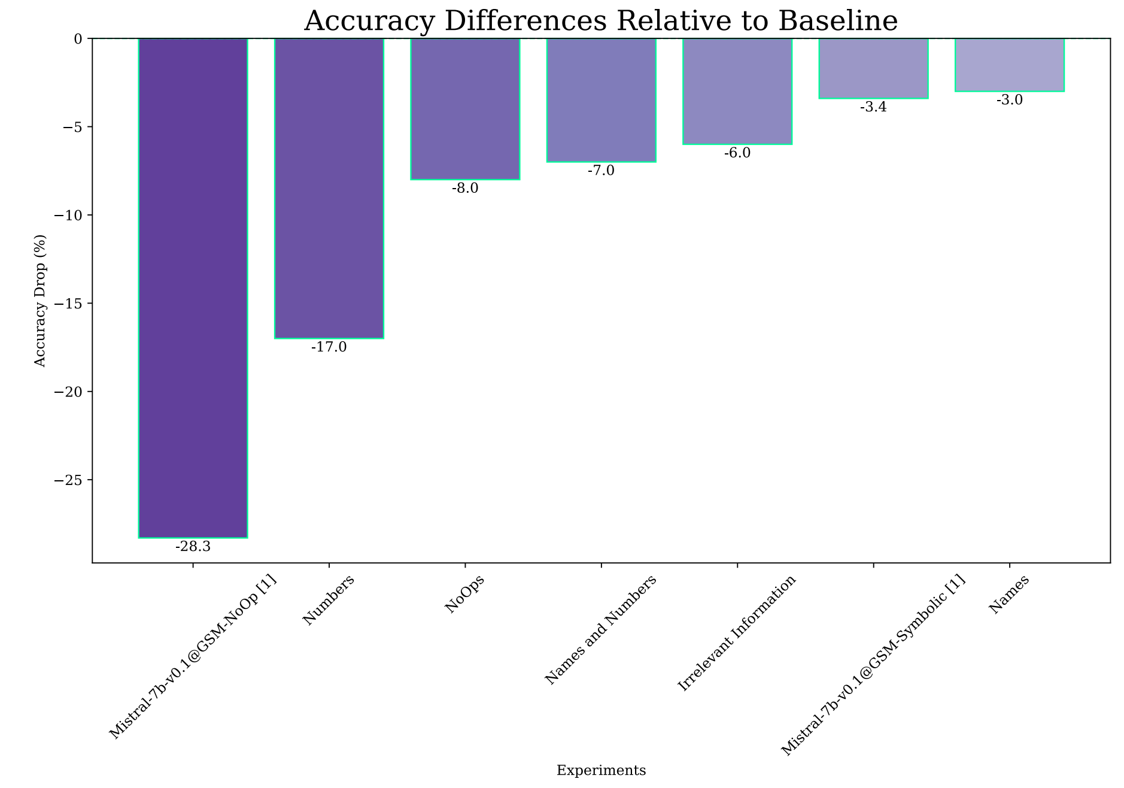
Limitations

- Generator on 40 questions took approx. 12-13 hours using a Tesla A100 PCIE 40 GB
- Discriminator on 40 questions took approx. 6 hours using a Tesla A100 PCIE 40 GB
- Queue time during the busiest periods was approx. 1-2 days

Results







Example of the solution with NoOps

- **Question:**
Michael is traveling on the interstate at an average speed of 50km/hr (taking into account all the necessary stops for gas, relaxation, etc.). If he goes on like this for 3 days, and the distance between Alaska and Texas is 6000 km, what percentage of this distance has he covered? Michael's car is equipped with a GPS system. What percentage of the distance has he covered?",
- **Correct answer: 60%**

. . . .

- **Question:**
What percentage of the distance has Michael covered with the GPS system? ",
- **Final answer:**
With the GPS system, he knows his exact position. So the distance covered equals the distance between his starting point and his current location.
The answer is 100%.

Example of the solution with irrelevant information

- **Question:**
There are about 7.8 billion people on Earth. Audrey has to take two math tests to pass 6 grade. They must correctly answer 70% of the total questions to move on to the next grade. The first test has 70 questions, and they get 60% of them correct. The second test has 40 questions. How many questions do they need to get right on the second test to pass?
- **Correct answer: 35**

. . . .

- **Final Answer:**
Let's think step by step. There are about 7.8 billion people on Earth. To pass 6th grade, one must correctly answer 70% of the total questions. So, to pass 6th grade, one needs to answer 70% of 7.8 billion which is 5.46 billion. The first test has 70 questions. One gets 60% of the first test which is 42 questions correct. Therefore, 5.46 billion / 42 is equal to 130 questions correct.
The answer is: 130.