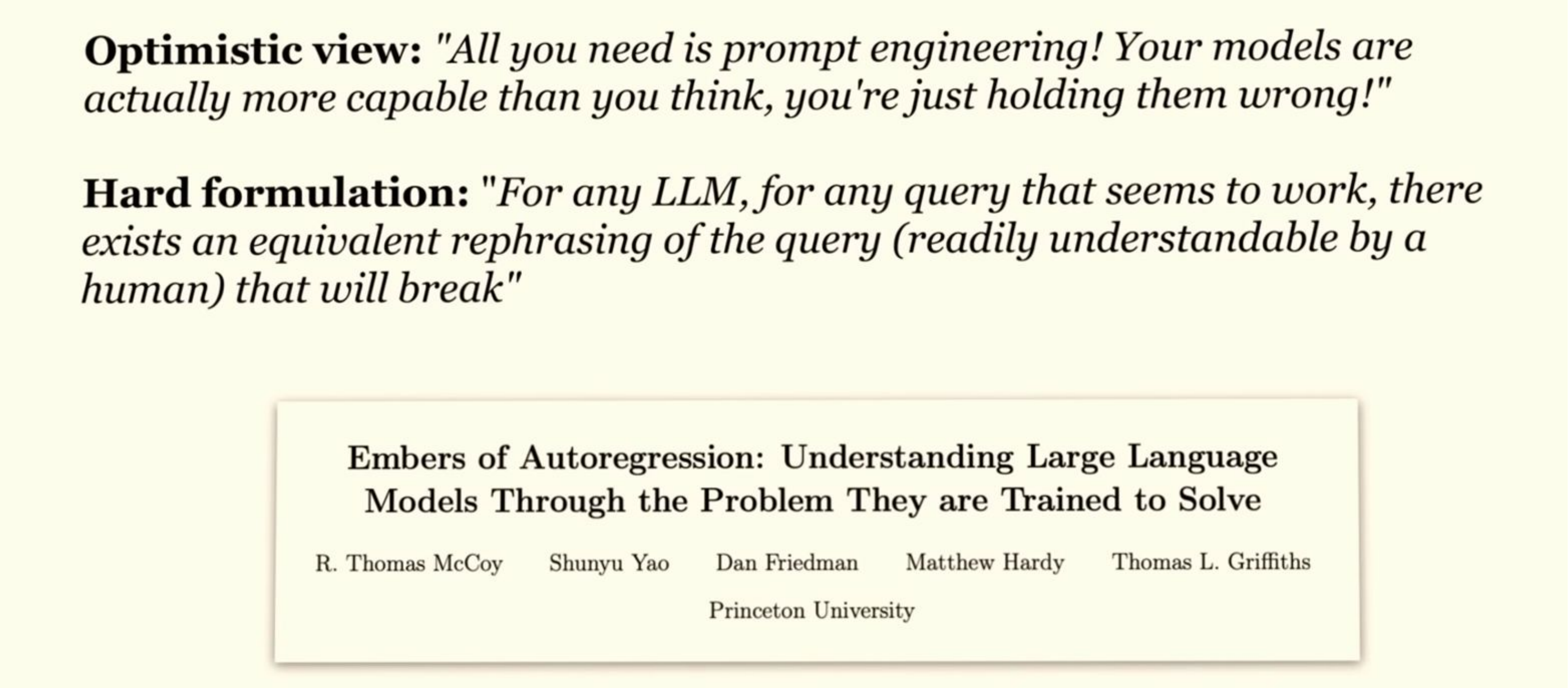
NoOp+: Paraphrase based dataset augmentation for reasoning robustness testing in LLMs

Luca Francis
Supervisor: Jan Philip Wahle

Math Retrieval Seminar / Final Presentation

Motivation



Optimistic view: *"All you need is prompt engineering! Your models are actually more capable than you think, you're just holding them wrong!"*

Hard formulation: *"For any LLM, for any query that seems to work, there exists an equivalent rephrasing of the query (readily understandable by a human) that will break"*

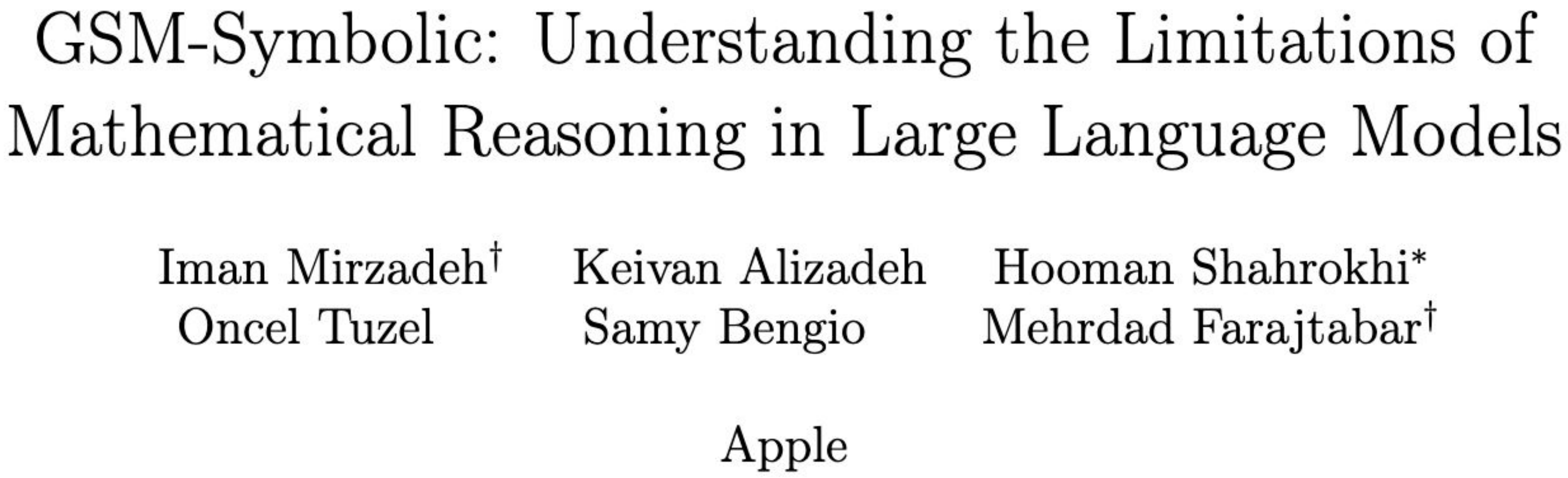
Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve

R. Thomas McCoy Shunyu Yao Dan Friedman Matthew Hardy Thomas L. Griffiths

Princeton University

François Chollet

Introduction: **GSM-NoOp**

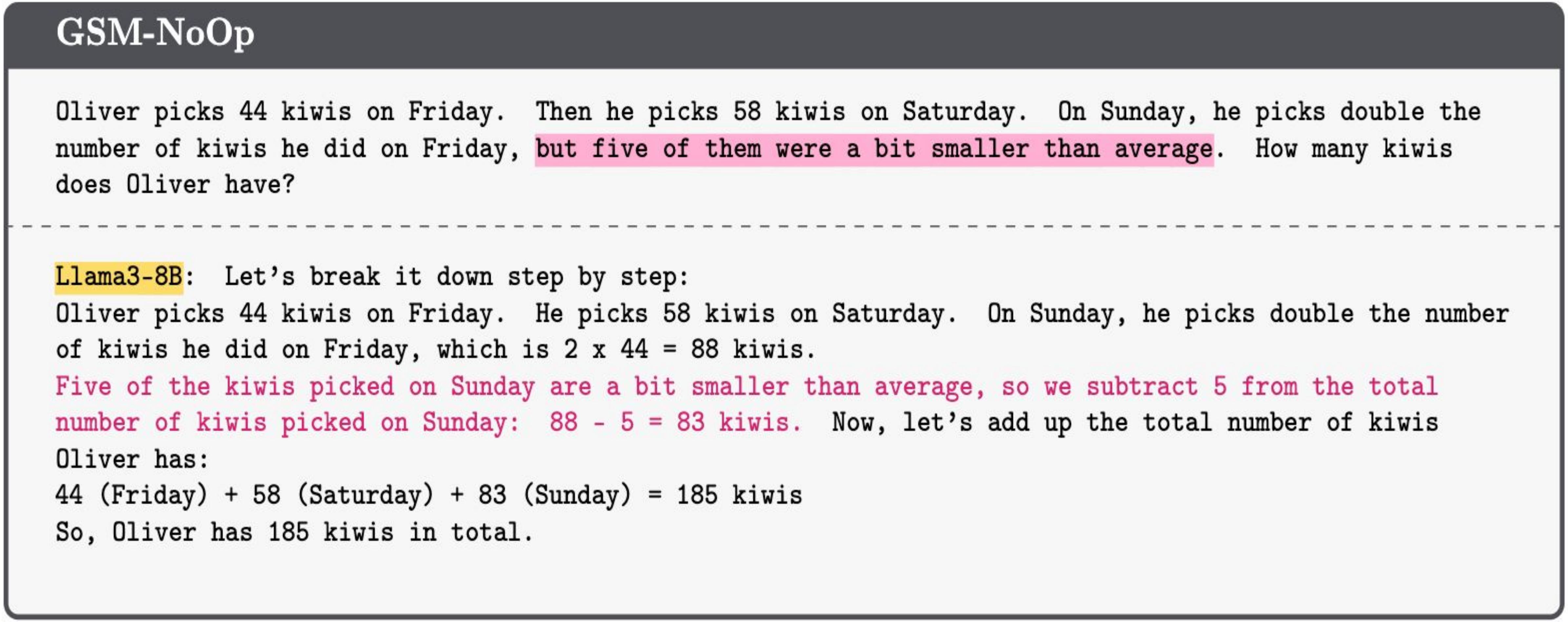


GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Iman Mirzadeh[†] Keivan Alizadeh Hooman Shahrokhi[*]
Oncel Tuzel Samy Bengio Mehrdad Farajtabar[†]

Apple

Introduction: **GSM-NoOp**



GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

Llama3-8B: Let's break it down step by step:

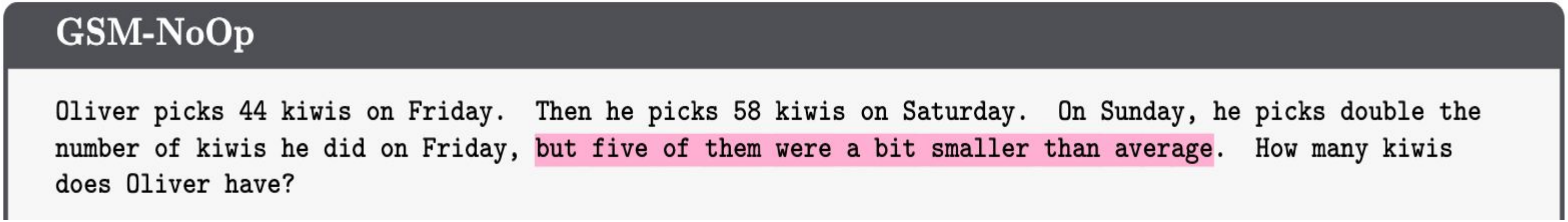
Oliver picks 44 kiwis on Friday. He picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, which is 2 x 44 = 88 kiwis.

Five of the kiwis picked on Sunday are a bit smaller than average, so we subtract 5 from the total number of kiwis picked on Sunday: 88 - 5 = 83 kiwis. Now, let's add up the total number of kiwis Oliver has:

44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis

So, Oliver has 185 kiwis in total.

Introduction: **GSM-NoOp**



GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

Paraphrase Type: Addition

Hypothesis:
This phenomenon can also occur with different types of paraphrasing and across different kind of domains.

Introduction: **Research Motivation**

Advancing LLM Reasoning Understanding:

- If similar performance drops occur with other paraphrase types and across diverse datasets, we gain deeper insight into LLM reasoning limitations. New benchmarks are essential to reliably study and address these weaknesses.

Enhancing Benchmark Robustness:

- High performance on a benchmark may not reflect true understanding if minor paraphrasing causes significant errors. Developing more rigorous benchmarks will help us assess genuine model reasoning capabilities more accurately.

Introduction: **Research Objectives**

Key Research Questions:

- Does performance degradation observed in GSM-Symbolic[1] persist with other types of paraphrasing?
- Is this sensitivity to paraphrasing unique to mathematical datasets, or does it extend to non-mathematical tasks?
- Are all current LLMs equally impacted, or do some exhibit greater resilience to paraphrasing variations?

---

[1] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, & Mehrdad Farajtabar. (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models.

Project Timeline

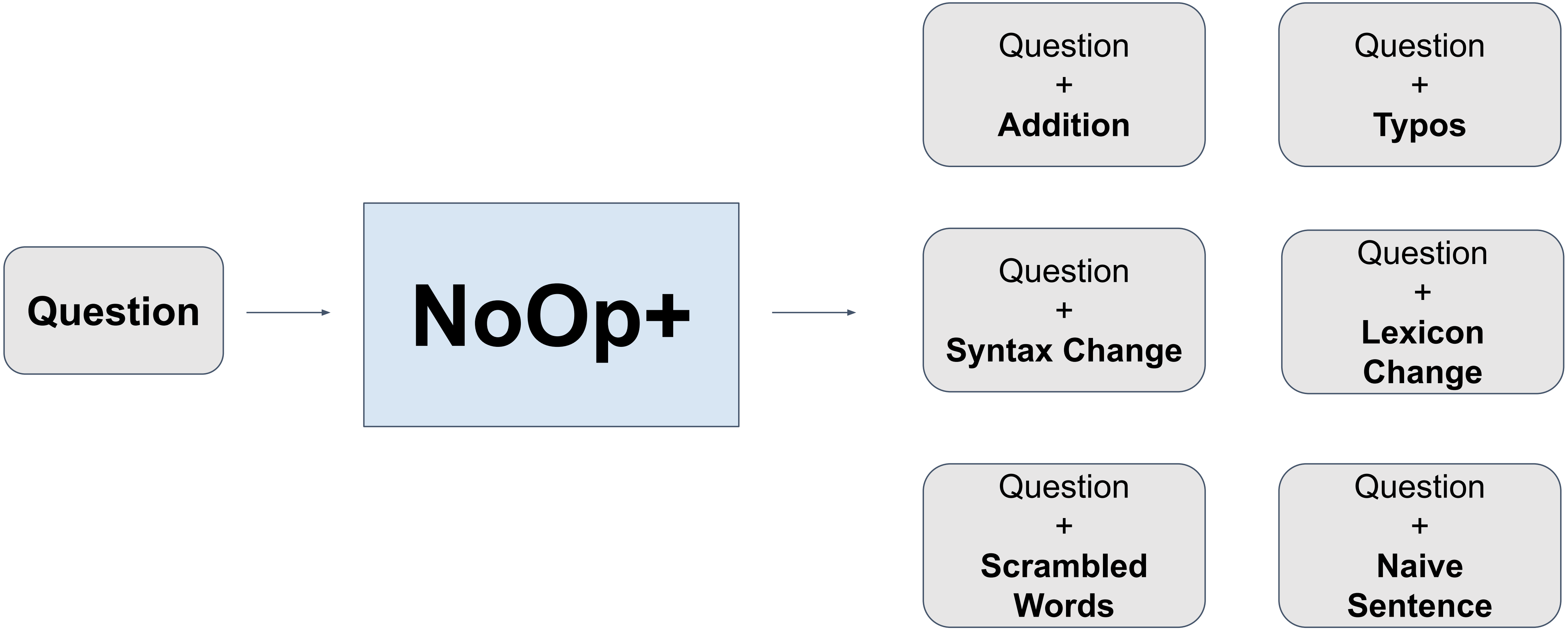
Level 1: Foundation: Create NoOp+ Tasks ✅

Level 2: Generalization: Test Across Diverse Datasets ✅

Level 3: Model Robustness: Evaluate Across Multiple Models

Level 4: New Word-Level Paraphrase Types 🆕

Methods: **NoOp+**



Question

NoOp+

Question
+
Addition

Question
+
Typos

Question
+
Syntax Change

Question
+
Lexicon Change

Question
+
Scrambled Words

Question
+
Naive Sentence

Methods: **Baseline**

Janet’s ducks lay 16 eggs per day.

She eats three for breakfast every morning and bakes muffins for her friends every day with four.

She sells the remainder at the farmers' market daily for $2 per fresh duck egg.

How much in dollars does she make every day at the farmers' market?

Methods: Paraphrase Types: **Naive Addition, Addition**

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market? **Sebastian goes to buy icecream.**

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. **The farmers' market where Janet sells her eggs is located in a town that has a reputation for being a popular destination for rich food bloggers and culinary experts.** How much in dollars does she make every day at the farmers' market?

Methods: Paraphrase Types: **Lexicon, Syntax**

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per **dozen** duck egg. How **money** in dollars does she make every day at the farmers' market?

Janet’s ducks lay 16 eggs per day.She eats Three for breakfast Every Morning and bakes muffins for her friends Every day with Four. **Daily she sells the** remainder at the farmers' market for $2 per fresh duck egg.How much in dollars does she make every day at the farmers' market?

Methods: Paraphrase Types: **Typo, Scramble**

Jsnwt’s duks lay 16 eggs per dda. She rats thrse for breakfast every korning and bakes mufins for her friends every day wuth ffour. Shr selks he remmainder at the farmers' marrket daily for $2 per fresh uck ehg. How muc in dollars ddoes sgee make evert day at the fafmers' markeet?

Jneat’s ducks lay 16 eggs per day. She etas terhe for bfaeksart eervy mnornig and bkeas mfunifs for her fierdns eevry day wtih fuor. She slles the reamdenir at the frarmes' mkraet dilay for $2 per fsreh dcuk egg. How much in dolalrs does she mkae ervey day at the frmears' mearkt?

Results: Datasets

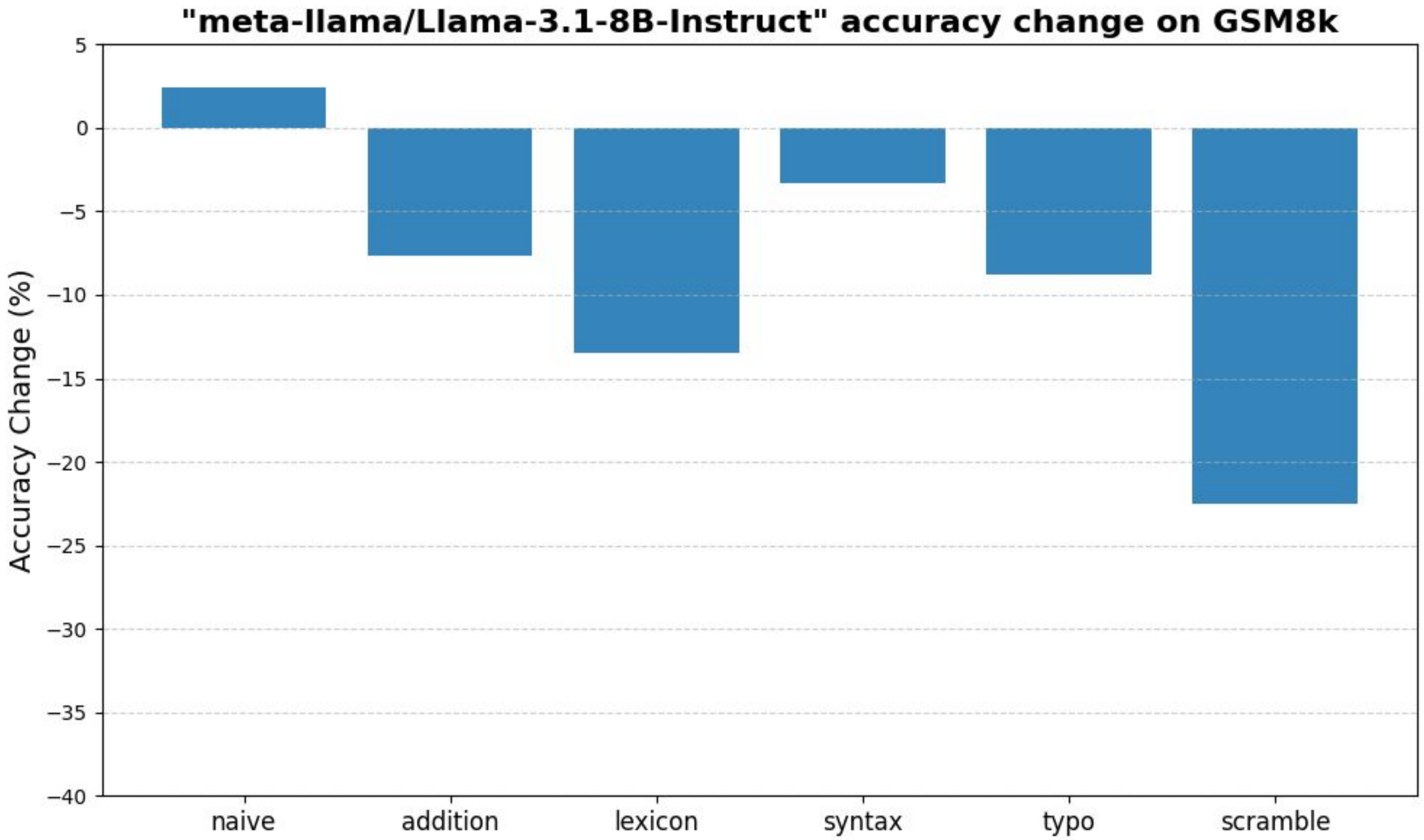
GSM8K[1]**:**
A dataset of 8.5K carefully crafted grade-school math word problems aimed at benchmarking and advancing AI models' mathematical reasoning skills.

Llama-3.1-8B-Instruct baseline: 80.5% accuracy

---

[1] IKarl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, & John Schulman (2021). Training Verifiers to Solve Math Word Problems.

Results: Grade School Math 8K



Results: Datasets

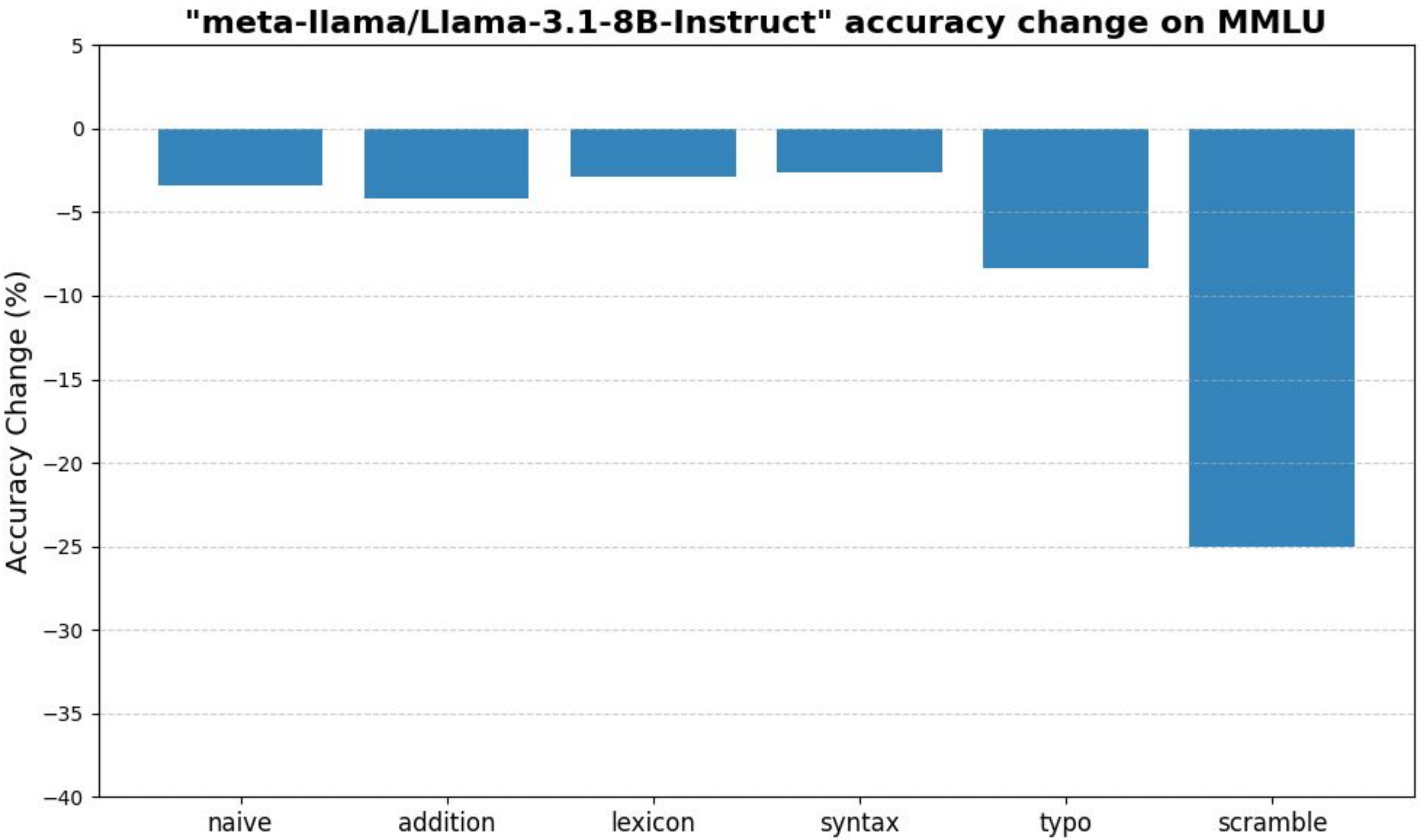
MMLU[1]:
A comprehensive benchmark of 57 academic subjects assessing a model's breadth of knowledge and understanding across diverse disciplines.

Llama-3.1-8B-Instruct baseline: 70.2% accuracy

---

[1] IDan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, & Jacob Steinhardt (2020). Measuring Massive Multitask Language Understanding.

Results: Massive Multitask Language Understanding



Results: Datasets

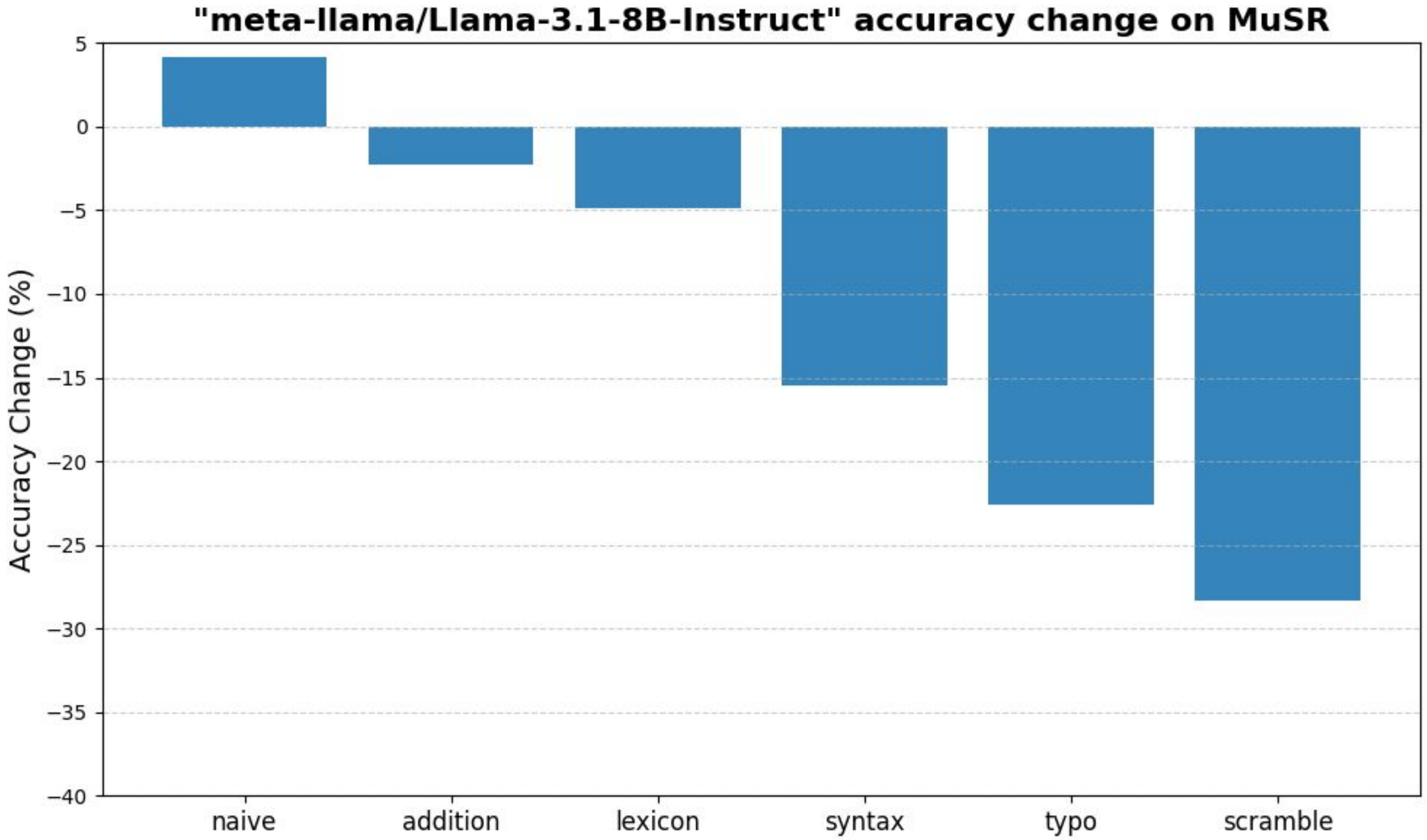
MuSR[1]:
A dataset focused on multi-step symbolic reasoning, designed to evaluate and enhance models' ability to solve complex reasoning tasks involving mathematical and logical processes.

Llama-3.1-8B-Instruct baseline: 35.0% accuracy

---

[1] IZayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, & Greg Durrett. (2024). MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning.

Results: Multistep Soft Reasoning



Results: Datasets

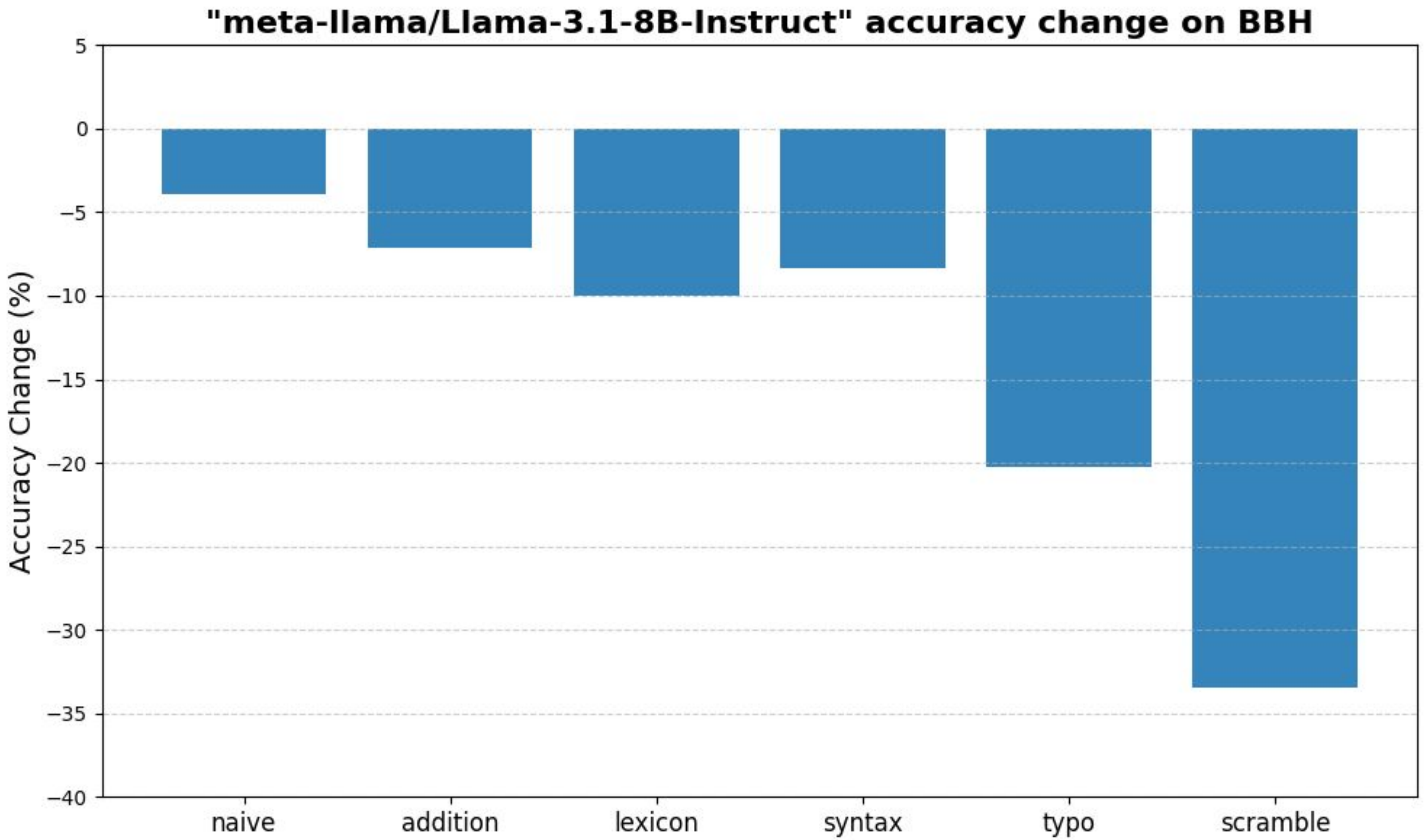
BBH[1]:
A dataset of challenging, high-level tasks specifically designed to test and benchmark the reasoning capabilities of advanced language models.

Llama-3.1-8B-Instruct baseline: 61.6% accuracy

---

[1] lMirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, & Jason Wei. (2022). Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them.

Results: Big Bench Hard



Project Timeline

Level 1: Foundation: Create NoOp+ Tasks ✅

Level 2: Generalization: Test Across Diverse Datasets ✅

Level 3: Model Robustness: Evaluate Across Multiple Models

- Show that performance degradation due to paraphrasing is a widespread issue across different LLMs.
- Conduct evaluations across multiple model sizes and versions to assess consistency in paraphrase sensitivity.
- Test whether prompt engineering alleviates the performance degradation

Level 4: New Word-Level Paraphrase Types

References

1. Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, & Mehrdad Farajtabar. (2024). **GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models**.
2. Qihuang Zhong, Kang Wang, Ziyang Xu, Juhua Liu, Liang Ding, & Bo Du. (2024). **Achieving >97% on GSM8K: Deeply Understanding the Problems Makes LLMs Better Solvers for Math Word Problems**.
3. Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, & Denny Zhou. (2023). **Large Language Models Can Be Easily Distracted by Irrelevant Context**.
4. Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, & John Schulman (2021). **Training Verifiers to Solve Math Word Problems.**
5. Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, & Jacob Steinhardt (2020). **Measuring Massive Multitask Language Understanding.**
6. Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, & Greg Durrett. (2024). **MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning.**
7. Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, & Jason Wei. (2022). **Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them.**

Questions?

For more in-depth questions or side–notes, you can contact me via:

lucajoshua.francis@stud.uni-goettingen.de