Creative Writer 🧙 or a Parrot 🦜? Automatically Evaluating LLM's Creativity/Originality is Possible!
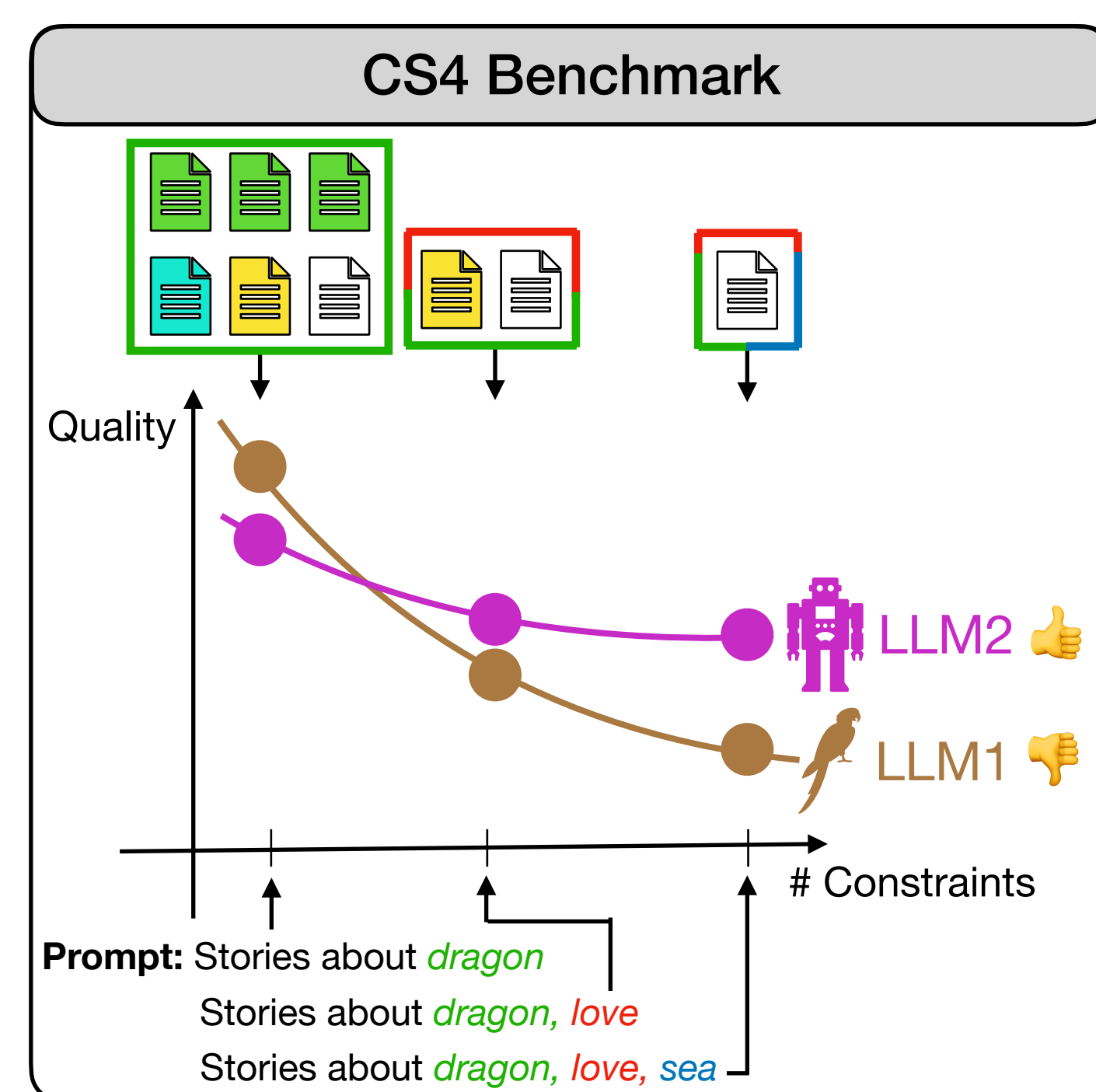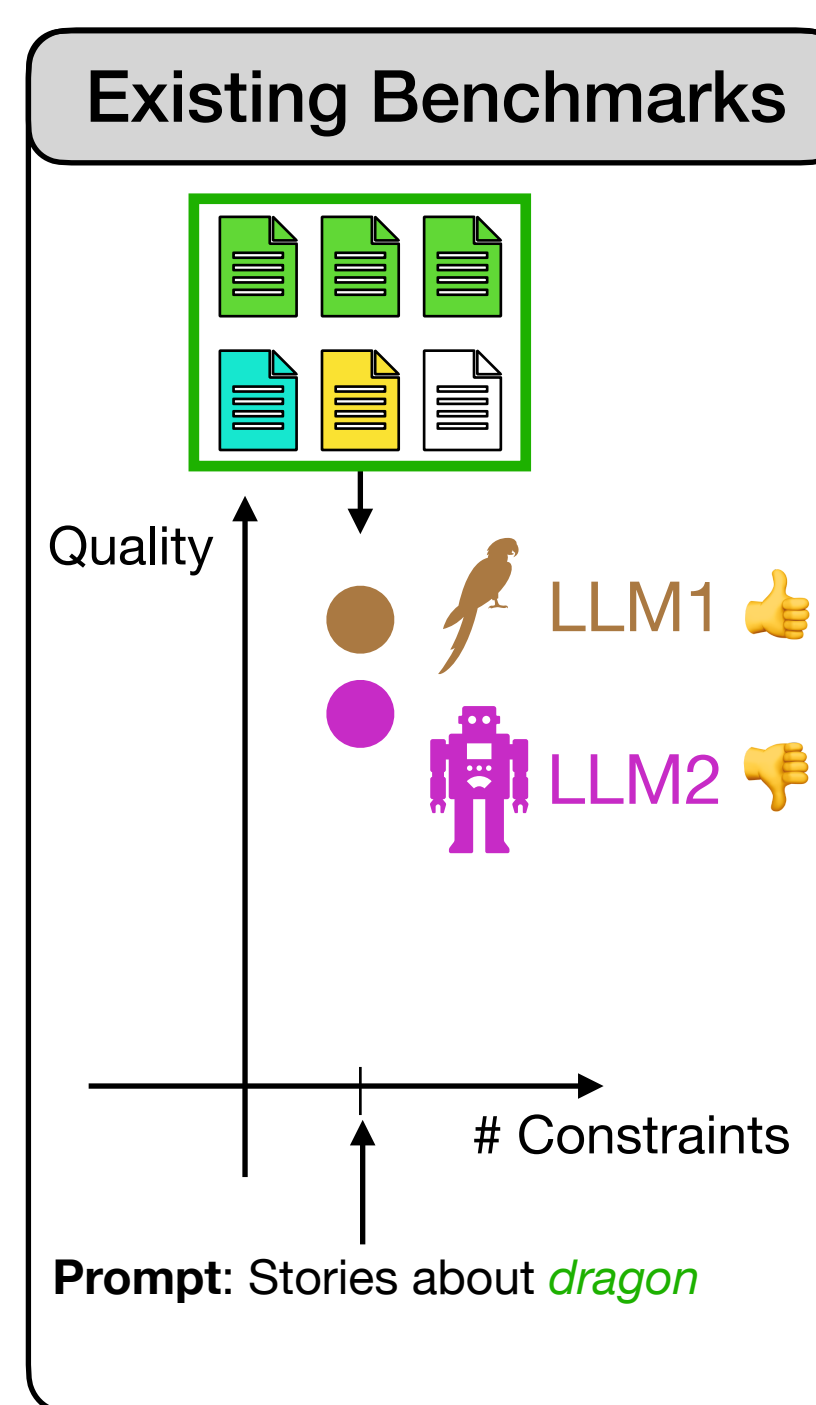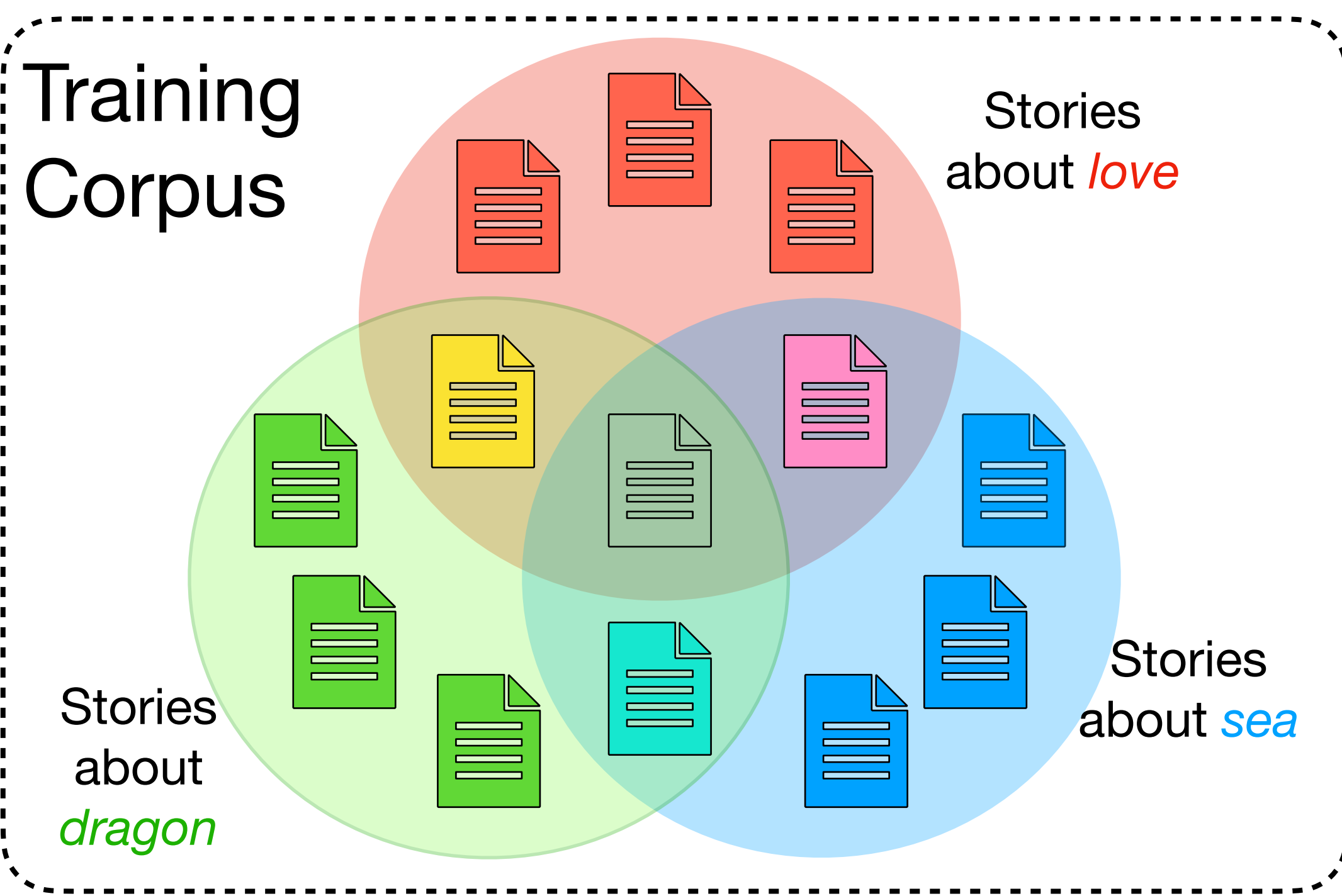
CS4: Measuring the Creativity of Large Language Models Automatically by Controlling the Number of Story–Writing Constraints

amazon

Anirudh Atmakuru*, Jatin Nainani*, Rohith Siddhartha Reddy Bheemreddy*, Anirudh Lakkaraju*, Zonghai Yao, Hamed Zamani, Haw-Shiuan Chang*

UMASS AMHERST

Introduction

By increasing the number of constraints in the prompt, we can increase the prompt specificity and hinder LLMs from retelling high-quality narratives in their training data.

If an LLM could output a bad story from more constraints, the LLM should be less creative to write original stories.

CS4 (Comparing the Skill of Creating Stories by Controlling the Synthesized Constraint Specificity) Benchmark
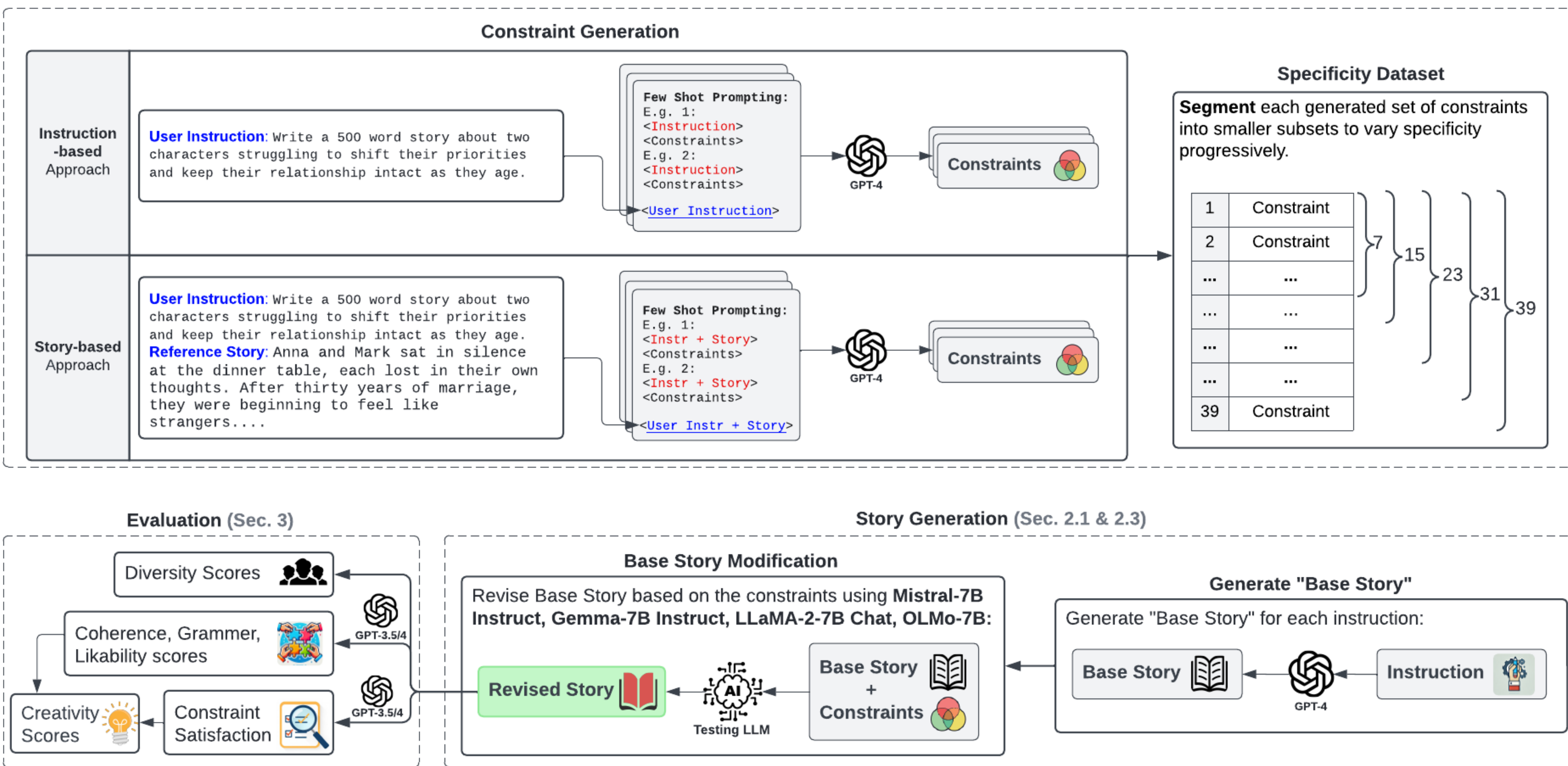
Example of Tradeoffs

Constraint A: "The main character enjoys reading at home."
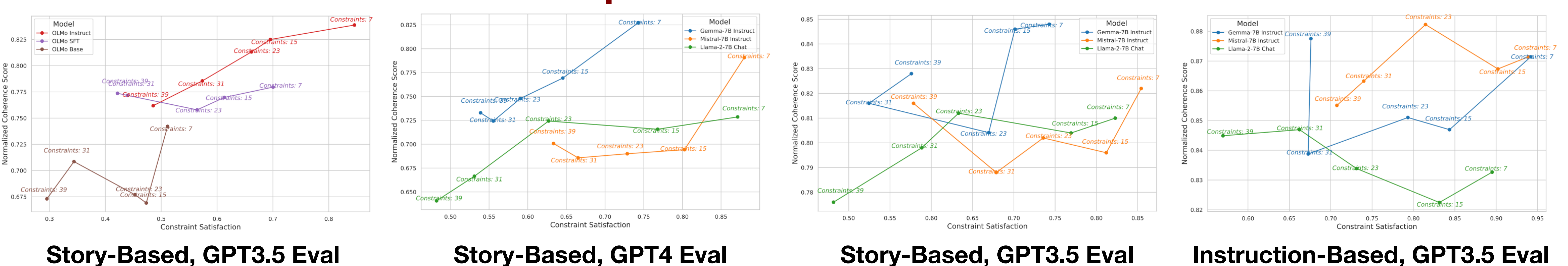Constraint B: "Something surprising happened at a school."

Coherent Story A (BUT Fails Constraint):
"Mary loves to read everyday. This makes her a really good student at school."

Constraint Satisfying Story B (BUT Incoherent):
"John enjoys reading at home. Then, something surprising happened at school!"

Experiment Results

Story-Based, GPT3.5 Eval — Story-Based, GPT4 Eval — Story-Based, GPT3.5 Eval — Instruction-Based, GPT3.5 Eval

Conclusions

- LHF can help LLMs select better stories from their training data but has limited influence in boosting LLMs' ability to produce creative stories that are unseen in the training corpora.

- Different LLMs perform very differently under different numbers of constraints and achieve different balances between the model's instruction-following ability and narrative coherence

Reference

[1] West, Peter, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang et al. "THE GENERATIVE AI PARADOX:"What It Can Create, It May Not Understand"." In The Twelfth International Conference on Learning Representations. 2023.

[2] Lu, Ximing, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger et al. "AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text." arXiv preprint arXiv:2410.04265 (2024).

[3] Lu, Yining, Dixuan Wang, Tianjian Li, Dongwei Jiang, and Daniel Khashabi. "Benchmarking Language Model Creativity: A Case Study on Code Generation." arXiv preprint arXiv:2407.09007 (2024).