Evaluating LLMs with Multiple Problems at once

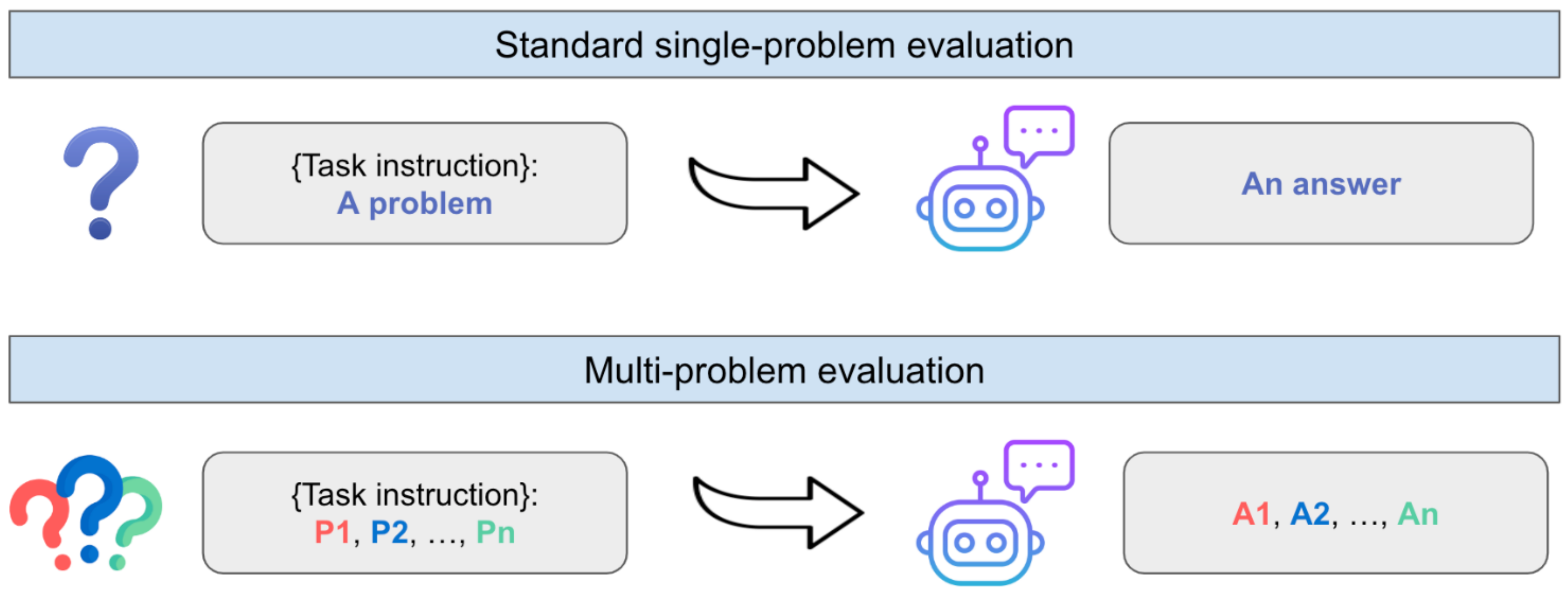
Zhengxiang Wang, Jordan Kodner, Owen Rambow
{first.last}@stonybrook.edu
Department of Linguistics & IACS, Stony Brook University

Stony Brook University

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

INTRODUCTION

- **Multi-Problem Prompting (MPP)**: A cost-efficient prompting technique that prompts multiple problems at once to avoid repeating a shared context
- **Multi-Problem Evaluation (MPE)**: An eval paradigm that evaluates via MPP an LLM's ability to handle multiple problems at once or in a single output
- **Motivation**: Provides a foundational insight into how LLMs operate over multi-problem inputs that can be sufficiently long and use information from individual problems contained within each multi-problem input.
- **MPE versus Single-Problem Evaluation (SPE)**: (1) Lesser Data Contamination Concerns; (2) Improved Controllability and Interpretability of Evaluation; (3) High Feasibility and Adaptability

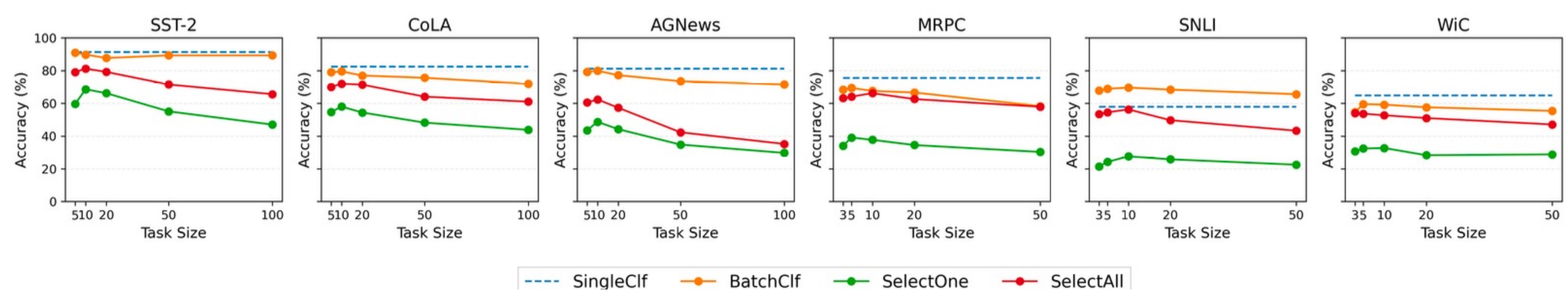


ZeMPE

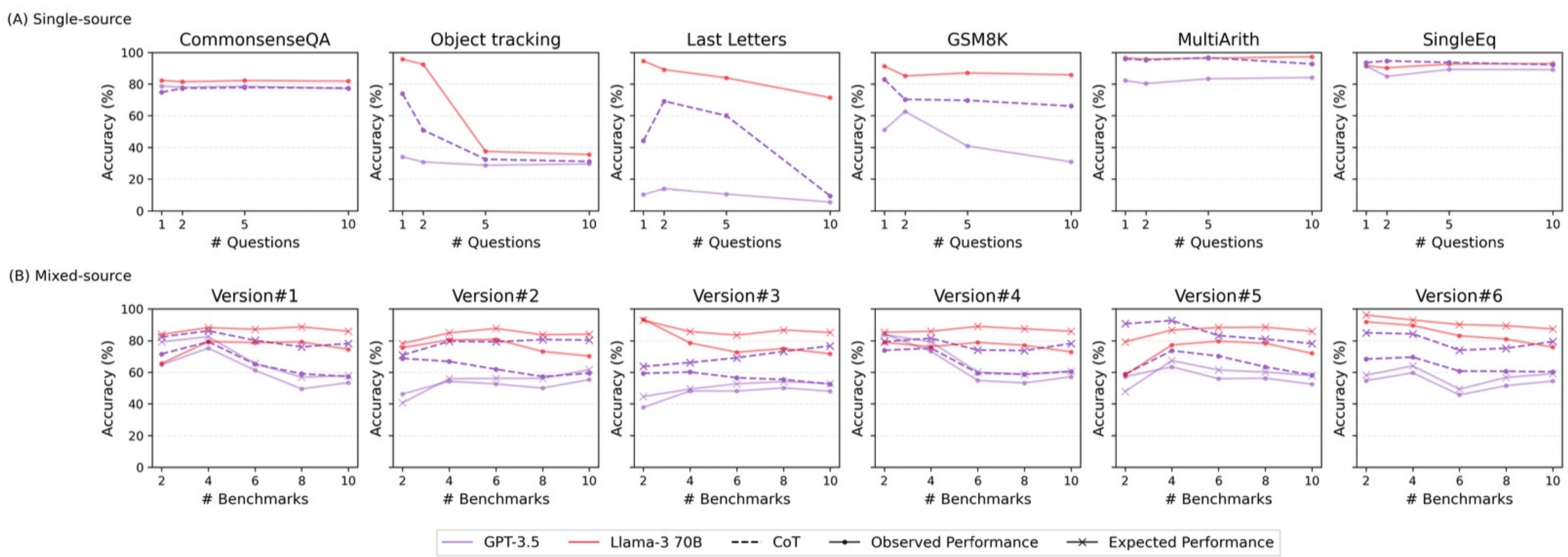
- **ZeMPE**: **Ze**ro-shot **M**ulti-**P**roblem **E**valuation, a benchmark comprising 53,1000 zero-shot multi-problem prompts
- **Classification-Related Tasks** : (1) SingleClf (Single Classification); (2) BatchClf (Batch Classification); (3) SelectOne (Index Selection One Label); (4) SelectAll (Index Selection All Labels)
- **Reasoning-Related Tasks**: (1) MultiReason^{SS} (single-source multi-problem reasoning) and (2) MultiReason^{MS} (mixed-source multi-problem reasoning)

Problem Type	Input/Output Format	Benchmark
Classification	Single-text input	SST-2 (Socher et al., 2013) CoLA (Warstadt et al., 2019) AGNews (Gulli, 2004)
	Text-pair input	MRPC (Dolan and Brockett, 2005) SNLI (Bowman et al., 2015) WiC (Pilehvar and Camacho-Collados, 2019)
Reasoning	Yes/no output	StrategyQA (Geva et al., 2021) Coin Flips (Wei et al., 2023)
	Multi-choice output	AQuA (Ling et al., 2017) CommonsenseQA (Talmor et al., 2019) Object tracking (Srivastava et al., 2023) Bigbench date (Srivastava et al., 2023)
	Free-response output	Last Letters (Wei et al., 2023) SVAMP (Patel et al., 2021) GSM8K (Roy and Roth, 2015) MultiArith (Patel et al., 2021) AddSub (Hosseini et al., 2014) SingleEq (Koncel-Kedziorski et al., 2015)

Experimental Results



- LLMs can handle multiple classifications at once under zero-shot with minimal performance loss.
- LLMs perform significantly worse on the selection tasks.



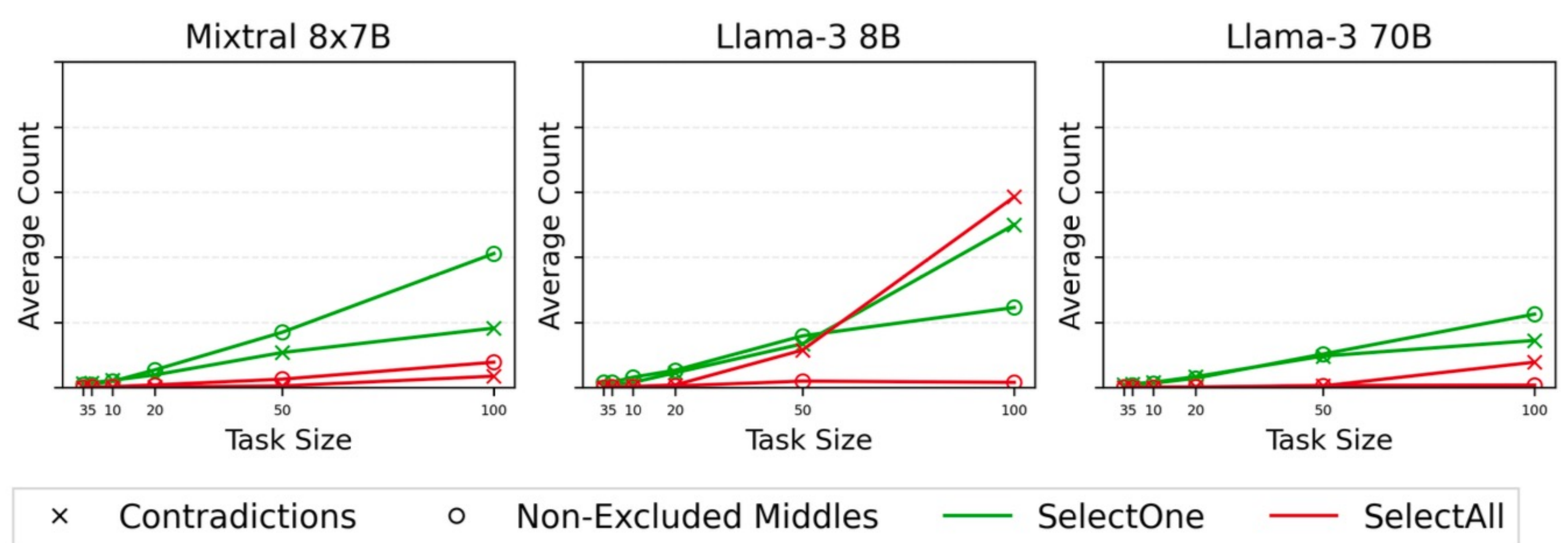
- LLMs can handle multiple reasoning problems at once when the problems are from the single source.
- When the reasoning problems are from mixed sources, LLMs perform worse than expected.
- Benefits of zero-shot-CoT prompting are transferrable under MPP.

Further Analyses

- Similar prediction and positional errors between SingleClf and BatchClf
- Why is SelectAll much harder than SelectOne. See right ➡️
- Exploring model-level factors that may enable MPP. See below ⬇️

	SingleClf	BatchClf	Avg # Answers
Llama-3 8B (Instruct)	80.5	79.4	5.0
GPT-3.5	84.2	79.6	5.0
Llama-3 8B (Base)	78.5	60.6	5.04
GPT-3 1.3B	63.0	0.0	0.03
GPT-3 175B	66.6	64.4	5.08
FLAN-T5-Large (0.78B)	76.0	NA	1.0
FLAN-T5-XL (3B)	80.2	NA	1.0
FLAN-T5-XXL (11B)	78.2	4.0	1.2

SingleClf and BatchClf (task size 5) accuracy (%) on CoLA



Conclusion

- LLMs are capable of handling multiple classification or reasoning problems from a single data source as well as handling them separately zero-shot.
- Two conditions are identified under which LLMs show consistent performance declines with MPP: (1) the two selection tasks; (2) mixed-source problems.
- We release a new MPE benchmark called ZeMPE to facilitate future MPE studies.