UC SANTA BARBARA

ArXiv Link

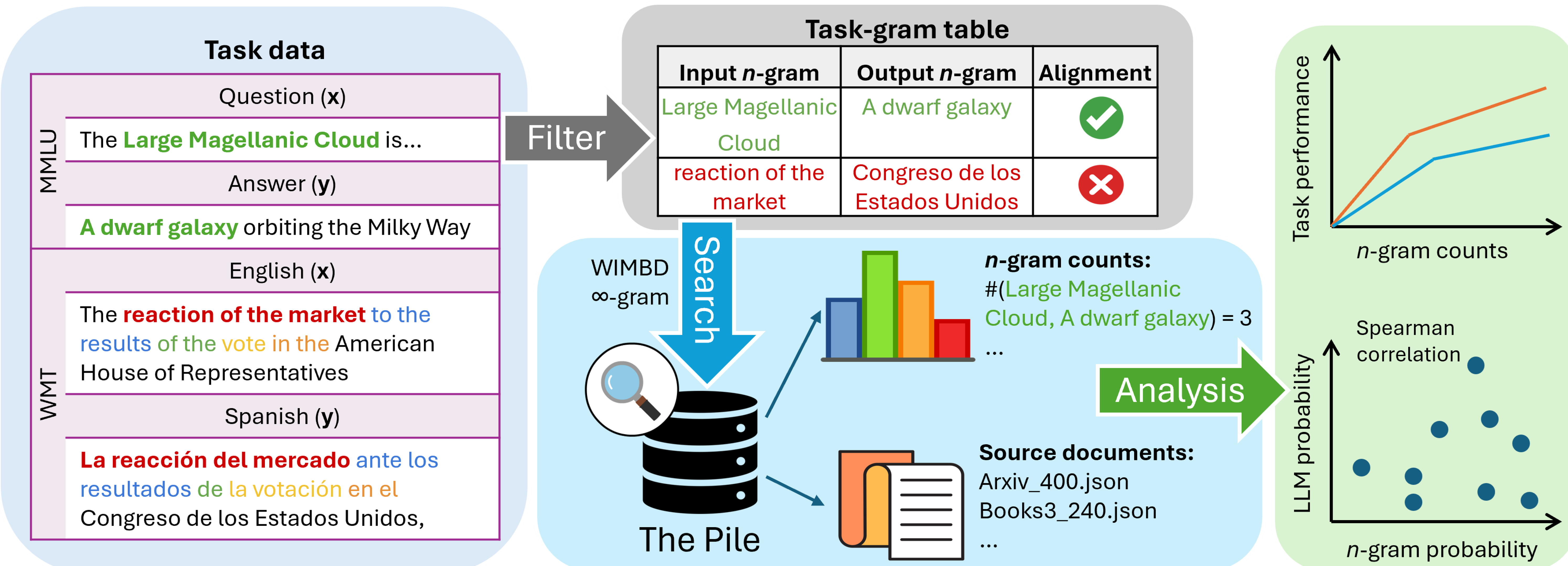
Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data

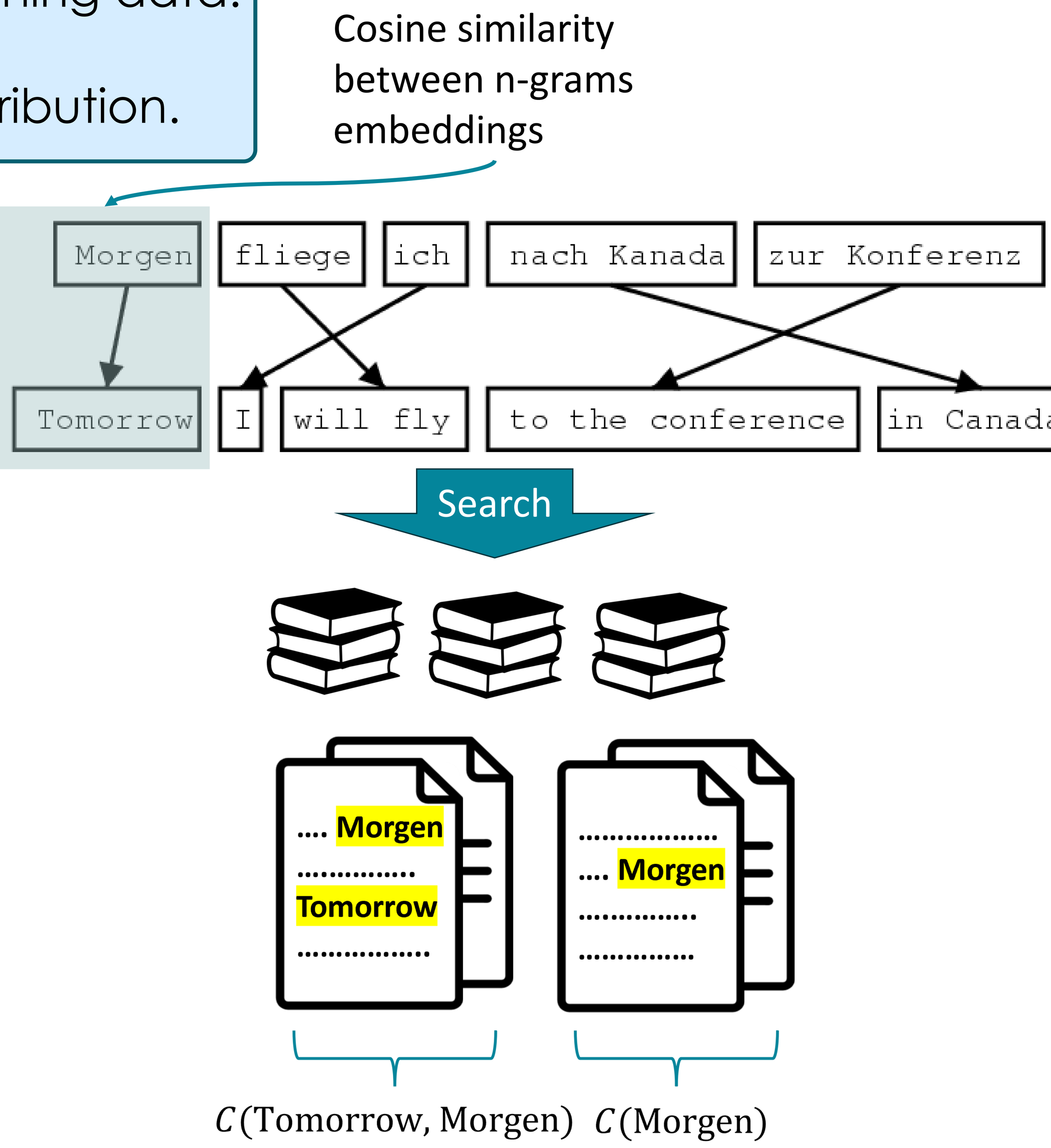
Xinyi Wang*[1], Antonis Antoniades*[1], Yanai Elazar[2,3], Alfonso Amayuelas[1], Alon Albalak[1], Kexun Zhang[4], William Yang Wang[1]
[1]University of California, Santa Barbara [2]Allen Institute for AI [3]University of Washington [4]Carnegie Mellon University
* Equal Contribution

Distributional Memorization: the correlation between the distribution of LLM outputs and the distribution of pretraining data.

Distributional Generalization: the divergence between the LLM's output distribution and the pretraining data distribution.

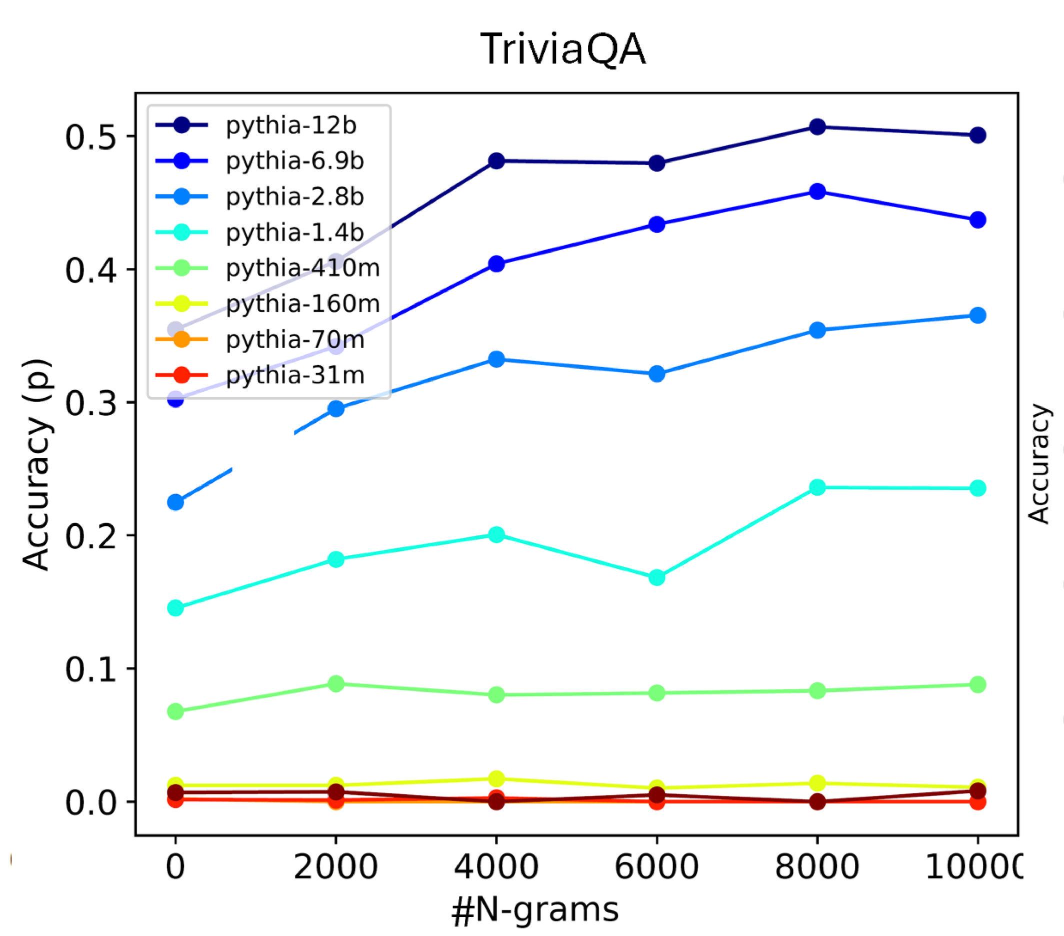
Tasks we used:
- **TriviaQA**: Commonsense Question Answering
- **WMT**: Translation
- **MMLU**: World knowledge understanding
- **GSM8K**: Math reasoning



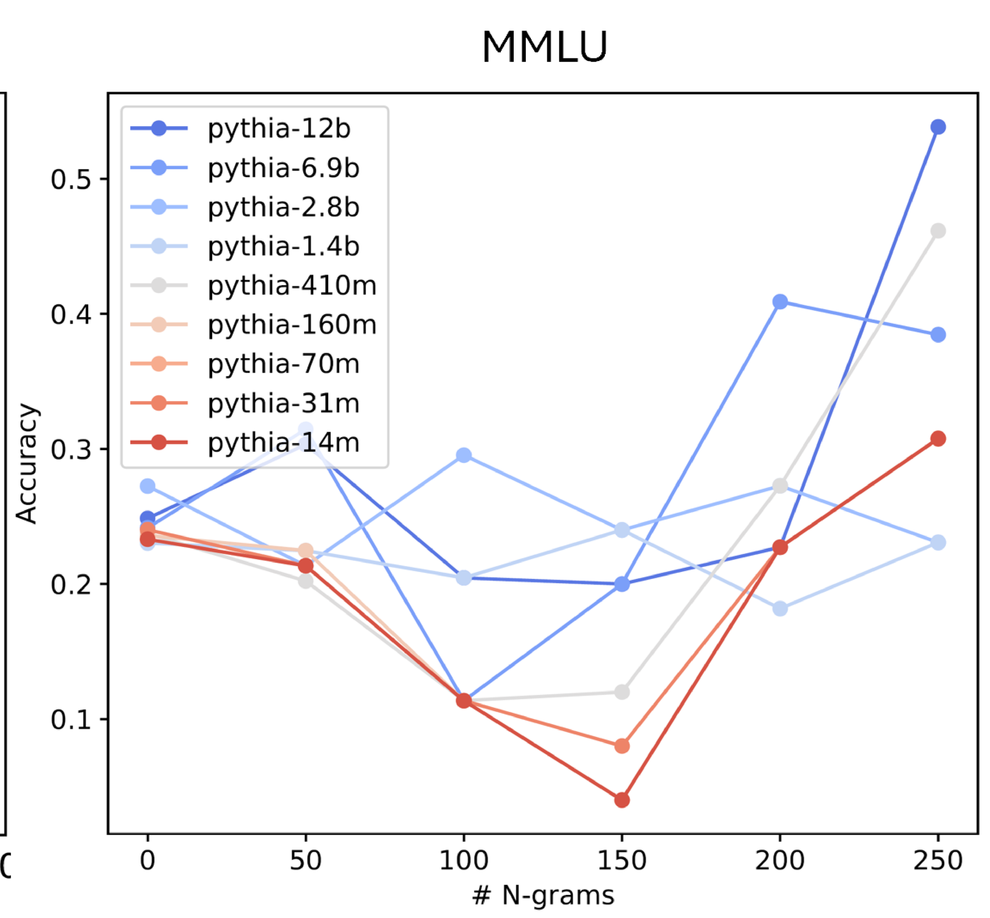


$$P_{data}(\text{Tomorrow}|\text{Morgen}) = \frac{C(\text{Tomorrow, Morgen})}{C(\text{Morgen})}$$

Task performance v.s. n-gram pair count



Knowledge intensive task: Common in pretraining data

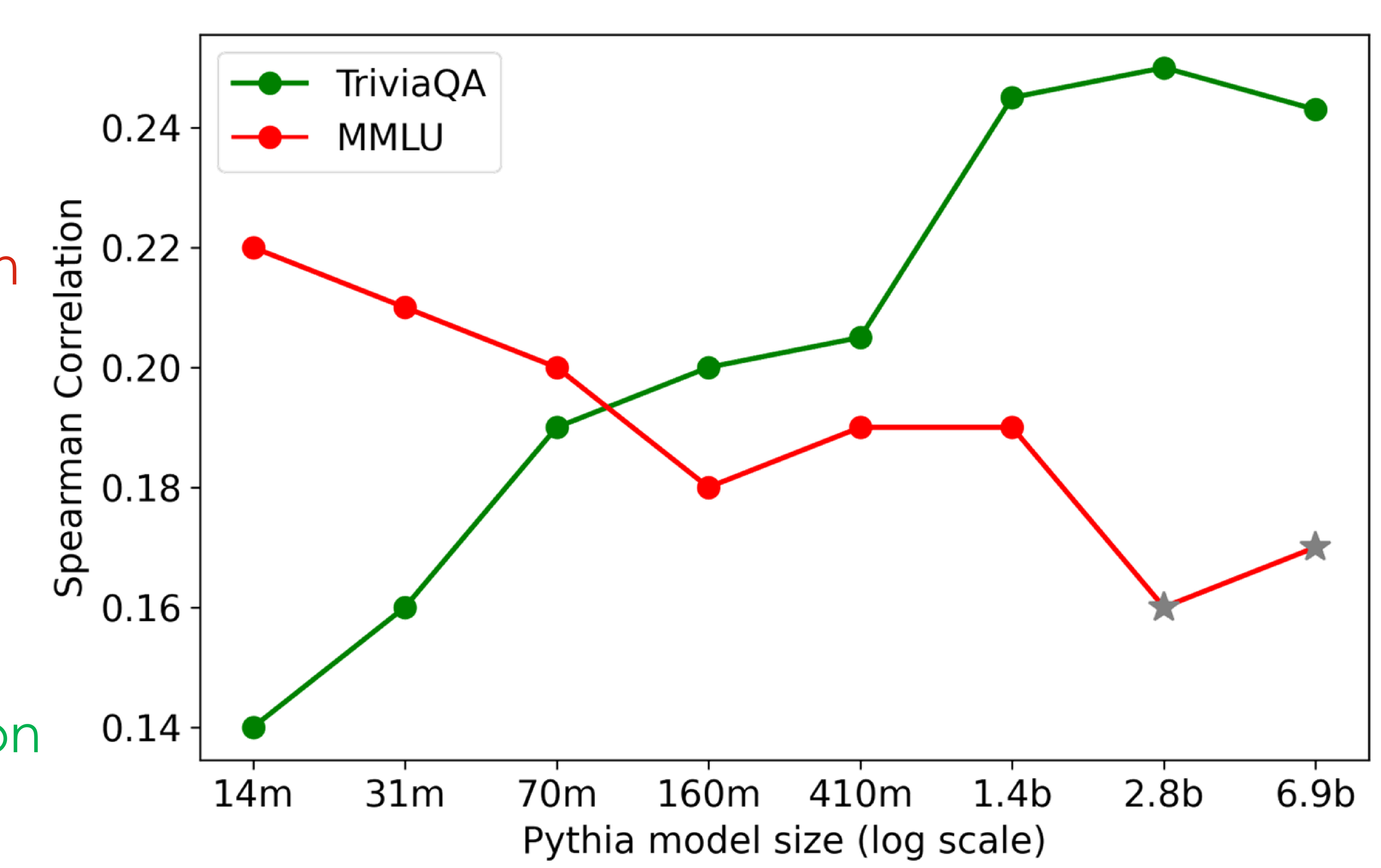


Reasoning intensive task: Rare in pretraining data

N-gram Distribution v.s. LLM Distribution

Reasoning intensive tasks rely more on generalization

Knowledge intensive task relies more on memorization



Generalization: *discourage* n-gram overlap between prompt and pretraining corpus.

Memorization: *encourage* n-gram overlap between prompt and pretraining corpus.

Practical implication: Prompt overlap with pretraining corpus

	TriviaQA		GSM8K	
	Memorization	Generalization	Memorization	Generalization
Pythia (6.9B)	**17%**	9%	2.6%	**2.8%**
Pythia-Instruct (6.9B)	**23.5%**	23.2%	6.3%	**7.3%**
Pythia (12B)	**28.7%**	23.2%	2.7%	**2.8%**
OLMo (7B)	**36.4%**	29.8%	2.5%	**3.1%**
OLMo-instruct (7B)	**29%**	10%	6.3%	**7.9%**

Table 1: Zero-shot accuracy on TriviaQA and GSM8K test set with memorization encouraged task prompt (maximize counts) and generalization encouraged task prompt (minimize counts).