Language Pre-training without Natural Language

Qian Liu (刘乾) @ Sea AI Lab
Bei Chen (陈蓓) @ Microsoft Research Asia

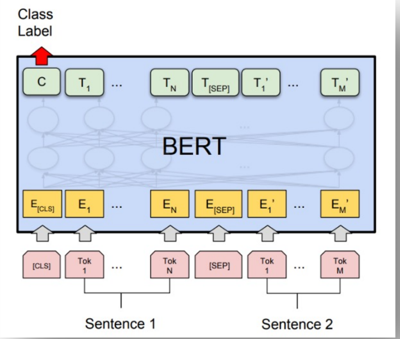
sigma

Current AI Paradigm: Language Models = SOTA

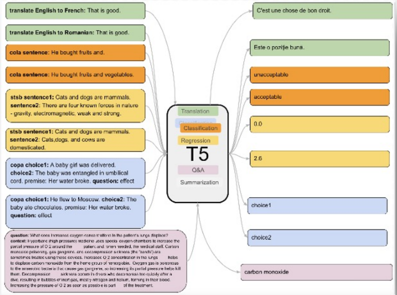
SuperGLUE GLUE Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

Leaderboard Version: **2.0**

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
+	3	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
+	4	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
+	5	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
	6	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
	7	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+	8	Infosys : DAWN : AI Research	RoBERTa-iCETS		77.4	84.7	88.2/91.6	85.8	78.4/37.5	82.9/82.4	83.8	69.1	65.1	35.2	93.8/68.8
+	9	Timo Schick	iPET (ALBERT) - Few-Shot (32 Examples)		75.4	81.2	79.9/88.8	90.8	74.1/31.7	85.9/85.4	70.8	49.3	88.4	36.2	97.8/57.9
	10	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	29.6	97.8/57.3
	11	Ben Mann	GPT-3 few-shot - OpenAI		71.8	76.4	52.0/75.6	92.0	75.4/30.5	91.1/90.2	69.0	49.4	80.1	21.1	90.4/55.3
	12	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4
			BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0	97.8/51.7
			Most Frequent Class		47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	0.0	100.0/50.0
			CBoW		44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	-0.4	100.0/50.0
			Outside Best		-	80.4	-	84.4	70.4/24.5	74.8/73.0	82.7	-	-	-	-
	-	Stanford Hazy Research	Snorkel [SuperGLUE v1.9]		-	-	88.6/93.2	76.2	76.4/36.3	-	78.9	72.1	72.6	47.6	-



BERT (Devlin et al., 2018)



T5 (Raffel et al., 2020)

Current AI Paradigm: Language Models = Human Parity

SQuAD2.0
The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799
4 Sep 11, 2020	EntitySpanFocus+AT (ensemble) *RICOH_SRCB_DML*	90.454	92.748

CoQA
A Conversational Question Answering Challenge

What is CoQA?

CoQA is a large-scale dataset for building **Co**nversational **Q**uestion **A**nswering systems. The goal of the CoQA challenge is to measure the ability of machines to understand a text passage and answer a series of interconnected questions that appear in a conversation. CoQA is pronounced as coca.

CoQA paper

CoQA contains 127,000+ questions with answers collected from 8000+ conversations. Each conversation is collected by pairing two crowdworkers to chat about a passage in the form of questions and answers. The unique features of CoQA include 1) the questions are conversational; 2) the answers can be free-form text; 3) each answer also comes with an evidence subsequence highlighted in the passage; and 4) the passages are collected from seven diverse domains. CoQA has a lot of challenging phenomena not present in existing reading comprehension datasets, e.g., coreference and pragmatic reasoning.

Download

Browse the examples in CoQA:

Browse CoQA

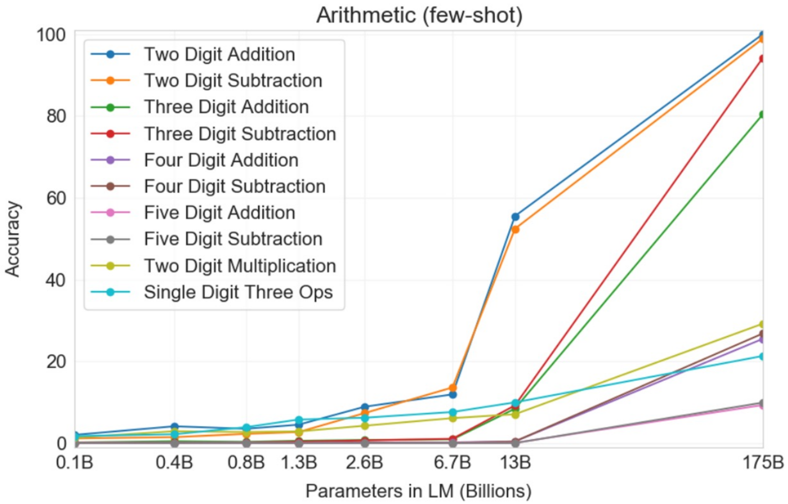
Leaderboard

Rank	Model	In-domain	Out-of-domain	Overall
	Human Performance *Stanford University* (Reddy & Chen et al. TACL '19)	89.4	87.4	88.8
1 Sep 05, 2019	RoBERTa + AT + KD (ensemble) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	91.4	**89.2**	**90.7**
1 Apr 22, 2020	TR-MT (ensemble) *WeChatAI*	**91.5**	88.8	**90.7**
2 Sep 05, 2019	RoBERTa + AT + KD (single model) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	90.9	**89.2**	90.4
3 Jan 01, 2020	TR-MT (ensemble) *WeChatAI*	91.1	87.9	90.2
4 Mar 29, 2019	Google SQuAD 2.0 + MMFT (ensemble) *MSRA + SDRG*	89.9	88.0	89.4
5 Dec 18, 2019	TR-MT (single model) *WeChatAI*	90.4	86.8	89.3
6 Sep 13, 2019	XLNet + Augmentation (single model) *Xiaoming*	89.9	86.9	89.0

Research Challenge: Reasoning

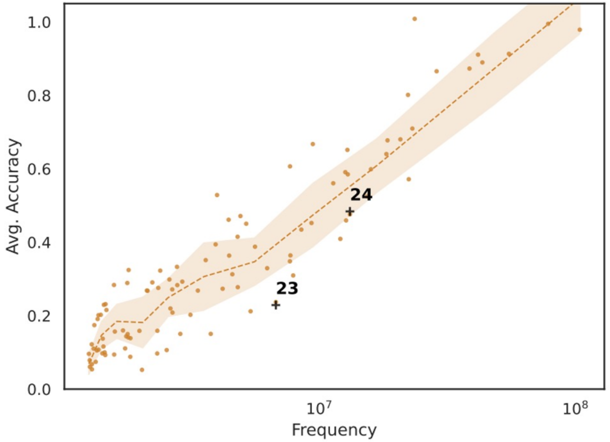
However, the reasoning capability is still the mysterious for language models — even for giant language models (e.g., GPT3).

Emergent Performance at 175B? No!



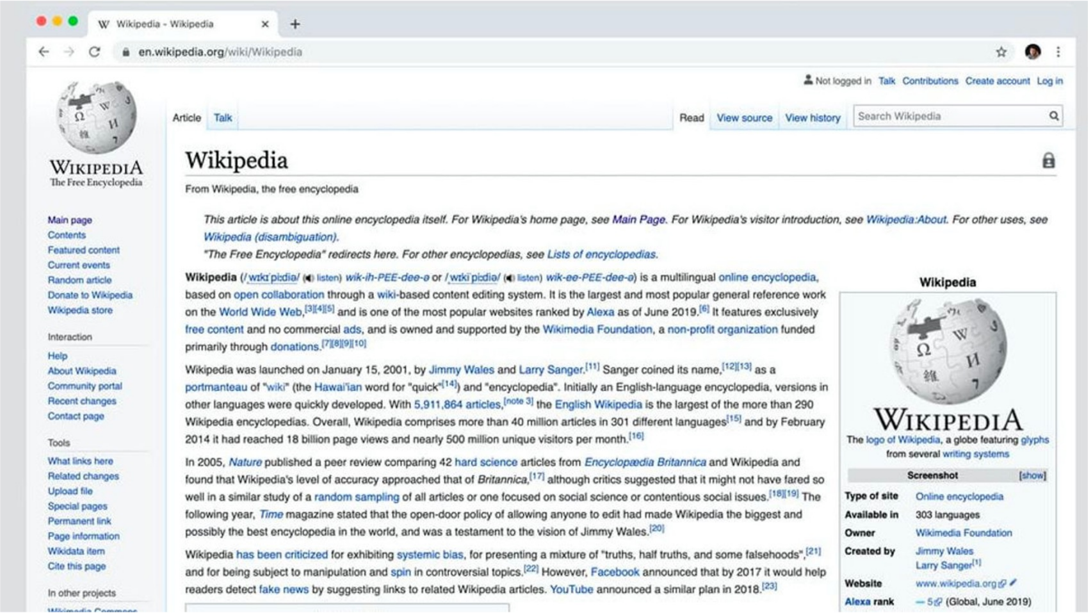
Reasoning, or correlation?

Q: What is 24 times 18? A: ___	*Model:* 432 ✓
Q: What is 23 times 18? A: ___	*Model:* 462 ✗



Research Challenge: Reasoning

However, it is difficult to obtain large amounts of clean natural language sentences containing clear evidence of reasoning.



Key Idea: Program as a Proxy

There are rich reasoning operations (e.g., sort) in the program execution process. Can we leverage programs instead of natural language sentences as pre-training corpus?

Program		Natural Language
sorted([1, -5, 10, 6], key=abs, reverse=True)	$\approx$	Given the list which contains 1, -5, 10 and 6, I want to order from high to low no matter what sign each number has, but keeping the sign

Key Idea: Program as a Proxy

There is a natural analogy between neural models and program executors!

Program

sorted([1, -5, 10, 6],
key=abs, reverse=True)

Program Executor

Natural Language

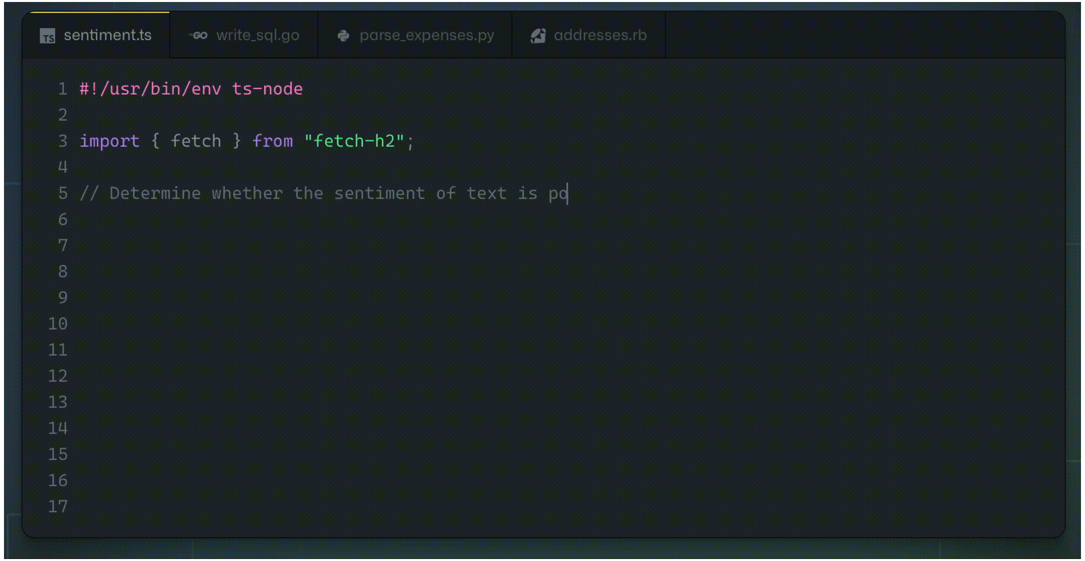
Given the list which contains 1, -5,
10 and 6, I want to order from high
to low no matter what sign each
number has, but keeping the sign

Neural Model

[10, 6, -5, 1]

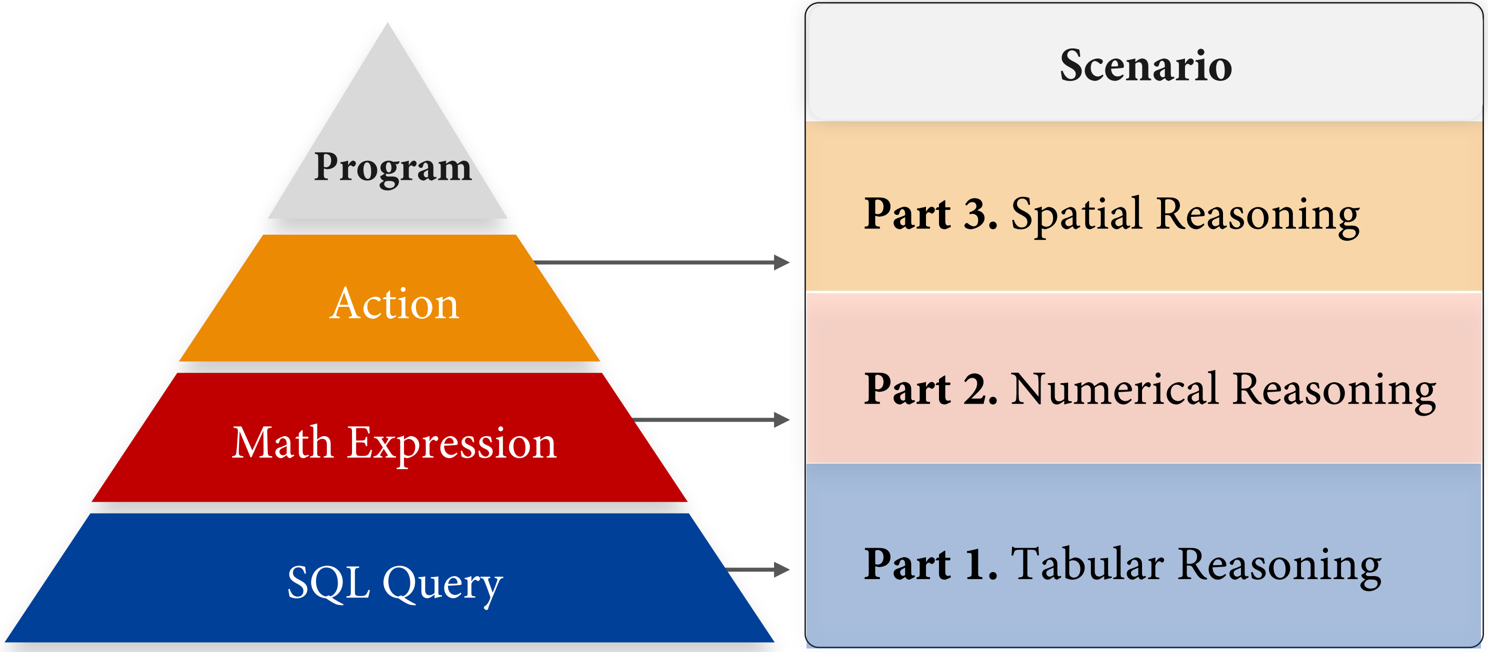
Method Comparison: Execution v.s. Generation

Recent language models can perform program generation, and the difference is that we leverage program execution for natural language reasoning beyond programs.



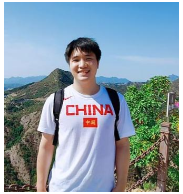
GitHub Copilot (2021)

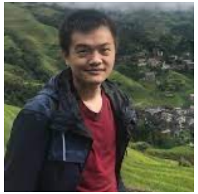
Overview: Tabular, Numerical and Spatial Reasoning



Part 1. SQL Query for Tabular Reasoning

TAPEX: Table Pre-training via Learning a Neural SQL Executor

Qian Liu[1] Bei Chen[2] Jiaqi Guo[3] Morteza Ziyadi[2] Zeqi Lin[2] Weizhu Chen[2] Jian-Guang Lou[2]

1 北京航空航天大學 BEIHANG UNIVERSITY 2 Microsoft 3 西安交通大学 XI'AN JIAOTONG UNIVERSITY

ICLR

10th International Conference on Learning Representations (ICLR 2022)

Background: Tabular Reasoning

City	Country	Nations	Year
Athens	Greece	14	1896
St. Louis	USA	12	1904
...	...	...	...
Athens	Greece	201	2004
Beijing	China	204	2008

Question

Greece held its last Summer Olympics in which year

Background: Tabular Reasoning

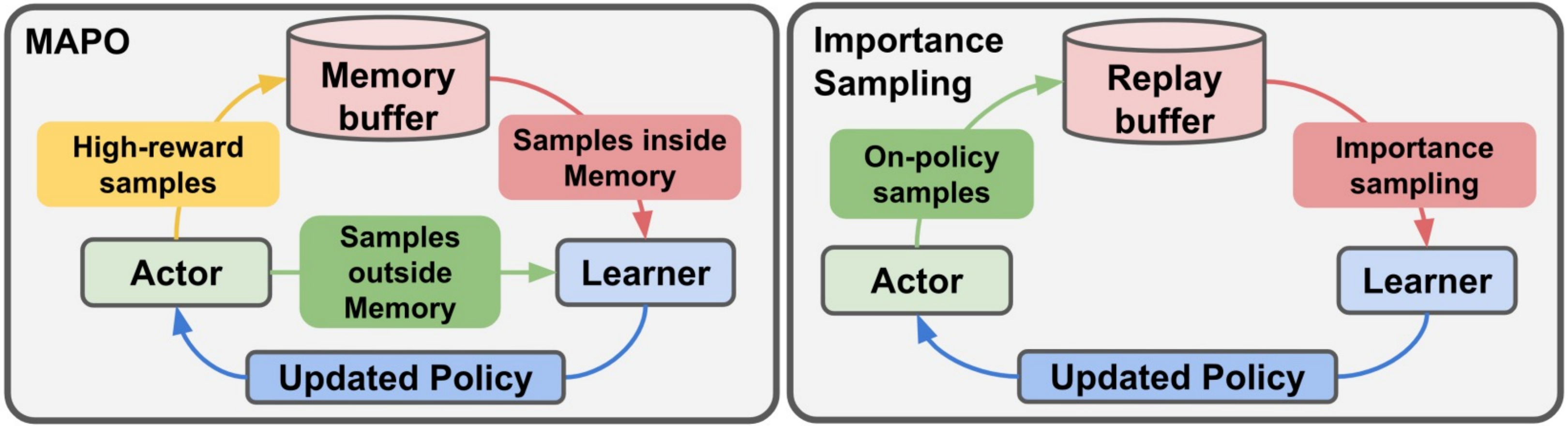
City	Country	Nations	Year
Athens	Greece	14	1896
St. Louis	USA	12	1904
…	…	…	…
Athens	Greece	201	2004
Beijing	China	204	2008

Question

Greece held its last Summer Olympics in which year

Previous Work: Reinforcement Learning

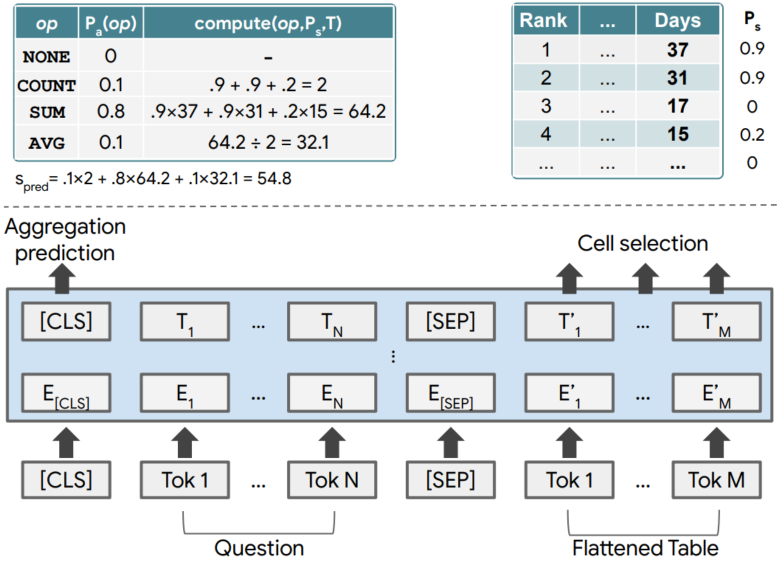
Obtain rewards by comparing execution results of sampled SQL queries with golden answers to train a text-to-SQL semantic parser. Hard to scale to complex scenarios.



[Chen et al. 2018]

Previous Work: Table Parsing

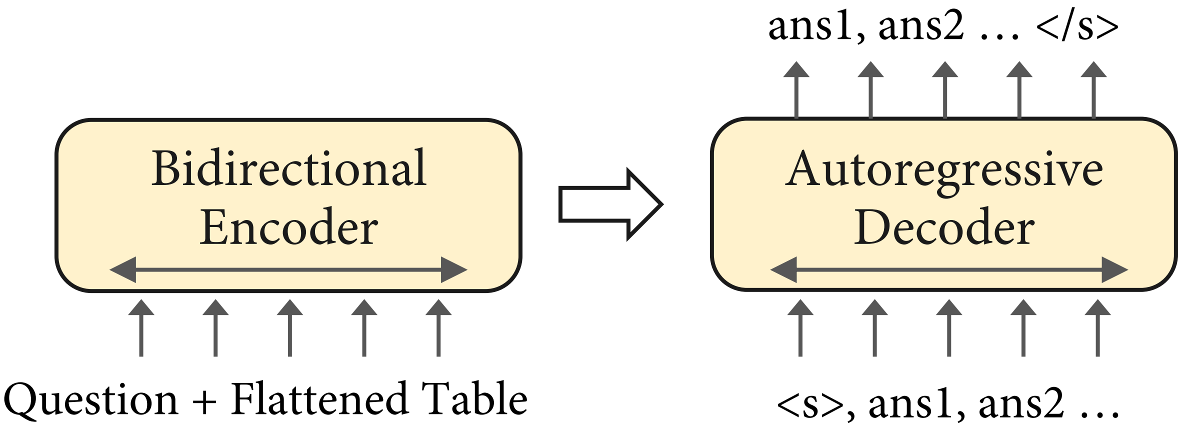
Predict answer by selecting table cell values and optionally applying an aggregation operator to the selected region. Flexibility is limited.



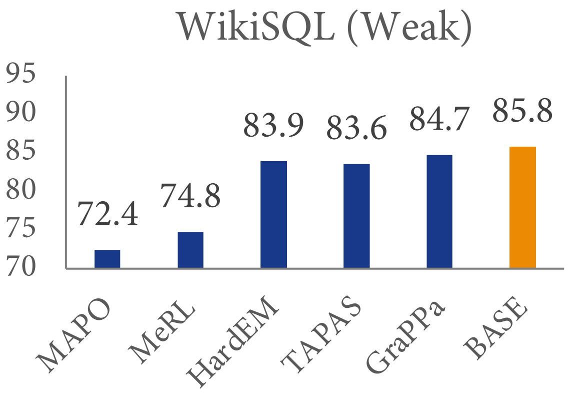
[Herzig et al. 2020]

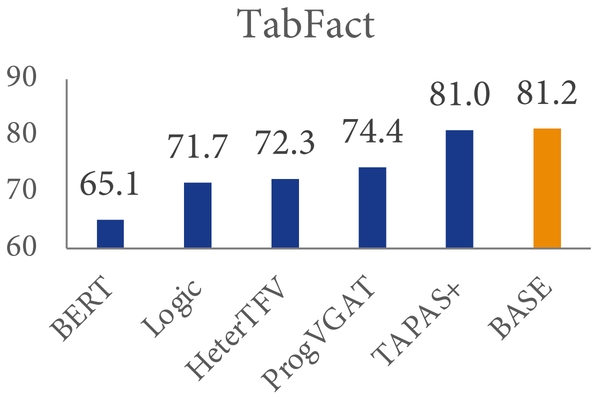
Preliminary: Generative Language Model

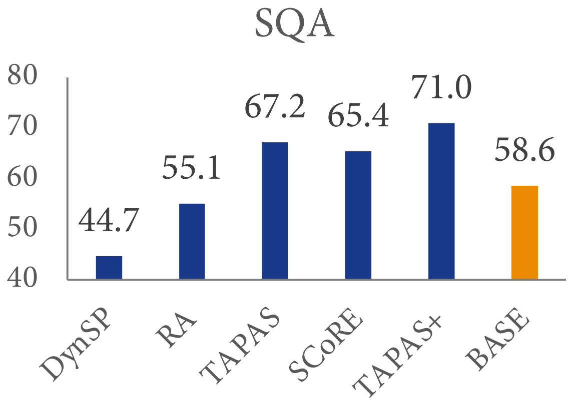
We formulate the task of table-based question answering as answer generation, and leverages generative language models (e.g., BART) to output autoregressively.

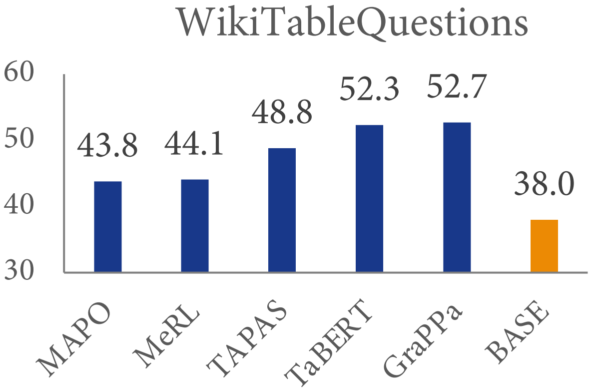


Preliminary Result: Models Are Data-Hungry



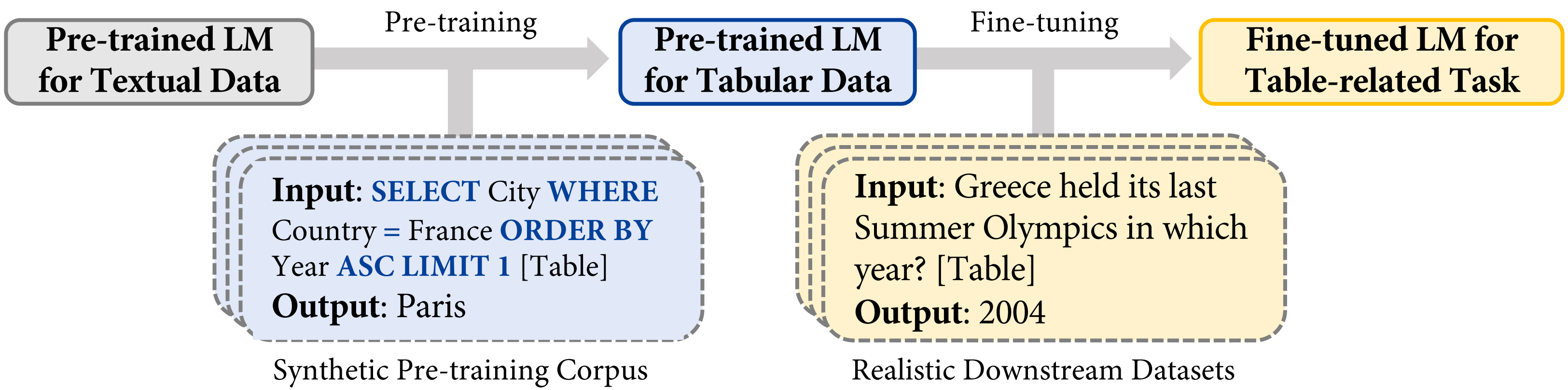






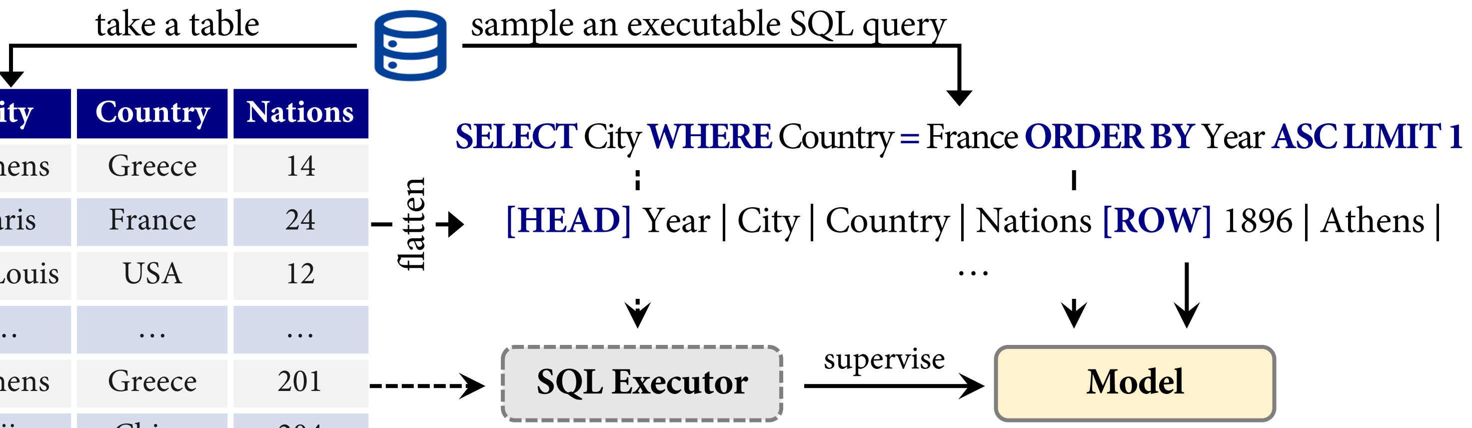
Method: SQL Execution Pre-training

Pre-training a model to mimic the behavior of a symbolic execution engine.



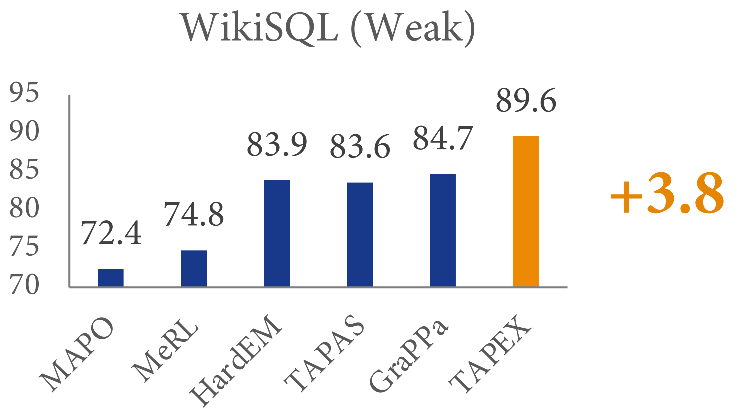
Method: SQL Execution Pre-training

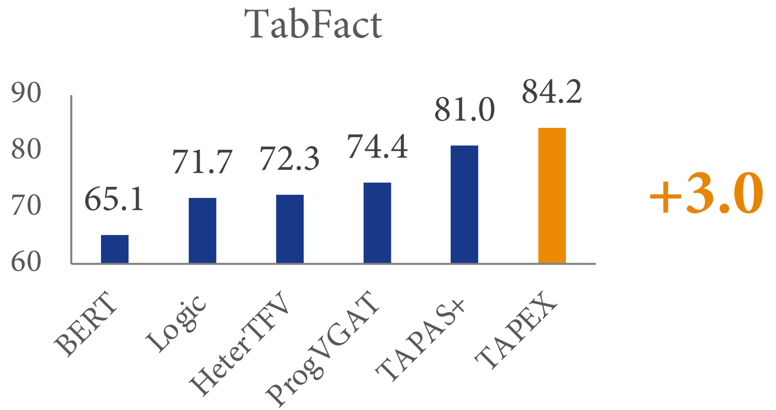
If we train a model to mimic the SQL query execution procedure over databases, we believe it learns programmatic reasoning from the execution engine.

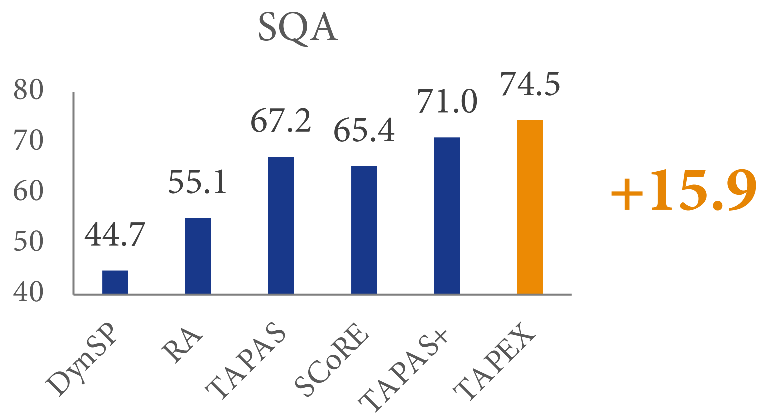


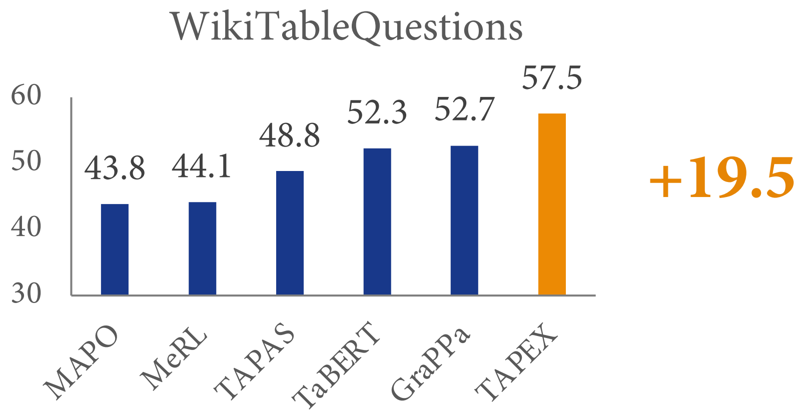
Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
…	…	…	…
2004	Athens	Greece	201
2008	Beijing	China	204

Experimental Result: Effective Pre-training

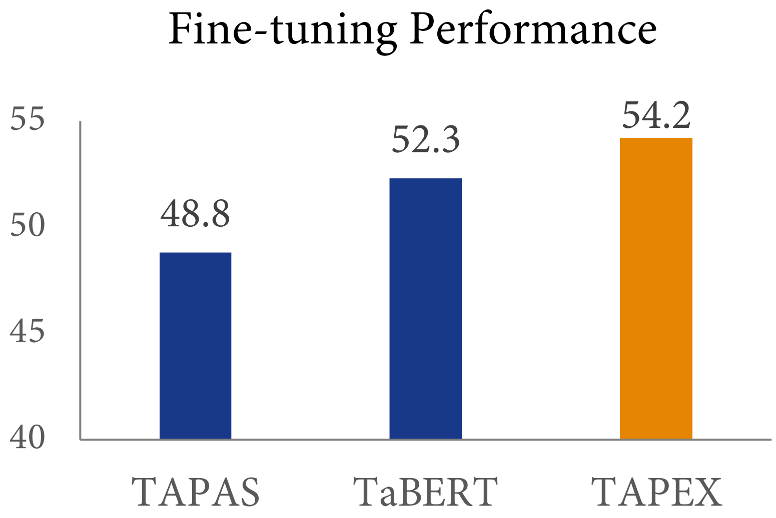


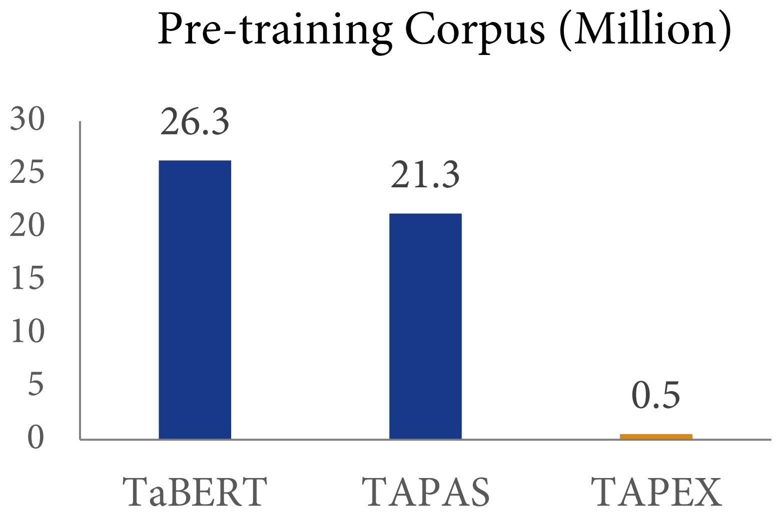






Experimental Result: Efficient Pre-training

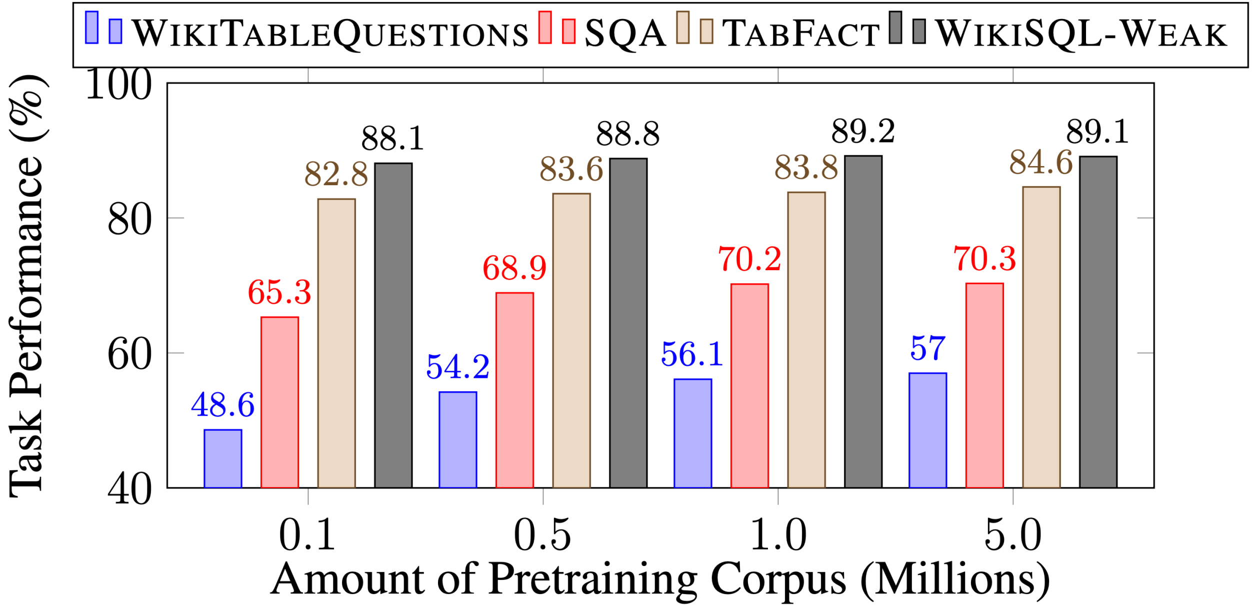




Compared with TaBERT, **2%** of corpus yields **2%** improvement!

Experimental Analysis: Larger is Better

Scaling up the pre-training corpus generally brings positive effects.



Experimental Analysis: Fine-grained Analysis

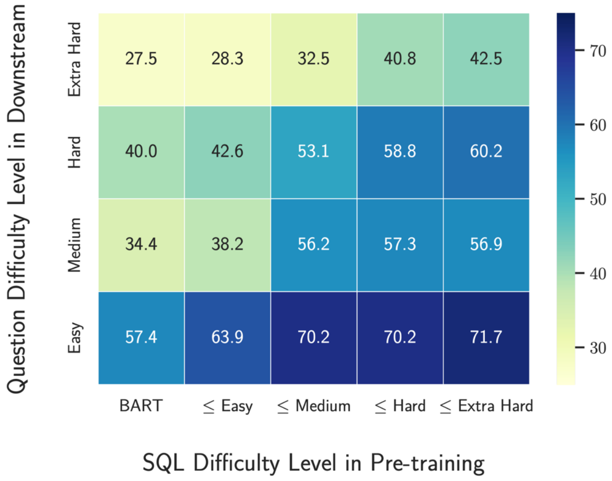
TAPEX significantly boosts the performance on all operators, implying that it does enhance BART's capabilities for joint reasoning over text and tables.

Operator	Example Question	BART	TAPEX
Select	What is **the years won** for each team?	41.3%	64.8% (+23.5%)
Filter	How long did **Taiki Tsuchiya** last?	40.1%	65.7% (+25.6%)
Aggregate	What is the **amount of** matches drawn?	26.9 %	57.4% (+30.5%)
Superlative	What was the **last** Baekje Temple?	46.3 %	64.3% (+18.0%)
Arithmetic	What is the **difference** between White voters and Black voters in 1948?	33.1 %	53.5% (+20.4%)
Comparative	Besides Tiger Woods, what other player won **between 2007 and 2009**?	30.0 %	55.9% (+25.9%)
Group	What was score **for each** winning game?	49.5 %	66.7% (+17.2%)

Experimental Analysis: Complexity

Adding simpler SQL queries can improve performance on harder questions.

Difficulty	Example SQL Query
Easy	SELECT Date SELECT COUNT (Canal) SELECT Name WHERE Age >= 28
Medium	SELECT Region ORDER BY ID DESC LIMIT 1 SELECT COUNT (Tornadoes) WHERE Date = 1965 SELECT District WHERE District != "Tikamgarh" AND Agg = 0
Hard	SELECT (SELECT COUNT(Distinct Area)) >= 5 SELECT COUNT (*) WHERE Result = "won" AND Year > 1987 SELECT Driver WHERE Manufacturer = "t-bird" ORDER BY Pos ASC LIMIT 1
Extra Hard	SELECT COUNT (*) WHERE Position = 1 AND Notes = "110 m hurdles" AND Year > 2008 SELECT Nation WHERE Nation != "Japan" AND Gold = (SELECT Gold WHERE Nation = "Japan") SELECT Tournament WHERE Tournament IN ("oldsmar", "los angeles") GROUP BY Tournament ORDER BY COUNT (*) DESC LIMIT 1



Experimental Analysis: Naturalness

However, replacing SQL with NL does not benefit the pre-training, because the translated NL sentences contain noise.

SQL Query	Translated NL Sentence	Faithfulness
SELECT Name WHERE Age >= 28	Who is at least 28 years old?	✓
SELECT MAX (Pick#)	What was the last pick in the 1989 major league baseball draft?	✗
SELECT Driver ORDER BY Pos DESC LIMIT 1	What driver came in last place?	✓
SELECT COUNT (Competition) WHERE Notes != 100	How many competitions have no notes?	✗
SELECT COUNT (*) WHERE Result = "won" AND Year > 1987	How many times did they win after 1987?	✓
SELECT MAX (Chart Position) – MIN (Chart Position) WHERE Release date = "july 21, 1995"	What is the difference between the chart position of july 21, 1995 and the chart position of july 22, 1995?	✗
SELECT Nation WHERE Nation != "Japan" AND Gold = (SELECT Gold WHERE Nation = "Japan")	Which other countries had the same number of gold medals as Japan?	✓
SELECT Incumbent Electoral History GROUP BY Incumbent Electoral History ORDER BY COUNT (*) DESC LIMIT 1	Who has held the office the most?	✗

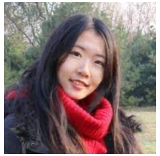
Take Away: Pre-training without Real Data

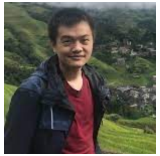
When performing continual pre-training, instead of mining a large noisy web corpus, we can also try to synthesize an accurate and small corpus.

Take Away: Pre-training without Language Modeling

When performing continual pre-training, instead of performing the general-purpose language modeling, we can also try to simulate the specialized skill.

POET: Reasoning Like Program Executor



*Xinyu Pi[1] *Qian Liu[2] Bei Chen[3] Morteza Ziyadi[3] Zeqi Lin[3] Yan Gao[3] Qiang Fu[3]


Jian-Guang Lou[3] Weizhu Chen[3]

1 UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
2 北京航空航天大學 BEIHANG UNIVERSITY
3 Microsoft

Background: Numerical Reasoning

Document

In **1517**, the seventeen-year-old King sailed to Castile. There, his Flemish court … In **May 1518,** Charles traveled to Barcelona in Aragon.

Question

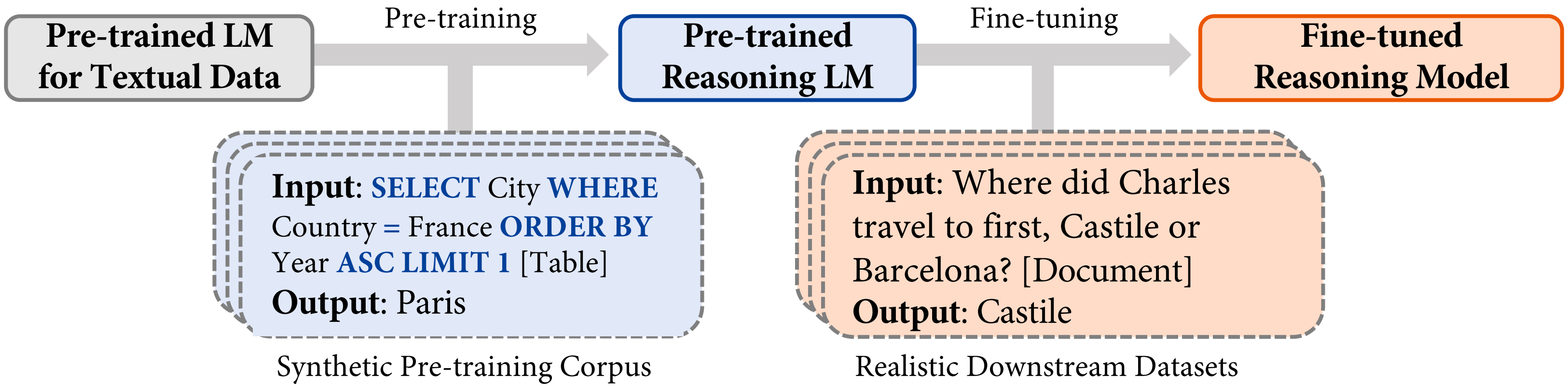
Where did Charles travel to first, Castile or Barcelona?

Answer

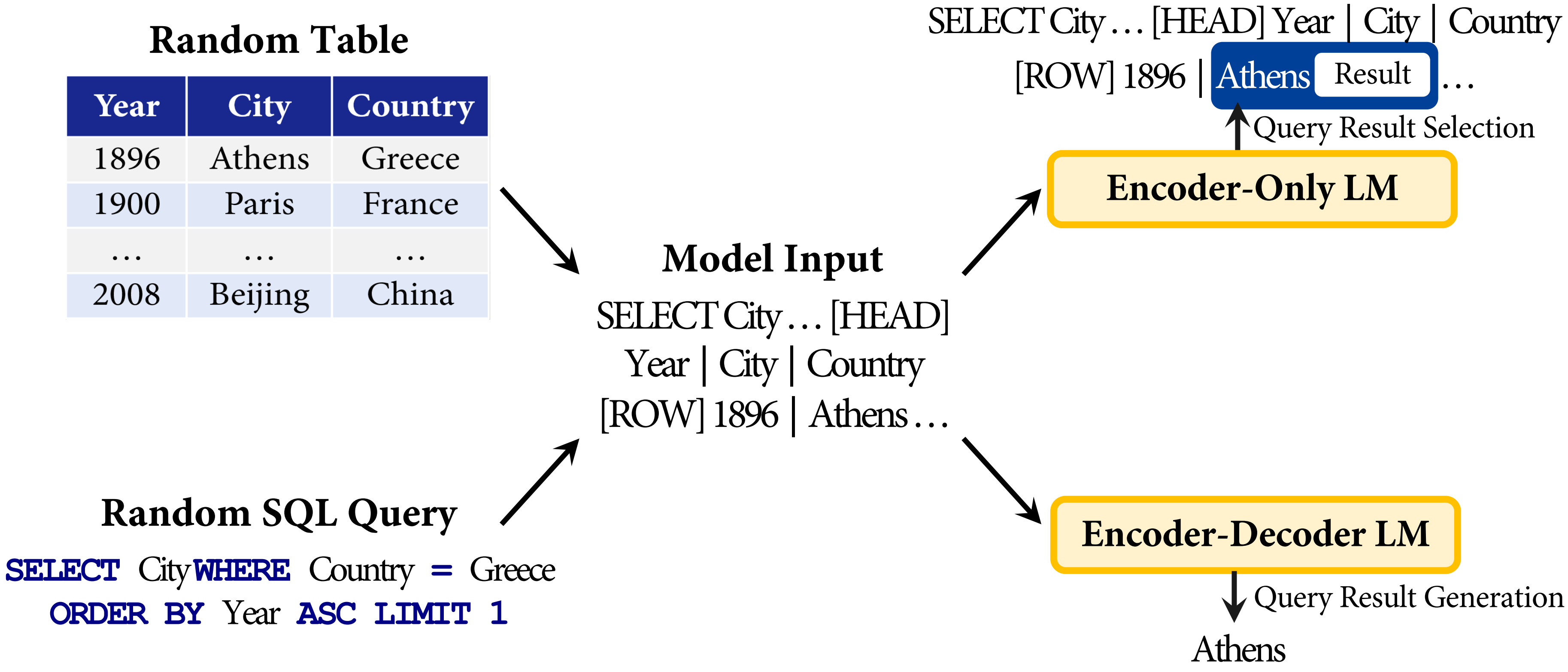
Castile

Method: SQL Execution Pre-training

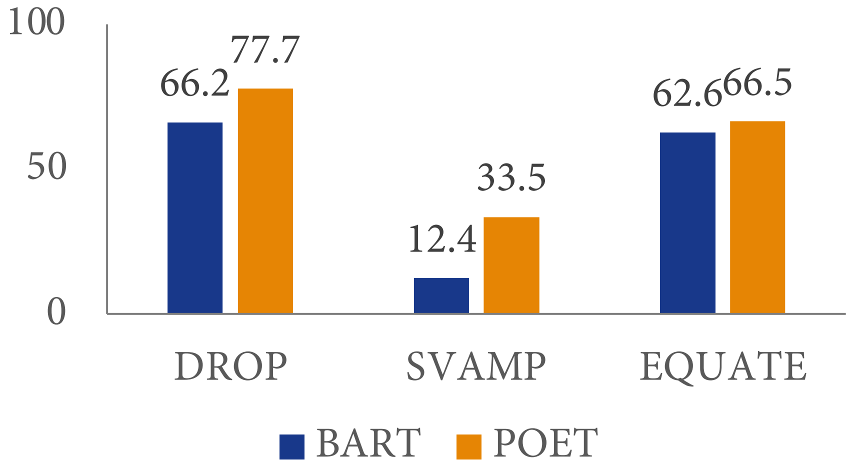
Since SQL queries involve rich numerical operations, we hope it can be leveraged to enhance the numerical reasoning capability of models on documents.

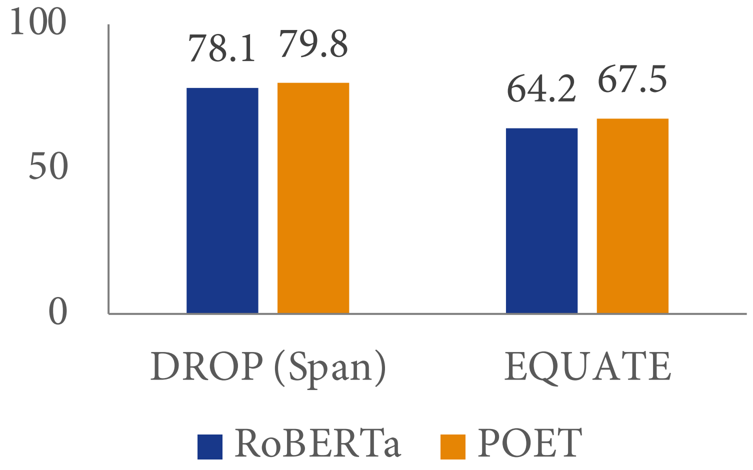


Method: SQL Execution for Different LMs



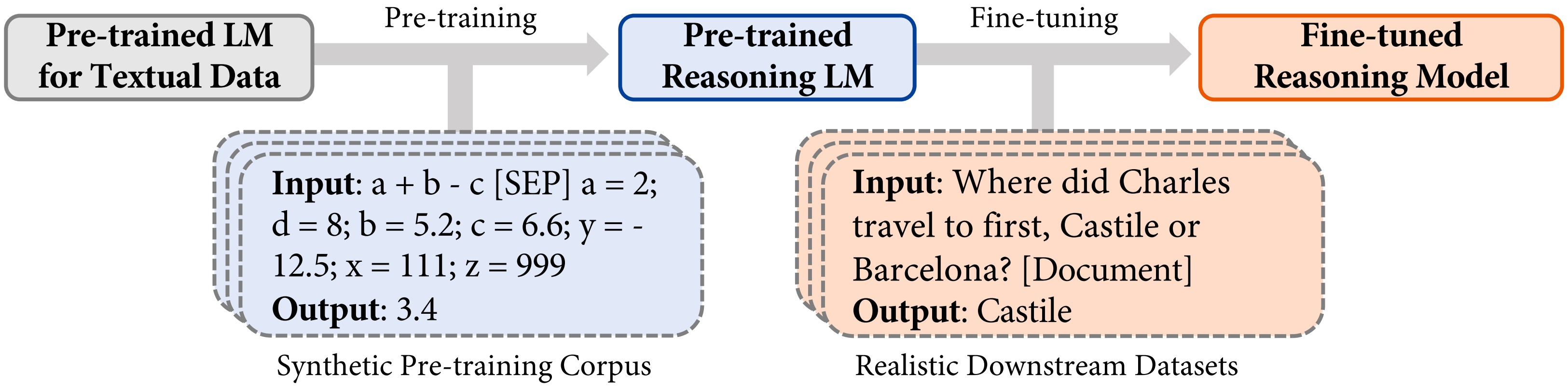
Experimental Result: Reasoning Transfer





Method: Math Expression Calculation

Observing the reasoning transfer from (SQL query, Database) to (Question, Passage), we propose a simplified method which leverages math expression for pre-training.

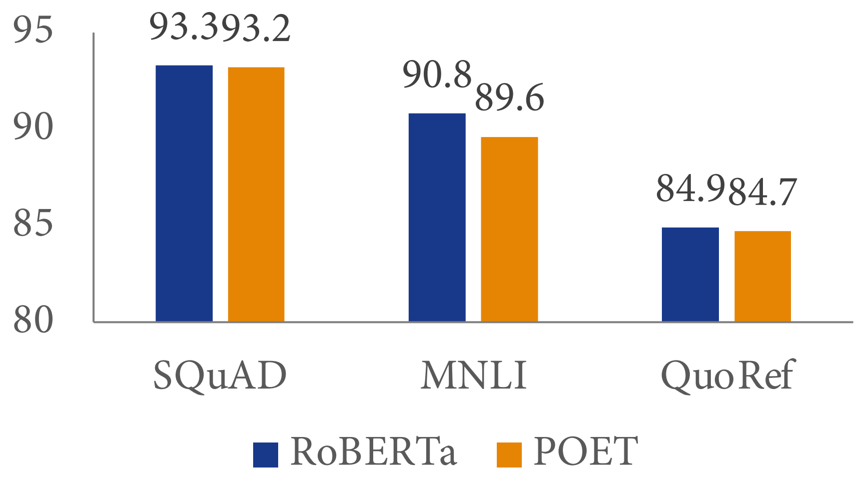


F1 on DROP dataset based on BART

69.2% **78.1%**

Experimental Analysis: Performance Hurt on Other Tasks?

Small (<1%). POET barely sacrifices the intrinsic understanding ability of language models.



Experimental Analysis: Benefit from Similarity of SQL to NL?

NO. Randomly mapping SQL keywords to the "strange" tokens still works well.

SELECT City FROM table ORDER BY Population DESC LIMIT 1 → ***Fine-tune*** → Which country contains the second largest part of the forest

77.7%

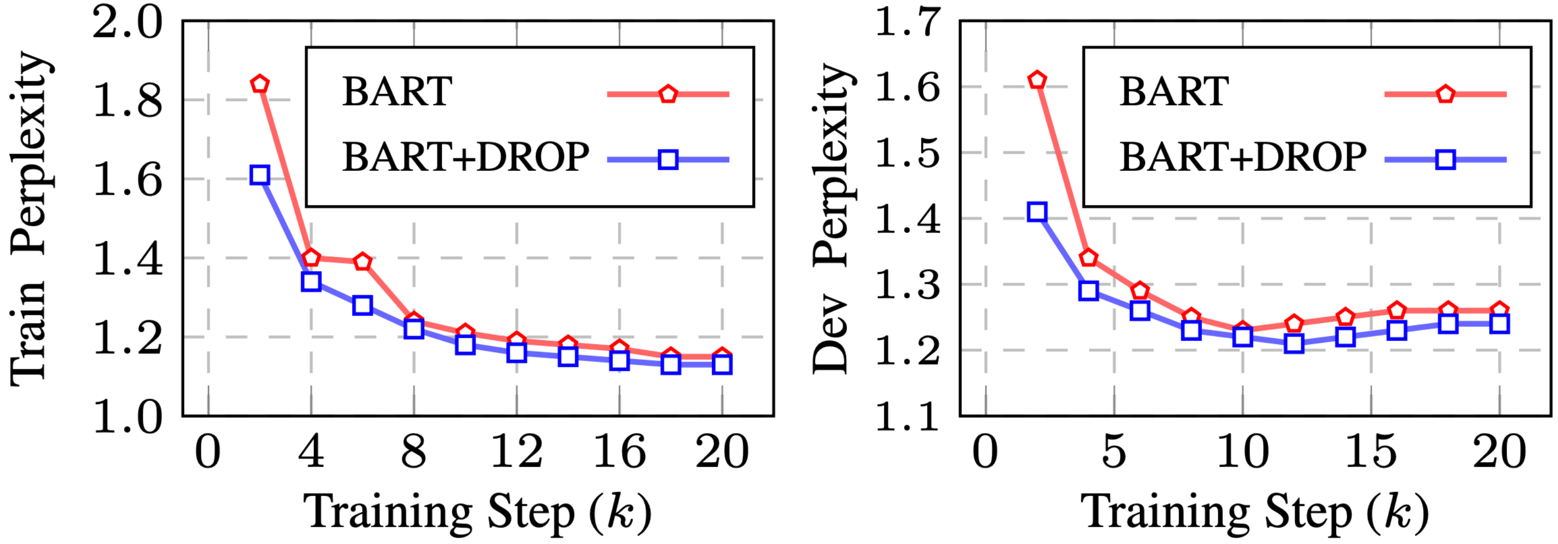
≈

unfocusedRange City guiIcon table externalToEVA awdownload Population ffffcc awdownloadclon 1 → ***Fine-tune*** → Which country contains the second largest part of the forest

76.9%

Experimental Analysis: **Pre-training on DROP Benefit SQL Execution?**

Yes. Pre-training on DROP leads to observably lower perplexity for SQL execution learning on both the train and dev sets.



Experimental Analysis: How Does it Work?

No answer. But we can get some insights from the following analogy.

Math Expression
$x + y - z$

Program

≈

Natural Language

Question
Where did Charles travel to first?

Variable Memory
$x = 152.0$; $y = 99.0$; $z = 70.3$;

Program Context

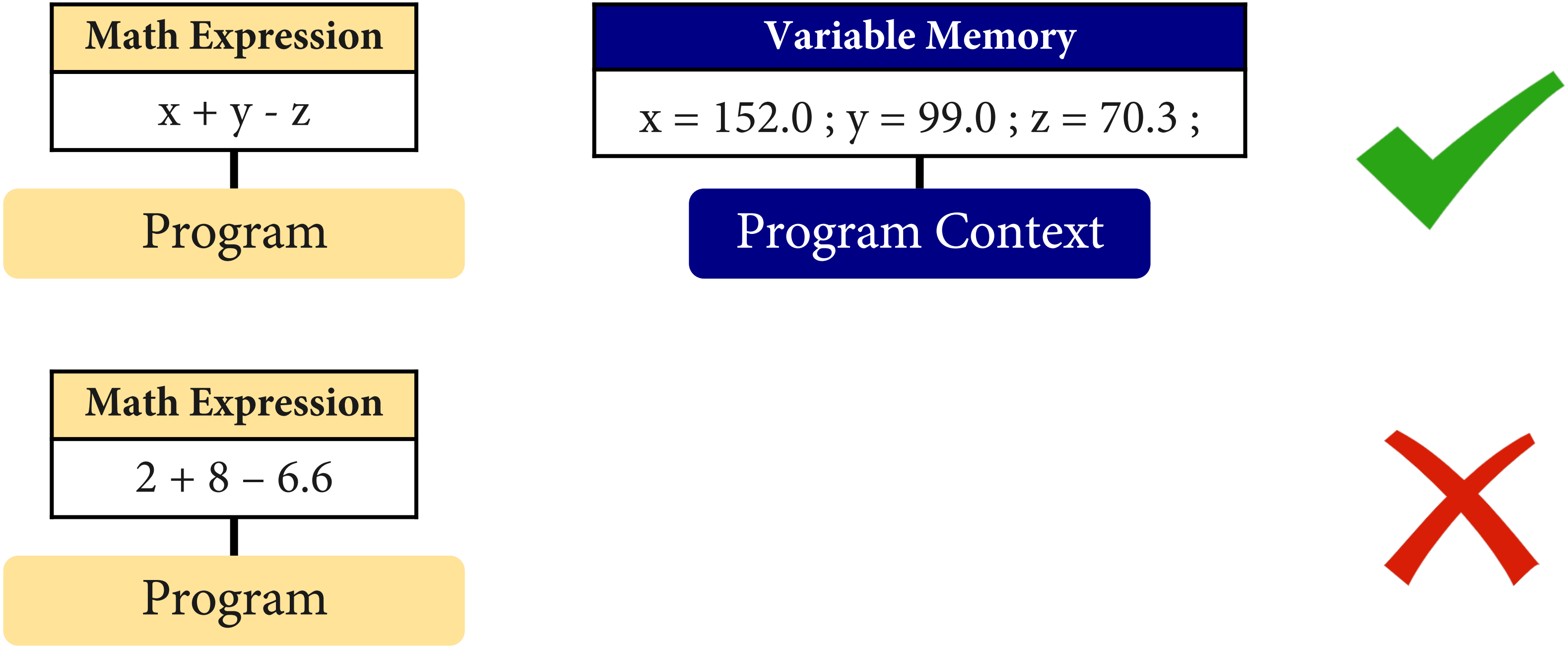
≈

Natural Context

Document
In 1517, the seventeen-year-old…

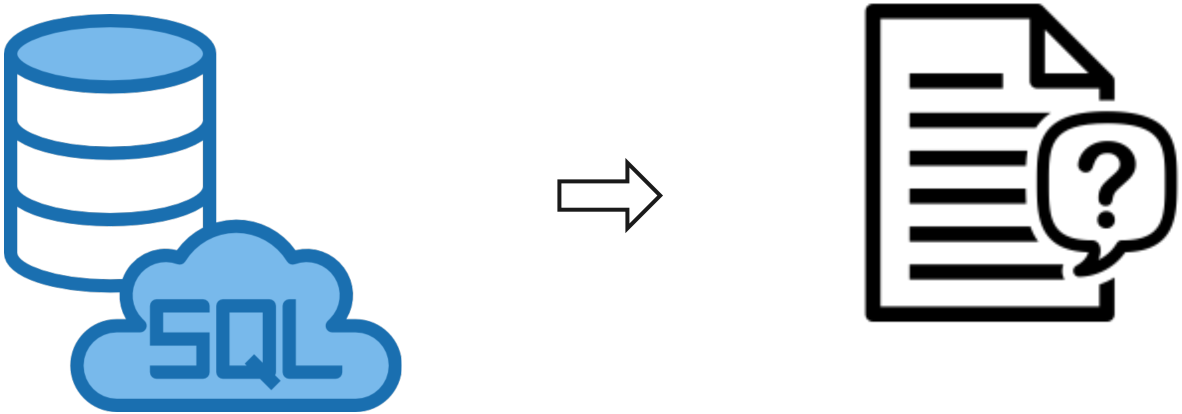
Experimental Analysis: How Does it Work?

Without program context, the pre-training cannot work well.



Take Away: Reasoning Transfer Occurs Across Modalities

Reasoning transfer occurs across modalities, and the analogy between pre-training and fine-tuning is important for the transference.



LEMON: Language-Based Environment Manipulation via Execution-Guided Pre-training


Qi Shi[1]


Qian Liu[2]


Bei Chen[3]


Yu Zhang[1]


Ting Liu[1]


Jian-Guang Lou[3]

1 哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY

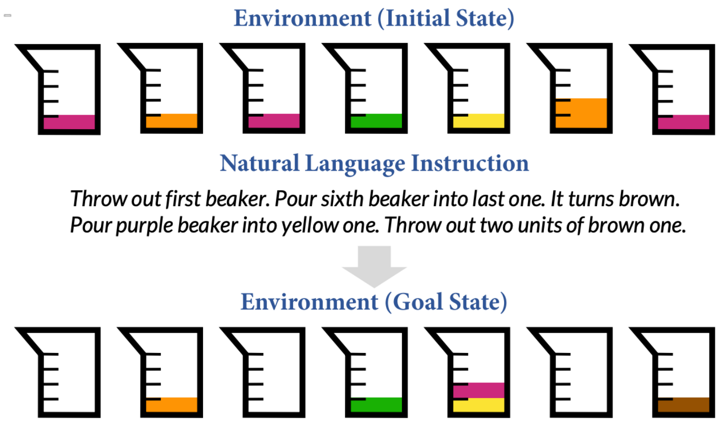
2 北京航空航天大學 BEIHANG UNIVERSITY

3 Microsoft

Background: Language-Based Environment Manipulation

Agents are required to manipulate the environments based on the natural language.

Instruction Following

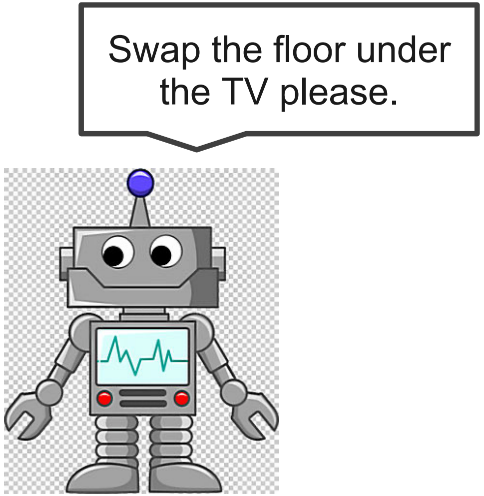


Procedural Text Understanding

		Participants:				
Paragraph (seq. of steps):		**water**	**light**	**CO2**	**mixture**	**sugar**
	state0	**soil**	**sun**	**?**	**-**	**-**
Roots absorb water from soil						
	state1	**roots**	**sun**	**?**	**-**	**-**
The water flows to the leaf.						
	state2	**leaf**	**sun**	**?**	**-**	**-**
Light from the sun and CO2 enter the leaf.						
	state3	**leaf**	**leaf**	**leaf**	**-**	**-**
The light, water, and CO2 combine into a mixture.						
	state4	**-**	**-**	**-**	**leaf**	**-**
Mixture forms sugar.						
	state5	**-**	**-**	**-**	**-**	**leaf**

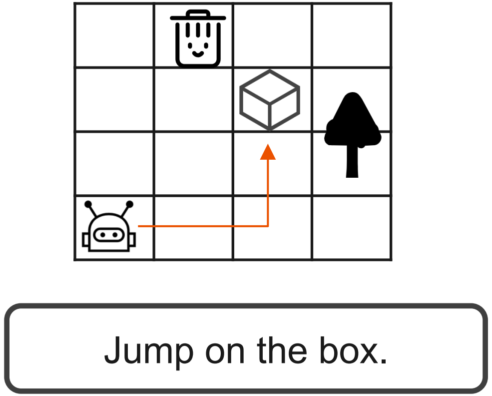
Time

Application: Language-Based Environment Manipulation



Agent Control

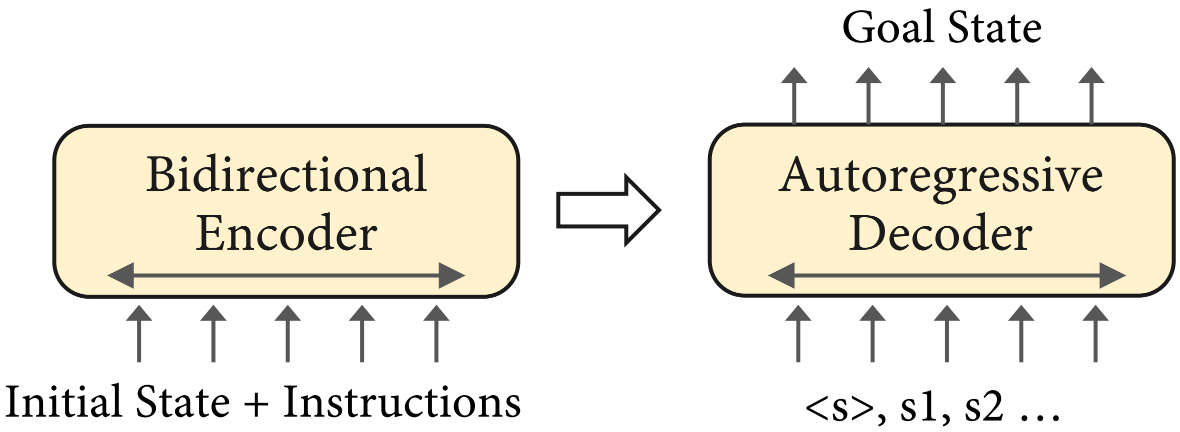
State Tracking



Virtual Interaction

Preliminary: Generative Language Model Again

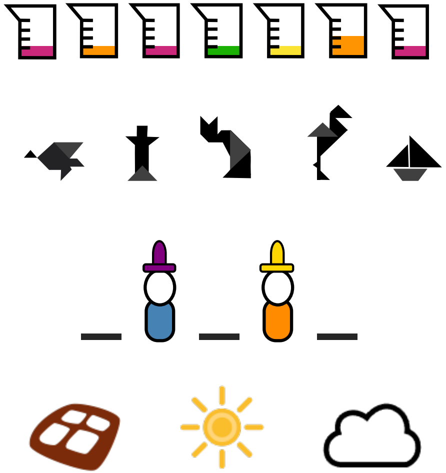
We formulate the task as a seq2seq paradigm, by leveraging generative PLMs (e.g., BART) to generate goal states directly.



Challenge: Spatial Reasoning

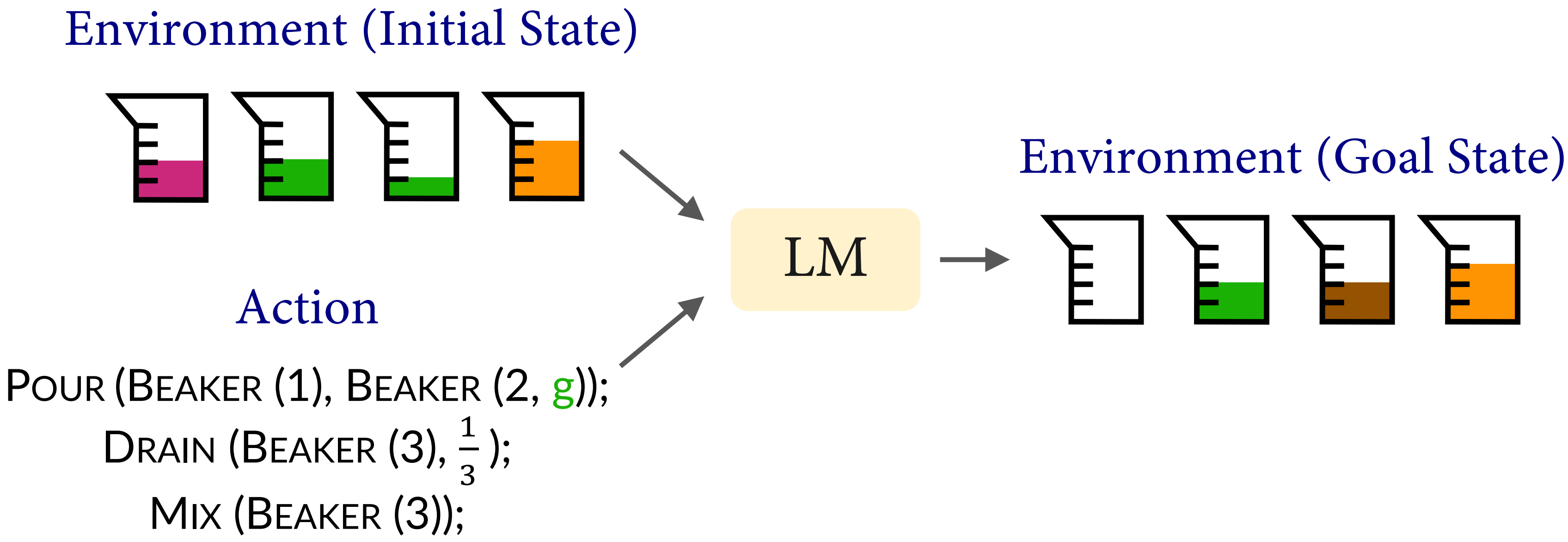
Since pre-trained language models does not observe environments before, it is difficult for them to perform accurate spatial reasoning.

 What are these?

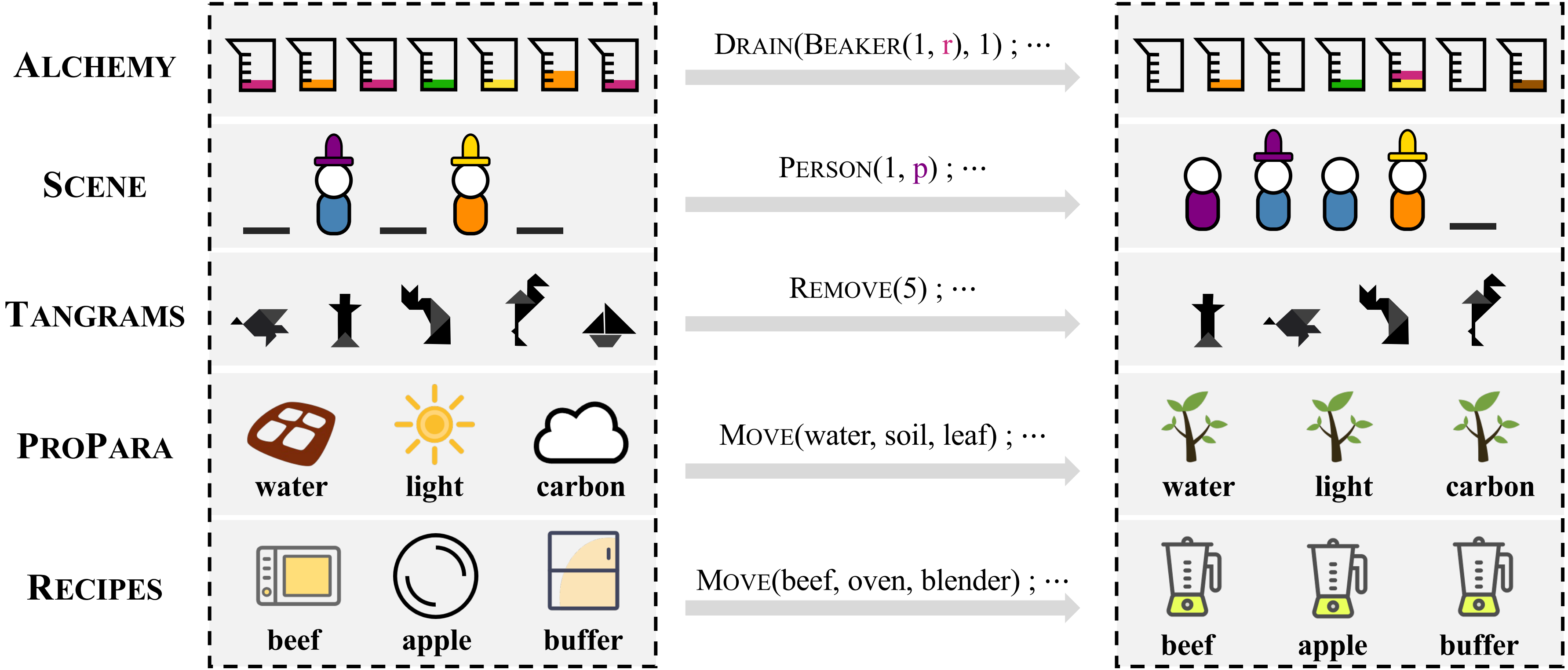


Motivation: Environment Exploration by Actions

Synthesizing diverse actions to drive LMs familiar with environments.



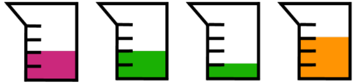
Method: Environment Exploration by Actions



Method: Environment Exploration by Actions

Pre-training

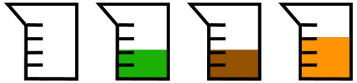
Environment (Initial State)



Program

POUR (BEAKER (1), BEAKER (2, g));
DRAIN (BEAKER (3), $\frac{1}{3}$);
MIX (BEAKER (3));

Environment (Goal State)



Fine-tuning

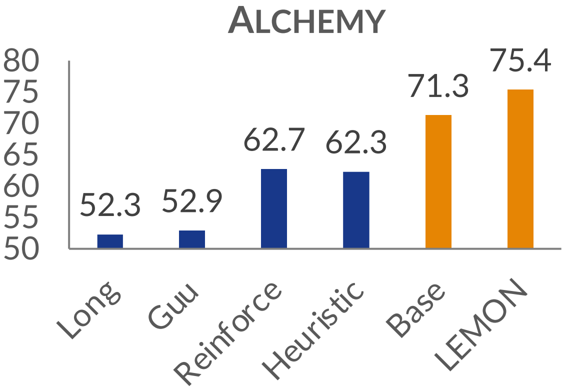
Environment (Initial State)

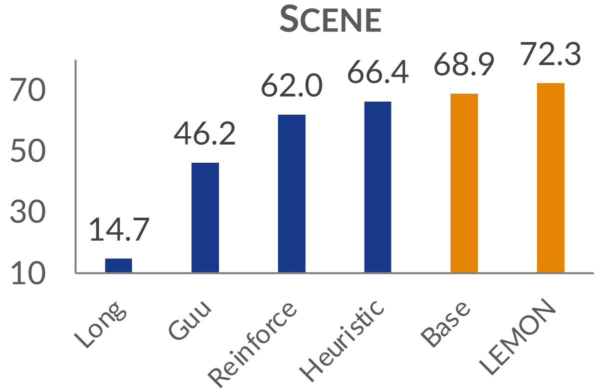
Natural Language Instruction

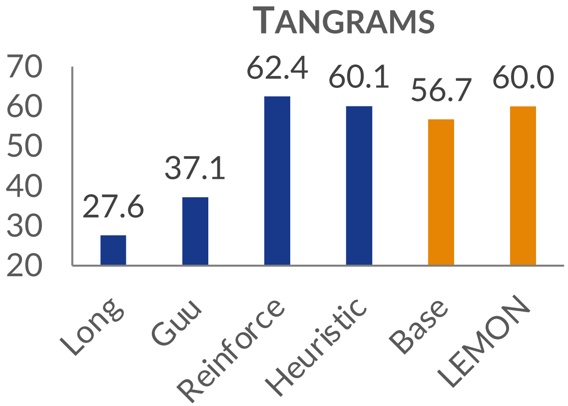
Throw out first beaker. Pour sixth beaker into last one. It turns brown. Pour purple beaker into yellow one. Throw out two units of brown one.

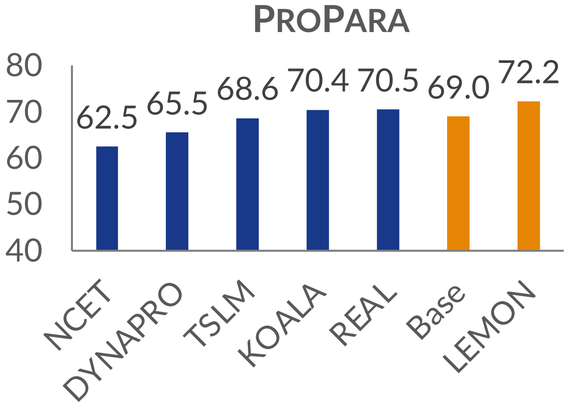
Environment (Goal State)

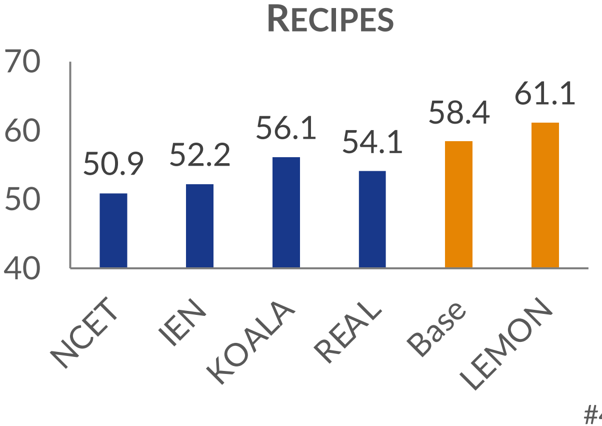
Experimental Result: SOTA on Five Benchmarks



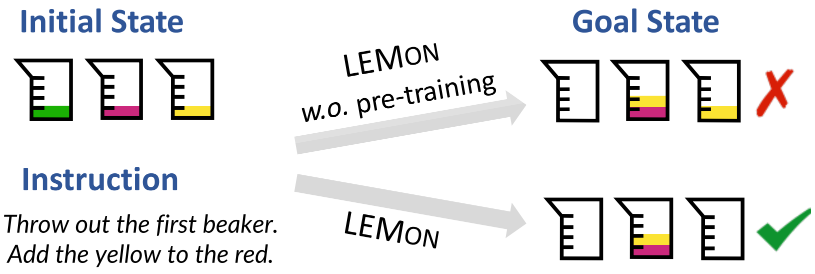




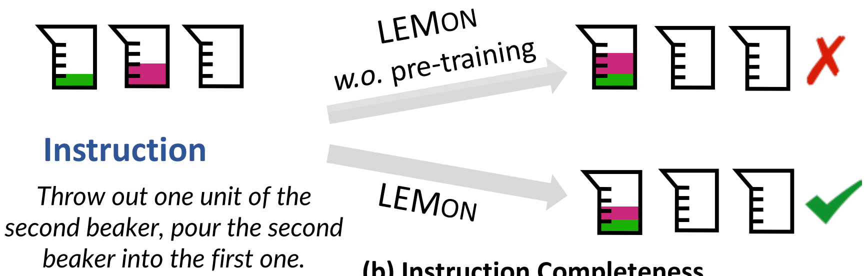




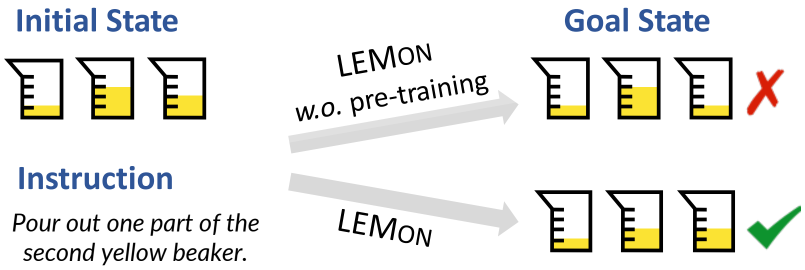
Experimental Analysis: What Does LEMON Learn?



(a) Operation Legality



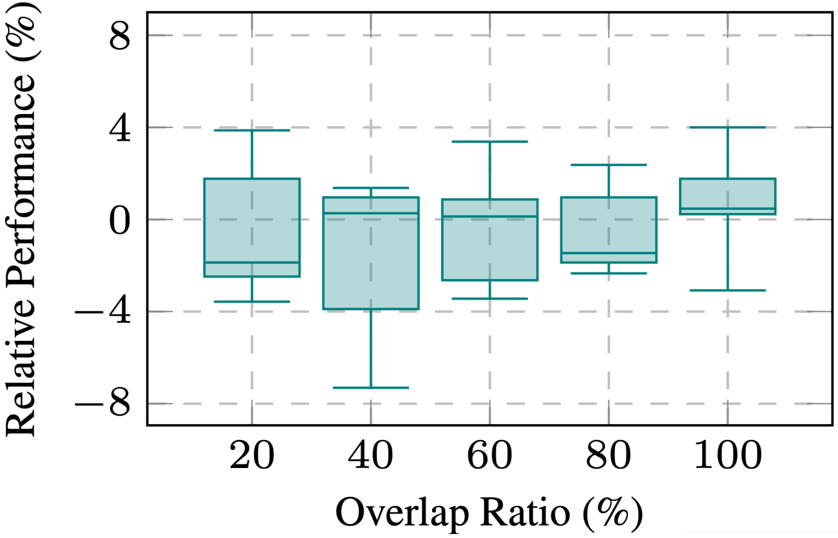
(b) Instruction Completeness



(c) Grounding Correctness

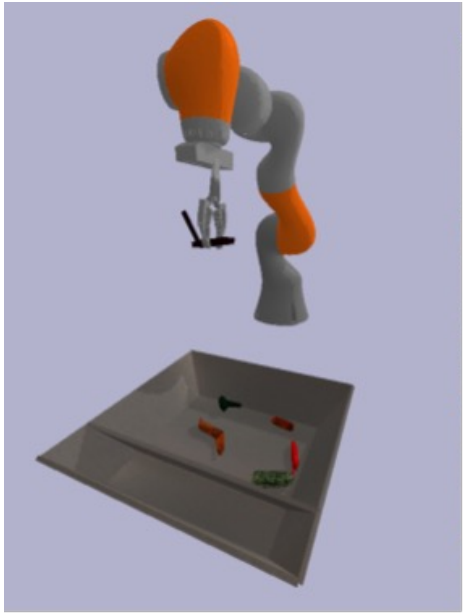
Experimental Analysis: Improvements from Leakage?

No. The box plot of the relative performance (vertical axis) with respect to the overlap ratio (horizontal axis) indicates the independence.

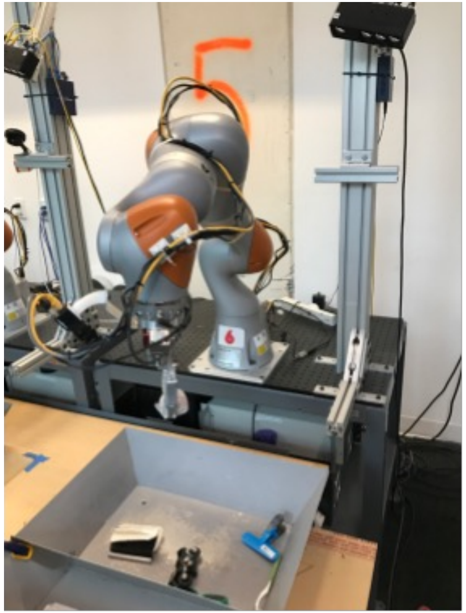


Take Away: Actions v.s. Simulation

Simulation to reality is a popular technique in autopilot. Actions can be regarded as kind of simulations which can facilitate the spatial reasoning in real space.



⇨



Thanks & QA

Qian Liu（刘乾）
Research Scientist @ Sea AI Lab