Language, Knowledge, and Reasoning: Semantic Parsing in the Era of Large Language Models

Rui Zhang, Assistant Professor, Penn State University
https://ryanzhumich.github.io/

CSE Faculty Lunch Seminar on Dec 1, 2022

About Myself and PhD Students in My Lab

- Penn State University (2020 -)
- Yale University (2017 - 2020)
- University of Michigan, Ann Arbor (2013 - 2017)
- Shanghai Jiao Tong University (2011 - 2013)

Yusen Zhang

https://yuszh.com/

Fall 2021 -

Research Area
- Summarization
- Semantic Parsing

Sarkar Das

https://sarathismg.github.io/

Spring 2021 -

Research Area
- Few-shot Learning
- Efficient NLP Methods
- Sequence Labeling

Nan Zhang

https://zn1010.github.io/

Fall 2020 -

Research Area
- Medical Summarization
- Faithful NLG

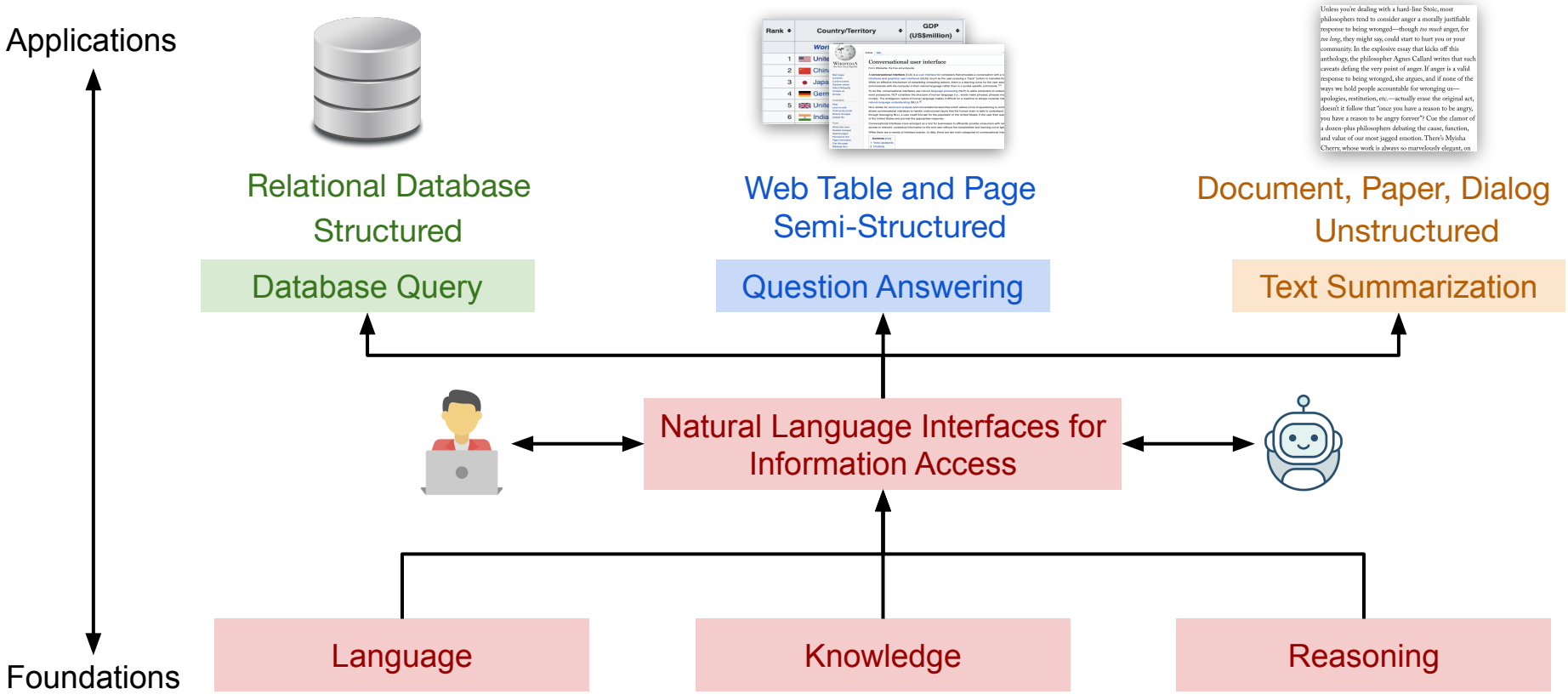
Haoran Zhang

https://windchimeran.github.io/

Spring 2022 -

Research Area
- Information Extraction
- Indirect Supervision

My Research: Natural Language Interfaces for Information and Knowledge



Semantic Parsing

Semantic parsing

From Wikipedia, the free encyclopedia

Semantic parsing is the task of converting a natural language utterance to a logical form: a machine-understandable representation of its meaning.

Natural Language Interfaces to Databases

Q: Which European countries have players who won the Australian Open at least 3 times?

Table 1: Matches

Id	Tourney	Year	Winner id	… …
1	Australian Open	2018	3	… …

Table 2: Ranking

Ranking	Points	Player id	Tours	… …
1	9,985	3	11	… …

Table 3: Players

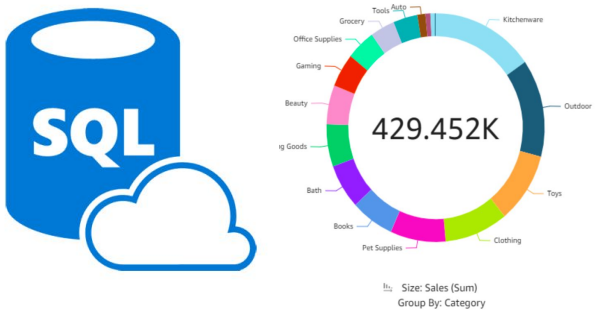
Id	Name	Nation	Continent	… …
1	Djokovic	Serbia	Europe	… …
2	Osaka	Japan	Asia	… …

Text-to-SQL Semantic Parsing

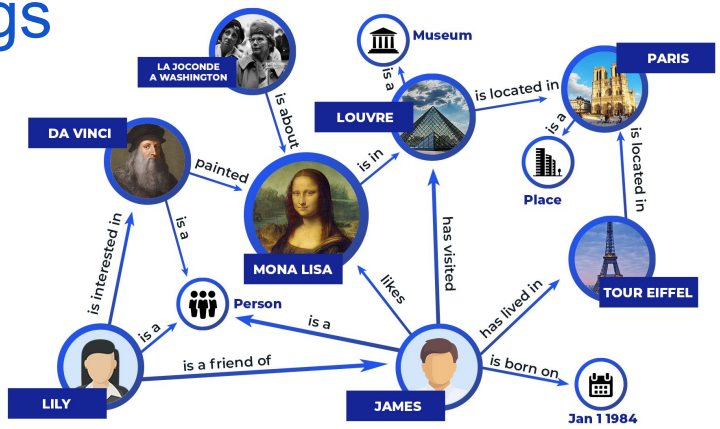
```
SELECT T1.nation
FROM players AS T1 JOIN matches AS T2
  ON T1.id = T2.winner_id
WHERE T2.Tourney = "Australian Open"
  AND T1.continent = "Europe"
GROUP BY T2.winner_id
HAVING COUNT(*) >= 3
```

Switzerland
Serbia
...

Semantic Parsing for Many Things



Text-to-SQL for Natural Language Interfaces to Databases



Question Answering over Knowledge Graphs

Instruction Following for Robotics

Problem

H-Index

Given a list of citations counts, where each citation is a nonnegative integer, write a function h_index that outputs the h-index. The h-index is the largest number h such that h papers have each least h citations.

Example:
Input: [3,0,6,1,4]
Output: 3

Generated Code

```
def h_index(counts):
  n = len(counts)
  if n > 0:
    counts.sort()
    counts.reverse()
    h = 0
    while (h < n and
        counts[h]-1>=h):
      h += 1
    return h
  else:
    return 0
```

Test Cases

```
Input:
[1,4,1,4,2,1,3,5,6]

Generated Code Output:
4                          ✓

Input:
[1000,500,500,250,100,
100,100,100,100,75,50,
30,20,15,15,10,5,2,1]

Generated Code Output:
15                         ✓
```

Language-to-Code Generation

Semantic Parsing History: From Leibniz to Symantec

Leibniz (1685) developed a **formal conceptual language**, the characteristica universalis, for use by an automated reasoner, the calculus ratiocinator.

Bertrand Russell and Alfred North Whitehead (Principia Mathematica, 1913) finalized the development of **modern first-order predicate logic (FOPC)**.

Richard Montague (1970) developed a formal method for **mapping natural language to FOPC** using Church's lambda calculus.

Bill Woods (1973) developed the **first NL database interface (LUNAR)** to answer scientists' questions about moon rocks 12 using a manually developed Augmented Transition Network (ATN) grammar.

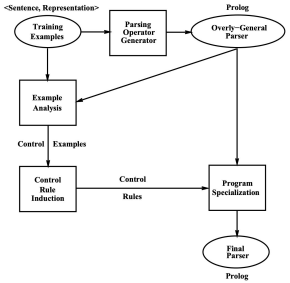
Dave Waltz (1975) developed the **next NL database interface (PLANES)** to query a database of aircraft maintenance for the US Air Force.

Gary Hendrix founded **Symantec ("semantic technologies") in 1982 to commercialize** NL database 14 interfaces based on manually developed semantic grammars, but they switched to other markets when this was not profitable.

Figure Adapted from Ray Mooney

Learning-based Semantic Parsing Research

Zelle and Mooney (1996)

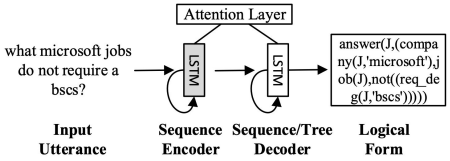


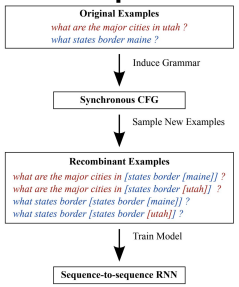
We now turn to issues of parsing and parameter estimation. Parsing under a PCCG involves computing the most probable logical form L for a sentence S,

$$\arg\max_L P(L|S;\bar{\theta}) = \arg\max_L \sum_T P(L,T|S;\bar{\theta})$$

Zettlemoyer and Collins (2005)

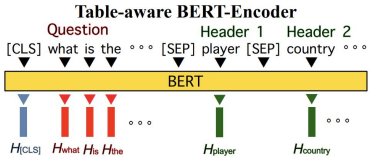
Dong and Lapata (2016)

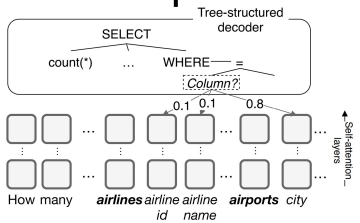




Jia and Liang (2016)

Hwang et al. (2019)

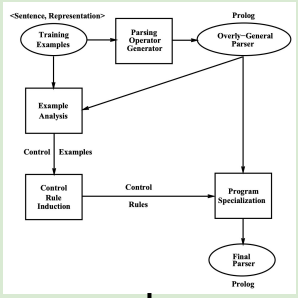




Wang et al. (2019)

Semantic Parsing Research: Paradigm Shifts

Zelle and Mooney (1996)

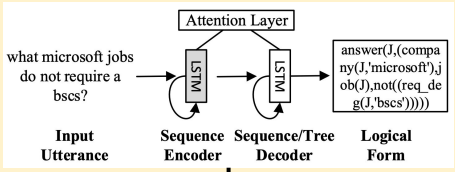


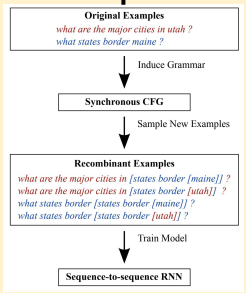
We now turn to issues of parsing and parameter estimation. Parsing under a PCCG involves computing the most probable logical form L for a sentence S,

$$\arg\max_L P(L|S;\bar{\theta}) = \arg\max_L \sum_T P(L,T|S;\bar{\theta})$$

Zettlemoyer and Collins (2005)

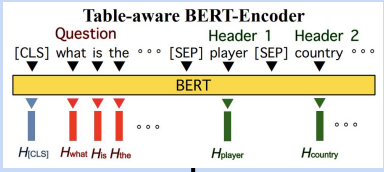
Dong and Lapata (2016)

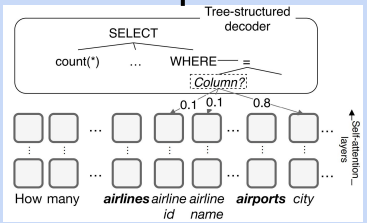




Jia and Liang (2016)

Hwang et al. (2019)





Wang et al. (2019)

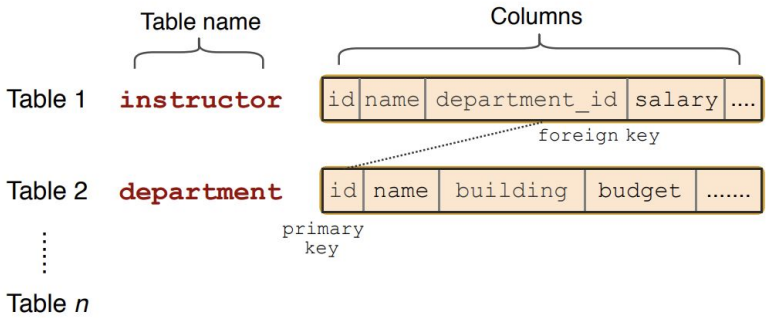
Non-Neural Network

E2E Neural Networks

Contextualized Embeddings and Pretrained Language Models

Delicate Data Curation

Annotators check database schema (e.g., database: `college`)



Annotators create:

Complex question: What are the name and budget of the departments with average instructor salary greater than the overall average?

Complex SQL:

```
SELECT T2.name, T2.budget
FROM instructor as T1 JOIN department as
T2 ON T1.department_id = T2.id
GROUP BY T1.department_id
HAVING avg(T1.salary) >
    (SELECT avg(salary) FROM instructor)
```

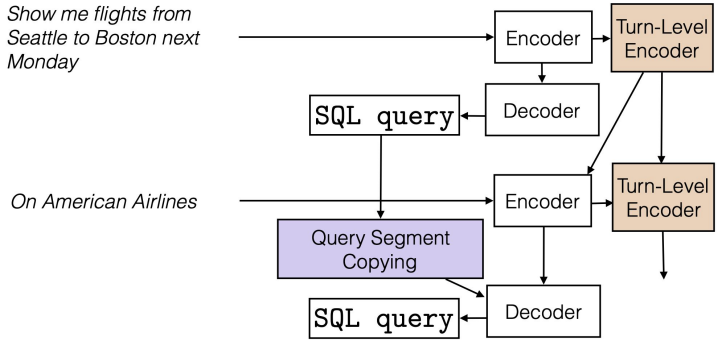
To address the need for a large and high-quality dataset for a new complex and cross-domain semantic parsing task, we introduce *Spider*, which consists of 200 databases with multiple tables, 10,181 questions, and 5,693 corresponding complex SQL queries, all written by 11 college students spending a total of 1,000 man-hours.

https://yale-lily.github.io/spider

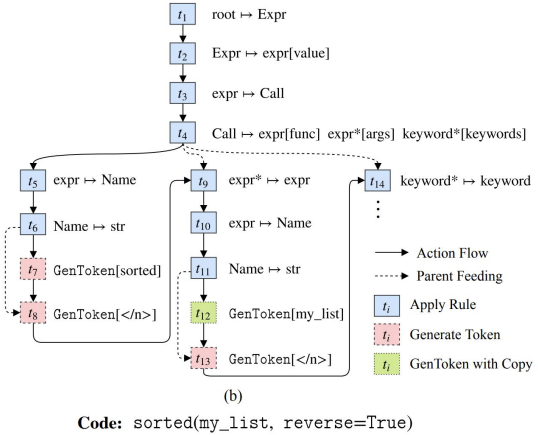
Overspecialized Model Design

Rules		Categories produced from logical form
Input Trigger	Output Category	$\arg\max(\lambda x.state(x) \wedge borders(x, texas), \lambda x.size(x))$
constant c	$NP : c$	$NP : texas$
arity one predicate p_1	$N : \lambda x.p_1(x)$	$N : \lambda x.state(x)$
arity one predicate p_1	$S\backslash NP : \lambda x.p_1(x)$	$S\backslash NP : \lambda x.state(x)$
arity two predicate p_2	$(S\backslash NP)/NP : \lambda x.\lambda y.p_2(y, x)$	$(S\backslash NP)/NP : \lambda x.\lambda y.borders(y, x)$
arity two predicate p_2	$(S\backslash NP)/NP : \lambda x.\lambda y.p_2(x, y)$	$(S\backslash NP)/NP : \lambda x.\lambda y.borders(x, y)$
arity one predicate p_1	$N/N : \lambda g.\lambda x.p_1(x) \wedge g(x)$	$N/N : \lambda g.\lambda x.state(x) \wedge g(x)$
literal with arity two predicate p_2 and constant second argument c	$N/N : \lambda g.\lambda x.p_2(x, c) \wedge g(x)$	$N/N : \lambda g.\lambda x.borders(x, texas) \wedge g(x)$
arity two predicate p_2	$(N\backslash N)/NP : \lambda x.\lambda g.\lambda y.p_2(x, y) \wedge g(x)$	$(N\backslash N)/NP : \lambda g.\lambda x.\lambda y.borders(x, y) \wedge g(x)$
an arg max / min with second argument arity one function f	$NP/N : \lambda g.\arg\max / \min(g, \lambda x.f(x))$	$NP/N : \lambda g.\arg\max(g, \lambda x.size(x))$
an arity one numeric-ranged function f	$S/NP : \lambda x.f(x)$	$S/NP : \lambda x.size(x)$

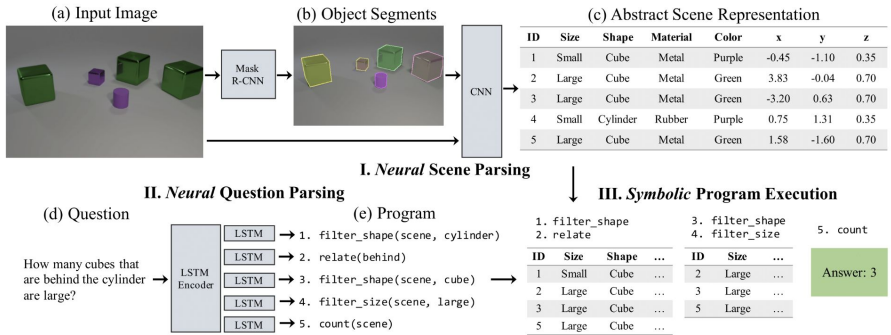
Hand-Crafted Rules and Features (Zettlemoyer and Collins, 2005)



Single Utterances vs Conversations (Suhr et al., 2018)



Customized Decoders to Follow Grammar (Yin et al., 2017)



Handling Different Forms of Data (Yi et al., 2018)

Pretrained Language Models are Getting Bigger

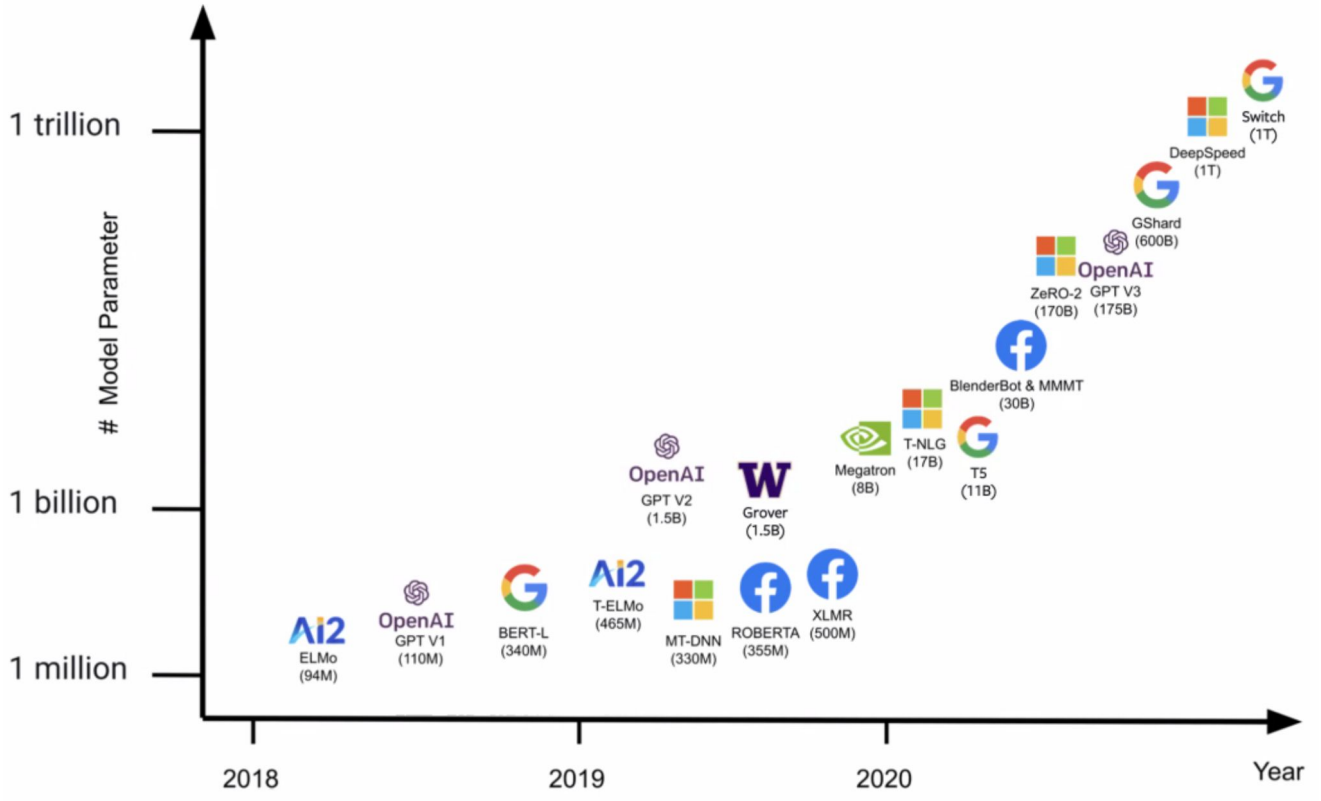


Figure Adapted from Myle Ott

Large Language Models Open a New Era for NLP Research

Classification

Classification

Classify items into categories via example.

Prompt

```
The following is a list of companies and the categories they fall into:

Apple, Facebook, Fedex

Apple
Category:
```

Sample response

```
Technology

Facebook
Category: Social Media

Fedex
Category: Delivery
```

Q&A

Answers | Generation | Conversation

Answer questions based on existing knowledge.

Prompt

```
I am a highly intelligent question answering bot. If you ask me a question that is rooted in truth, I will give you the answer. If you ask me a question that is nonsense, trickery, or has no clear answer, I will respond with "Unknown".

Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: Which party did he belong to?
A: He belonged to the Republican Party.

Q: What is the square root of banana?
A: Unknown

Q: How does a telescope work?
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

Q: Where were the 1992 Olympics held?
A: The 1992 Olympics were held in Barcelona, Spain.

Q: How many squigs are in a bonk?
A: Unknown

Q: Where is the Valley of Kings?
A:
```

Sample response

```
The Valley of Kings is located in Luxor, Egypt.
```

TL;DR summarization

Transformation | Generation

Summarize text by adding a 'tl;dr:' to the end of a text passage. It shows that the API understands how to perform a number of tasks with no instructions.

Prompt

```
A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.[1] Neutron stars are the smallest and densest stellar objects, excluding black holes and hypothetical white holes, quark stars, and strange stars.[2] Neutron stars have a radius on the order of 10 kilometres (6.2 mi) and a mass of about 1.4 solar masses. [3] They result from the supernova explosion of a massive star, combined with gravitational collapse, that compresses the core past white dwarf star density to that of atomic nuclei.

Tl;dr
```

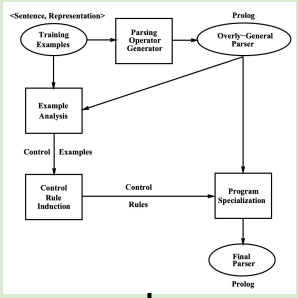
Sample response

```
A neutron star is the collapsed core of a massive supergiant star. These ultra-dense objects are incredibly fascinating due to their strange properties and their potential for phenomena such as extreme gravitational forces and a strong magnetic field.
```

https://beta.openai.com/examples

Focus on this Talk: How Large Language Models Help Semantic Parsing?

Zelle and Mooney (1996)

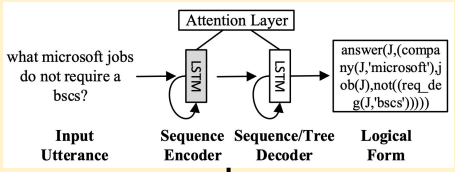


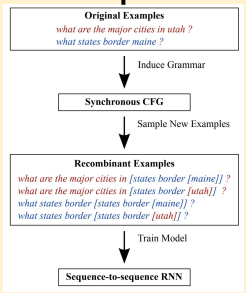
We now turn to issues of parsing and parameter estimation. Parsing under a PCCG involves computing the most probable logical form L for a sentence S,

$$\arg\max_L P(L|S;\bar{\theta}) = \arg\max_L \sum_T P(L,T|S;\bar{\theta})$$

Zettlemoyer and Collins (2005)

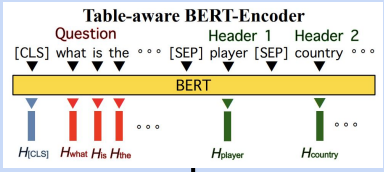
Dong and Lapata (2016)

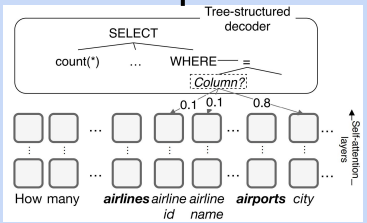




Jia and Liang (2016)

Hwang et al. (2019)





Wang et al. (2019)

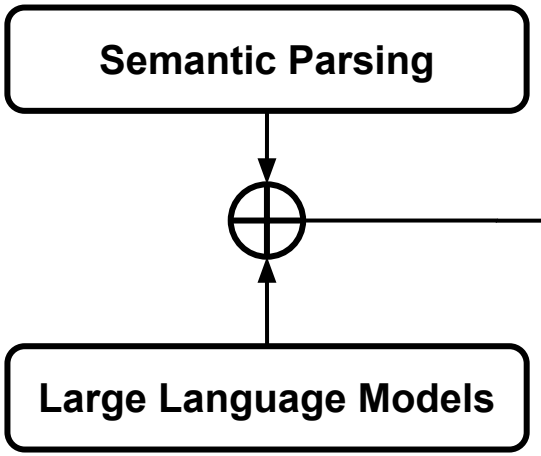
Non-Neural Network

E2E Neural Networks

Contextualized Embeddings and Pretrained Language Models

LLMs

Large Language Models Help Semantic Parsing in Three Ways



Unified Semantic Parsing Framework

UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models, EMNLP 2022

Multilingual Semantic Parsing

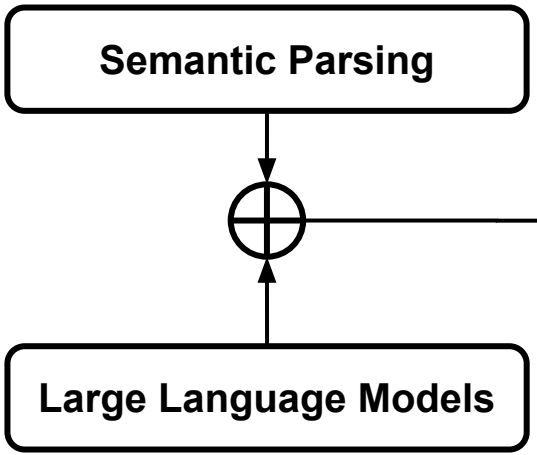
XSemPLR: Benchmarking Cross-Lingual Semantic Parsing in Multiple Languages and Meaning Representations, work in progress 2022

XRICL: Cross-lingual Retrieval-Augmented In-Context Learning for Cross-lingual Text-to-SQL Semantic Parsing, Findings of EMNLP 2022

Semantic Parsing for Numerical and Logical Reasoning

MultiHiertt: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data, ACL 2022

FOLIO: Natural Language Reasoning with First-Order Logic, arXiv 2022



Unified Semantic Parsing Framework

UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models, EMNLP 2022

Multilingual Semantic Parsing

XSemPLR: Benchmarking Cross-Lingual Semantic Parsing in Multiple Languages and Meaning Representations, work in progress 2022

XRICL: Cross-lingual Retrieval-Augmented In-Context Learning for Cross-lingual Text-to-SQL Semantic Parsing, Findings of EMNLP 2022

Semantic Parsing for Numerical and Logical Reasoning

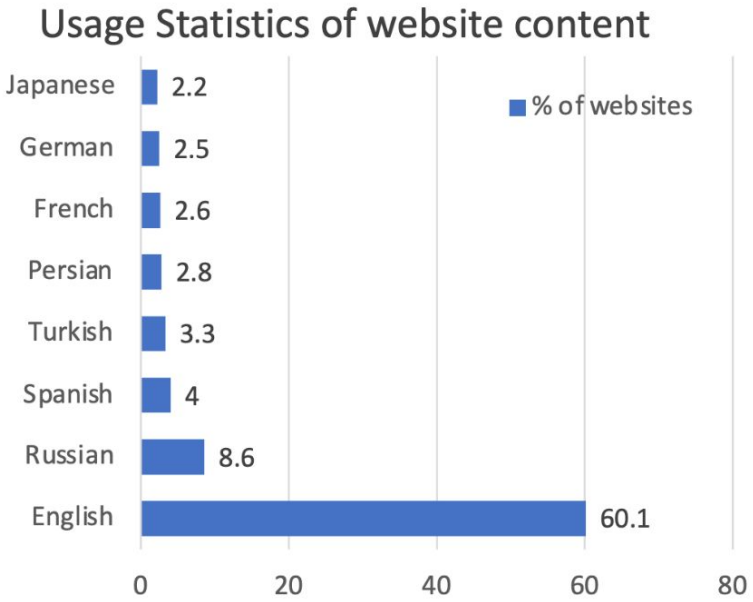
MultiHiertt: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data, ACL 2022

FOLIO: Natural Language Reasoning with First-Order Logic, arXiv 2022

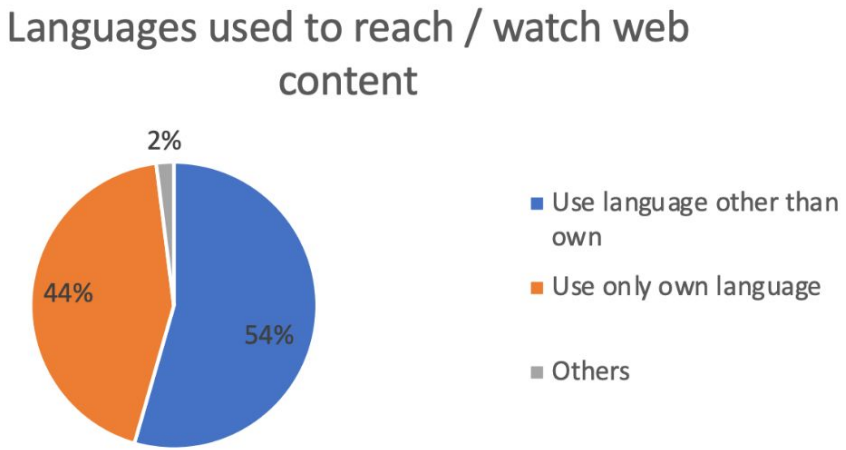
XSemPLR: Benchmarking Cross-Lingual Semantic Parsing in Multiple Languages and Meaning Representations

Zhang et al., Work in Progress

Multilingual Information Access

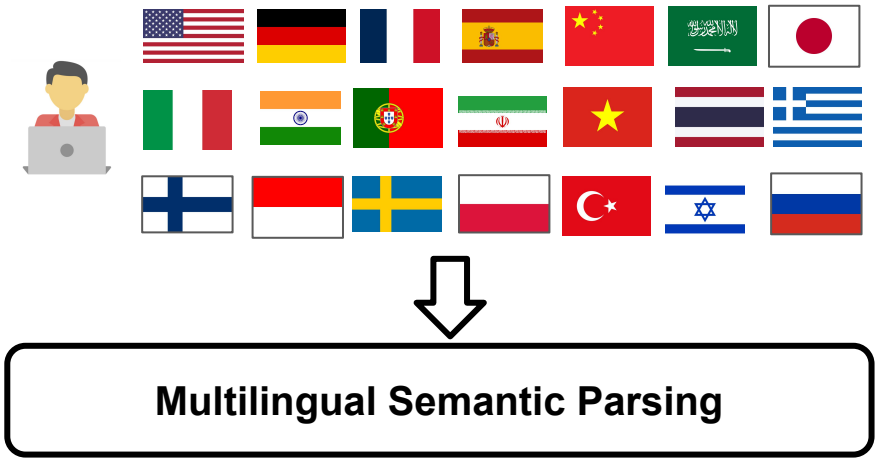


(Data from Technology report: Usage statistics of content languages for websites)



User language preferences online (EU)

Multilingual Semantic Parsing

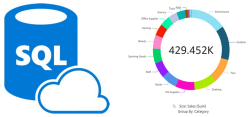


Multilingual Semantic Parsing

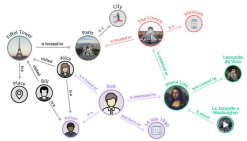
Multilingual Semantic Parsing

SQL, SPARQL, GraphQL, Lambda-Calculus
Functional Query Language, Prolog
Intent and Slot, ThingTalk Query Language

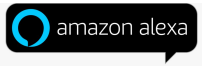
Databases



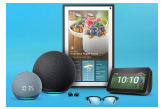
Knowledge Graphs



Virtual Assistants



Smart Home Devices

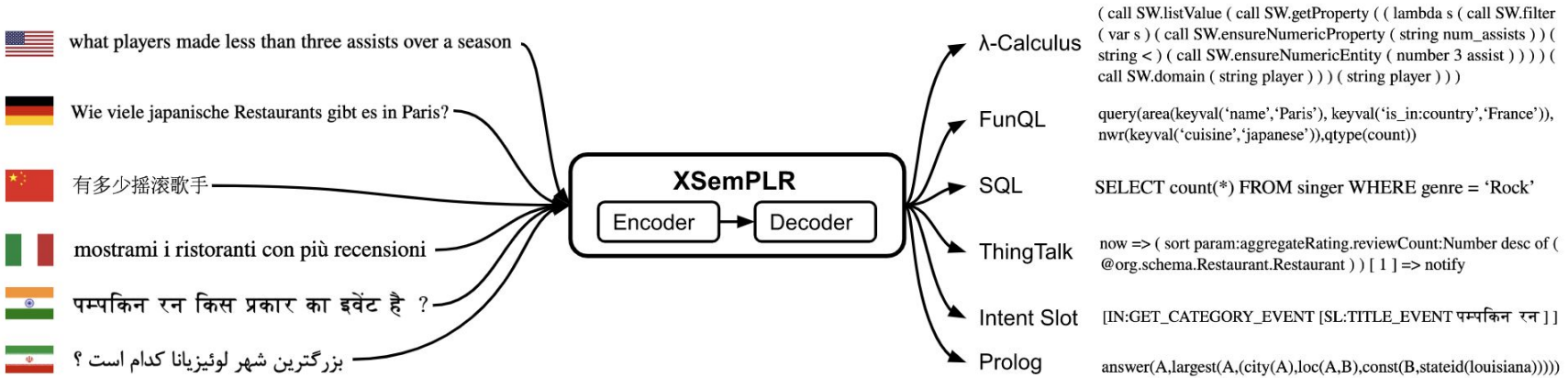


Human-Robot Interaction

Code Generation

Multiple Input Languages, Multiple Output Meaning Representations

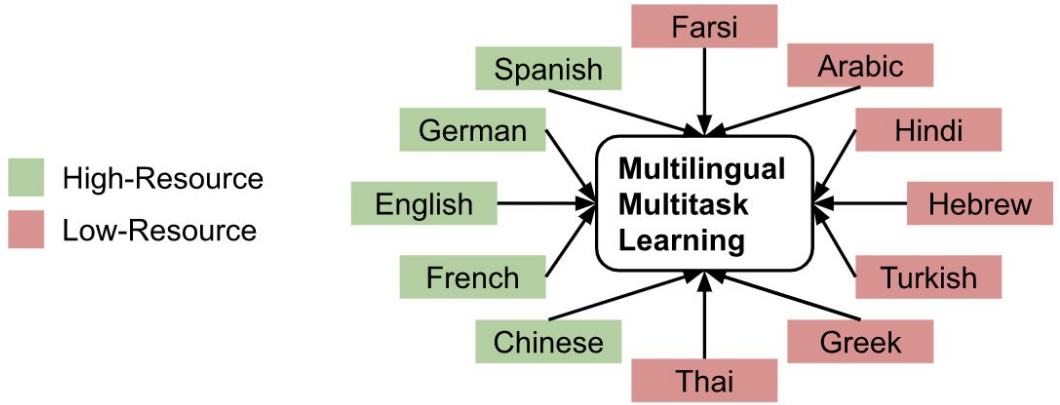
Task	Dataset	Meaning Representation	Language	Executable	Domain	Train	Dev	Test
NLI for Databases	ATIS	SQL	7	yes	1	4303	481	444
NLI for Databases	GeoQuery	SQL,Lambda,FunQL,Prolog	8	yes	1	548	49	277
NLI for Databases	Spider	SQL	3	yes	138	8095	1034	–
NLI for Databases	NLmaps	Functional Query Language	2	yes	1	1500	–	880
QA on Knowledge Graph	Overnight	Lambda Calculus	3	yes	8	8754	2188	2740
QA on Knowledge Graph	MCWQ	SPARQL	4	yes	1	4006	733	648
QA on Web	Schema2QA	ThingTalk Query Language	11	yes	2	8932	–	971
Task-Oriented DST	MTOP	Hierarchical Intent and Slot	6	no	11	5446	863	1245
Code Generation	MCoNaLa	Python	4	yes	open	2379	–	1788



mT5 is the Best

	ATIS	GeoQuery	Spider	NLmaps	Overnight	MCWQ	Schema2QA	MTOP	Average
Translate-Test									
mT5	44.50	65.91	45.26	66.36	59.69	19.85	3.18*	29.78*	50.26
In-language Monolingual									
LSTM	35.00	60.26	11.54	68.60	15.10	10.38	36.80	63.40	37.64
mBERT+PTR	30.63	82.40	40.40	83.82	57.47	23.46	52.53	75.41	55.77
XLM-R+PTR	31.31	**85.79**	47.30	85.17	59.10	23.53	62.37	80.36	58.59
mBART	41.93	63.40	33.31	83.19	59.60	30.02	50.35	75.76	54.70
mT5	**53.15**	81.05	**53.14**	**91.65**	**66.29**	**30.15**	**65.16**	**81.83**	**65.30**

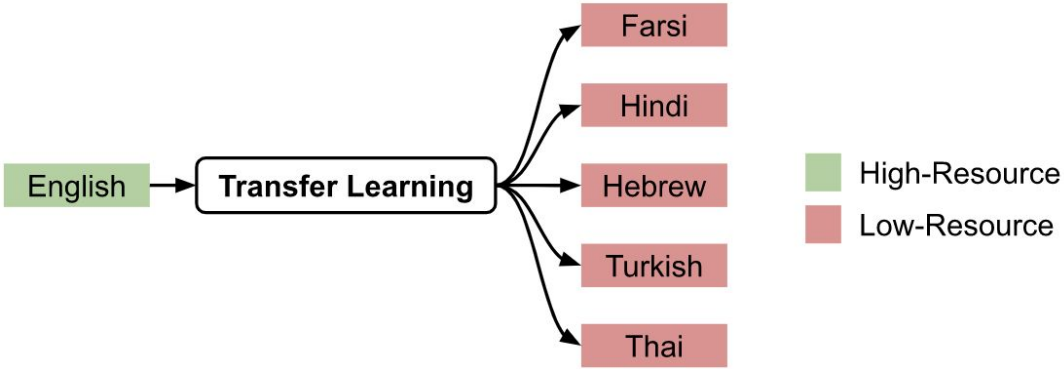
Multilingual Multitask Learning



Multilingual Multitask Learning Helps

	ATIS	GeoQuery	Spider	NLmaps	Overnight	MCWQ	Schema2QA	MTOP	Average
Translate-Test									
mT5	44.50	65.91	45.26	66.36	59.69	19.85	3.18*	29.78*	50.26
In-language Monolingual									
LSTM	35.00	60.26	11.54	68.60	15.10	10.38	36.80	63.40	37.64
mBERT+PTR	30.63	82.40	40.40	83.82	57.47	23.46	52.53	75.41	55.77
XLM-R+PTR	31.31	**85.79**	47.30	85.17	59.10	23.53	62.37	80.36	58.59
mBART	41.93	63.40	33.31	83.19	59.60	30.02	50.35	75.76	54.70
mT5	**53.15**	81.05	**53.14**	**91.65**	**66.29**	**30.15**	**65.16**	**81.83**	**65.30**
In-language Monolingual Few-Shot									
mT5	22.26	7.48	25.57	26.93	9.17	0.77	22.61	61.90	22.09
In-language Multilingual									
mT5	54.45	82.04	–	–	–	–	60.92	82.95	70.09

Transfer Learning



Need Better Zero/Few-shot Learning for Other Languages

	ATIS	GeoQuery	Spider	NLmaps	Overnight	MCWQ	Schema2QA	MTOP	Average
Translate-Test									
mT5	44.50	65.91	45.26	66.36	59.69	19.85	3.18*	29.78*	50.26
In-language Monolingual									
LSTM	35.00	60.26	11.54	68.60	15.10	10.38	36.80	63.40	37.64
mBERT+PTR	30.63	82.40	40.40	83.82	57.47	23.46	52.53	75.41	55.77
XLM-R+PTR	31.31	**85.79**	47.30	85.17	59.10	23.53	62.37	80.36	58.59
mBART	41.93	63.40	33.31	83.19	59.60	30.02	50.35	75.76	54.70
mT5	**53.15**	81.05	**53.14**	**91.65**	**66.29**	**30.15**	**65.16**	**81.83**	**65.30**
In-language Monolingual Few-Shot									
mT5	22.26	7.48	25.57	26.93	9.17	0.77	22.61	61.90	22.09
In-language Multilingual									
mT5	54.45	82.04	–	–	–	–	60.92	82.95	70.09
Cross-lingual Zero-Shot Transfer									
mT5	31.85	39.40	41.93	34.89	52.68	4.06	44.04	50.18	37.38
Cross-lingual Few-Shot Transfer									
mT5	49.57	68.18	–	–	–	–	59.24	74.83	62.96

XRICL: Cross-lingual Retrieval-Augmented In-Context Learning for Cross-lingual Text-to-SQL Semantic Parsing

Peng Shi, Rui Zhang, He Bai, Jimmy Lin

Findings of EMNLP 2022

In-Context Learning:
Learning from Examples without Parameter Updates

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

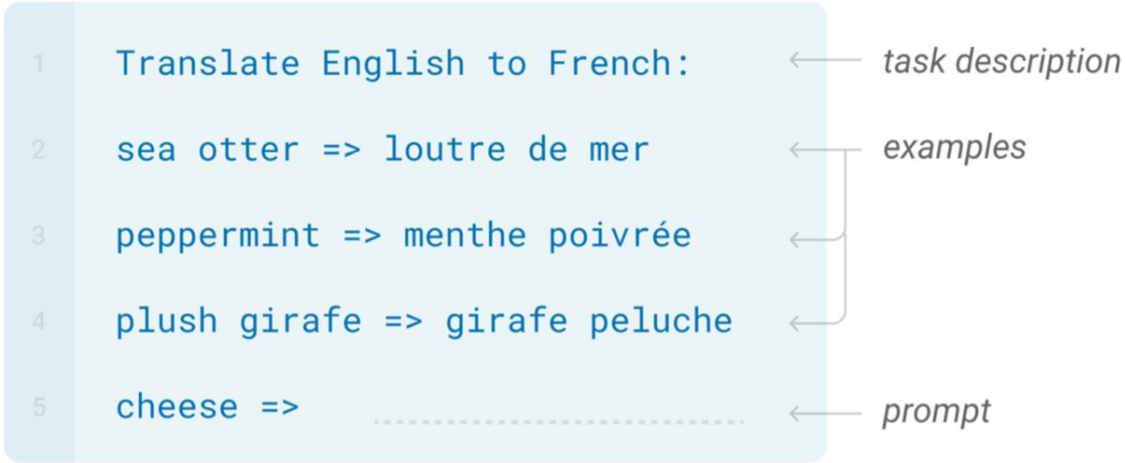
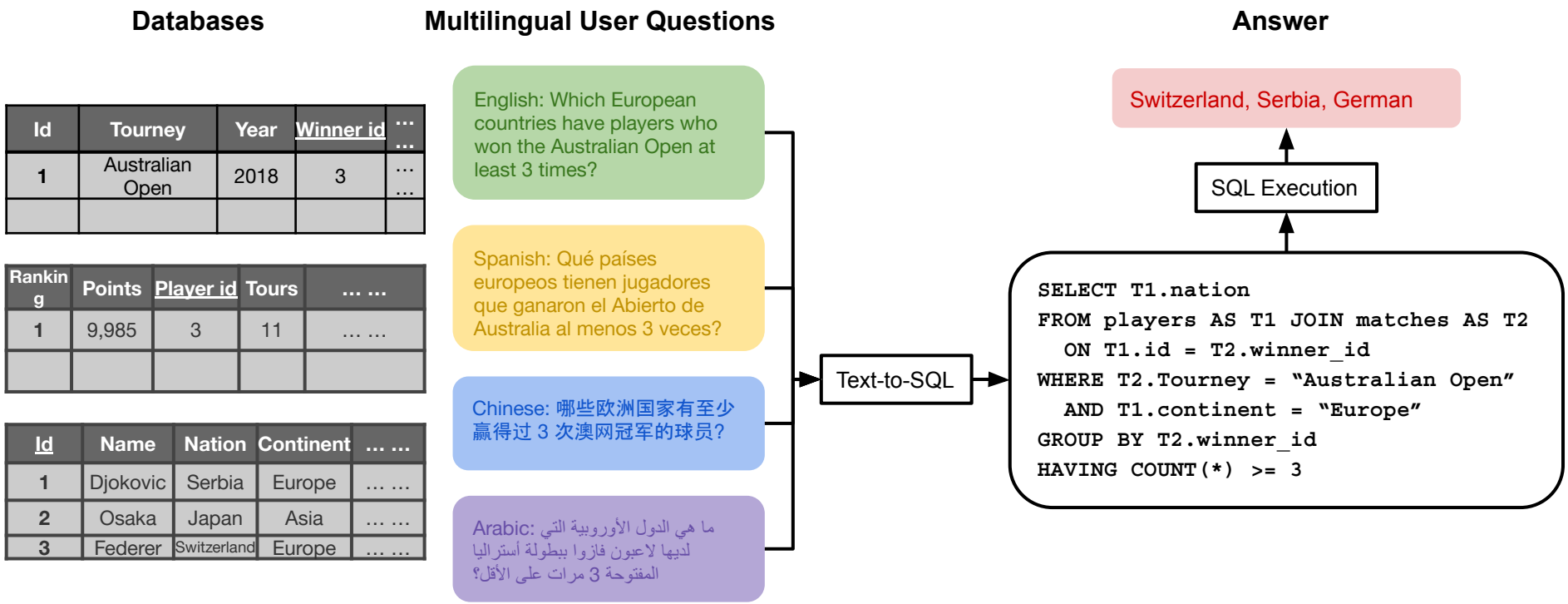


Figure from http://ai.stanford.edu/blog/in-context-learning/

Our Task: Cross-lingual Text-to-SQL



Our Goal

Our Goal

1. Use In-Context Learning with LLMs for Cross-lingual Semantic Parsing
2. Have only English Annotations of Text-to-SQL pairs.

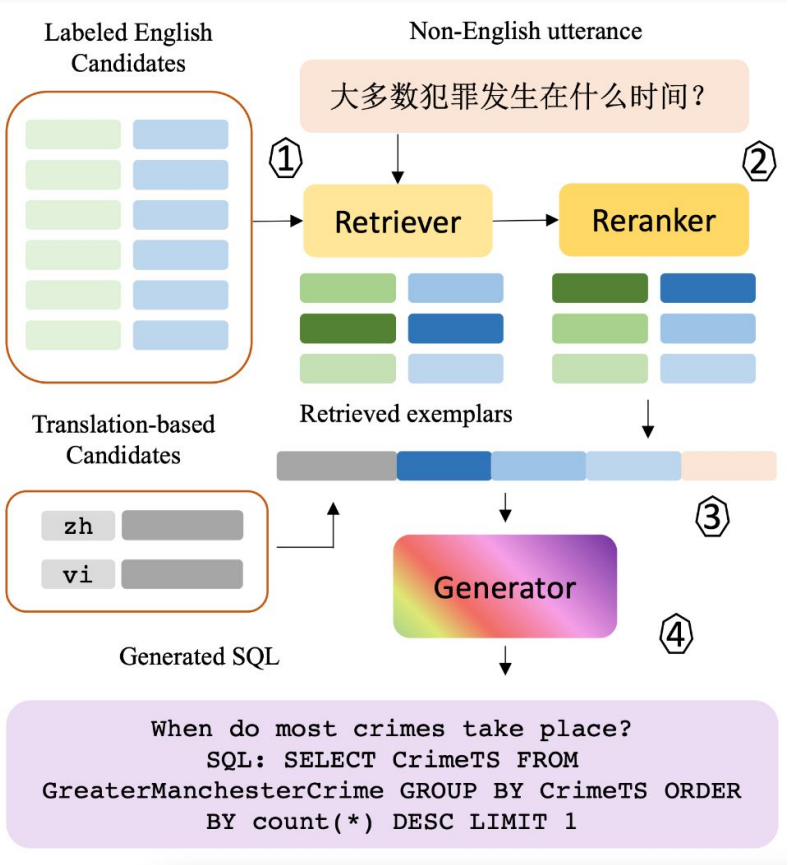
Our Goal, Challenges, and Solutions

Our Goal

1. Use In-Context Learning with LLMs for Cross-lingual Semantic Parsing
2. Have only English Annotations of Text-to-SQL pairs.

Challenges		Solutions
1. Find most relevant English examples.	→	Cross-lingual Retrieval
2. Facilitate translation.	→	Translation-based Prompt

XRICL Framework



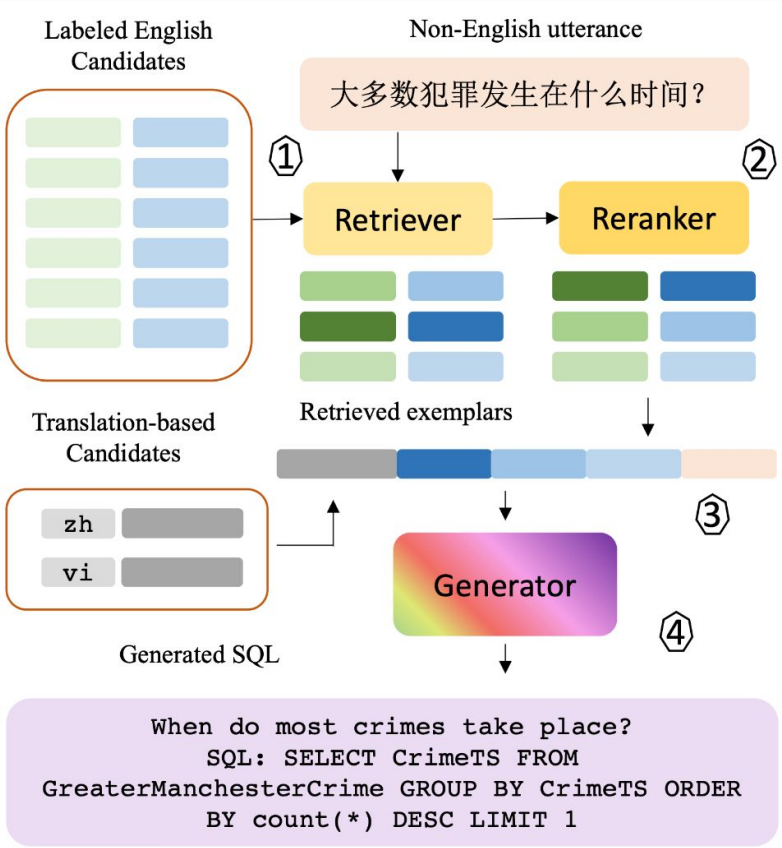
(1) *Cross-lingual Exemplar Retrieval*: Retrieve a list of N English exemplars that are relevant to the input non-English example x.

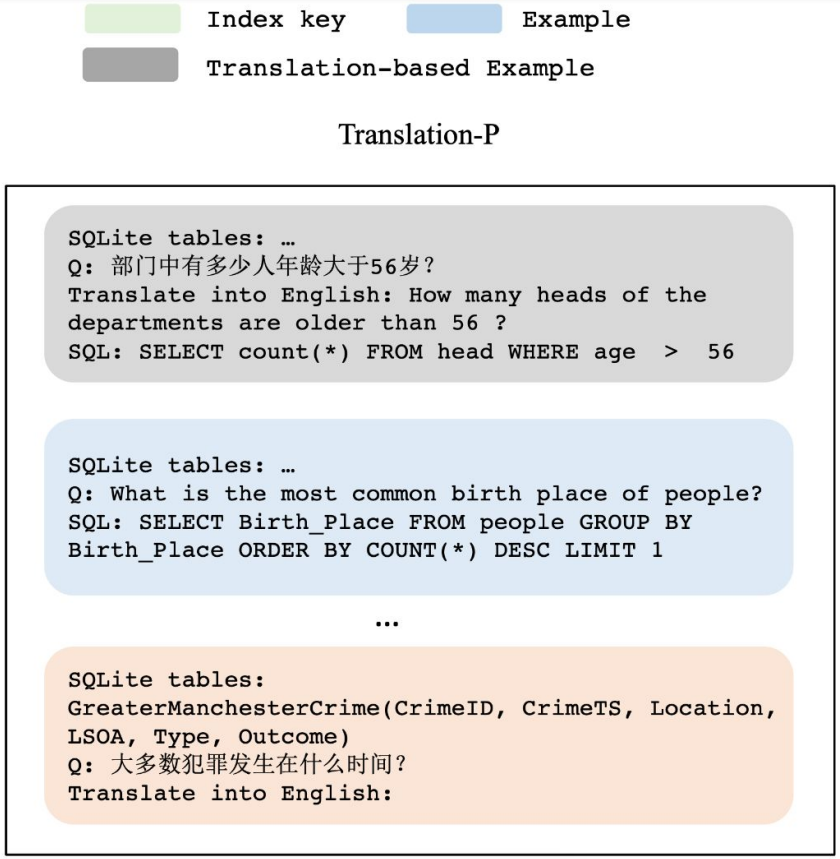
(2) *Exemplar Reranking*: Rerank the retrieved N exemplars and use the top K exemplars to construct prompts.

(3) *Prompt Construction with Translation as Chain of Thought*: Construct a prompt consisting of the translation exemplar as a chain of thought, the selected K exemplars, and the input example.

(4) *Inference*: Feed the prompt into a pre-trained language model to generate SQL.

Prompt Examples in XRICL





Two New Benchmarks

XSpider: English, Chinese, Vietnamese, Farsi, Hindi

XKaggle-DBQA: English, Chinese, Farsi, Hindi

Cross-lingual Exemplar Retriever

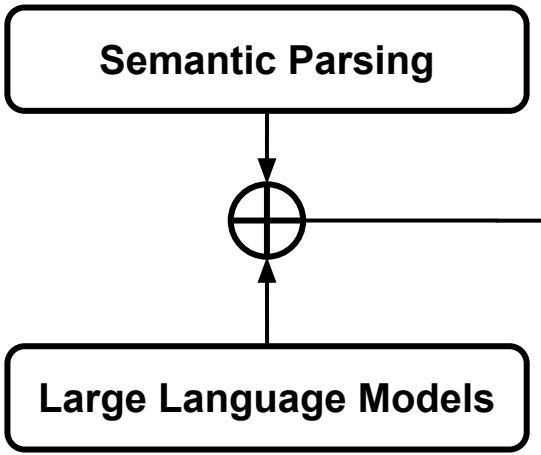
Model	zh-full	zh		vi		fa		hi	
	EM	EM	TS	EM	TS	EM	TS	EM	TS
(1) mT5 zero-shot	39.7	47.9	48.4	42.1	40.1	41.3	39.5	**41.2**	**39.7**
(2) mUSE	38.4	43.0	46.8	31.8	33.4	28.9	31.1	22.2	23.7
(3) mSBERT	37.9	41.3	47.1	34.6	33.5	29.3	31.8	22.0	22.3
(4) mT5-encoder	44.4	48.1	51.4	41.3	39.5	38.4	38.5	28.6	27.0
(5) DE-Retriever	46.0	50.4	53.9	42.2	40.7	38.2	40.0	29.9	27.9
(6) DE-R^2	46.4	52.1	55.3	**44.4**	41.9	40.0	40.6	30.0	28.2

Model	zh	fa	hi
(1) mT5 zero-shot	9.7	8.1	7.6
(2) mUSE	20.7	12.4	16.2
(3) mSBERT	14.7	13.0	11.9
(4) mT5-Encoder	22.2	16.8	16.2
(5) DE-Retriever	26.5	18.4	16.8
(6) DE-R^2	27.0	18.4	17.8

Translation-Augmented Prompts

Model	zh-full	zh		vi		fa		hi	
	EM	EM	TS	EM	TS	EM	TS	EM	TS
(1) mT5 zero-shot	39.7	47.9	48.4	42.1	40.1	41.3	39.5	**41.2**	**39.7**
(2) mUSE	38.4	43.0	46.8	31.8	33.4	28.9	31.1	22.2	23.7
(3) mSBERT	37.9	41.3	47.1	34.6	33.5	29.3	31.8	22.0	22.3
(4) mT5-encoder	44.4	48.1	51.4	41.3	39.5	38.4	38.5	28.6	27.0
(5) DE-Retriever	46.0	50.4	53.9	42.2	40.7	38.2	40.0	29.9	27.9
(6) DE-R^2	46.4	52.1	55.3	**44.4**	41.9	40.0	40.6	30.0	28.2
(7) + Translation-P	**47.4**	**52.7**	**55.7**	43.7	**43.6**	**43.2**	**45.1**	**32.6**	**32.4**

Model	zh	fa	hi
(1) mT5 zero-shot	9.7	8.1	7.6
(2) mUSE	20.7	12.4	16.2
(3) mSBERT	14.7	13.0	11.9
(4) mT5-Encoder	22.2	16.8	16.2
(5) DE-Retriever	26.5	18.4	16.8
(6) DE-R^2	27.0	18.4	17.8
(7) + Translation-P	**28.1**	**20.0**	**19.5**



Unified Semantic Parsing Framework

UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models, EMNLP 2022

Multilingual Semantic Parsing

XSemPLR: Benchmarking Cross-Lingual Semantic Parsing in Multiple Languages and Meaning Representations, work in progress 2022

XRICL: Cross-lingual Retrieval-Augmented In-Context Learning for Cross-lingual Text-to-SQL Semantic Parsing, Findings of EMNLP 2022

Semantic Parsing for Numerical and Logical Reasoning

MultiHiertt: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data, ACL 2022

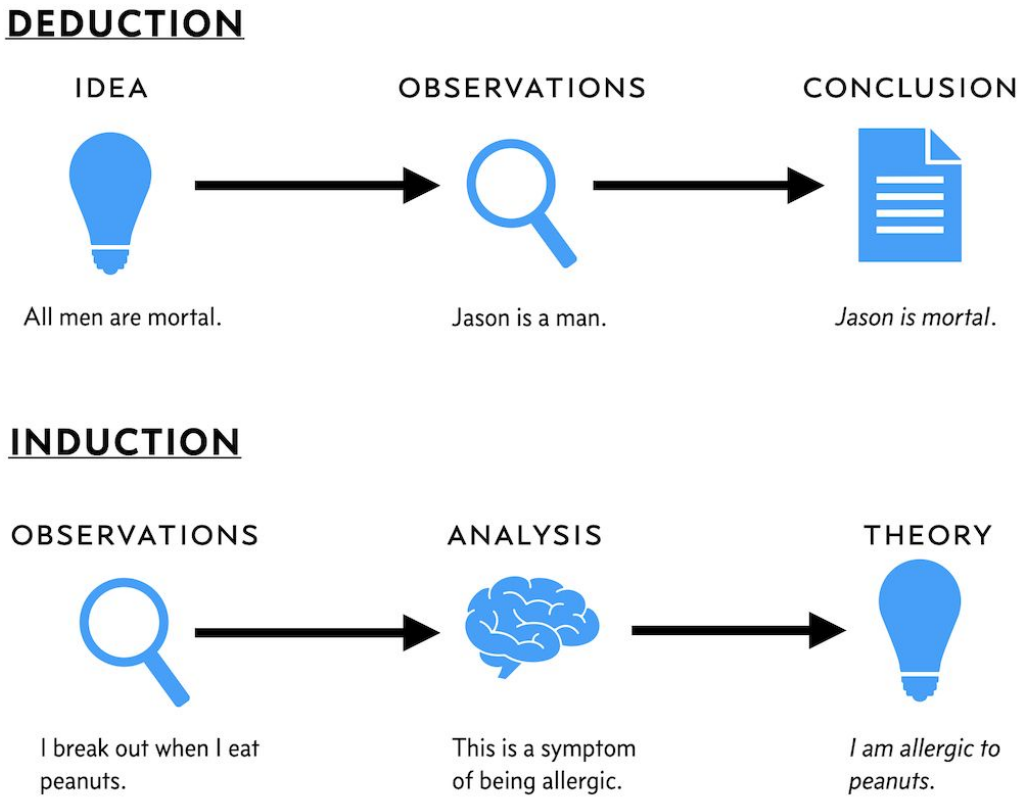
FOLIO: Natural Language Reasoning with First-Order Logic, arXiv 2022

FOLIO: Natural Language Reasoning with First-Order Logic

Han et al., arXiv 2022

https://github.com/Yale-LILY/FOLIO

Logical Reasoning: Deductive vs Inductive



Syllogism

Major premise: All men are mortal.

Minor premise: Socrates is a man.

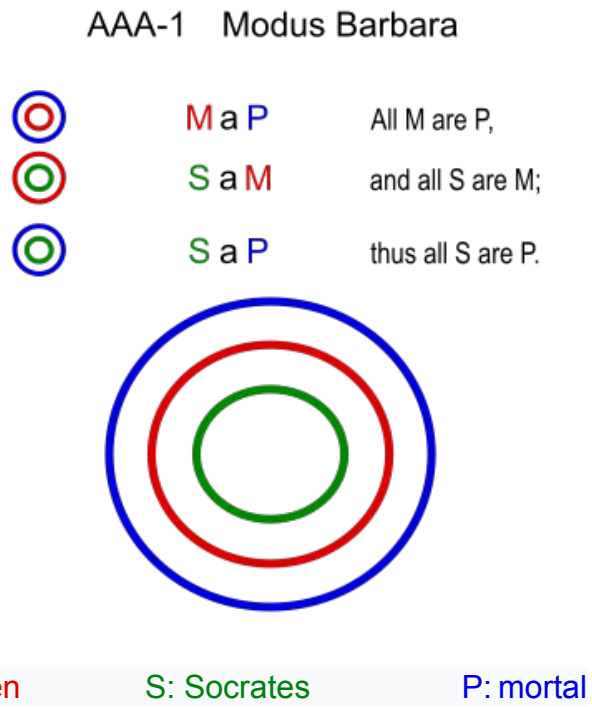
Conclusion: Therefore, Socrates is mortal.

Syllogism by Venn Diagram

Major premise: All men are mortal.

Minor premise: Socrates is a man.

Conclusion: Therefore, Socrates is mortal.



Syllogism by First-Order Logic

Major premise:	All men are mortal.	$\forall x\ \text{Men}(x) \Rightarrow \text{Mortal}(x)$
Minor premise:	Socrates is a man.	$\text{Men}(\text{socrates})$
Conclusion:	Therefore, Socrates is mortal.	$\text{Mortal}(\text{socrates})$

FOLIO: A New Dataset for Natural Language Reasoning with First-Order Logic

A FOLIO example based on the Wild Turkey Wikipedia page: https://en.wikipedia.org/wiki/Wild_turkey

NL premises

1. There are six types of wild turkeys: Eastern wild turkey, Osceola wild turkey, Gould's wild turkey, Merriam's wild turkey, Rio Grande wild turkey, and the Ocellated wild turkey.
2. Tom is not an Eastern wild turkey.
3. Tom is not an Osceola wild turkey.
4. Tom is also not a Gould's wild turkey, or a Merriam's wild turkey, or a Rio Grande wild turkey.
5. Tom is a wild turkey.

NL Conclusions -> Labels

A. Tom is an Ocellated wild turkey. -> True
B. Tom is an Eastern wild turkey. -> False
C. Joey is a wild turkey. -> Unknown

FOL Premises

1. $\forall x(\text{WildTurkey}(x) \rightarrow (\text{Eastern}(x) \vee \text{Osceola}(x) \vee \text{Goulds}(x) \vee \text{Merriams}(x) \vee \text{Riogrande}(x) \vee \text{Ocellated}(x)))$
2. $\neg(\text{WildTurkey}(tom) \wedge \text{Eastern}(tom))$
3. $\neg(\text{WildTurkey}(tom) \wedge \text{Osceola}(tom))$
4. $\text{WildTurkey}(tom) \rightarrow \neg(\text{Goulds}(tom) \vee \text{Merriams}(tom) \vee \text{Riogrande}(tom))$
5. $\text{WildTurkey}(tom)$

FOL conclusions -> Labels

A. $\text{Ocellated}(tom)$ -> True
B. $\text{Eastern}(tom)$ -> False
C. $\text{WildTurkey}(joey)$ -> Unknown

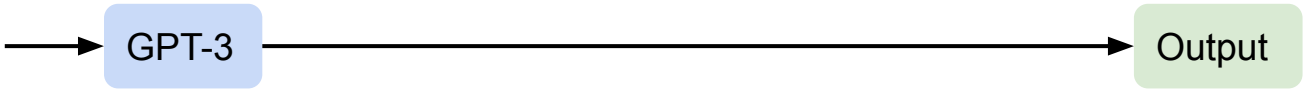
FOLIO is Different

Dataset	Size	Reasoning	Text Source	Real-World Example	# Premises for Conclusions	Vocab	# Distinct AST
CLUTTER	6k	Inductive	Synthetic	×	×	-	×
RECLOR	6k	Mixed forms	GMAT, LSAT exams	✓	×	-	×
LogiQA	8.6k	Mixed forms	NCSE exams	✓	×	-	×
RuleTaker	500k	Deductive	Synthetic	×	$1 \sim 5$	101	48
ProofWriter	500k	Deductive	Synthetic	×	$1 \sim 5$	101	48
LogicNLI	20k	FOL	Synthetic	×	$1 \sim 5$	1077	30
FOLIO (ours)	1,435	FOL	**Expert-written, Real-world**	✓	**$1 \sim 8$**	**4351**	**76**

Source	#Stories	#Premises	#Conclusions	NL		FOL							
				Vocab	#Words	$\forall$	$\exists$	$\neg$	$\wedge$	$\vee$	$\rightarrow$	$\leftrightarrow$	$\oplus$
WikiLogic	304	1353	753	3250	8.50	860	376	374	1256	136	749	21	32
HybLogic	183	1054	682	1902	11.52	793	42	669	363	246	924	0	245
Total	487	2407	1435	4351	9.86	1653	418	1043	1619	382	1673	21	277

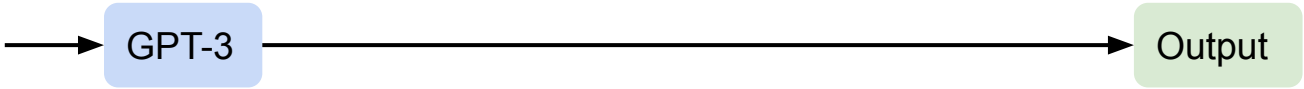
Large Language Models as Soft Logic Reasoners

[8 NL-Label Examples] Premises: All men are mortal. Socrates is a man. Conclusion: Therefore, Socrates is mortal. True, False, or Unknown?

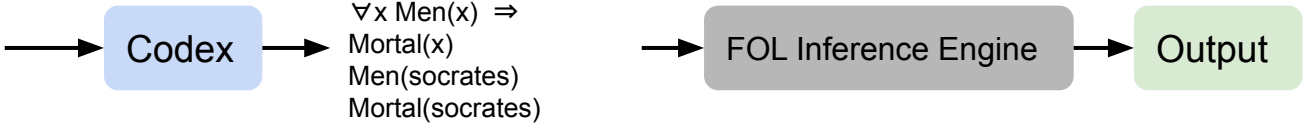


Large Language Models as FOL Semantic Parsers

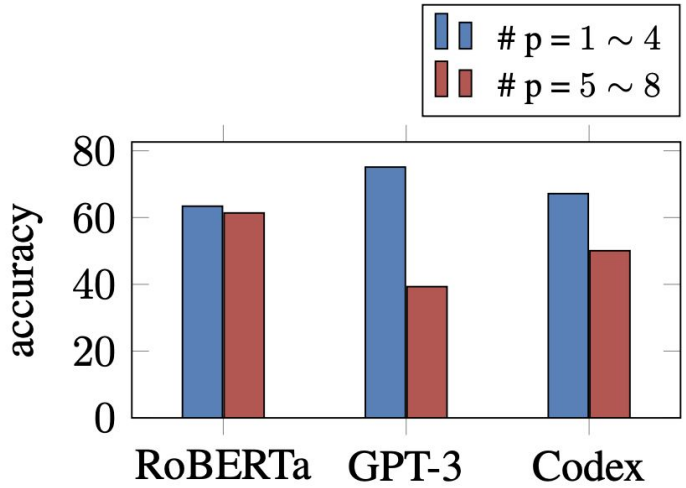
[8 NL-Label Examples] Premises: All men are mortal. Socrates is a man. Conclusion: Therefore, Socrates is mortal. True, False, or Unknown?



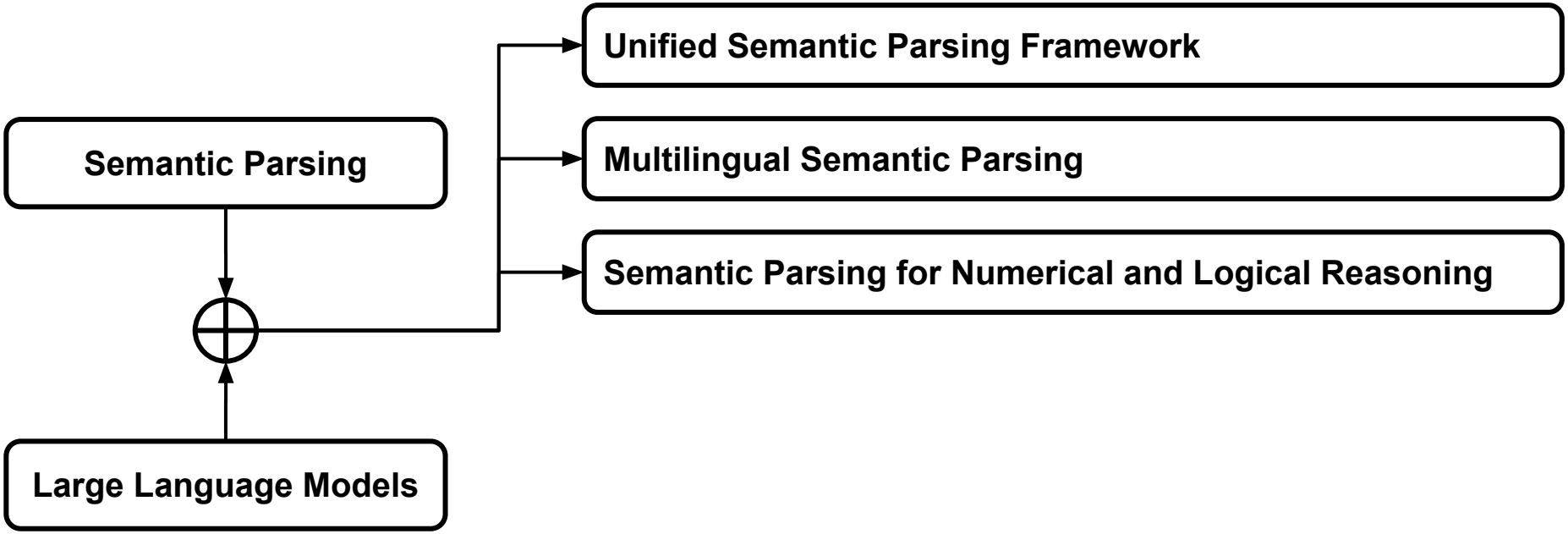
[8 NL-FOL Examples] Premises: All men are mortal. Socrates is a man. Conclusion: Therefore, Socrates is mortal. FOL is



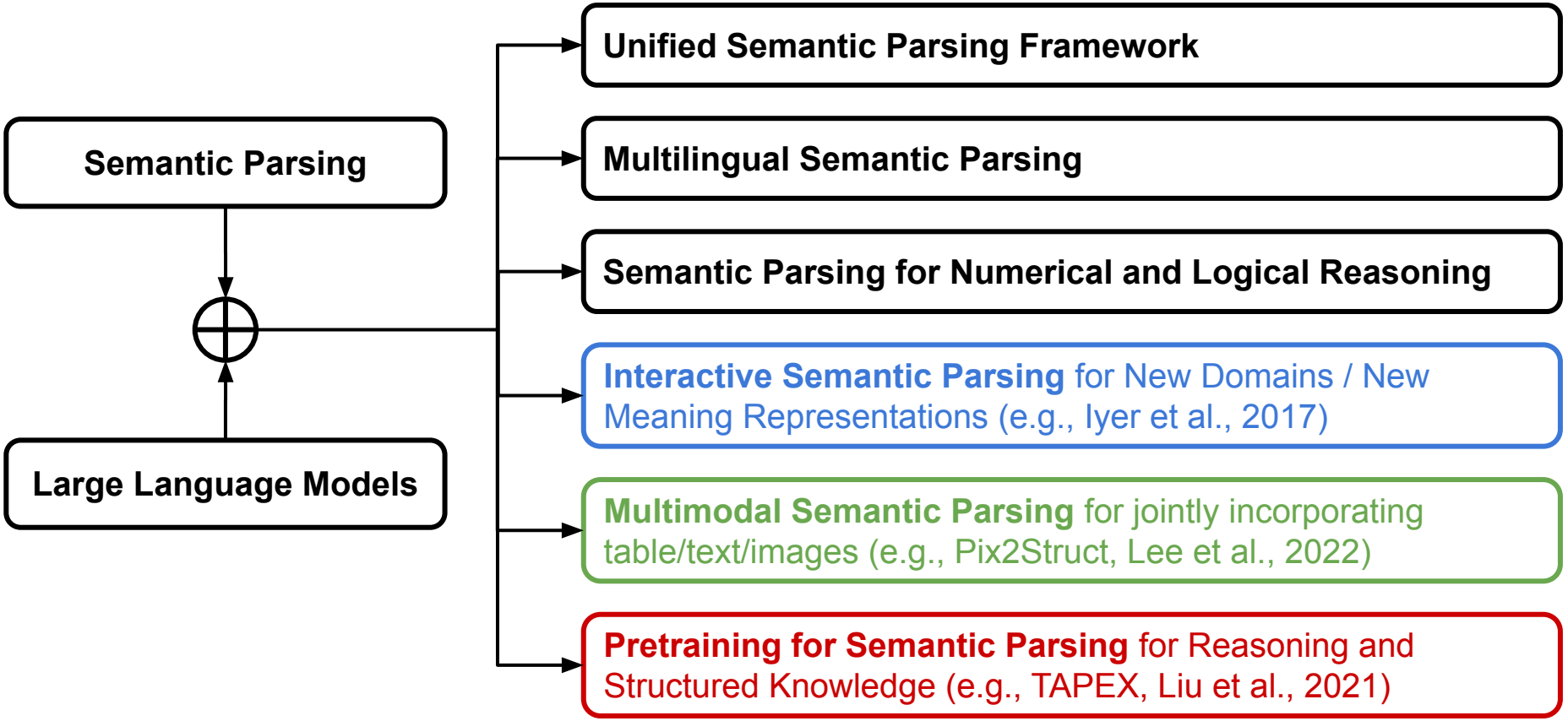
FOLIO is Challenging for Large Language Models



Conclusions



Future Directions



Thanks! Any Questions?