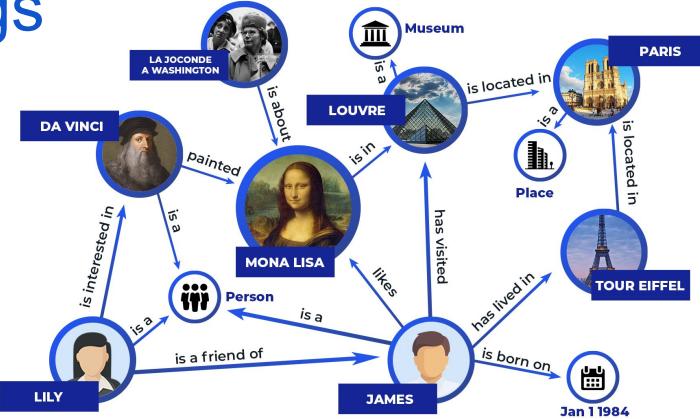


Semantic Parsing in the Era of Large Language Models

Rui Zhang, Assistant Professor, Penn State University
UPenn CLunch on Oct 31, 2022Happy Halloween! 🎃



Semantic Parsing for Many Things



Text-to-SQL for Natural Language Interfaces to Databases



Instruction Following for Robotics

Question Answering over Knowledge Graphs

Problem	Generated Code	Test Cases
<p>H-Index</p> <p>Given a list of citations counts, where each citation is a nonnegative integer, write a function <code>h_index</code> that outputs the h-index. The h-index is the largest number h such that h papers have each least h citations.</p> <p>Example: Input: [3,0,6,1,4] Output: 3</p>	<pre>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</pre>	<p>Input: [1,4,1,4,2,1,3,5,6]</p> <p>Generated Code Output: 4 ✓</p> <p>Input: [1000,500,500,250,100, 100,100,100,75,50, 30,20,15,15,10,5,2,1]</p> <p>Generated Code Output: 15 ✓</p>

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.



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



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



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



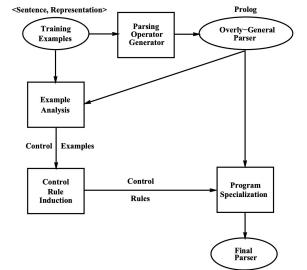
Bill Woods (1973) developed the **first NL database interface (LUNAR)** to answer scientists' questions about moon rocks 12 using a manually developed Augmented Transition Network (ATN) grammar.



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.

Learning-based Semantic Parsing Research

Zelle and Mooney (1996)

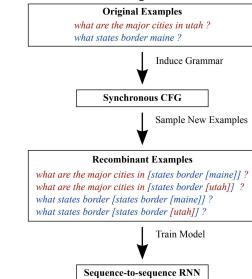
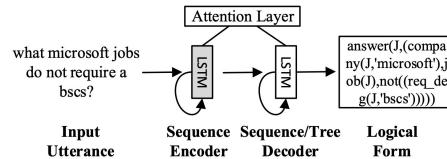


We now turn to issues of parsing and parameter estimation. Parsing under a PCCG involves computing the most probable logical form L for a sentence S ,

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

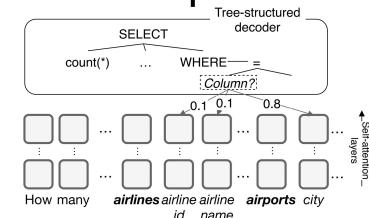
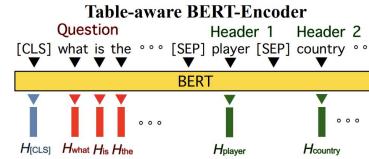
Zettlemoyer and Collins (2005)

Dong and Lapata (2016)



Jia and Liang (2016)

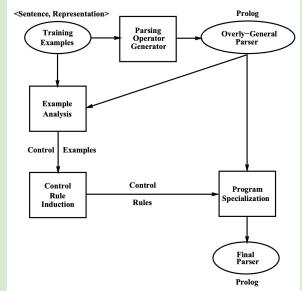
Hwang et al. (2019)



Wang et al. (2019)

Semantic Parsing Research: Paradigm Shifts

Zelle and Mooney (1996)



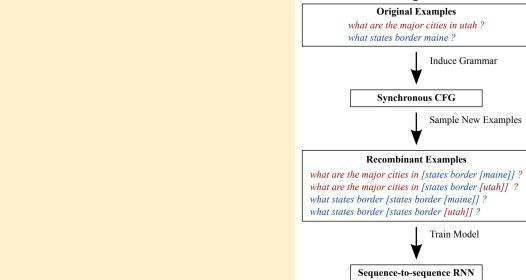
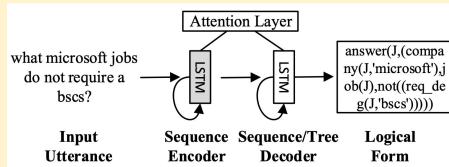
We now turn to issues of parsing and parameter estimation. Parsing under a PCCG involves computing the most probable logical form L for a sentence S ,

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

Zettlemoyer and Collins (2005)

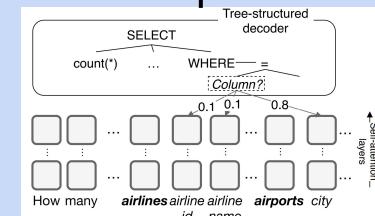
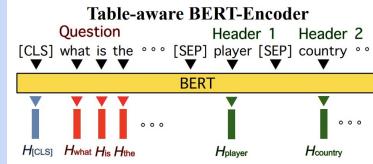
Non-Neural Network

Dong and Lapata (2016)



E2E Neural Networks

Hwang et al. (2019)

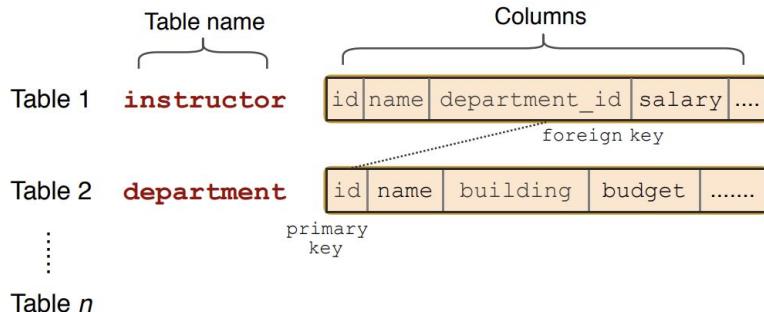


Wang et al. (2019)

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

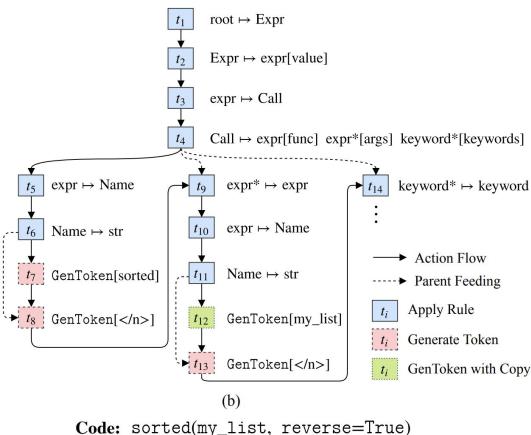
```
SELECT T2.name, T2.budget
FROM instructor as T1 JOIN department as
T2 ON T1.department_id = T2.id
GROUP BY T1.department_id
HAVING avg(T1.salary) >
(SELECT avg(salary) FROM instructor)
```

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.

Overspecialized Model Design

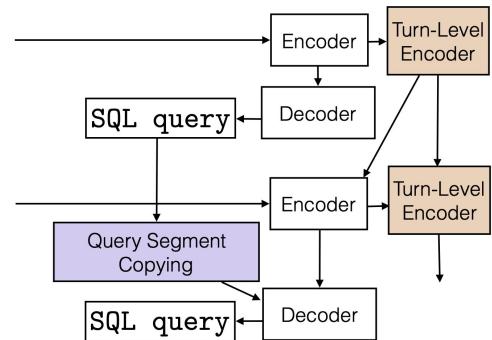
Rules		Categories produced from logical form
Input Trigger	Output Category	$\arg \max(\lambda x. state(x) \wedge borders(x, \text{texas}), \lambda x. size(x))$
constant c	$NP : c$	$NP : \text{texas}$
arity one predicate p_1	$N : \lambda x. p_1(x)$	$N : \lambda x. state(x)$
arity one predicate p_1	$S \setminus NP : \lambda x. p_1(x)$	$S \setminus NP : \lambda x. state(x)$
arity two predicate p_2	$(S \setminus NP) / NP : \lambda x. \lambda y. p_2(x, y)$	$(S \setminus NP) / NP : \lambda x. \lambda y. borders(x, y)$
arity two predicate p_2	$(S \setminus NP) / NP : \lambda x. \lambda y. p_2(x, y)$	$(S \setminus NP) / NP : \lambda x. \lambda y. borders(x, y)$
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	$N / N : \lambda g. \lambda x. p_2(x, c) \wedge g(x)$	$N / N : \lambda g. \lambda x. borders(x, \text{texas}) \wedge g(x)$
arity two predicate p_2	$(N \setminus N) / NP : \lambda x. \lambda y. p_2(x, y) \wedge g(x)$	$(N \setminus N) / NP : \lambda g. \lambda x. \lambda y. borders(x, y) \wedge g(x)$
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)

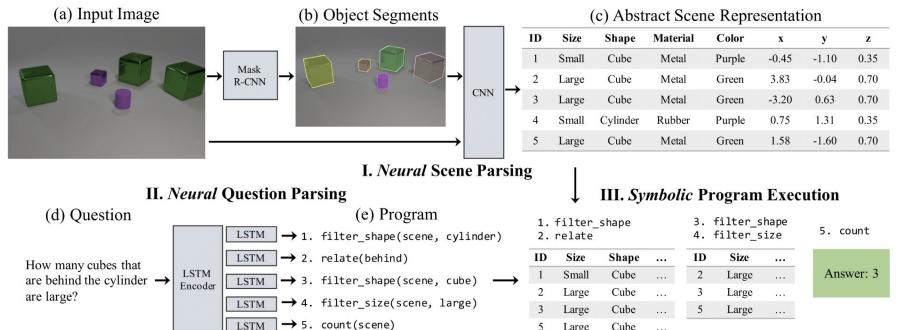


Customized Decoders to Follow Grammar (Yin et al., 2017)

Show me flights from Seattle to Boston next Monday



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



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

Pretrained Language Models are Getting Bigger

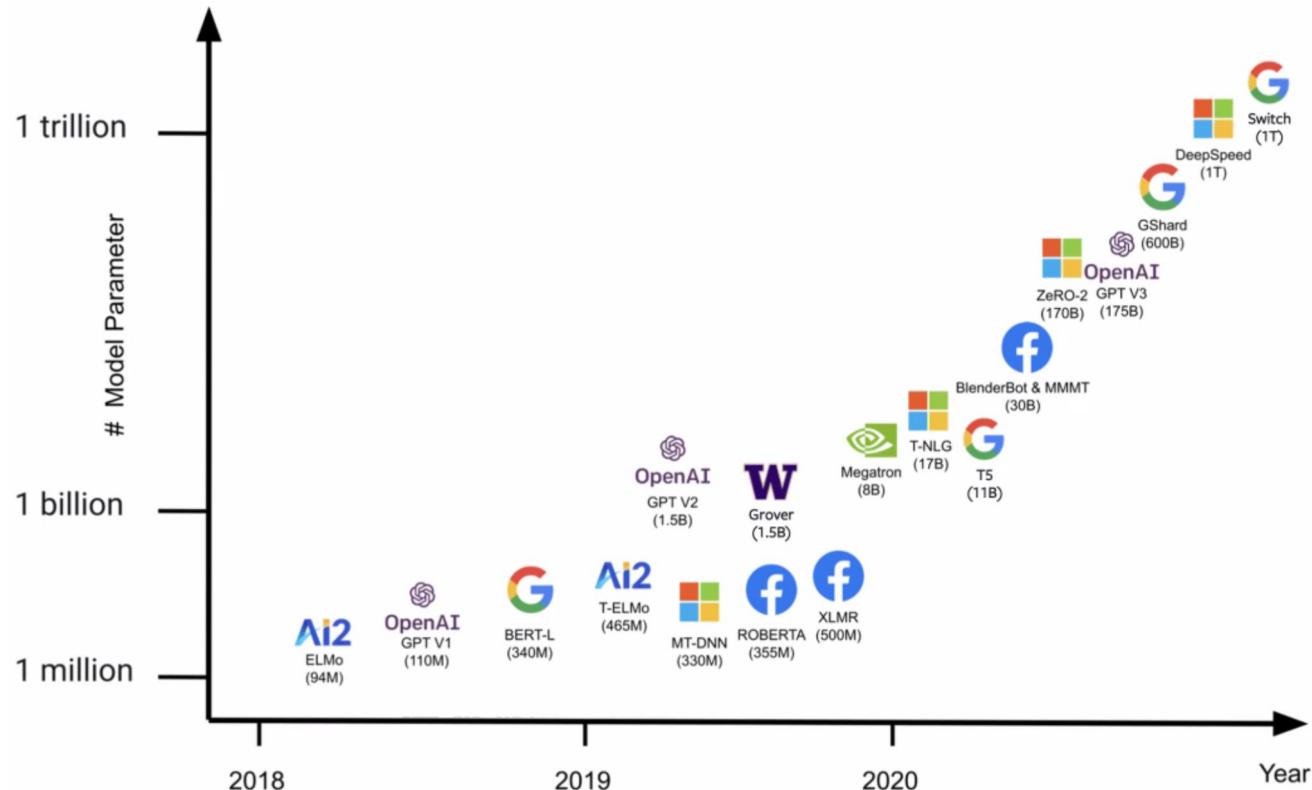


Figure Adapted from Myle Ott

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 squids 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.

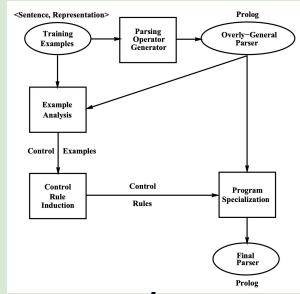
Tl;dr

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.

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

Zelle and Mooney (1996)



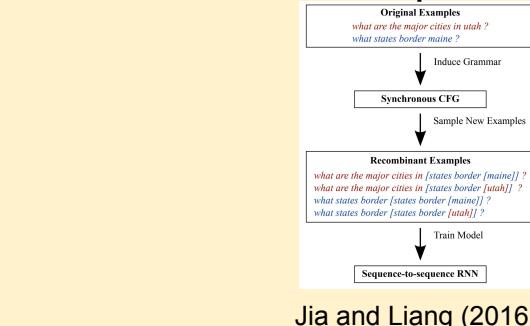
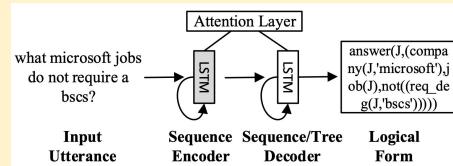
We now turn to issues of parsing and parameter estimation. Parsing under a PCCG involves computing the most probable logical form L for a sentence S ,

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

Zettlemoyer and Collins (2005)

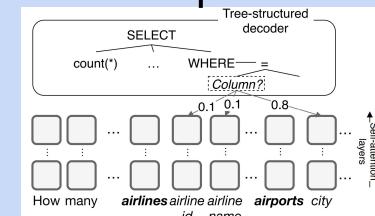
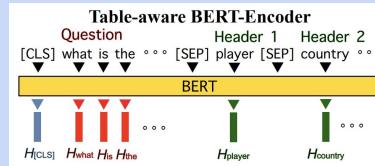
Non-Neural Network

Dong and Lapata (2016)



E2E Neural Networks

Hwang et al. (2019)

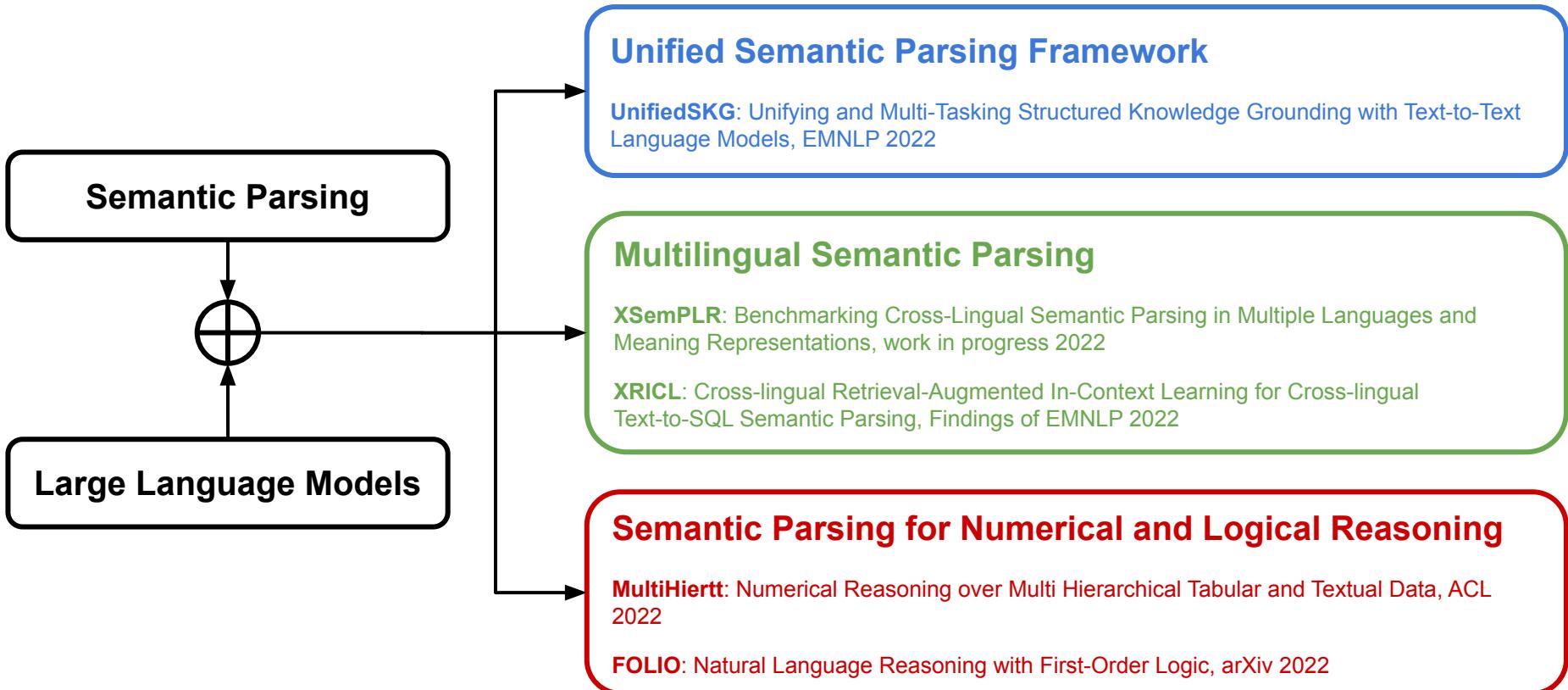


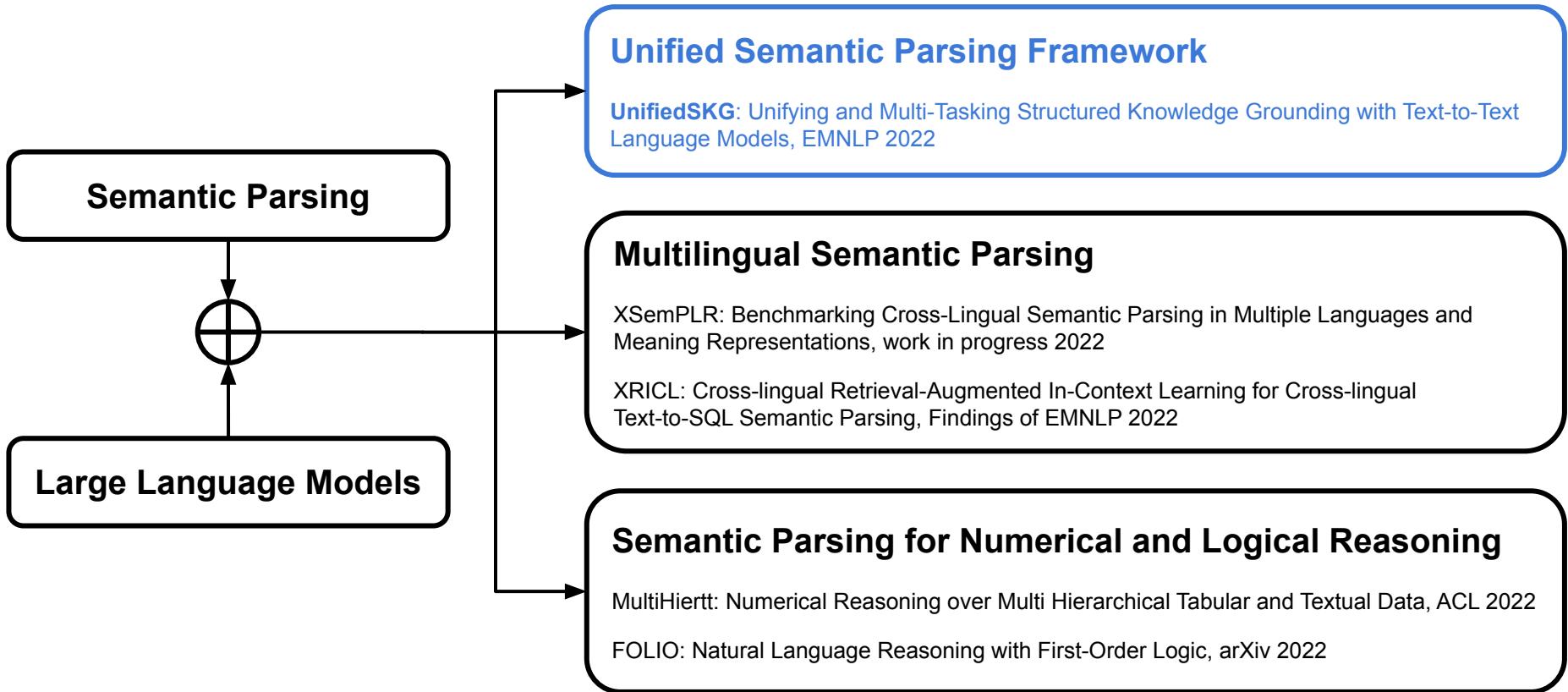
Wang et al. (2019)

Contextualized Embeddings and Pretrained Language Models

LLMs

Large Language Models Help Semantic Parsing in Three Ways





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

Xie et al., EMNLP 2022

<https://github.com/hkunlp/unifiedskg>

Many NLP Tasks need Structured Knowledge

Semantic Parsing

Which players did win the Australian Open?

Question Answering

Greece held its last Summer Olympics in which year?

Data-to-Text Generation

Describe the table result.

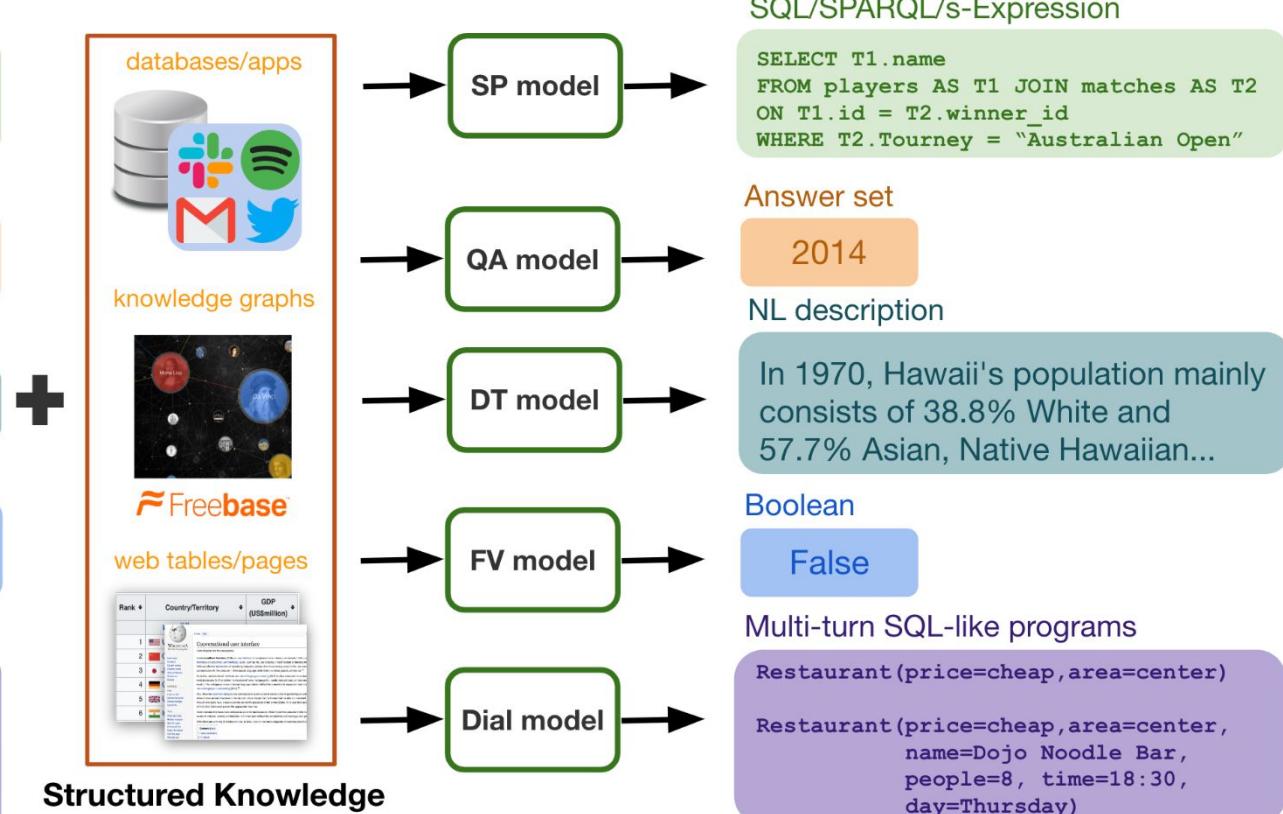
Fact Verification

Canada obtained 3 more gold medals than Mexico.

Dialogs

I am looking for a cheap restaurant in the city center.

Book a table for 8 at 18:30 on Thursday.



Unified Structured Knowledge Grounding

Semantic Parsing

Which players did win the Australian Open?

Question Answering

Greece held its last Summer Olympics in which year?

Data-to-Text Generation

Describe the table result.

Fact Verification

Canada obtained 3 more gold medals than Mexico.

Dialogs

I am looking for a cheap restaurant in the city center.

Book a table for 8 at 18:30 on Thursday.



Structured Knowledge

SQL/SPARQL/s-Expression

```
SELECT T1.name  
FROM players AS T1 JOIN matches AS T2  
ON T1.id = T2.winner_id  
WHERE T2.Tourney = "Australian Open"
```

Answer set

2014

NL description

In 1970, Hawaii's population mainly consists of 38.8% White and 57.7% Asian, Native Hawaiian...

Boolean

False

Multi-turn SQL-like programs

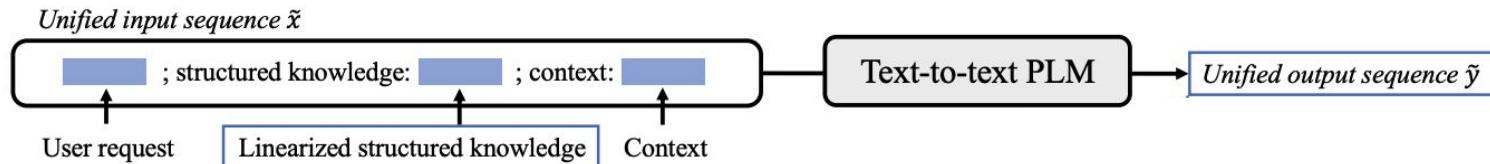
```
Restaurant(price=cheap,area=center)
```

```
Restaurant(price=cheap,area=center,  
name=Dojo Noodle Bar,  
people=8, time=18:30,  
day=Thursday)
```

Unify 21 Tasks Across 6 Task Families

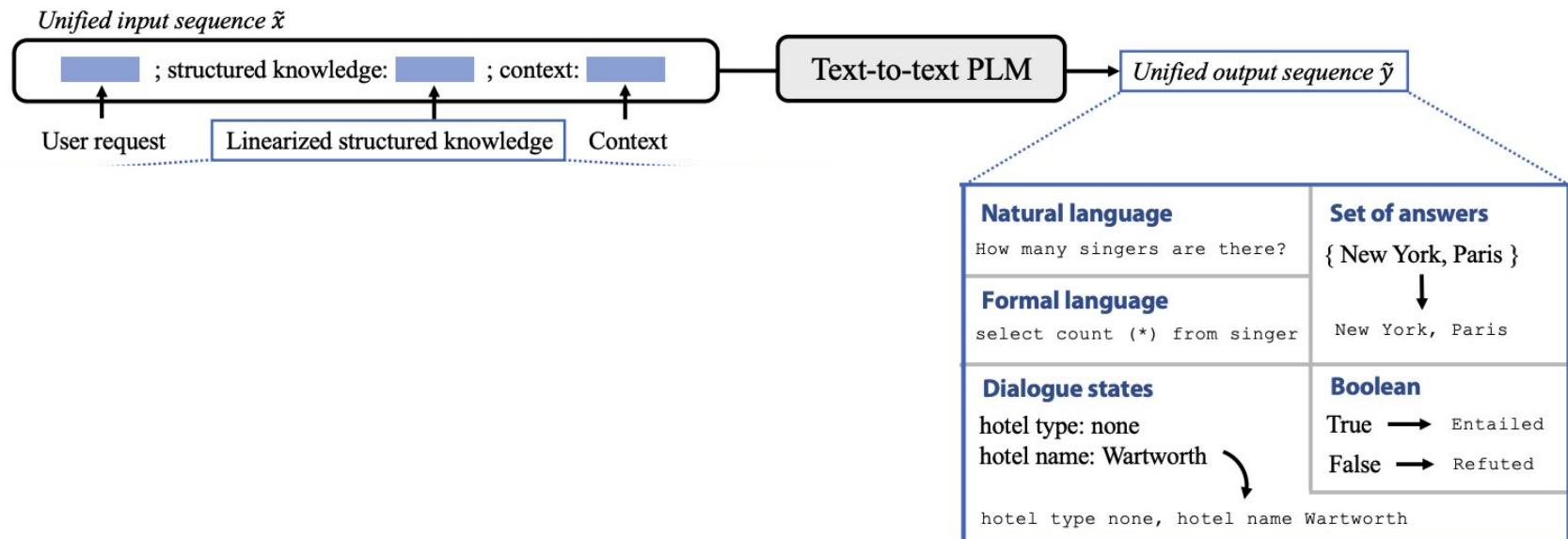
Task Family	Task	Knowledge Input	User Input	Output
<i>Semantic Parsing</i>	Spider (Yu et al., 2018)	Database	Question	SQL
	GrailQA (Gu et al., 2021)	Knowledge Graph	Question	s-Expression
	WebQSP (Yih et al., 2016)	Knowledge Graph	Question	s-Expression
	MTOP (Li et al., 2021)	API Calls	Question	TOP Representation
<i>Question Answering</i>	WikiSQL (Zhong et al., 2017)	Table	Question	Answer
	WikiTQ (Pasupat and Liang, 2015)	Table	Question	Answer
	CompWebQ (Talmor and Berant, 2018)	Knowledge Graph	Question	Answer
	HybridQA (Chen et al., 2020c)	Table + Text Passage	Question	Answer
	MultiModalQA (Talmor et al., 2021)	Table + Text + Image	Question	Answer
	FeTaQA (Nan et al., 2021a)	Table	Question	Free-Form Answer
<i>Data-to-Text</i>	DART (Nan et al., 2021b)	Triple	None	Text
	ToTTo (Parikh et al., 2020)	Highlighted Table	None	Text
<i>Conversational</i>	MultiWoZ (Budzianowski et al., 2018)	Ontology	Dialog	Dialog State
	KVRET (Eric et al., 2017)	Table	Dialog	Response
	SParC (Yu et al., 2019b)	Database	Multi turn	SQL
	CoSQL (Yu et al., 2019a)	Database	Dialog	SQL
	SQA (Iyyer et al., 2017)	Table	Multi turn	Answer
<i>Fact Verification</i>	TabFact (Chen et al., 2020b)	Table	Statement	Boolean
	FEVEROUS (Aly et al., 2021)	Table + Text	Statement	Boolean
<i>Formal-Language-to-Text</i>	SQL2Text (Shu et al., 2021)	Optional Database	SQL	Text
	Logic2Text (Chen et al., 2020d)	Table Schema	Python-like program	Text

Unified Architecture using Text-to-Text Pretrained Language Models



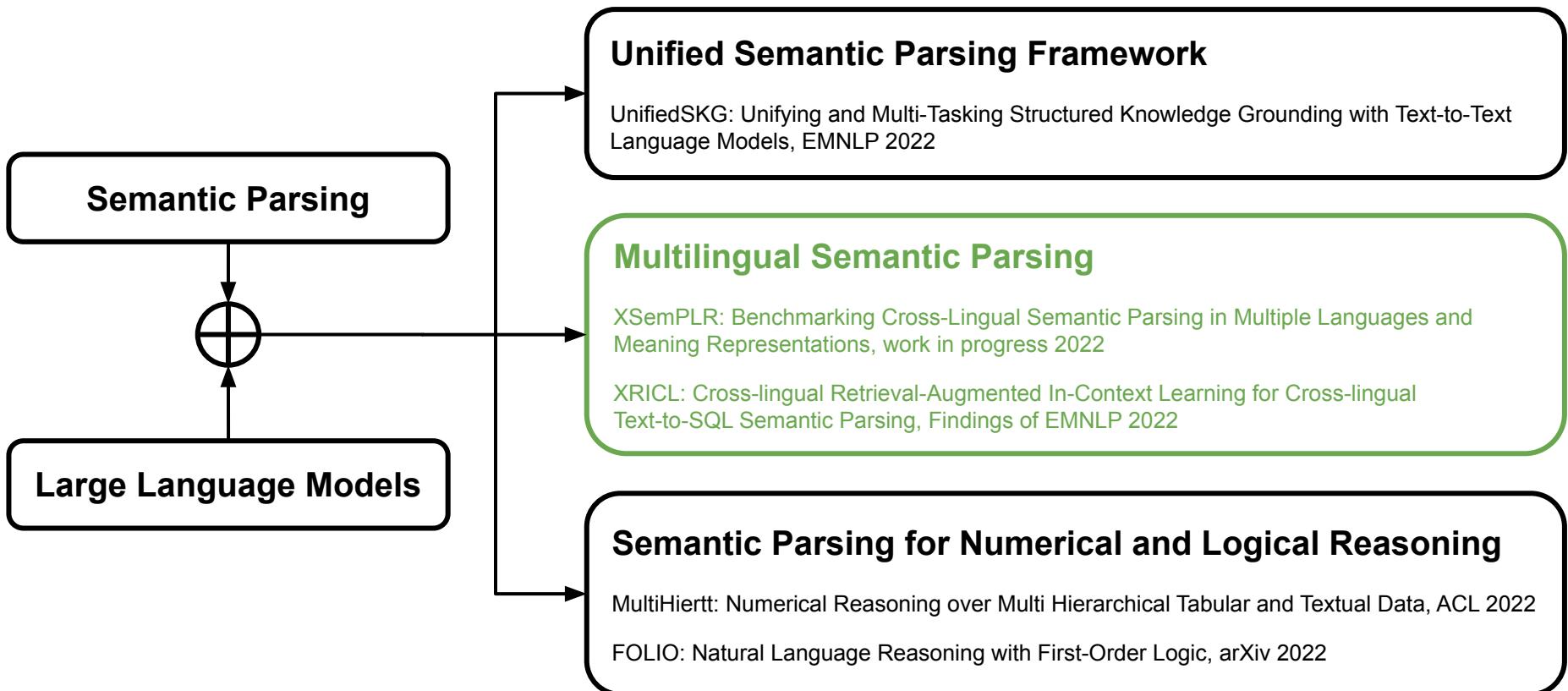
Tables		Triples	Knowledge graph	Ontology
player	no.	col : player no. row 1 : antonio 21 row 2 : voshon 2	(Mars Hill College, joined, 1973) (Mars Hill College, location, Mars Hill)	Cut Bank film.cinematography Ben Richardson Cut Bank film.language English
antonio	21	Mars Hill College : joined : 1973 Mars Hill College : location : Mars Hill	Ben Richardson English ↓	hotel-pricerange from {cheap, expensive dontcare} hotel-parking from {free, no, yes, dontcare}

Unified Architecture using Text-to-Text Pretrained Language Models



One Architecture, Multiple SOTA

	Metric	T5-base	T5-large	T5-3B	Previous sota (w/o extra)
Spider (dev.)	Match	58.12	66.63	71.76	75.5⁺ (Scholak et al., 2021)
GrailQA	Match	62.39	67.30	70.11	83.8⁺ (Ye et al., 2021b)
WebQSP	F1	78.83	79.45	80.70	83.6⁺ (Ye et al., 2021b)
MTOP	Match	85.49	86.17	86.78	86.36 (Pasupat et al., 2021)
WikiTQ	Acc	35.76	43.22	49.29	44.5 (Wang et al., 2019)
WikiSQL	Acc	82.63	84.80	85.96	85.8 (Liu et al., 2021)
CompWebQ	Acc	68.43	71.38	73.26	70.4 [†] (Das et al., 2021)
HybridQA (dev.)	Acc	54.07	56.95	59.41	60.8 [†] (Eisenschlos et al., 2021)
MultiModalQA (dev.)	F1	75.51	81.84	85.28	82.7 (Yoran et al., 2021)
FeTaQA	BLEU	29.91	32.45	33.44	30.54 (Nan et al., 2021a)
DART	BLEU	46.22	46.89	46.66	46.89 (Nan et al., 2021b)
ToTTo (dev.)	BLEU	48.29	48.95	48.95	48.95 (Kale and Rastogi, 2020)
MultiWoZ2.1	Joint Acc	54.64	54.45	55.42	60.61[*] (Dai et al., 2021)
KVRET	Micro F1	66.45	65.85	67.88	63.6 (Gou et al., 2021)
SParC (dev.)	Match	50.54	56.69	61.51	54.1 (Hui et al., 2021)
CoSQL (dev.)	Match	42.30	48.26	54.08	56.9⁺ (Scholak et al., 2021)
SQA	Overall Acc	52.91	61.28	62.37	58.6 (Liu et al., 2021)
TabFact	Acc	76.13	80.85	83.68	74.4 (Yang et al., 2020)
FEVEROUS (dev.)	Acc	75.05	79.81	82.40	82.38 (Aly et al., 2021)
SQL2Text	BLEC	93.52	93.68	94.78	93.7 (Shu et al., 2021)
Logic2Text	BLEC	90.66	90.57	91.39	88.6 (Shu et al., 2021)

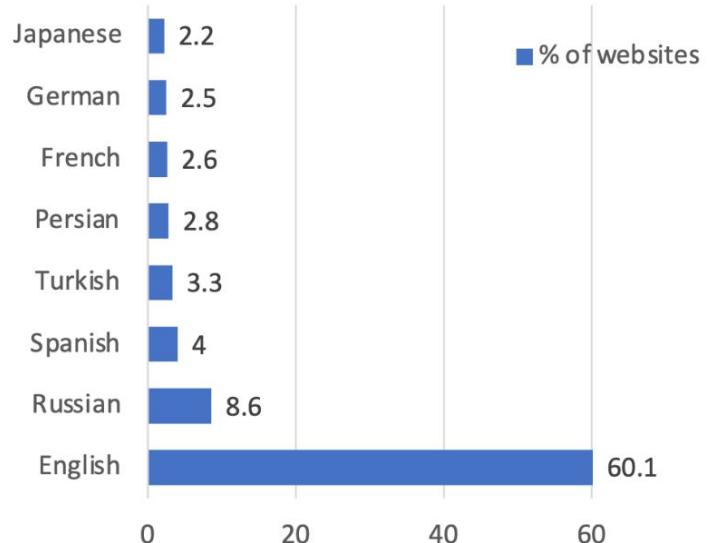


XSemPLR: Benchmarking Cross-Lingual Semantic Parsing in Multiple Languages and Meaning Representations

Zhang et al., Work in Progress

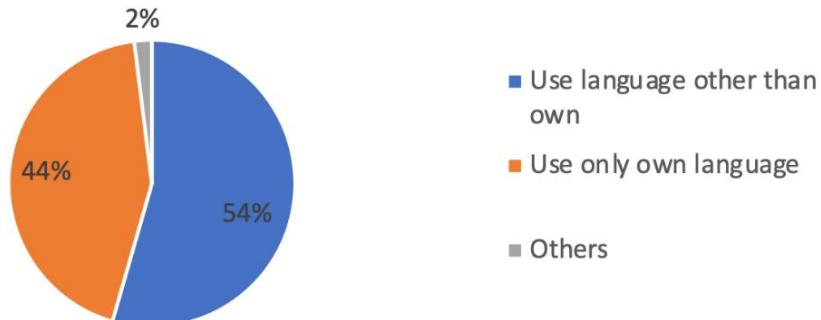
Multilingual Information Access

Usage Statistics of website content



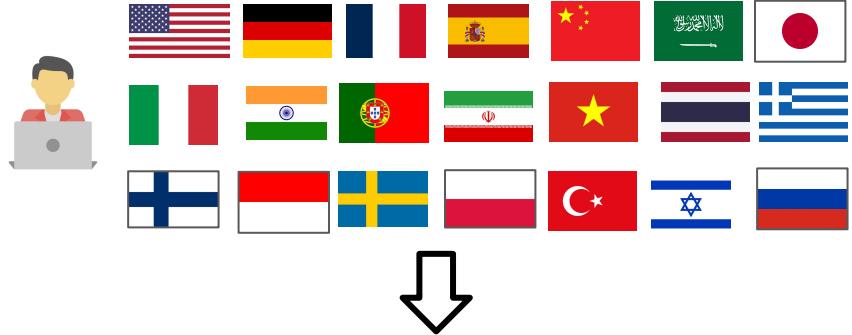
(Data from Technology report: [Usage statistics of content languages for websites](#))

Languages used to reach / watch web content

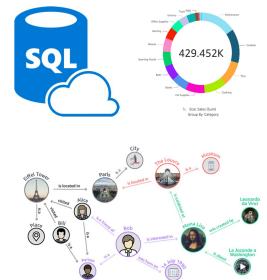
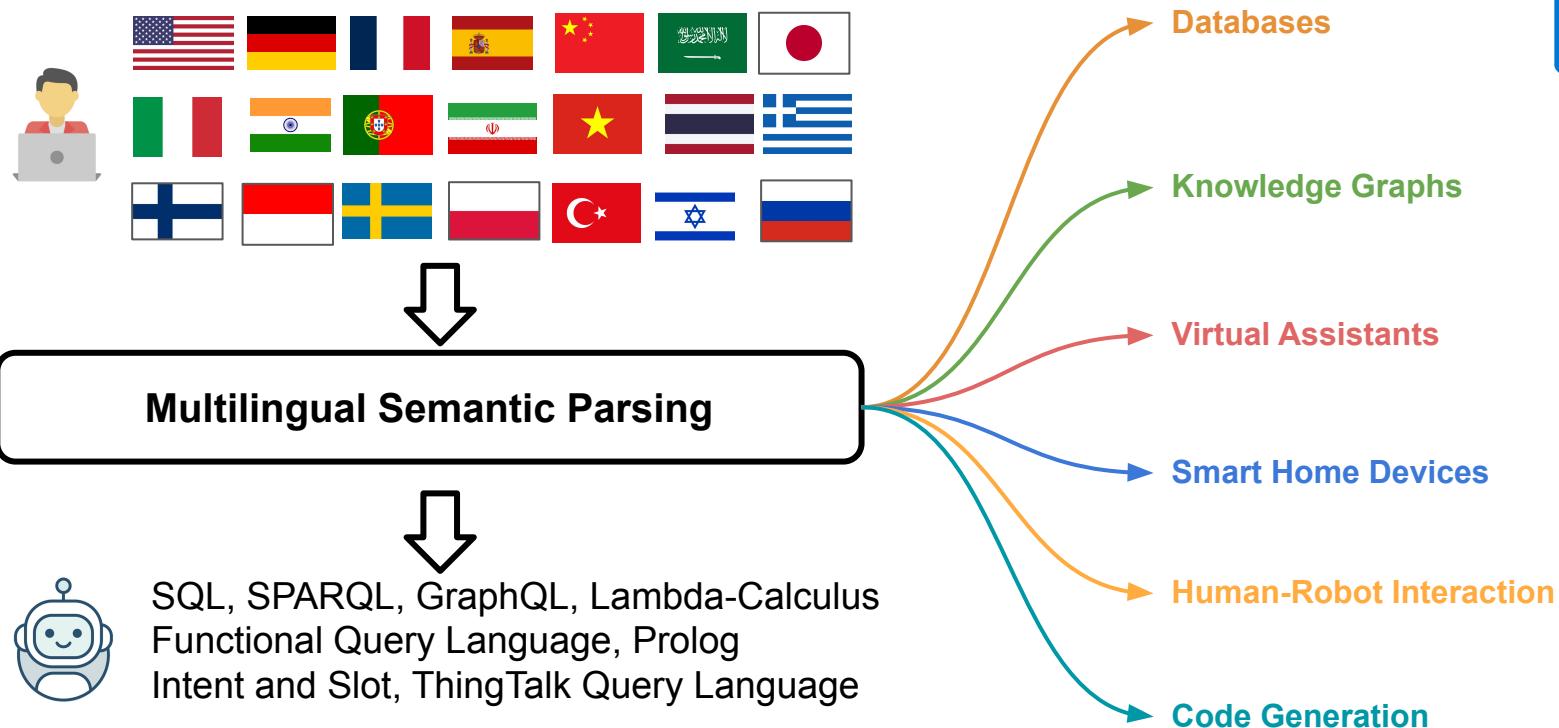


User language preferences online (EU)

Multilingual Semantic Parsing

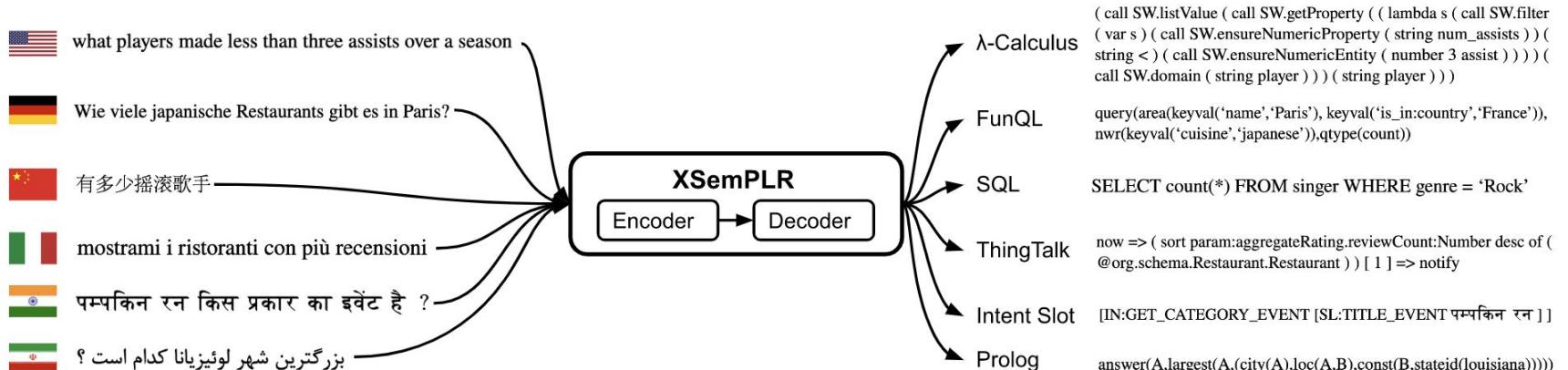


Multilingual Semantic Parsing



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

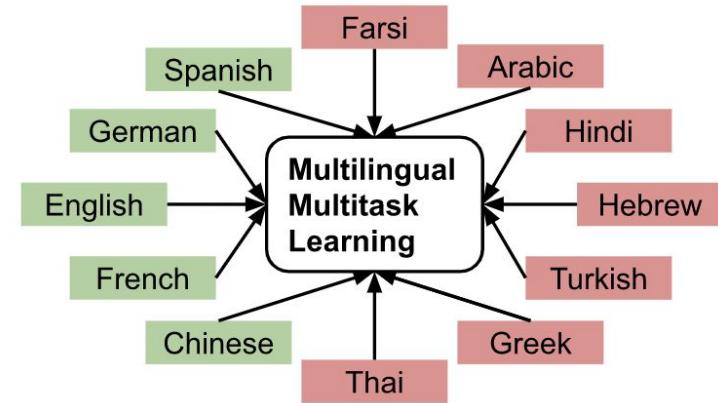


mT5 is the Best

	ATIS	GeoQuery	Spider	NLmaps	Overnight	MCWQ	Schema2QA	MTOP	Average
<i>Translate-Test</i>									
mT5	44.50	65.91	45.26	66.36	59.69	19.85	3.18*	29.78*	50.26
<i>In-language Monolingual</i>									
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

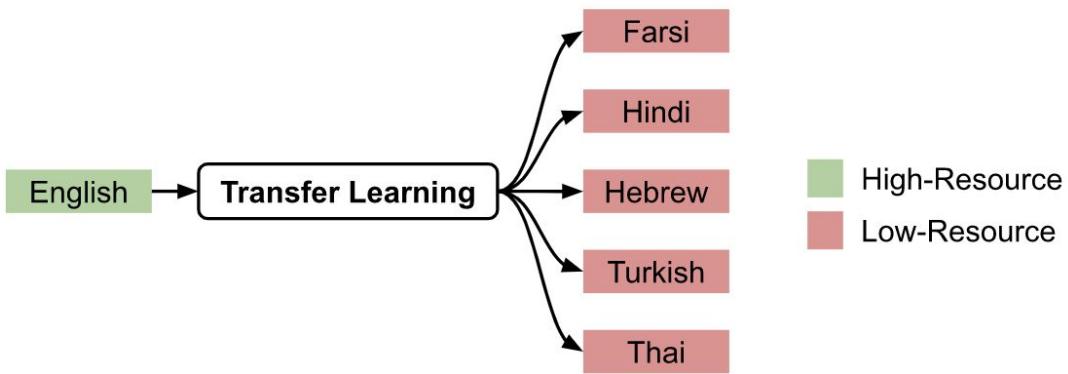
High-Resource
Low-Resource



Multilingual Multitask Learning Helps

	ATIS	GeoQuery	Spider	NLmaps	Overnight	MCWQ	Schema2QA	MTOP	Average
<i>Translate-Test</i>									
mT5	44.50	65.91	45.26	66.36	59.69	19.85	3.18*	29.78*	50.26
<i>In-language Monolingual</i>									
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
<i>In-language Monolingual Few-Shot</i>									
mT5	22.26	7.48	25.57	26.93	9.17	0.77	22.61	61.90	22.09
<i>In-language Multilingual</i>									
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
<i>Translate-Test</i>									
mT5	44.50	65.91	45.26	66.36	59.69	19.85	3.18*	29.78*	50.26
<i>In-language Monolingual</i>									
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
<i>In-language Monolingual Few-Shot</i>									
mT5	22.26	7.48	25.57	26.93	9.17	0.77	22.61	61.90	22.09
<i>In-language Multilingual</i>									
mT5	54.45	82.04	–	–	–	–	60.92	82.95	70.09
<i>Cross-lingual Zero-Shot Transfer</i>									
mT5	31.85	39.40	41.93	34.89	52.68	4.06	44.04	50.18	37.38
<i>Cross-lingual Few-Shot Transfer</i>									
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.

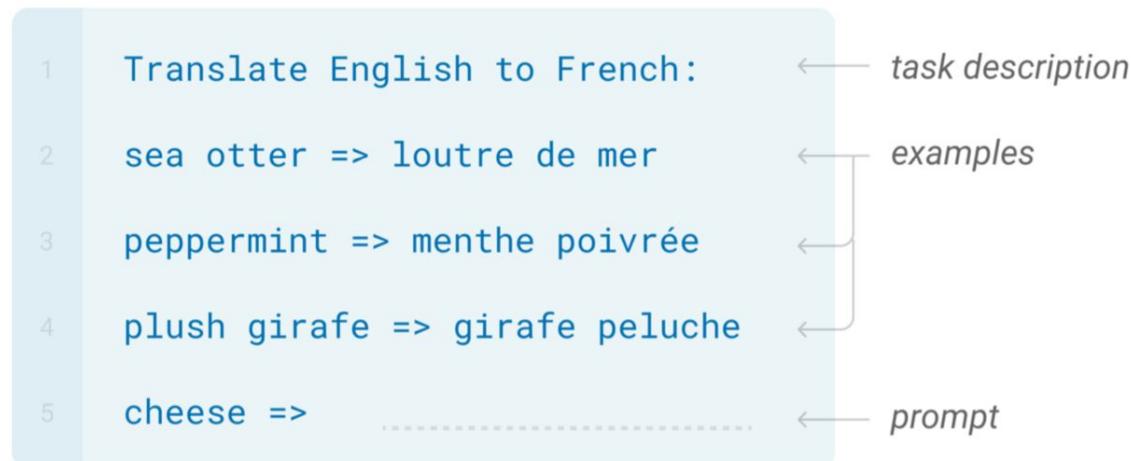
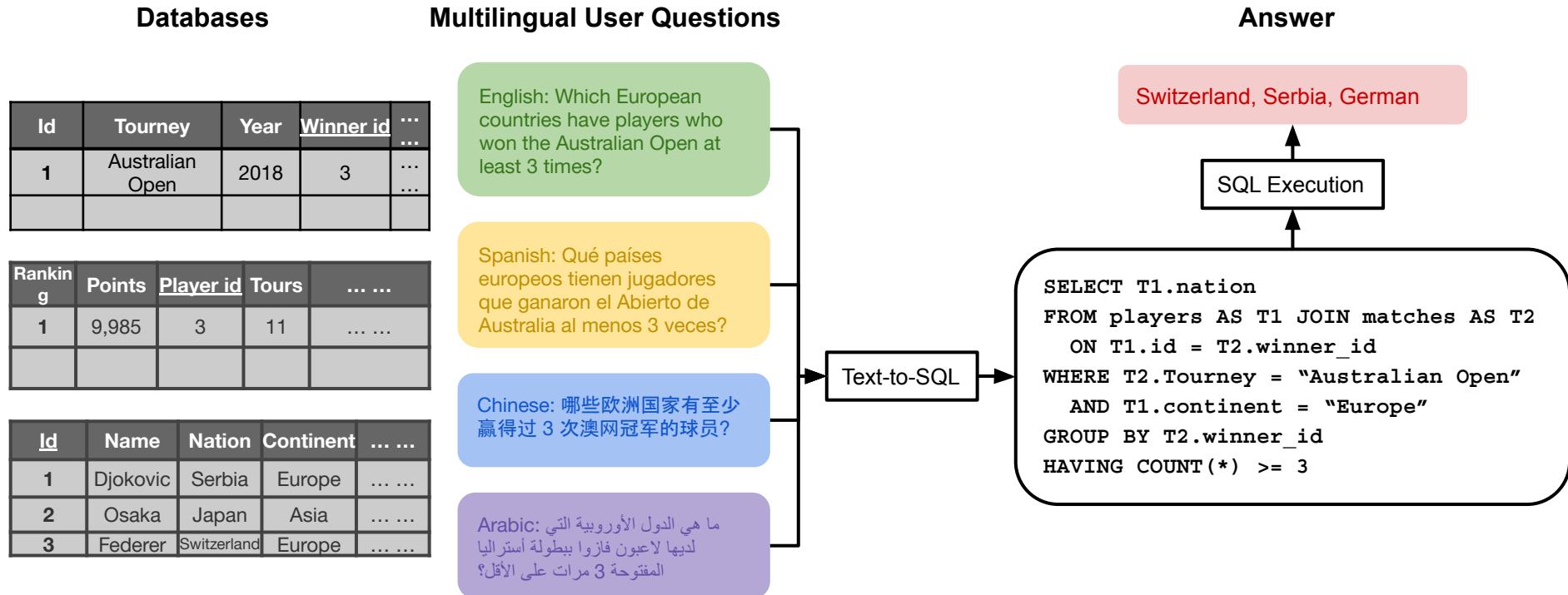


Figure from <http://ai.stanford.edu/blog/in-context-learning/>

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.

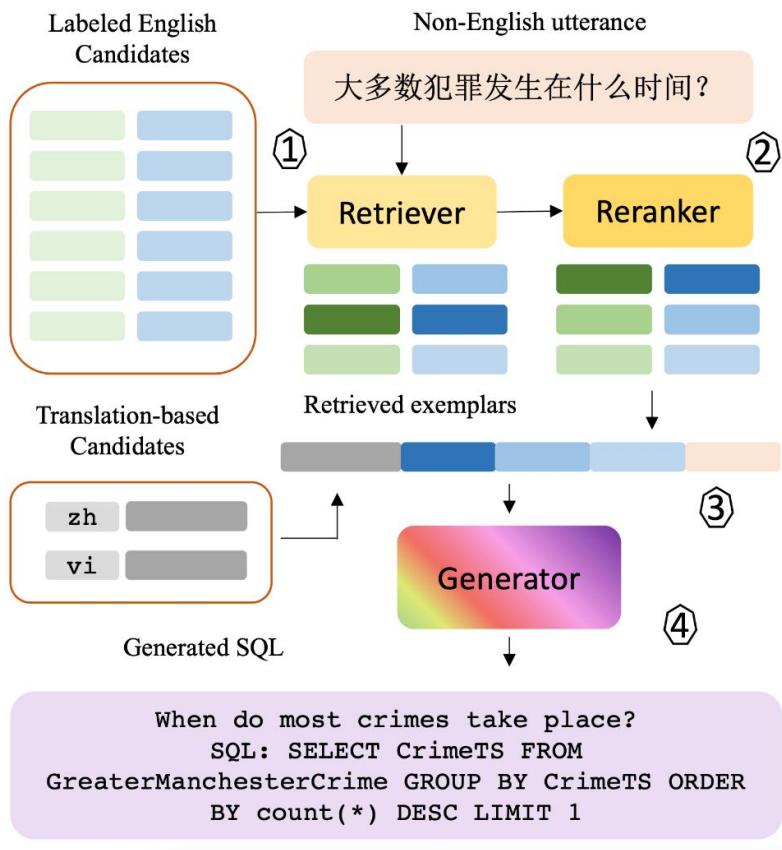
Challenges

1. Find most relevant English examples.
2. Facilitate translation.

Solutions

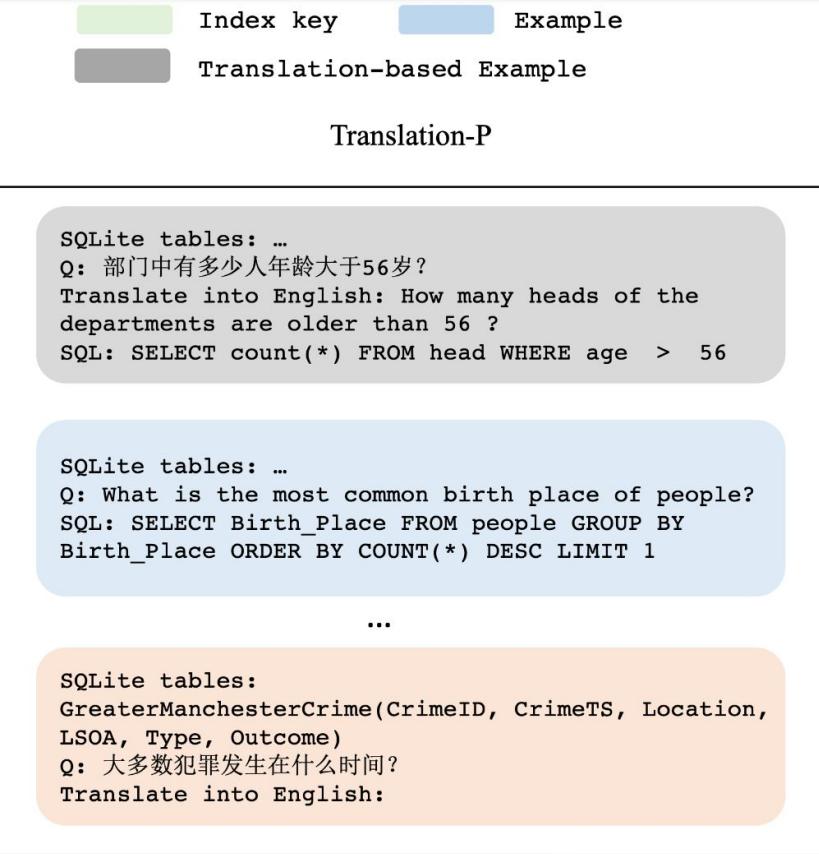
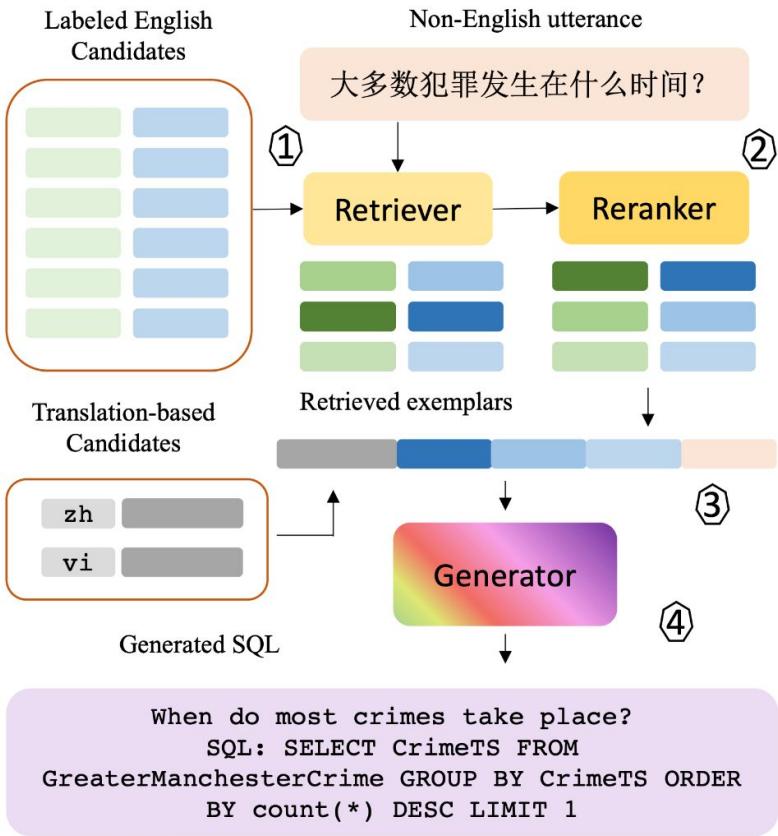
- Cross-lingual Retrieval
→ Translation-based Prompt

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

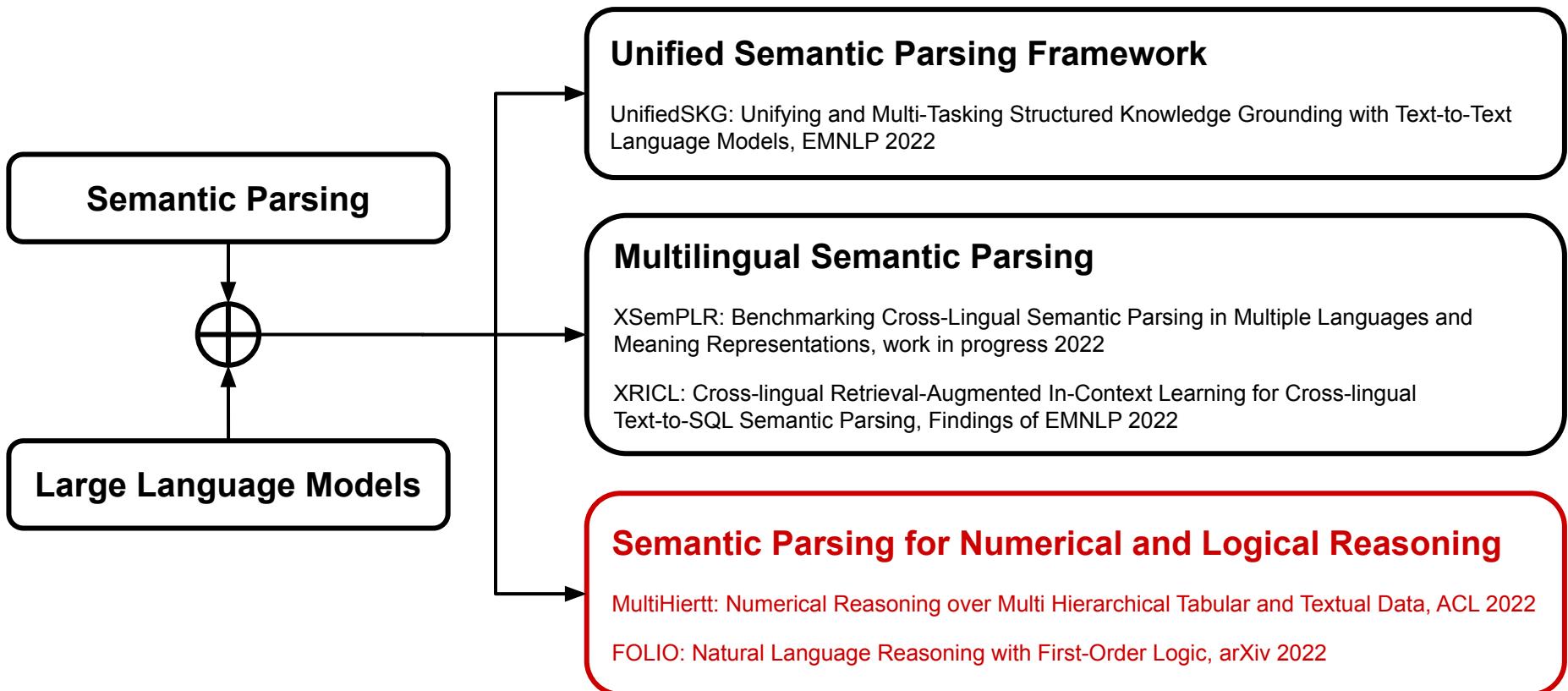
Model	zh-full		zh		vi		fa		hi	
	EM	TS	EM	TS	EM	TS	EM	TS	EM	TS
(1) mT5 zero-shot	39.7	47.9	48.4	42.1	40.1	41.3	39.5	41.2	39.7	
(2) mUSE	38.4	43.0	46.8	31.8	33.4	28.9	31.1	22.2	23.7	
(3) mSBERT	37.9	41.3	47.1	34.6	33.5	29.3	31.8	22.0	22.3	
(4) mT5-encoder	44.4	48.1	51.4	41.3	39.5	38.4	38.5	28.6	27.0	
(5) 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	TS	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



MultiHierTT: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data

Yilun Zhao, Yunxiang Li, Chenying Li, Rui Zhang

<https://github.com/psunlpgroup/MultiHierTT>

ACL 2022

Numerical Reasoning over Text

Problem	
<p>Gwen was organizing her book case making sure each of the shelves had exactly 9 books on it. She has 2 types of books - mystery books and picture books. If she had 3 shelves of mystery books and 5 shelves of picture books, how many books did she have total?</p>	
Solution	Expression Tree of Solution
$(3 + 5) \times 9 = 72$	<pre>graph TD; x((x)) --> plus((+)); x --> nine((9)); plus --> three((3)); plus --> five((5))</pre>

(Roy and Roth, EMNLP 2015)

Numerical Reasoning over Tables

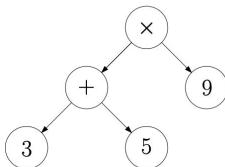
Problem

Gwen was organizing her book case making sure each of the shelves had exactly 9 books on it. She has 2 types of books - mystery books and picture books. If she had 3 shelves of mystery books and 5 shelves of picture books, how many books did she have total?

Solution

$$(3 + 5) \times 9 = 72$$

Expression Tree of Solution

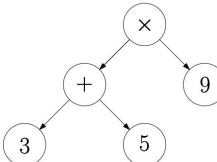


(Roy and Roth, EMNLP 2015)

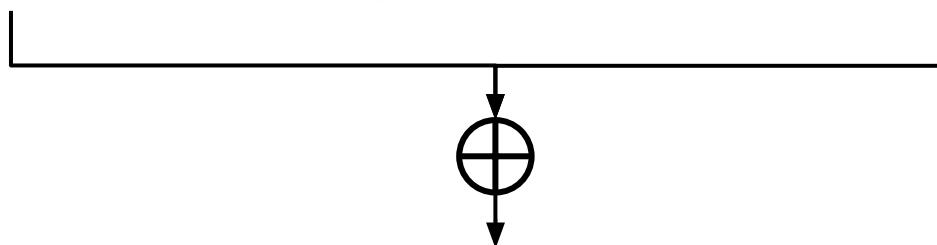
	A	B	C	D
1	Table 2: Decomposition of changes in participation rates from 1996 to 2016, men	Both	Men	Women
2				percent
3				
4	Actual			
5	1996	23.8	32.2	16.6
6	2007	33.3	40.1	27.3
7	2016	37.7	43.5	32.4
8	2016 Counterfactual			
9	With 1996 age structure only	35.9	42.6	30.1
10	With 1996 education only	30.6	37.7	24.3
11	With 1996 family structure only	33.7	39.2	28.5
12	With 1996 age, family and education structure	31.6	39.1	25.4
13	What percentage of overall change in participation rates among women was caused by compositional effects?			
14	=1 - (D12 - D5) / (D7 - D5)			

(Cheng, ACL 2022)

Our Research Question: Numerical Reasoning from Both?

Problem	
<p>Gwen was organizing her book case making sure each of the shelves had exactly 9 books on it. She has 2 types of books - mystery books and picture books. If she had 3 shelves of mystery books and 5 shelves of picture books, how many books did she have total?</p>	
Solution	Expression Tree of Solution
$(3 + 5) \times 9 = 72$	

A	B	C	D
1 Table 2: Decomposition of changes in participation rates from 1996 to 2016, men	Both	Men	Women
2			percent
3			
4 Actual			
5 1996	23.8	32.2	16.6
6 2007	33.3	40.1	27.3
7 2016	37.7	43.5	32.4
8 2016 Counterfactual			
9 With 1996 age structure only	35.9	42.6	30.1
10 With 1996 education only	30.6	37.7	24.3
11 With 1996 family structure only	33.7	39.2	28.5
12 With 1996 age, family and education structure	31.6	39.1	25.4
13 What percentage of overall change in participation rates among women was caused by compositional effects?			
14 $=1 - (D12 - D5) / (D7 - D5)$			



Numerical Reasoning over Tabular and Textual Data

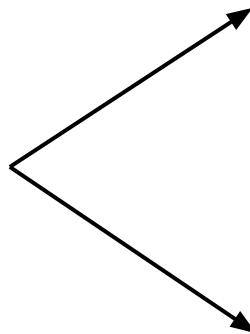
Question: In 2018, what was the total Expenses in the segment that involves F-35 volume?

Answer: 11685 millions

Numerical expression:

$$9889 + 1796$$

Both Tables and Text



Document:

(... abbreviate...)

The following table presents product and service sales and operating expenses by segment (dollar in millions):

Segment	Year Ended December 31			
	2018	2017	Sales	Expenses
Innovation Systems				
Product	2,894	2,582	—	—
Service	382	351	—	—
Aerospace Systems				
Product	11,087	9,889	10,064	8,988
Service	2,009	1,796	2,067	1,854
Mission Systems				
Product	7,329	6,335	7,012	6,088
Service	4,380	3,854	4,458	3,940
Technology Service				
Product	485	450	391	360
Service	3,812	3,404	4,296	3,878

Product sales for 2018 increased \$4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of \$2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

(... abbreviate...)

The table below reconciles funds provided to each segment (dollar in millions):

Segment	2018		% Change
	Funded	Funded	
Innovation Systems	5,928	—	—
Aerospace Systems	11,448	9,560	19.7 %
Mission Systems	9,676	9,277	4.3 %
Technology Services	2,883	2,792	3.3 %

Approximately \$26.6 billion of the \$53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019. (... abbreviate...)

Question: In 2018, what was the total Expenses in the segment that involves F-35 volume?

Answer: 11685 millions

Numerical expression:

$$9889 + 1796$$

Complex Table Hierarchy

Document:

(... abbreviate...)

The following table presents product and service sales and operating expenses by segment (dollar in millions):

Segment	Year Ended December 31			
	2018	2017	Sales	Expenses
Innovation Systems				
Product	2,894	2,582	—	—
Service	382	351	—	—
Aerospace Systems				
Product	11,087	9,889	10,064	8,988
Service	2,009	1,796	2,067	1,854
Mission Systems				
Product	7,329	6,335	7,012	6,088
Service	4,380	3,854	4,458	3,940
Technology Service				
Product	485	450	391	360
Service	3,812	3,404	4,296	3,878

Product sales for 2018 increased \$4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of \$2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

(... abbreviate...)

The table below reconciles funds provided to each segment (dollar in millions):

Segment	2018		% Change
	Funded	Funded	
Innovation Systems	5,928	—	—
Aerospace Systems	11,448	9,560	19.7 %
Mission Systems	9,676	9,277	4.3 %
Technology Services	2,883	2,792	3.3 %

Approximately \$26.6 billion of the \$53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019. (... abbreviate...)

Question: In 2018, what was the total Expenses in the segment that involves F-35 volume?

Answer: 11685 millions

Numerical expression:

$$9889 + 1796$$

Fact Retrieval Across Tables and Text

Document:

(... abbreviate...)

The following table presents product and service sales and operating expenses by segment (dollar in millions):

Segment	Year Ended December 31			
	2018	2017	Sales	Expenses
Innovation Systems				
Product	2,894	2,582	—	—
Service	382	351	—	—
Aerospace Systems				
Product	11,087	9,889	10,064	8,988
Service	2,009	1,796	2,067	1,854
Mission Systems				
Product	7,329	6,335	7,012	6,088
Service	4,380	3,854	4,458	3,940
Technology Service				
Product	485	450	391	360
Service	3,812	3,404	4,296	3,878

Product sales for 2018 increased \$4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of \$2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

(... abbreviate...)

The table below reconciles funds provided to each segment (dollar in millions):

Segment	2018		% Change
	Funded	Funded	
Innovation Systems	5,928	—	—
Aerospace Systems	11,448	9,560	19.7 %
Mission Systems	9,676	9,277	4.3 %
Technology Services	2,883	2,792	3.3 %

Approximately \$26.6 billion of the \$53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019. (... abbreviate...)

Question: In 2018, what was the total sales increase in the segment with most funds in 2017?

Answer: 965

Numerical expression: $(11087 - 10064) + (2009 - 2067)$

Fact Retrieval Across Different Tables

Document:

(... abbreviate...)

The following table presents product and service sales and operating expenses by segment (dollar in millions):

Segment	Year Ended December 31		
	2018	2017	
Sales	Expenses	Sales	Expenses
Innovation Systems			
Product	2,894	2,582	—
Service	382	351	—
Aerospace Systems			
Product	11,087	9,889	10,064
Service	2,009	1,796	2,067
Mission Systems			
Product	7,329	6,335	7,012
Service	4,380	3,854	4,458
Technology Service			
Product	485	450	391
Service	3,812	3,404	4,296
			3,878

Product sales for 2018 increased \$4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of \$2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

(... abbreviate...)

The table below reconciles funds provided to each segment (dollar in millions):

Segment	2018		2017
	Funded	Funded	% Change
Innovation Systems	5,928	—	—
Aerospace Systems	11,448	9,560	19.7 %
Mission Systems	9,676	9,277	4.3 %
Technology Services	2,883	2,792	3.3 %

Approximately \$26.6 billion of the \$53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019. (... abbreviate...)

Question: In 2018, what was the total sales increase in the segment with most funds in 2017?

Answer: 965

Numerical expression: $(11087 - 10064) + (2009 - 2067)$



Multi-hop Reasoning by Generating Arithmetic Equations

Document:

(... abbreviate...)

The following table presents product and service sales and operating expenses by segment (dollar in millions):

Segment	Year Ended December 31		
	2018	2017	
Sales	Expenses	Sales	Expenses
Innovation Systems			
Product	2,894	2,582	—
Service	382	351	—
Aerospace Systems			
Product	11,087	9,889	10,064
Service	2,009	1,796	2,067
Mission Systems			
Product	7,329	6,335	7,012
Service	4,380	3,854	4,458
Technology Service			
Product	485	450	391
Service	3,812	3,404	4,296
			3,878

Product sales for 2018 increased \$4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of \$2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

(... abbreviate...)

The table below reconciles funds provided to each segment (dollar in millions):

Segment	2018		2017
	Funded	Funded	% Change
Innovation Systems	5,928	—	—
Aerospace Systems	11,448	9,560	19.7 %
Mission Systems	9,676	9,277	4.3 %
Technology Services	2,883	2,792	3.3 %

Approximately \$26.6 billion of the \$53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019. (... abbreviate...)

MT2Net Model

Whole Document containing Multiple hierarchical tables and paragraphs

The following table presents product and service sales and operating expenses by segment (dollar in millions):

Segment	Year Ended December 31			
	2018	2017	Sales	Expenses
Innovation Systems				
Product	2,894	2,582	—	—
Service	382	351	—	—
Aerospace Systems				
Product	11,087	9,889	10,064	8,988
Service	2,009	1,796	2,067	1,854
Mission Systems				
Product	7,329	6,335	7,012	6,088
Service	4,380	3,854	4,458	3,940
Technology Service				
Product	485	450	391	360
Service	3,812	3,404	4,296	3,878

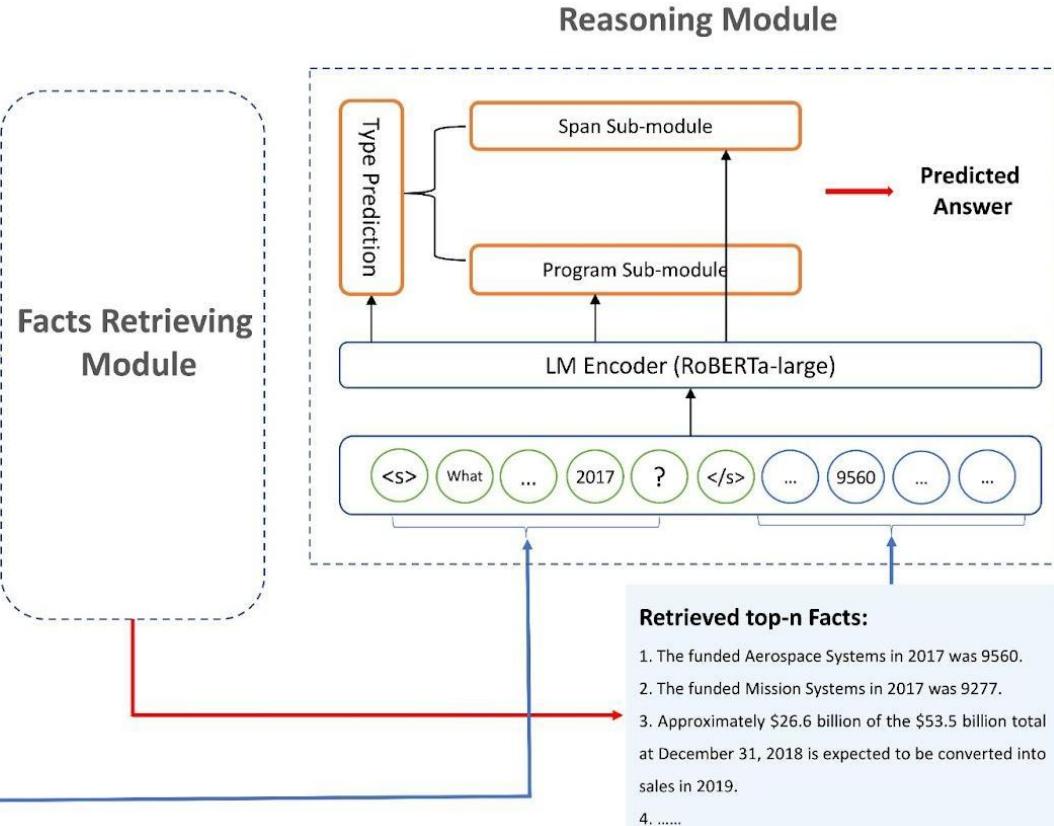
Product sales for 2018 increased \$4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of \$2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

Segment	2018		2017	
	Funded	Funded	% Change	
Innovation Systems	5,928	—	—	—
Aerospace Systems	11,448	9,560	19.7 %	
Mission Systems	9,676	9,277	4.3 %	
Technology Services	2,883	2,792	3.3 %	

Approximately \$26.6 billion of the \$53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019.

Question

What was the total sales increase in the segment with most funds in 2017?



More Challenging with More Tables and Reasoning Steps

Performance Breakdown	EM	F ₁
Regarding supporting facts coverage		
text-only questions	49.26	53.29
table-only questions	36.77	38.55
w/ \geq 2 tables	24.32	24.96
table-text questions	33.04	35.15
w/ \geq 2 tables	21.04	23.36
Regarding numerical reasoning steps		
1 step	43.62	47.80
2 steps	34.67	37.91
3 steps	22.43	24.57
> 3 steps	15.14	17.19
Full Results	36.22	38.43

FOLIO: Natural Language Reasoning with First-Order Logic

Han et al., arXiv 2022

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

Logical Reasoning: Deductive vs Inductive

DEDUCTION

IDEA



All men are mortal.

OBSERVATIONS



Jason is a man.

CONCLUSION



Jason is mortal.

INDUCTION

OBSERVATIONS



I break out when I eat peanuts.

ANALYSIS



This is a symptom of being allergic.

THEORY



I am allergic to peanuts.

Syllogism

Major premise: All men are mortal.

Minor premise: Socrates is a man.

Conclusion: Therefore, Socrates is mortal.

Syllogism by Venn Diagram

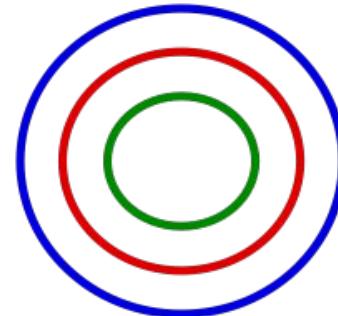
AAA-1 Modus Barbara

Major premise: All men are mortal.

Minor premise: Socrates is a man.

Conclusion: Therefore, Socrates is mortal.

- M a P All M are P,
- S a M and all S are M;
- S a P thus all S are P.



M: men

S: Socrates

P: 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.

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

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 conclusions -> Labels

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

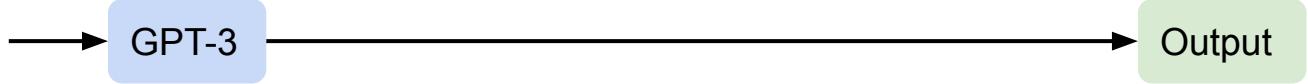
FOLIO is Different

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 ~ 5	101	48
ProofWriter	500k	Deductive	Synthetic	✗	1 ~ 5	101	48
LogicNLI	20k	FOL	Synthetic	✗	1 ~ 5	1077	30
FOLIO (ours)	1,435	FOL	Expert-written, Real-world	✓	1 ~ 8	4351	76

Source	#Stories	#Premises	#Conclusions	NL		FOL							
				Vocab	#Words	∀	∃	¬	∧	∨	→		
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

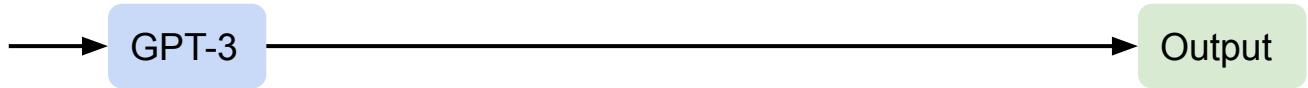
Large Language Models as Soft Logic Reasoners

[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

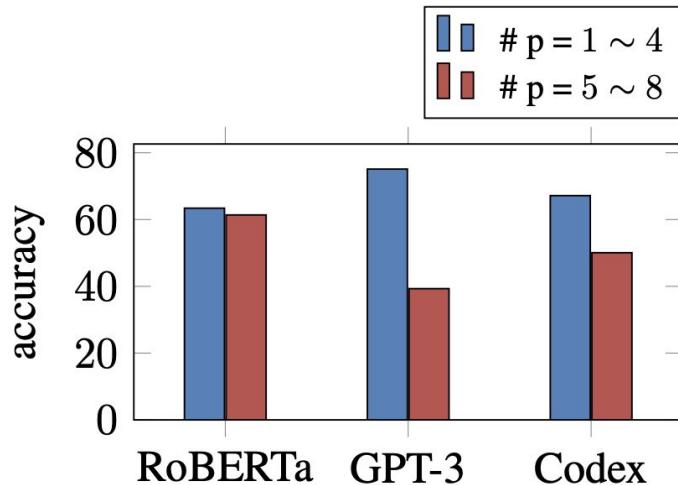
[8 NL-Label Examples] Premises: All men are mortal. Socrates is a man. Conclusion: Therefore, Socrates is mortal. True, False, or Unknown?



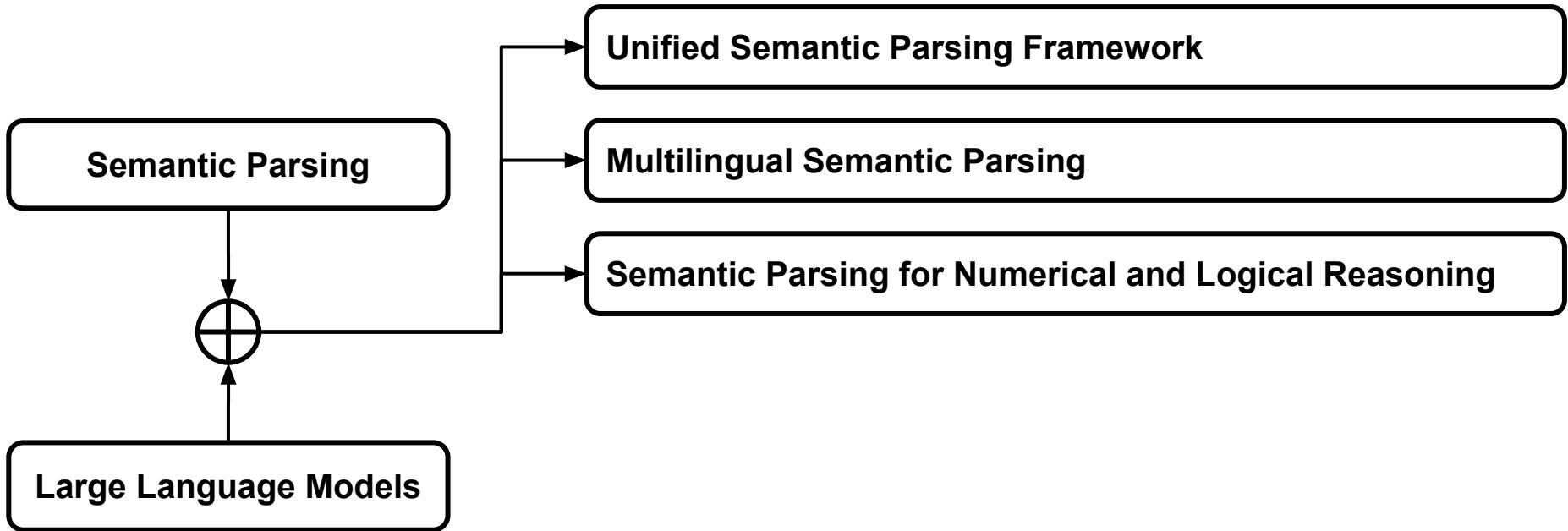
[8 NL-FOL Examples] Premises: All men are mortal. Socrates is a man. Conclusion: Therefore, Socrates is mortal. FOL is



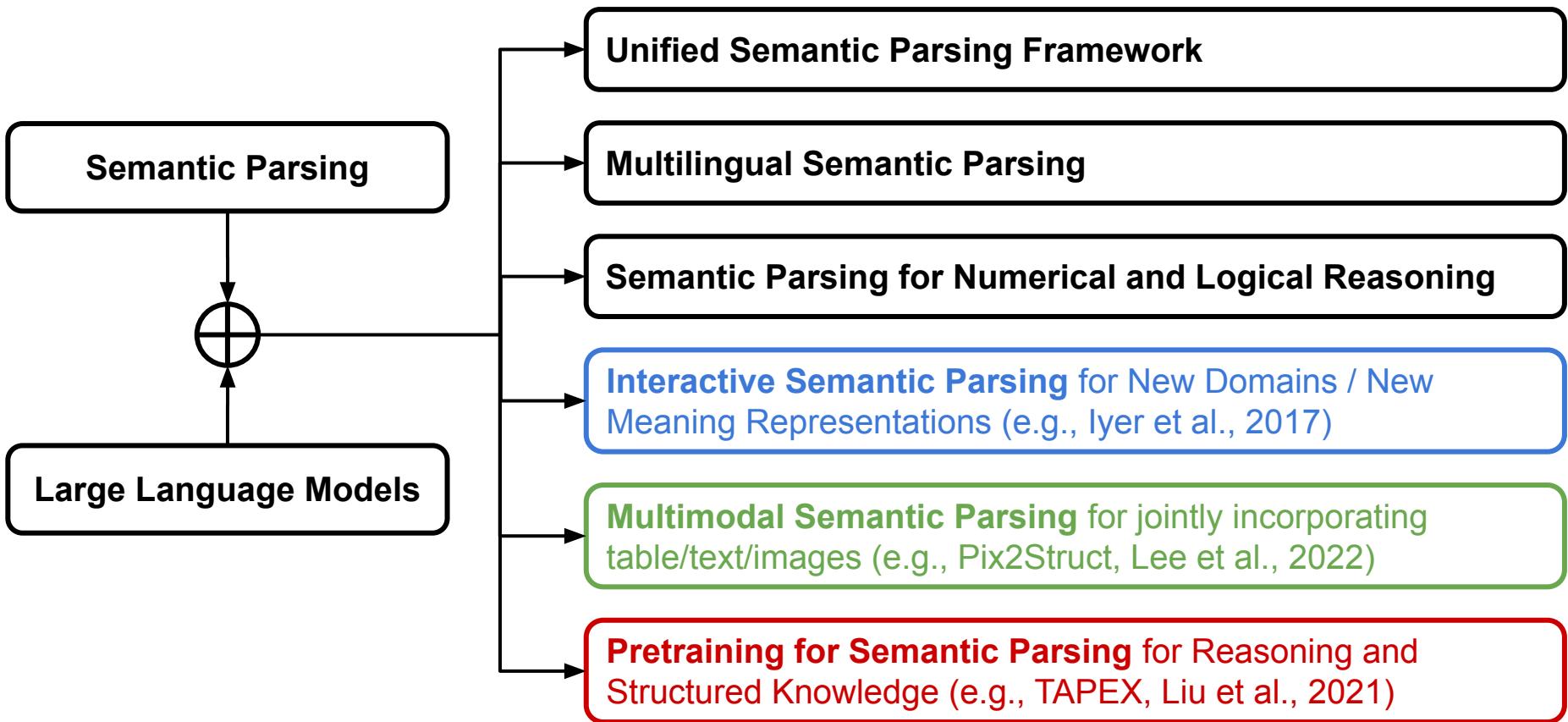
FOLIO is Challenging for Large Language Models



Conclusions



Future Directions



Thanks! Any Questions?

