Text-Table Understanding and Text2SQL

Jingfeng Yang

The Multimodality World

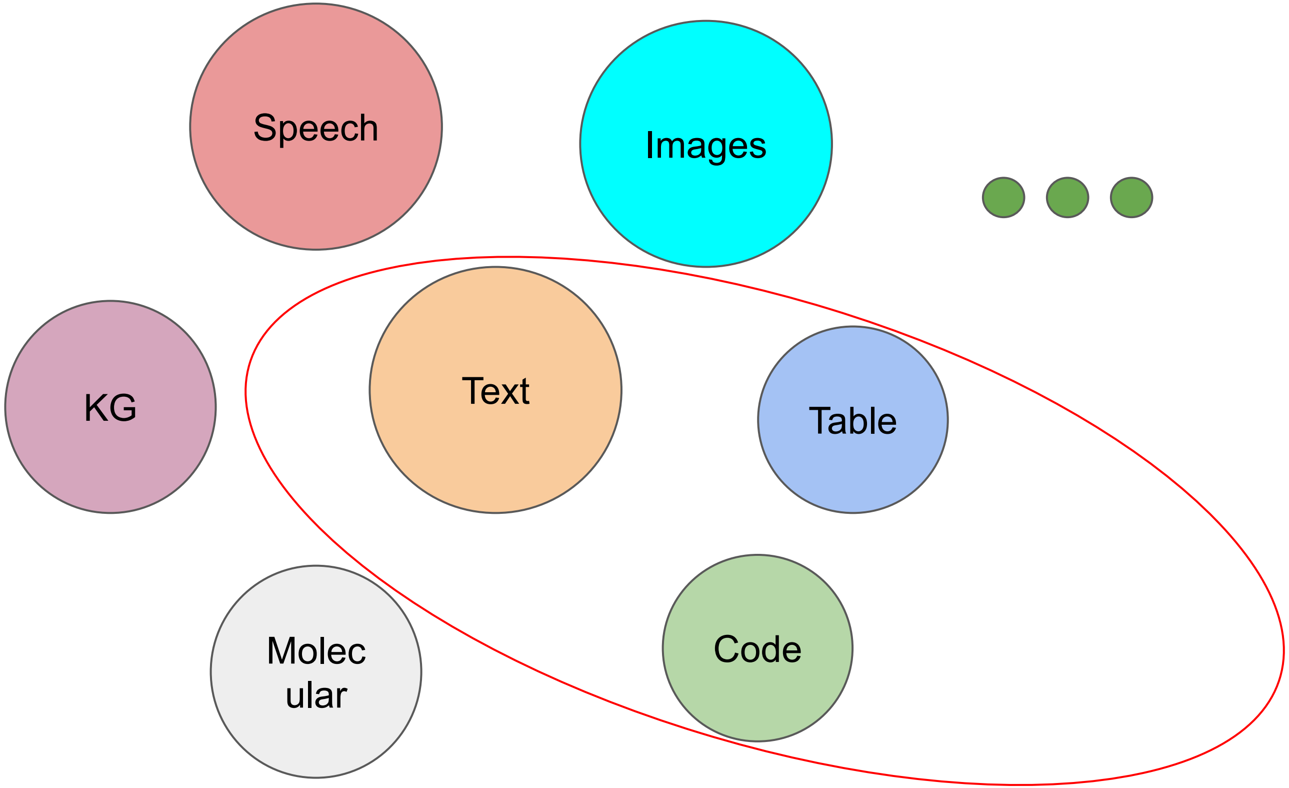


Table-Text Understanding

Original intent:
What super hero from Earth appeared most recently?

1. Who are all of the super heroes?

2. Which of them come from Earth?

3. Of those, who appeared most recently?

Legion of Super Heroes Post-*Infinite Crisis*

Character	*First Appeared*	*Home World*	*Powers*
Night Girl	2007	Kathoon	Super strength
Dragonwing	2010	Earth	Fire breath
Gates	2009	Vyrga	Teleporting
XS	2009	Aarok	Super speed
Harmonia	2011	Earth	Elemental

Sequential QA dataset (SQA) (Iyyer et al., 2017)

Two Fashions of Table-Text Understanding

- Given table-text pairs, a model directly outputs labels or answers.
    - How to better encode table-text pairs? (ACL 2022)

- A model first transforms texts to Code (SQL), and then execute SQL queries on tables to get labels or answers.
    - How to better transform texts to SQL ? (NAACL 2022 Findings)

How to better encode table-text pairs?

TableFormer: Robust Transformer Modeling for Table-Text Encoding

Jingfeng Yang[*] **Aditya Gupta**[†] **Shyam Upadhyay**[†]
Luheng He[†] **Rahul Goel**[†] **Shachi Paul**[†]
[*]Georgia Institute of Technology
[†]Google Assistant
jingfengyangpku@gmail.com
tableformer@google.com

ACL 2022 (Oral)

Recent Approaches to Table-Text Modeling

- General Recipe
  - Step 1: Pretraining on text-table pairs
    - Pretraining on existing table-text corpus (Wikipedia, ToTTo etc.):
      - TaBERT (Yin et al., 2020)
      - TAPAS (Herzig et al., 2020)
      - StruG (Deng et al., 2021)
    - Data augmentation for pretraining
      - Intermediate pretraining (Eisenschlos et al., 2020)
      - GRAPPA (Yu et al., 2021)
      - TaPEx (Liu et al. 2022)
  - Step 2: Fine-tuning on specific dataset (e.g. SQA)

Problem 1: Non-Robust Modeling

Question: Of all song lengths, which one is the longest?
Gold Answer: 5:02

Title	Producers	Length
Screwed Up	Mr. Lee	5:02
Smile	Sean T	4:32
Ghetto Queen	I.N.F.O. & NOVA	5:00

Problem 1: Non-Robust Modeling

Question: Of all song lengths, which one is the longest?
Gold Answer: 5:02
TAPAS Predicted Answer: 5:00

Title	Producers	Length
Screwed Up	**Mr. Lee**	**5:02**
Smile	Sean T	4:32
Ghetto Queen	**I.N.F.O. & NOVA**	**5:00**

Problem 1: Non-Robust Modeling

Question: Of all song lengths, which one is the longest?
Gold Answer: 5:02
TAPAS Predicted Answer: 5:00

Title	Producers	Length
Screwed Up	**Mr. Lee**	**5:02**
Smile	Sean T	4:32
Ghetto Queen	**I.N.F.O. & NOVA**	**5:00**

TAPAS Predicted Answer After Perturbation: 5:02

Title	Producers	Length
Smile	Sean T	4:32
Ghetto Queen	I.N.F.O. & NOVA	5:00
Screwed Up	**Mr. Lee**	**5:02**

Model is not robust to row/column order changes!

Accuracy drops from 66.8 to 60.5 on SQA dataset after perturbation.

Problem 2: Lack of Structural Biases

Question: Which nation received 2 silver medals?
Gold Answer: Spain, Ukraine
TAPAS Predicted Answer: Spain

Nation	Gold	Silver	Bronze
Great Britain	2	1	2
Spain	1	2	0
Norway	1	0	0
Ukraine	0	2	0

Problem 2: Lack of Structural Biases

Question: Which nation received 2 silver medals?
Gold Answer: Spain, Ukraine
TAPAS Predicted Answer: Spain

Nation	Gold	**Silver**	Bronze
Great Britain	2	1	2
Spain	1	**2**	0
Norway	1	0	0
Ukraine	0	**2**	0

Identify “Silver” column and “2” cells in this column

Problem 2: Lack of Structural Biases

Question: Which nation received 2 silver medals?
Gold Answer: Spain, Ukraine
TAPAS Predicted Answer: Spain

Nation	Gold	**Silver**	Bronze
Great Britain	2	1	2
Spain	1	**2**	0
Norway	1	0	0
Ukraine	0	**2**	0

Output contents of the same rows in “Nation” column

TableFormer

Robust Table+Text Modeling

Table-Text (Relative) Attention Bias Types

Question: Which nation received 2 silver medals?

Relative Attention:

Nation	Silver
Spain	2
Norway	0
Ukraine	2

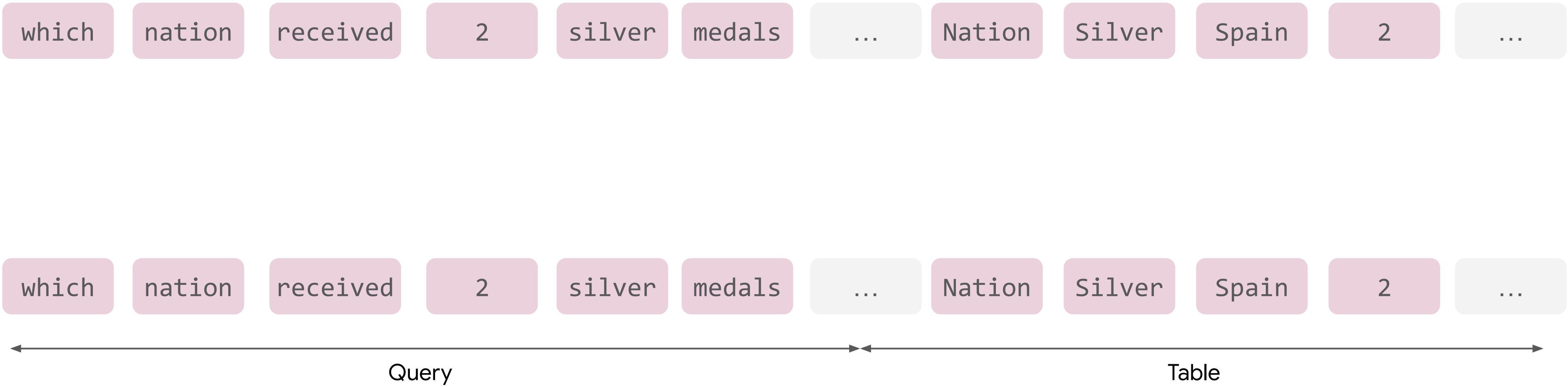


Table-Text (Relative) Attention Bias Types

Question: Which nation received 2 silver medals?

Nation	Silver
Spain	2
Norway	0
Ukraine	2

Relative Attention:

- **Header to Sentence**

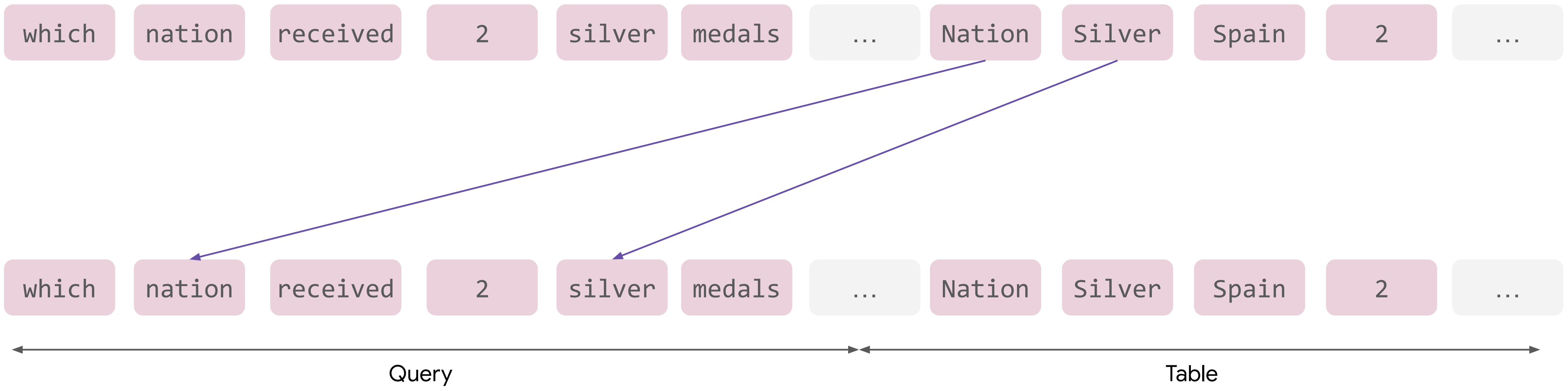


Table-Text (Relative) Attention Bias Types

Question: Which nation received 2 silver medals?

Nation	Silver
Spain	2
Norway	0
Ukraine	2

Relative Attention:

- Header to Sentence
- Cell to Sentence

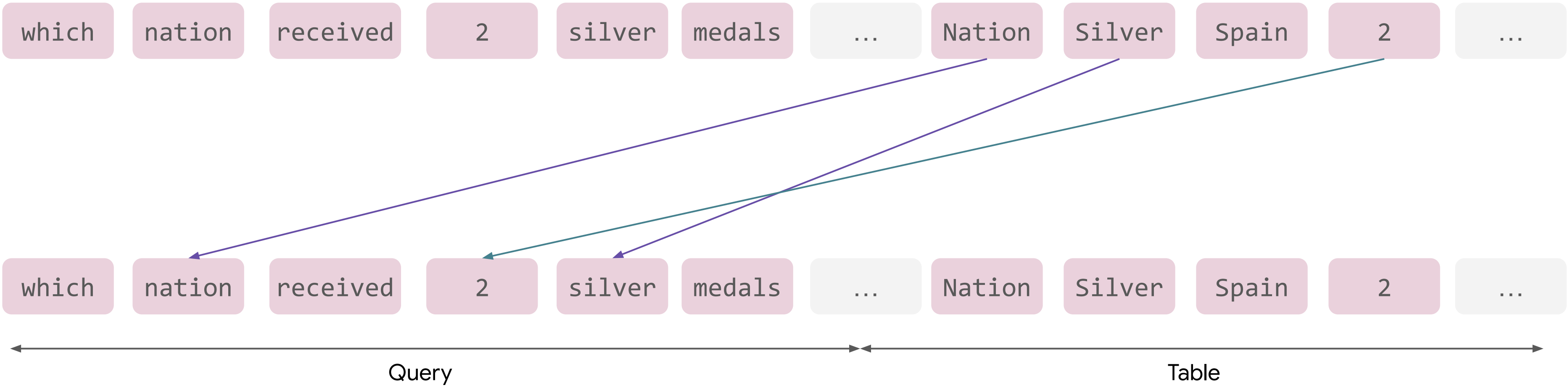


Table-Text (Relative) Attention Bias Types

Question: Which nation received 2 silver medals?

Nation	Silver
Spain	2
Norway	0
Ukraine	2

Relative Attention:

- **Header to Sentence**
- **Cell to Sentence**
- **Cell to Column Header**

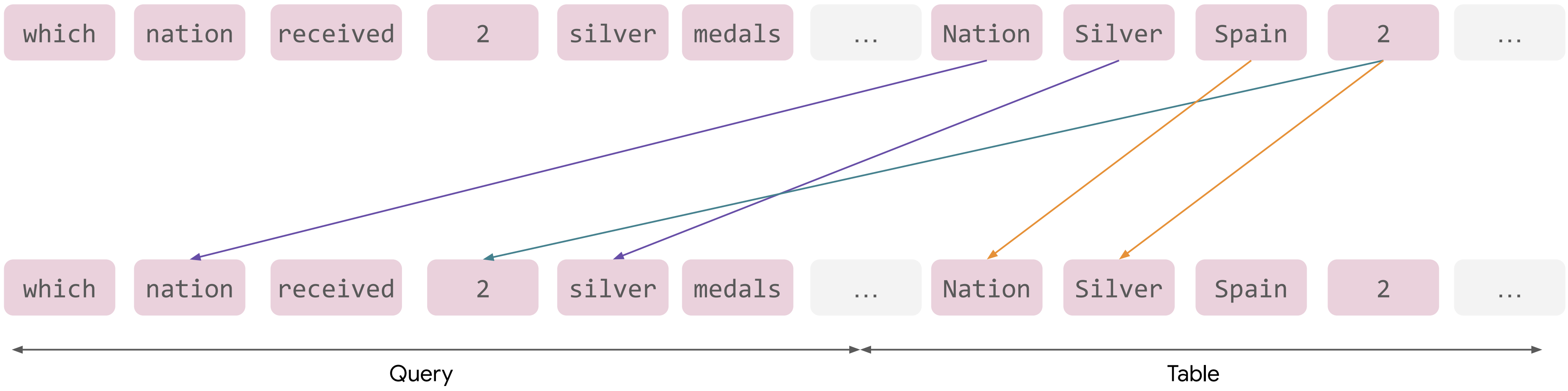


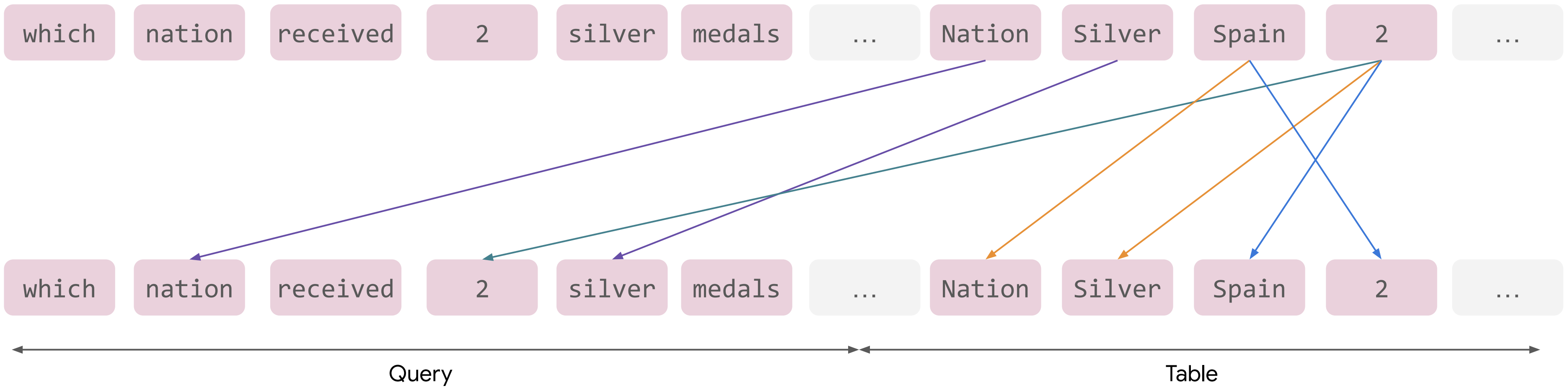
Table-Text (Relative) Attention Bias Types

Question: Which nation received 2 silver medals?

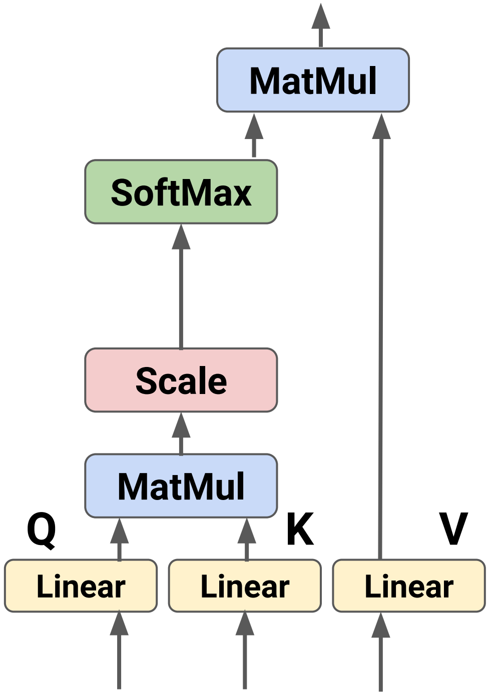
Nation	Silver
Spain	2
Norway	0
Ukraine	2

Relative Attention:

- Header to Sentence
- Cell to Sentence
- Cell to Column Header
- Same Row
- ...



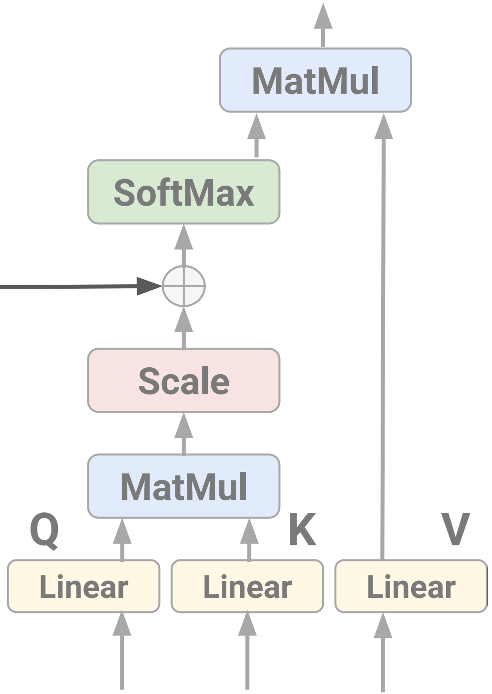
Transformer (Vaswani et al. 2017)



$$\mathrm{Attn}(H) = \mathrm{softmax}(\frac{QK^\top}{\sqrt{d_K}})V$$

TableFormer (our work)

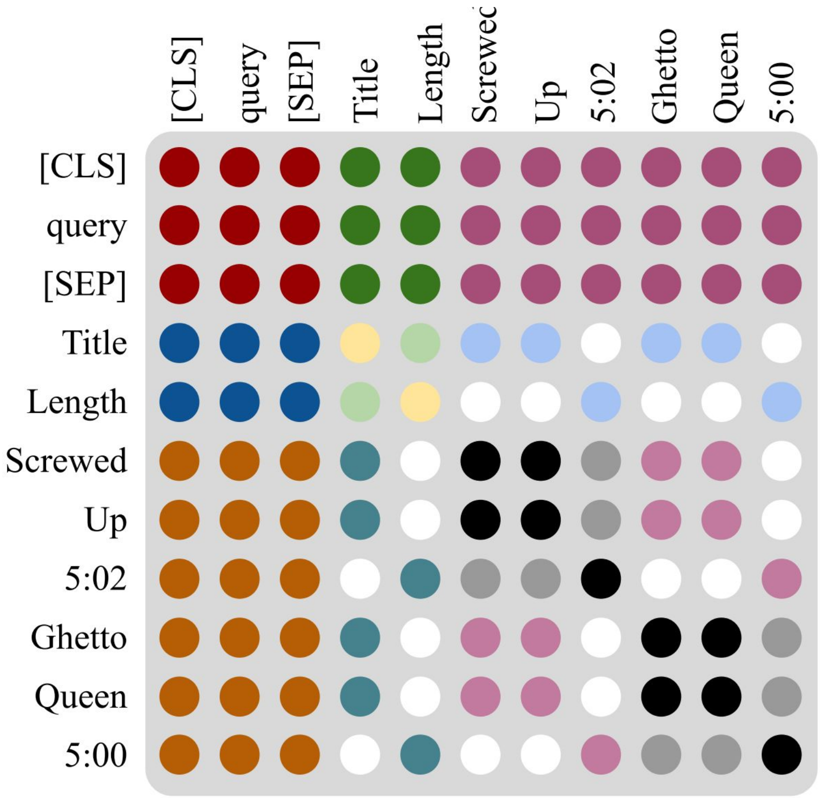
$\hat{A}$ **Learnable Table-Text Attention Bias Matrix (13 types of attention biases)**



$$\mathrm{Attn}(H) = \mathrm{softmax}(\frac{QK^\top}{\sqrt{d_K}})V$$

$$\boxed{\bar{A} = \frac{QK^\top}{\sqrt{d_K}}, \quad A = \bar{A} + \hat{A}}$$

Table-Text (Relative) Attention Bias Types



	Attention Bias Type
	header to sentence
	cell to sentence
	cell to its column header
	same row bias
	same column bias
...	...

TAPAS Input

Table:

Title	Length
Screwed Up	5:02
Ghetto Queen	5:00

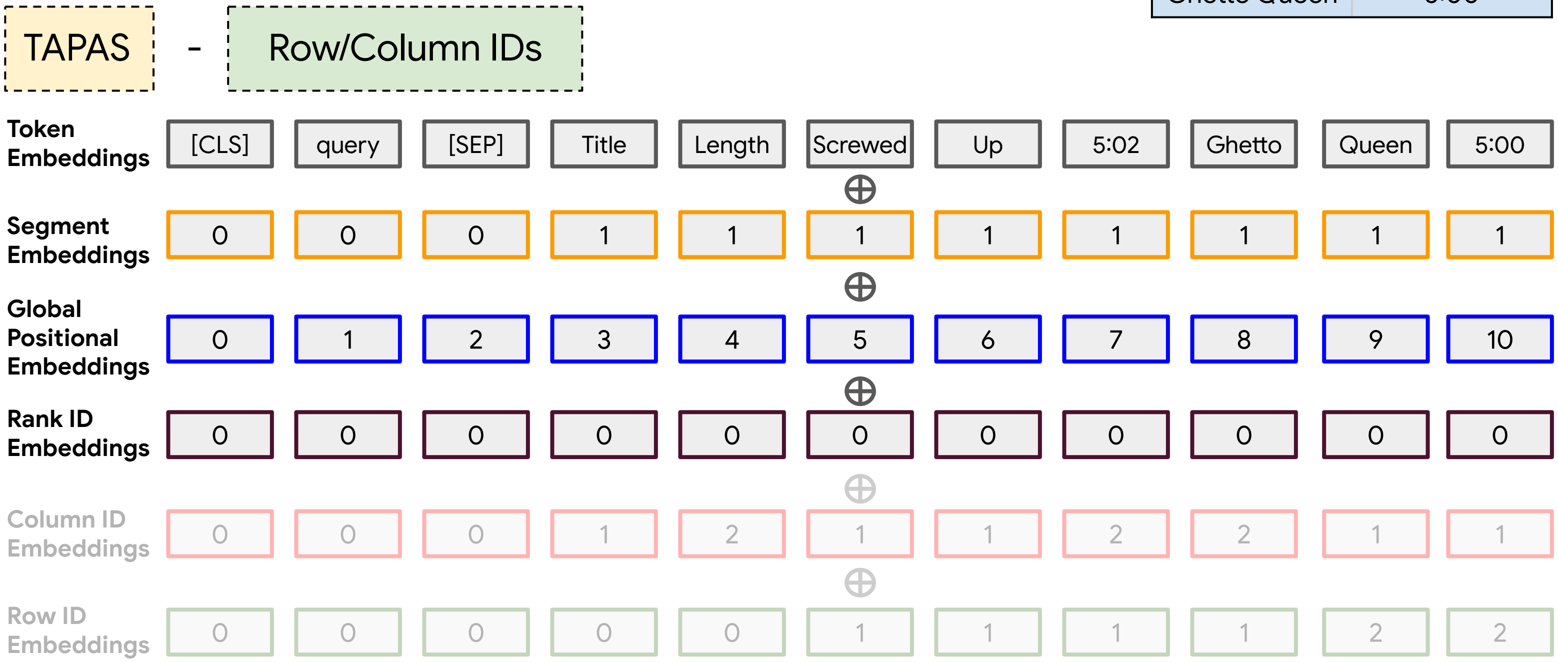
TAPAS

Token Embeddings	[CLS]	query	[SEP]	Title	Length	Screwed	Up	5:02	Ghetto	Queen	5:00
Segment Embeddings	0	0	0	1	1	1	1	1	1	1	1
Global Positional Embeddings	0	1	2	3	4	5	6	7	8	9	10
Rank ID Embeddings	0	0	0	0	0	0	0	0	0	0	0
Column ID Embeddings	0	0	0	1	2	1	1	2	2	1	1
Row ID Embeddings	0	0	0	0	0	1	1	1	1	2	2

TableFormer Input

Table:

Title	Length
Screwed Up	5:02
Ghetto Queen	5:00



TableFormer Input

Table:

Title	Length
Screwed Up	5:02
Ghetto Queen	5:00

TAPAS - Row/Column IDs + Per Cell Positional IDs

Token Embeddings	[CLS]	query	[SEP]	Title	Length	Screwed	Up	5:02	Ghetto	Queen	5:00
Segment Embeddings	0	0	0	1	1	1	1	1	1	1	1
Per-Cell Positional Embeddings	0	1	2	0	0	0	1	0	0	1	0
Rank ID Embeddings	0	0	0	0	0	0	0	0	0	0	0
Column ID Embeddings	0	0	0	1	2	1	1	2	2	1	1
Row ID Embeddings	0	0	0	0	0	1	1	1	1	2	2

Results

Experimental Setup

1. Reasoning Tasks
   a. Wikipedia Table based QA
   b. Table and Text Entailment

2. Evaluation Settings and Metrics
   a. Accuracy in Standard Evaluation
   b. Accuracy in Perturbation Evaluation: Randomly shuffle rows and columns of tables on test set without changing table contents
   c. Variation Percentage (VP) after Perturbation:

$$VP = \frac{\text{\# incorrect predictions that were corrected + \# correct predictions that became incorrect}}{\text{\# total}}$$

Table-based Sequential QA: SQA (Iyyer et al., 2017)

Original intent:
What super hero from Earth appeared most recently?

1. Who are all of the super heroes?

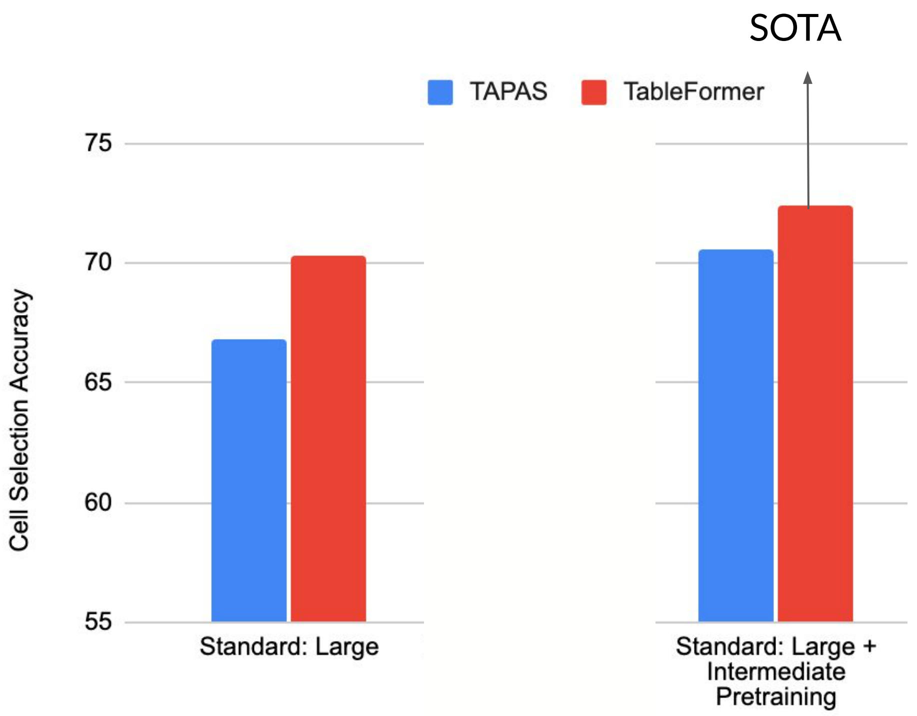
2. Which of them come from Earth?

3. Of those, who appeared most recently?

Legion of Super Heroes Post-*Infinite Crisis*

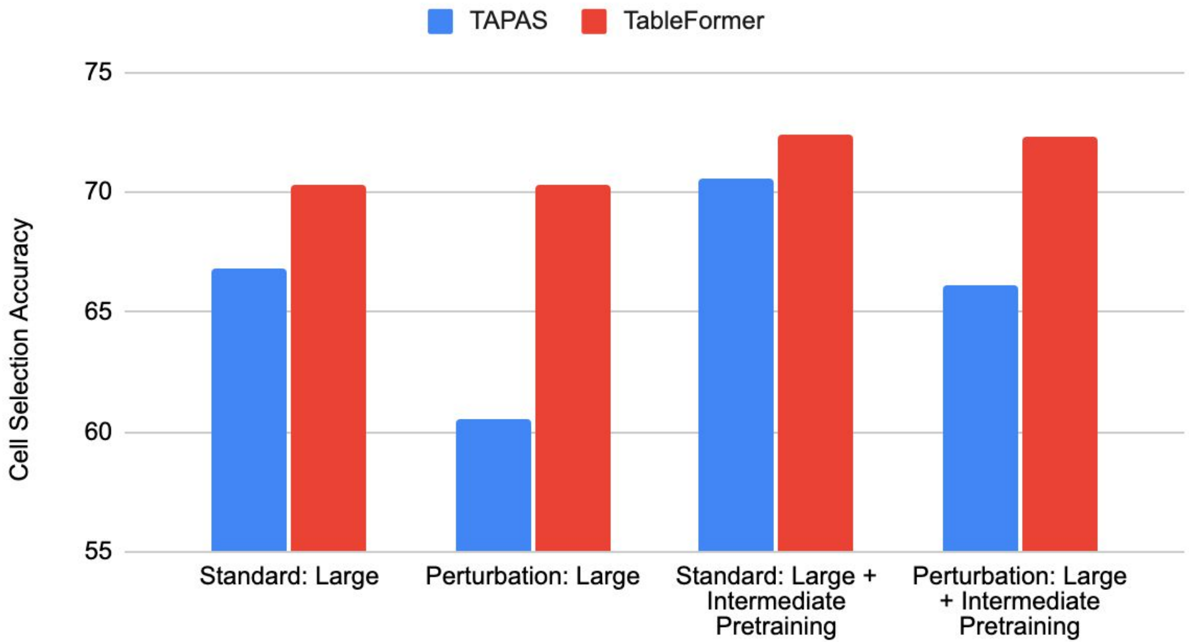
Character	*First Appeared*	*Home World*	*Powers*
Night Girl	2007	Kathoon	Super strength
Dragonwing	2010	Earth	Fire breath
Gates	2009	Vyrga	Teleporting
XS	2009	Aarok	Super speed
Harmonia	2011	Earth	Elemental

Results on SQA (Table-based Sequential QA)



Better overall performance with new SoTA!

Results on SQA (Table-based Sequential QA)



Invariant to perturbations which affect previous approaches!

Results on SQA (Instance-level Robustness)

Variation Percentage (VP) after Perturbation

$$\text{VP} = \frac{\texttt{\# incorrect predictions that were corrected + \# correct predictions that became incorrect}}{\texttt{\# total}}$$

	TAPAS	TableFormer
Large	15.1%	0.0%
Large + Intermediate Pretraining	10.8%	0.0%

TableFormer prediction is strictly robust to perturbations in the instance level!

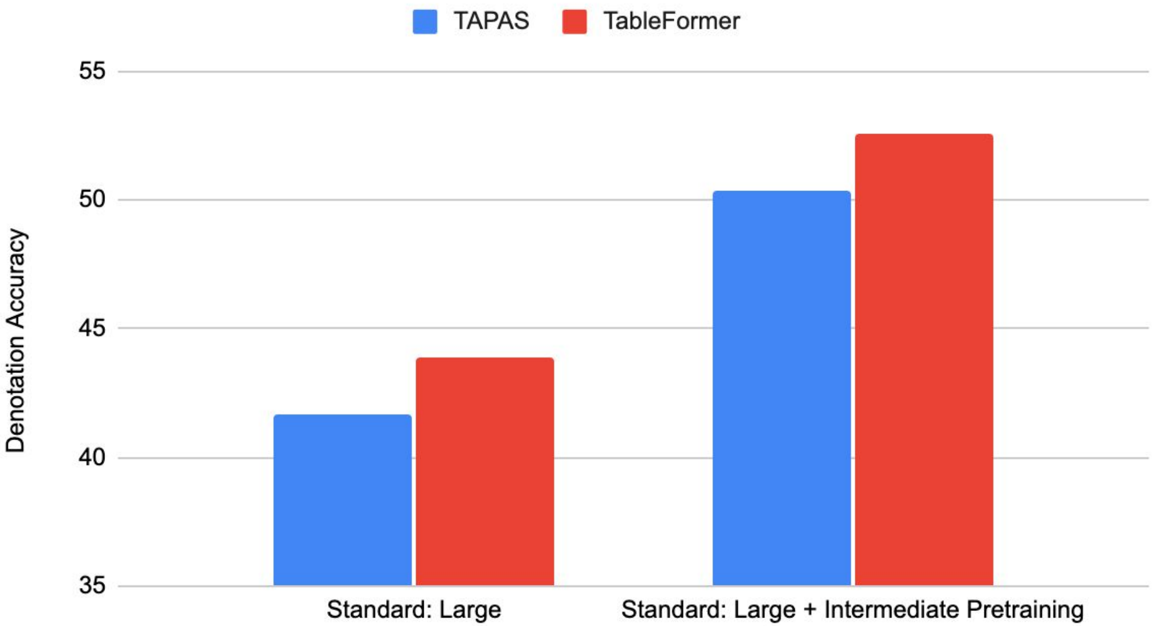
Table-based Complex QA: WikiTableQuestions (Pasupat et al., 2015)

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
2012	London	UK	204

x_1: *"Greece held its last Summer Olympics in which year?"*
y_1: {2004}

x_2: *"In which city's the first time with at least 20 nations?"*
y_2: {Paris}

Results on WTQ (Table-based Complex QA)



Better overall performance

Table-Text Entailment: TabFact (Chen et al., 2020)

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

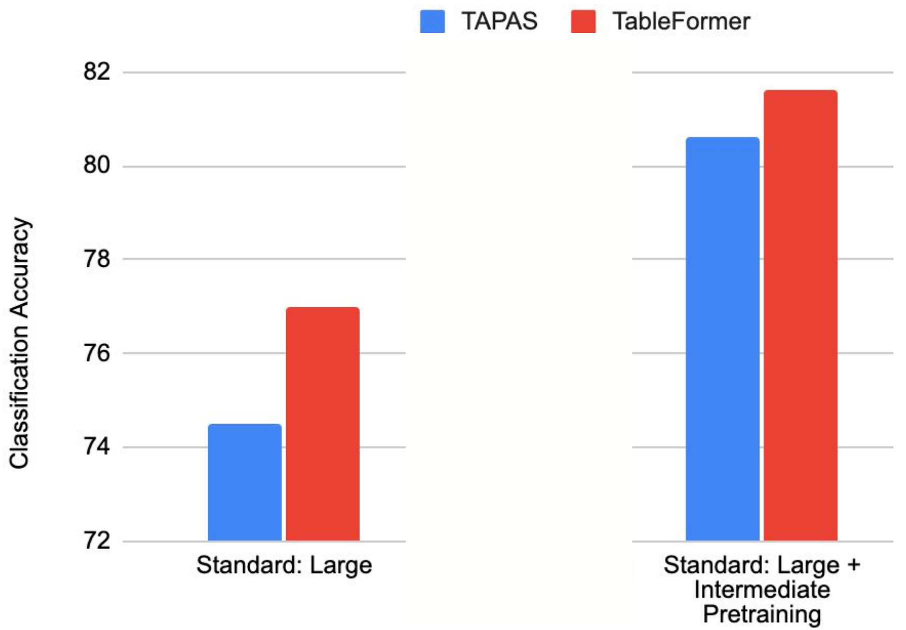
Entailed Statement

1. John E. Moss and Phillip Burton are both re-elected in the house of representative election.
2. John J. Mcfall is unopposed during the re-election.
3. There are three different incumbents from democratic.

Refuted Statement

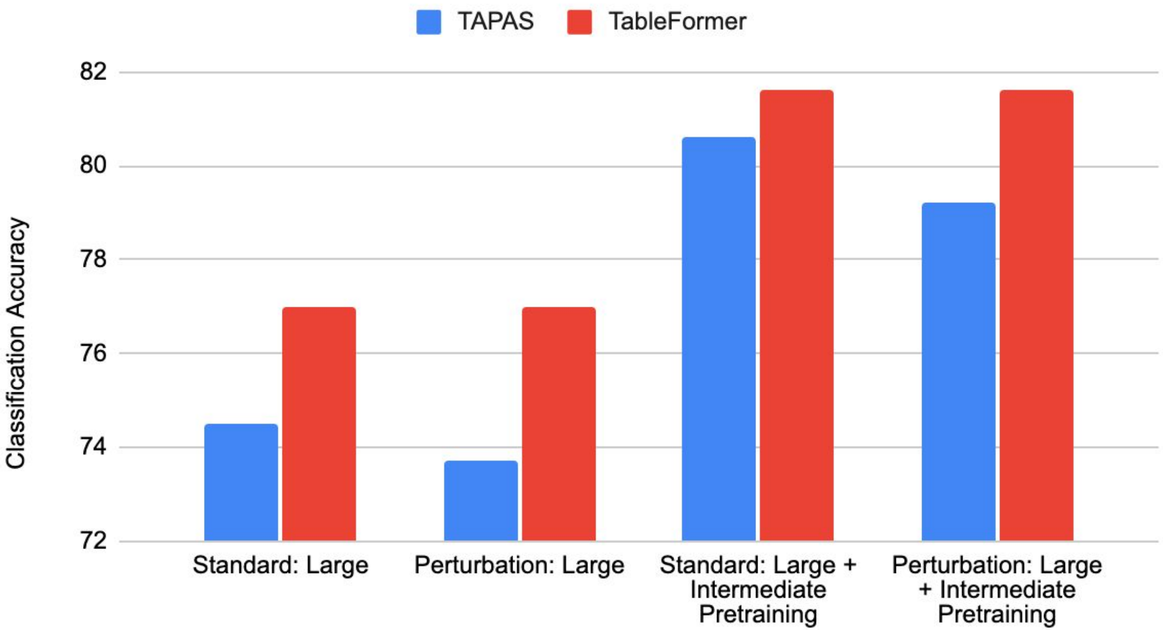
1. John E. Moss and George Paul Miller are both re-elected in the house of representative election.
2. John J. Mcfall failed to be re-elected though being unopposed.
3. There are five candidates in total, two of them are democrats and three of them are republicans.

Results on TabFact (Table-Text Entailment)



Better overall performance on wide range of tasks

Results on TabFact (Table-Text Entailment)



Invariant to perturbations which affect previous approaches!

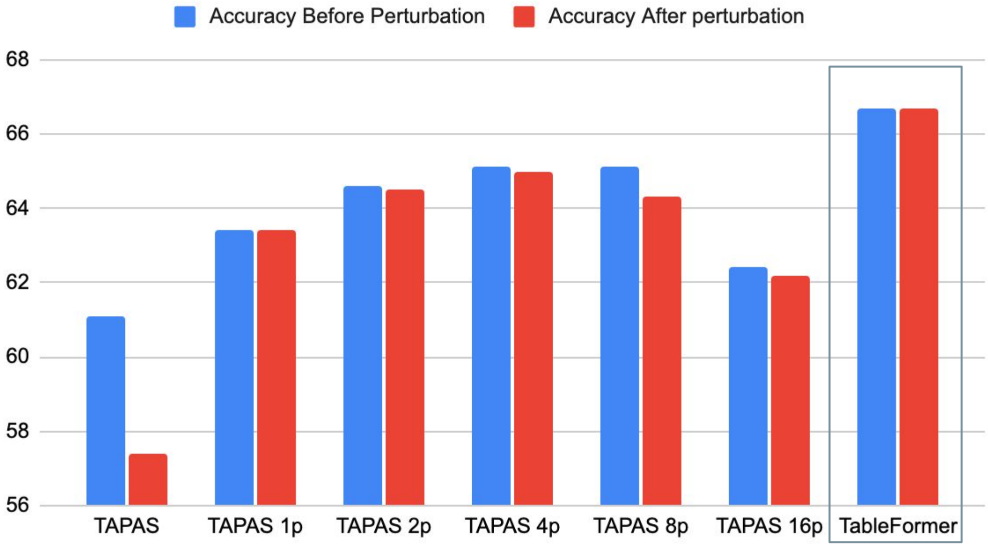
Model Size Comparison

	Number of parameters
TAPAS Base	110 M
TableFormer Base	110 M - 2*512*768 + 12*12*13 = 110 M **- 0.8 M + 0.002 M**
TAPAS Large	340 M
TableFormer Large	340 M - 2*512*1024 + 24*16*13 = 340 M **- 1.0 M + 0.005M**

Better Performance with even fewer parameters!

TableFormer v.s. Perturbed Data Augmentation

Experiment: Augment training data using {1, 2, 4, 8, 16} perturbations



Perturbed data augmentation can improve robustness to some extent, but the performance is still worse than TableFormer.

TableFormer v.s. Perturbed Data Augmentation

Experiment: Augment training data using {1, 2, 4, 8, 16} perturbations

Model	Variation Percentage
TAPAS	14.0%
TAPAS 1p	9.9%
TAPAS 2p	8.4%
TAPAS 4p	8.1%
TAPAS 8p	7.2%
TAPAS 16p	7.0%
TableFormer	**0.0%**

TableFormer has strict robustness in the instance level, while perturbed data augmentation do not have such a guarantee.

TableFormer Attention Bias Ablation Study

SQA dev result	ALL	SEQ
TableFormer base	**62.1**	**38.4**
- same row bias	**32.1**	**2.8**
- same column info	54.5	29.3
- cell to its column header	60.7	36.6
- cell to sentence	60.5	36.4
- header to sentence	61.1	36.3

Same row and column biases are the most important to encode table structures.
Cell/header to sentence biases could help better table-text alignment.

TableFormer Takeaways

- Structural attention biases in TableFormer help understand tables with relative attention and smaller model size.

- Current table encoding methods are not robust to table row and column order perturbation, while TableFormer is guaranteed to be robust to such perturbation.

- TableFormer has advantages over augmenting training data by row and column perturbation.

How to better transform texts to SQL ?

SeqZero: Few-shot Compositional Semantic Parsing with Sequential Prompts and Zero-shot Models

Jingfeng Yang[†] Haoming Jiang[†] Qingyu Yin[†]
Danqing Zhang[†] Bing Yin[†] Diyi Yang[‡]
[†] Amazon
[‡] Georgia Institute of Technology
{jingfe, jhaoming, qingyy, danqinz, alexbyin}@amazon.com
dyang888@gatech.edu

NAACL 2022 Findings

What's the major problem of Seq2Seq Semantic Parsing?

Semantic Parsing: Natural Language utterance -> Formal Language utterance (e.g. SQL Query)

Problem: Compositional Genarlization

Training Example 1:
Natural: How many people live in Chicago ?
Formal (SQL): SELECT city.population FROM city WHERE city.city_name = "Chicago"

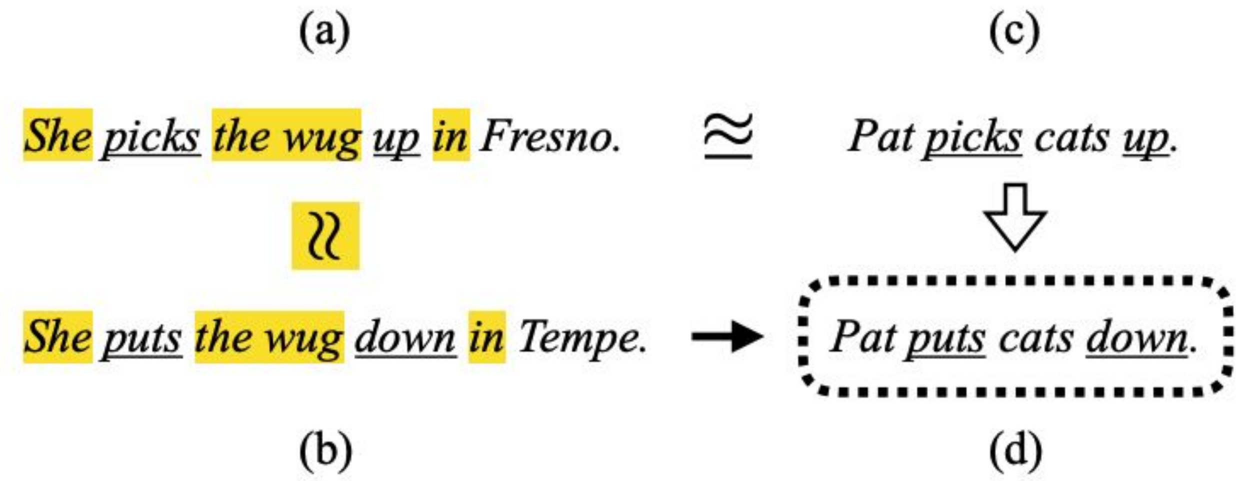
Training Example 2:
Natural: Give me the state that borders Utah .
Formal (SQL): SELECT border_info.border FROM border_info WHERE boder_info.state_name = "Utah"

Test Example:
Natural: How many people live in Utah ?
Formal (FunQL): SELECT state.population FROM state WHERE state.state_name = "Utah"

Examples are from GeoQuery dataset.

What is Compositional Generalization?

Compositional generalization is the ability to generalize systematically to a new data distribution by combining known components



Andreas J. Good-enough compositional data augmentation. ACL 2020.

Compositional Generalization Beyond Language

TEXT DESCRIPTION

An astronaut   Teddy bears   A bowl of soup

riding a horse   lounging in a tropical resort in space   playing basketball with cats in space

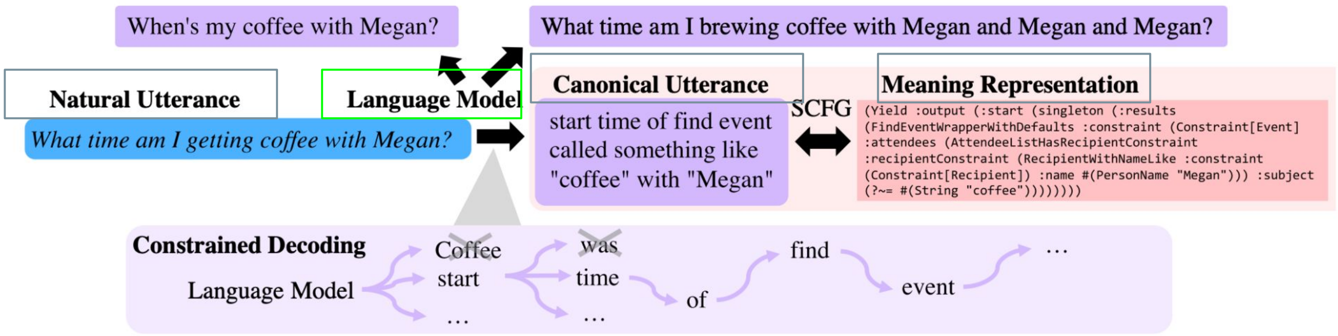
in a photorealistic style   in the style of Andy Warhol   as a pencil drawing

→

DALL·E 2

Prior Work: Semantic Parsing via Paraphrasing (SPP) and LMs

- Schucher et al., 2021, Shin et al., 2021



Natural Utterance -> Canonical Utterance -> Formal Language Utterance

Pretrained Language Models   Rules or Grammar

Problem 1: Lengthy and Complex Output

The canonical utterance is lengthy and complex due to compositional structure of the formal languages, which is still hard for LMs

Solution: Decompose the problem into a sequence of sub-problems, and the LMs only need to make a sequence of short prompt-based predictions.

Problem 2: Spurious Biases in Compositional Generalization

Question:

how many people live in **Utah ?**

Gold SQL:

SELECT **state** . population FROM **state**
WHERE **state** . **state**_name = ”Utah”

Finetuned BART Predicted SQL:

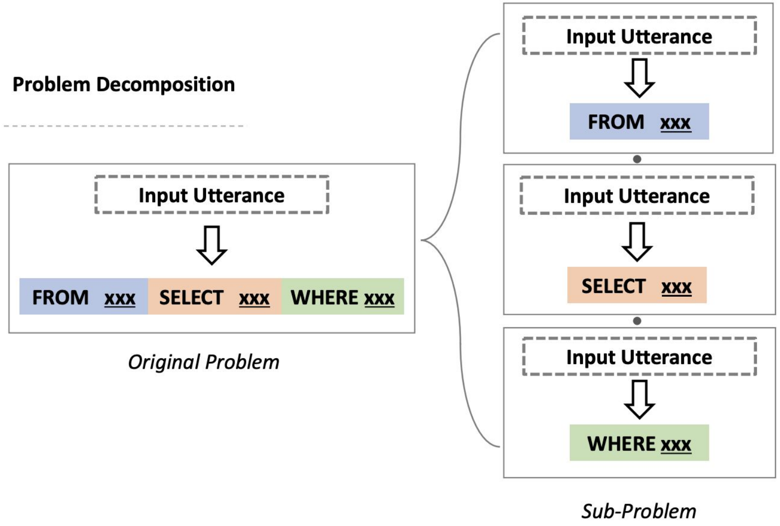
SELECT **city** . population FROM **city**
WHERE **city** . **city**_name = ”Utah"

Figure 1: Finetuned BART’s OOD generalization errors due to overfitting the spurious biases.

Solution:

- Ensemble of
  - Pertained models: better out-of-distribution (OOD) generalizability.
  - Fine-tuned models: better in-distribution generalizability.
- Has both advantages and avoids overfitting.

Problem Decomposition and Sequential Prompt Filling



Each sub-problem is finished by filing in a prompt by a LM.

Ensemble of Few-shot and Zero-shot Models

Constrained rescaling of zero-shot models:

Probability of zero-shot LM

Rescaled probability of zero-shot LM

$$P_{\theta_{i,z}}(w|x) = \frac{\mathbb{1}(w \in V_i(x)) P_{\theta_0}(w|x)}{\sum_{w_j \in V_i(x)} P_{\theta_0}(w_j|x)},$$

Allowed vocabulary given prefix

Ensemble:

$$P_{\theta_i} = \gamma_i P_{\theta_{i,f}} + (1 - \gamma_i) P_{\theta_{i,z}},$$

Final probability

Probability of few-shot LM

Overview of SeqZero

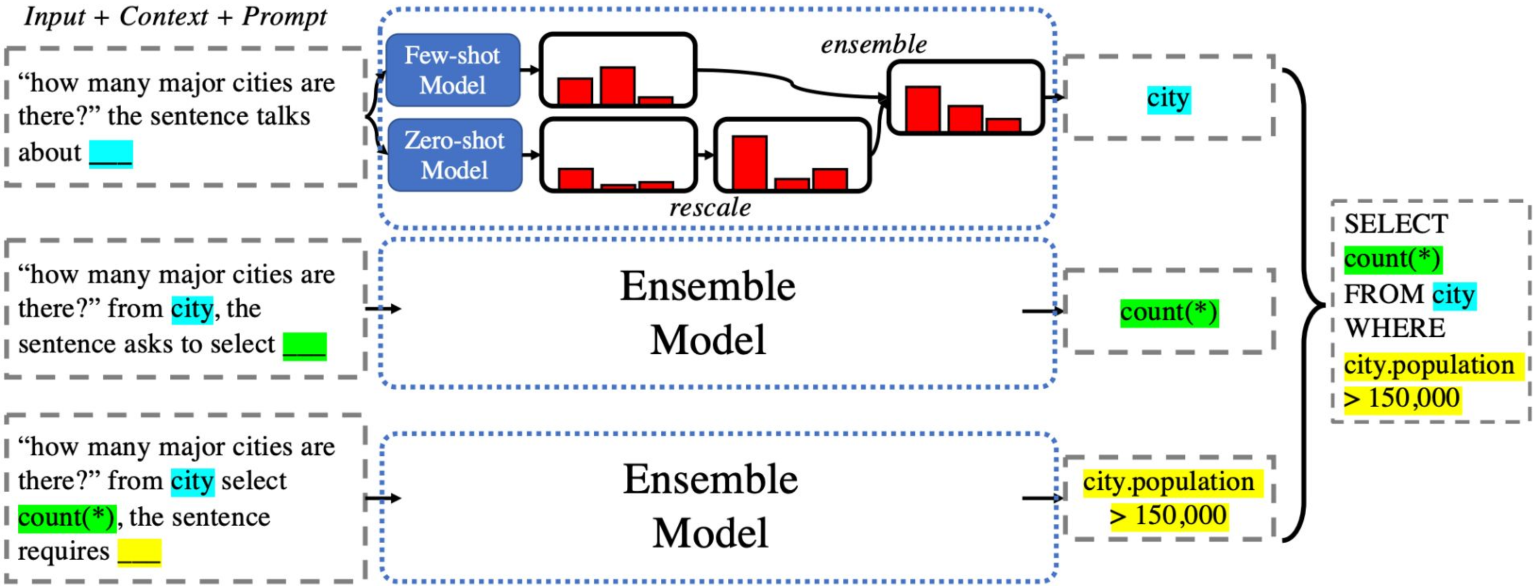


Figure 3: Pipeline of sequential prompt filling and SQL generation on GeoQuery. Note that, the scale of the prediction probability of the zero-shot model is very small before rescaling.

Dataset and Evaluation

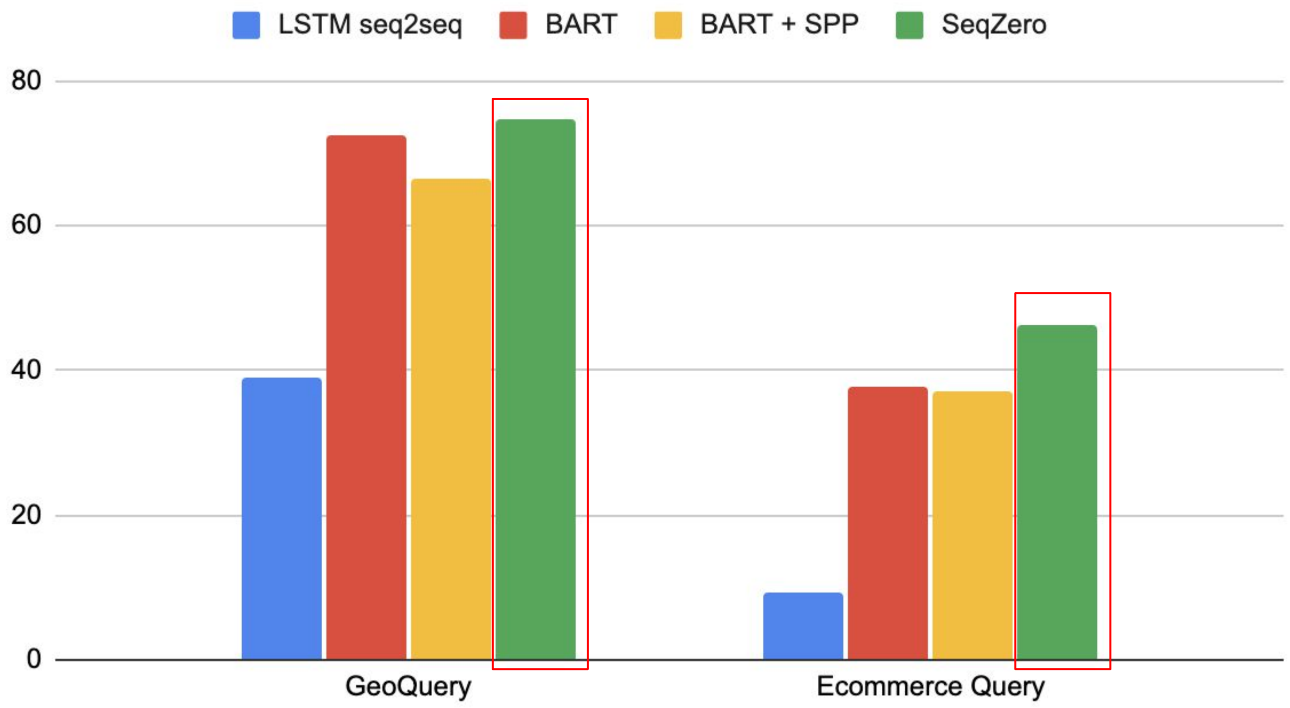
- Dataset:
  - GeoQuery Compositional Split
  - EcommerceQuery Compositional Split

Test Example:
Natural: petrol trimmer over 100 dollar
Formal (SQL): SELECT * FROM ASINs WHERE Maching Algorithm("petrol trimmer") == True and Price > 100

    - In training set, there are "Price <" and "Size >" combinations, but no "Price >" combination.

- Evaluation Metric:
  - Exact Match (Whole SQL utterance accuracy)

SeqZero Outperforms all Baselines



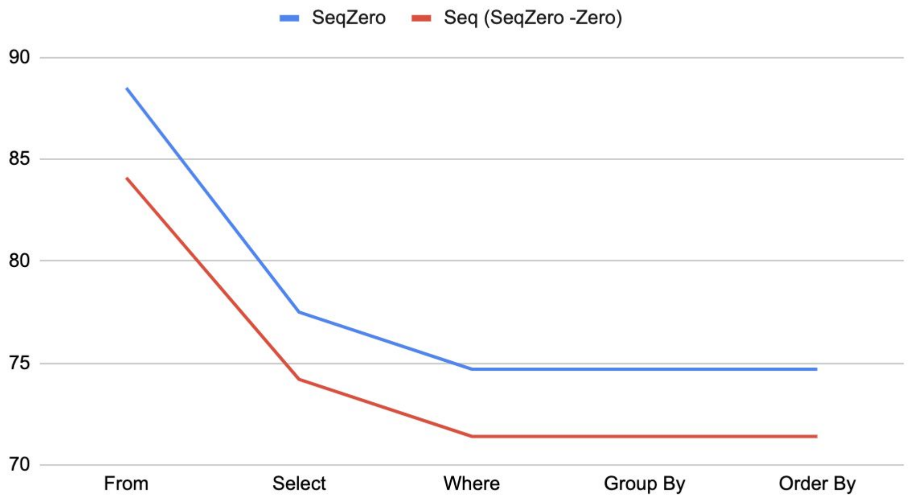
Effect of Zero-shot Models and Sequential Prompts

Method	GeoQuery	EcoQuery
SeqZero	**74.7**	**46.2**
−Seq	74.2	44.5
−Zero	71.4	37.7

Table 2: Ablation study of SeqZero.

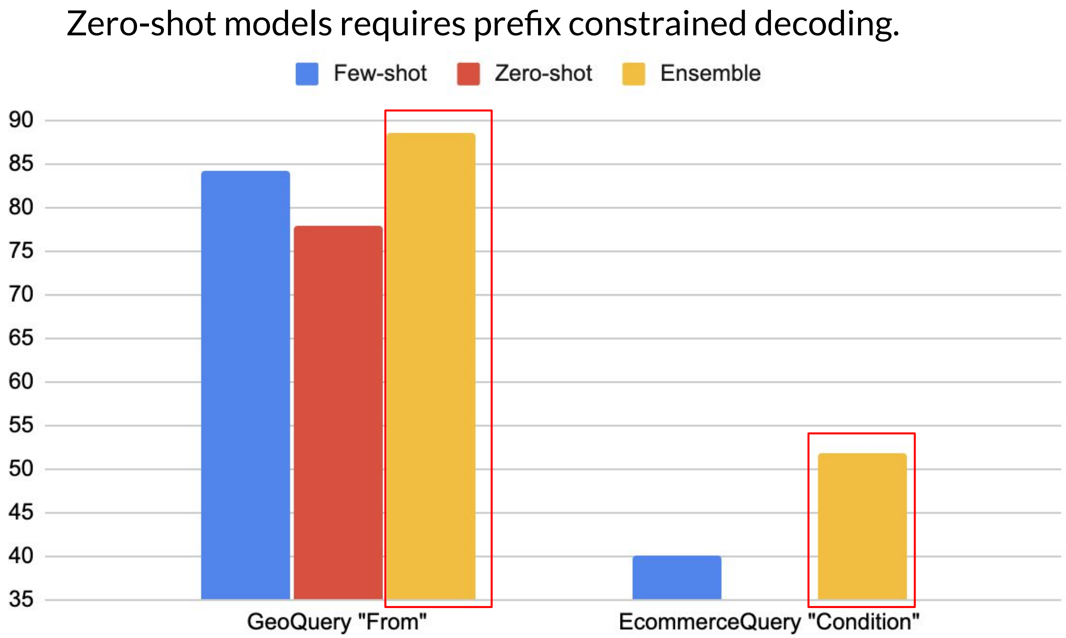
- Without the help of zero-shot models, the performance decreases a lot.
- Without sequential prompts, it's hard to design specific prompts for subproblems and mine knowledge from zero-shot (pretrained) models.

Analysis of Sequential Prompt Based Models



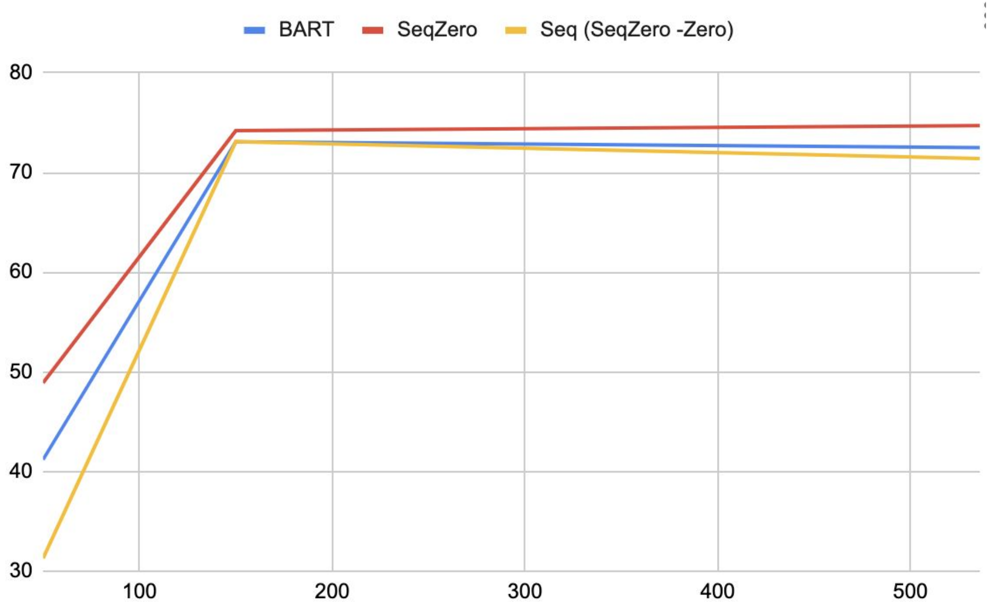
Ensemble of Zero-shot model in SeqZero boosts performance on the "FROM" clause, thus significantly reduces the error propagation, leading to better performance on all clauses.

Zero-shot, Few-shot models, and Their Ensemble



Ensemble of Zero-shot (Pretrained) and Few-shot (Finetuned) models has better performance because it achieves much better compositionally OOD generalization while maintaining in-distribution generalizability.

Few-shot Settings



Before certain point, SeqZero has larger improvement with more examples. Increasing training examples with the same templates enhances overfitting of seq2seq models, leading to larger gap between SeqZero and others.

SeqZero Takeaways

- Problem decomposition and sequential prompts enables flexible prompt designing.

- Ensemble of zero-shot (pretrained) and few-shot (finetuned) models achieves better compositional OOD generalizability, while maintaining in-distribution generalizability.

- Constrained rescaling is important for ensemble of zero-shot and few-shot models to work in the generation task.

Recent Work of Table Understanding and Semantic Parsing (Large LM Era and In-context Learning)

Chain-of-Thought Prompting & Least-to-Most Prompting

Think of semantic parsing as Chain-of-Thought for Question Answering, then sequential prompting in our SeqZero is least-to-most prompting. Our work was earlier than least-to-most prompting and at the same time as Chain-of-Thought prompting.

Semantic Parsing Results:

Prompting method	code-davinci-002	code-davinci-001	text-davinci-002*
Standard prompting	16.7	0.4	6.0
Chain-of-Thought	16.2	0.0	0.0
Least-to-Most	**99.7**	60.7	76.0

Table 9: Accuracies (%) of different prompting methods on the test set of SCAN under the length-based split. The results of text-davinci-002 are based on a random subset of 100 commands.

Wei J, Wang X, Schuurmans D, et al. Chain of thought prompting elicits reasoning in large language models[J]. arXiv preprint arXiv:2201.11903, 2022.

Zhou D, Schärli N, Hou L, et al. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models[J]. arXiv preprint arXiv:2205.10625, 2022.

Adapting Chain-of-thought Prompting for Table Reasoning

Type	Model	Test EM
Train	Pasupat and Liang (2015)	37.1
Train	Zhang et al. (2017)	43.7
Train	Liang et al. (2018)	43.7
Train	Agarwal et al. (2019)	44.1
Train	Wang et al. (2019)	44.5
PT + FT	Herzig et al. (2020)	48.8
PT + FT	Yu et al. (2021)	**52.7**
1-shot	Direct Prediction	24.5
2-shot	Direct Prediction	26.8
1-shot	Chain of Thoughts	41.8
2-shot	Chain of Thoughts	**42.4**

Table 1: Experimental Results on WikiTableQuestions. PT means pre-training and FT means fine-tuning.

Chen W. Large Language Models are few (1)-shot Table Reasoners[J]. arXiv preprint arXiv:2210.06710, 2022.

LM-based Decomposition and Sequential Least-to-Most Prompting for Semantic Parsing

	MCD1	MCD2	MCD3	Ave.
Fully Supervised				
T5-base (Herzig et al., 2021)	58.5	27.0	18.4	34.6
T5-large (Herzig et al., 2021)	65.1	32.3	25.4	40.9
T5-3B (Herzig et al., 2021)	65.0	41.0	42.6	49.5
HPD (Guo et al., 2020)	79.6	59.6	67.8	69.0
T5-base + IR (Herzig et al., 2021)	85.8	64.0	53.6	67.8
T5-large + IR (Herzig et al., 2021)	88.6	79.2	72.7	80.2
T5-3B + IR (Herzig et al., 2021)	88.4	85.3	77.9	83.9
LeAR (Liu et al., 2021)	91.7	89.2	91.7	90.9
Prompting				
(Ours) Dynamic Least-to-Most	**94.3**	**95.3**	**95.5**	**95.0**

Table 1: Test accuracy across the MCD splits for the CFQ dataset.

Drozdov A, Schärli N, Akyürek E, et al. Compositional semantic parsing with large language models[J]. arXiv preprint arXiv:2209.15003, 2022.

Large LM (GPT-3 Codex) Decomposition to Functions

Method	Dev.	Test
Finetuned		
T5-3B (Xie et al., 2022)	51.9	50.6
Tapex (Liu et al., 2021)	60.4	59.1
TaCube (Zhou et al., 2022)	61.1	61.3
OmniTab (Jiang et al., 2022)	-	63.3
Without Finetuning		
Codex end-to-end QA	50.5	48.7
Codex SQL†	60.2	61.1
Codex BINDER † (Ours)	**65.0**	**64.6**

Table 1: WIKITQ execution accuracy on development and test sets. † denotes a symbolic method that outputs intermediate languages.

Cheng Z, Xie T, Shi P, et al. Binding Language Models in Symbolic Languages[J]. arXiv preprint arXiv:2210.02875, 2022.

In-context Learning v.s. Fine-tuning v.s. Prompt Tuning for Semantic Parsing

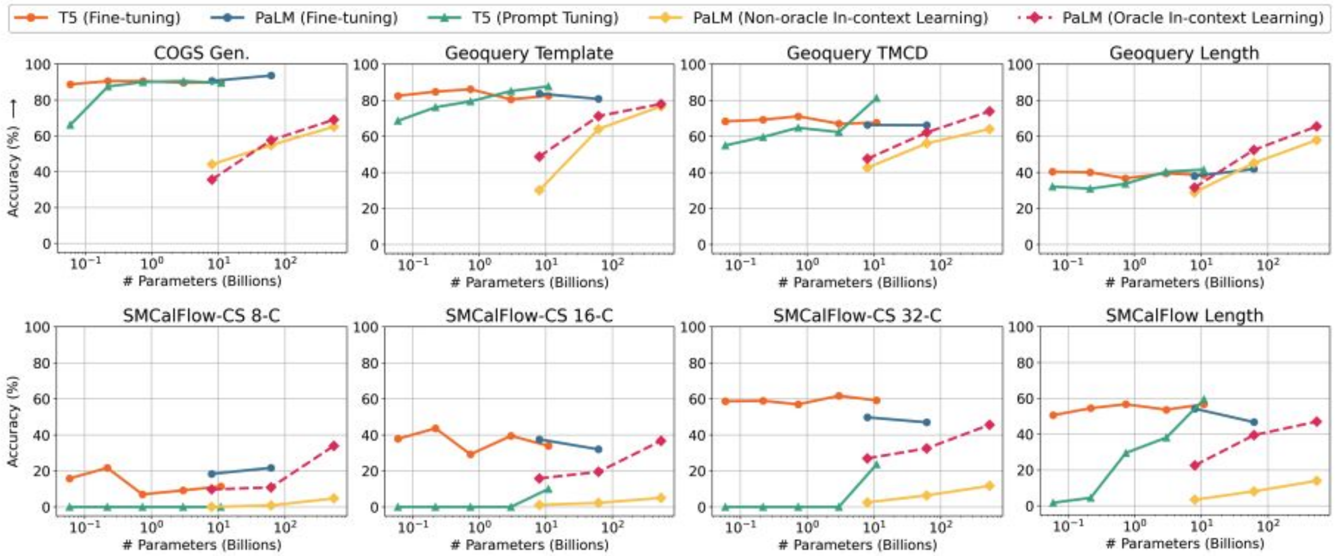


Figure 2: Scaling curves for different datasets and splits using different training schemes. Note that the in-context learning with an oracle retriever (dashed) cannot be compared directly with other methods as it has access to the gold output.

Qiu L, Shaw P, Pasupat P, et al. Evaluating the Impact of Model Scale for Compositional Generalization in Semantic Parsing[J]. arXiv preprint arXiv:2205.12253, 2022.

Conclusions / Questions

- Are inductive biases (e.g. Attention Biases in TableFormer) still useful in the future with even larger models?

- In-context learning is probably an alternative to our ensemble method in SeqZero, in order to have better compositional generalizability, because it avoids fine-tuning models to overfitting spurious biases as indicated by SeqZero.

- In large LM and in-context learning era, compositional generalization could be potentially somehow solved, but still with our proposed idea of sequential prompting (least-to-most prompting).