Meeting PRIN - 03/07/2024

Federico Belotti (federico.belotti@unimib.it)

INSID&S Lab
Department of Informatics,
Systems and Communication
Università degli Studi di
Milano-Bicocca

Last Meeting: Entity Linking in Tables a Quick Recap

1. **Entity Linking (EL) in tables**
   - Reconciling cells in tables with entities from a background KB
   - Part of a broader set of semantic tasks on tabular data
     - Semantic annotations: EL, Column Type Annotation, Column Property Annotation
     - More tasks LMs:
   - Typically consists of two steps
     - Candidate entity retrieval, entity disambiguation
2. **Unimib previous work**
   - Alligator: feature-based ML approach tested on several benchmark datasets (SemTab Challenge)
3. **Approaches based on LM**
   - Encoder-based
     - TURL: several classification-based tasks for tabular data (BERT)
     - UNICORN: ~ multi-task cross-encoder for matching-related tasks (BERT)
   - Autoregressive LLMs
     - TableLlama: several tasks (LLama2)

One Relevant Question by SEBD Reviewers

- LLM for data preparation
  - Too resource-demanding for “discounted DQ”?
- As of today: we don’t know :)
  - High training costs and resource-greedy inference (e.g., 7B for TableLlama)
  - Higher generalization? Less need for in-domain fine-tuning?
  - Several trade offs to consider…
  - … we need to better understand the performance, advantages and pitfalls of LLMs for data preparation tasks (@UNIMIB: EL and other semantic tasks on table)

Last Meeting Open Questions: EL in Tables

1. **In-depth comparison of feature-based, encoder-based, LLM-based EL in tables**
    - **Single, unified and standard benchmark for those methods is missing!**
        - How this models perform on SemTab Challange datasets?
    - **Hard to make them comparable**
        - Every method has its own test case given the same test dataset
    - **How they perform in a in-domain (ID), out-of-domain (OOD) and moderately-OOD settings?**
        - w/ and w/o fine-tuning

Submitted a paper that we can share

2. **Aspects of EL in tables not covered by top LLM-based approaches**
    - **How they perform w.r.t. NIL entities?**
        - What they tell us about NIL entities?
3. **Possible directions for future research**
    - **Can we reduce the memory/time consumption of LLM-based method while maintaining performance?**
        - Fine-tune Phi-3-mini (3B) seems promising
    - **Can those method be fine-tuned with Human-In-The-Loop feedbacks? If so, how?**
        - TableLlama is the most suitable (RLHF)
        - Human-In-The-Loop from imitation
    - **What about improving candidate retrieval?**

Last meeting - Open Questions

1. **Single, unified and standard benchmark for those methods is missing!**
    - How this models perform on SemTab Challange datasets?
2. **Hard to make them comparable**
    - Every method has its own test case given the same test dataset
3. **How they perform in a zero-shot setting?**
4. **How they perform w.r.t. NIL entities?**
    - What they tell us about NIL entities?
5. **Can Meta-Learning be applied to learn new matching tasks with few examples?**
    - Unicorn is the most suitable
6. **Can those method be fine-tuned with Human-In-The-Loop feedbacks? If so, how?**
    - TableLlama is the most suitable (RLHF)
7. **What about candidate generation?**

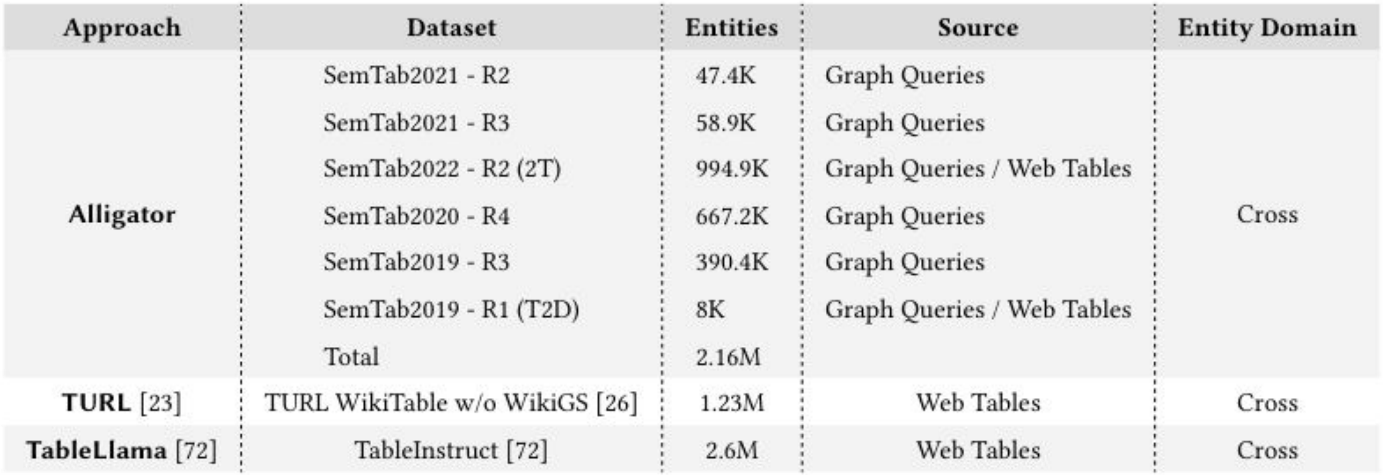
1: TURL vs TableLlama vs Alligator

Objectives

1. Asses performances on a shared benchmark (w/ and w/ fine-tuning)
2. Understand capabilities of LLM-based STI approaches on tables from different dataset
3. Test the generalisation capability of LLM-based approaches after fine-tuning in MOOD settings
4. Compare the approaches in inference time and occupied memory to get insights about their applicability on application domains where processing of large tables may be require

Pre-train data per approach

Table 1: Statistics of the datasets used to pre-train the considered approaches



Approach	Dataset	Entities	Source	Entity Domain
Alligator	SemTab2021 - R2	47.4K	Graph Queries	Cross
	SemTab2021 - R3	58.9K	Graph Queries	
	SemTab2022 - R2 (2T)	994.9K	Graph Queries / Web Tables	
	SemTab2020 - R4	667.2K	Graph Queries	
	SemTab2019 - R3	390.4K	Graph Queries	
	SemTab2019 - R1 (T2D)	8K	Graph Queries / Web Tables	
	Total	2.16M		
TURL [23]	TURL WikiTable w/o WikiGS [26]	1.23M	Web Tables	Cross
TableLlama [72]	TableInstruct [72]	2.6M	Web Tables	Cross

Fine-tuning and eval data per approach

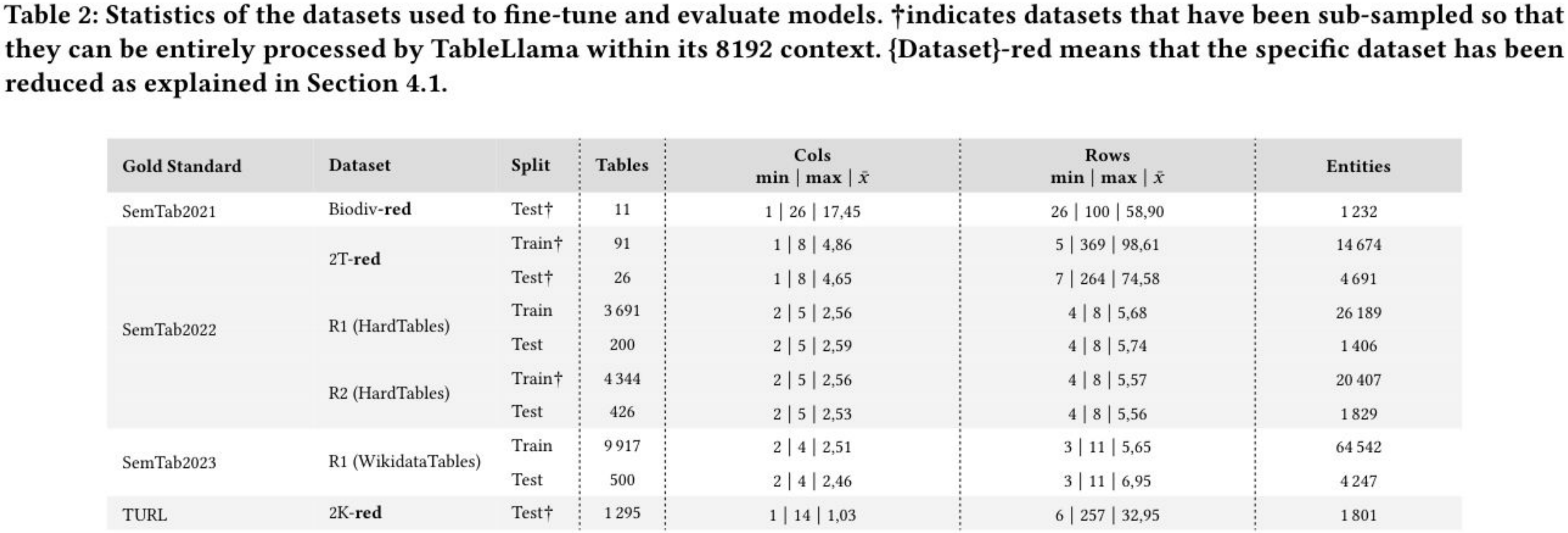


Table 2: Statistics of the datasets used to fine-tune and evaluate models. †indicates datasets that have been sub-sampled so that they can be entirely processed by TableLlama within its 8192 context. {Dataset}-red means that the specific dataset has been reduced as explained in Section 4.1.

Gold Standard	Dataset	Split	Tables	Cols min | max | $\bar{x}$	Rows min | max | $\bar{x}$	Entities
SemTab2021	Biodiv-**red**	Test†	11	1 | 26 | 17,45	26 | 100 | 58,90	1 232
SemTab2022	2T-**red**	Train†	91	1 | 8 | 4,86	5 | 369 | 98,61	14 674
		Test†	26	1 | 8 | 4,65	7 | 264 | 74,58	4 691
	R1 (HardTables)	Train	3 691	2 | 5 | 2,56	4 | 8 | 5,68	26 189
		Test	200	2 | 5 | 2,59	4 | 8 | 5,74	1 406
	R2 (HardTables)	Train†	4 344	2 | 5 | 2,56	4 | 8 | 5,57	20 407
		Test	426	2 | 5 | 2,53	4 | 8 | 5,56	1 829
SemTab2023	R1 (WikidataTables)	Train	9 917	2 | 4 | 2,51	3 | 11 | 5,65	64 542
		Test	500	2 | 4 | 2,46	3 | 11 | 6,95	4 247
TURL	2K-**red**	Test†	1 295	1 | 14 | 1,03	6 | 257 | 32,95	1 801

Test data categorization

- **In-Domain**: a test set Y for an approach A is considered “IN” for A if A has been trained on a dataset X generated from the same data source and covering similar domain(s) as Y
- **Out-Of-Domain:** a test set Y for an approach A is considered “OOD” if A has been trained on a set of data X generated from a different data source and covering different domain(s) as Y
- **Moderately-OOD:** a test set Y for an approach A is considered “MOOD” for A if A has been trained on a set of data X generated from the same data source but covering different domain(s) as Y or viceversa, i.e., if A has been trained on a set of data X generated from a different source but covering similar domain(s) as X

Test data categorization

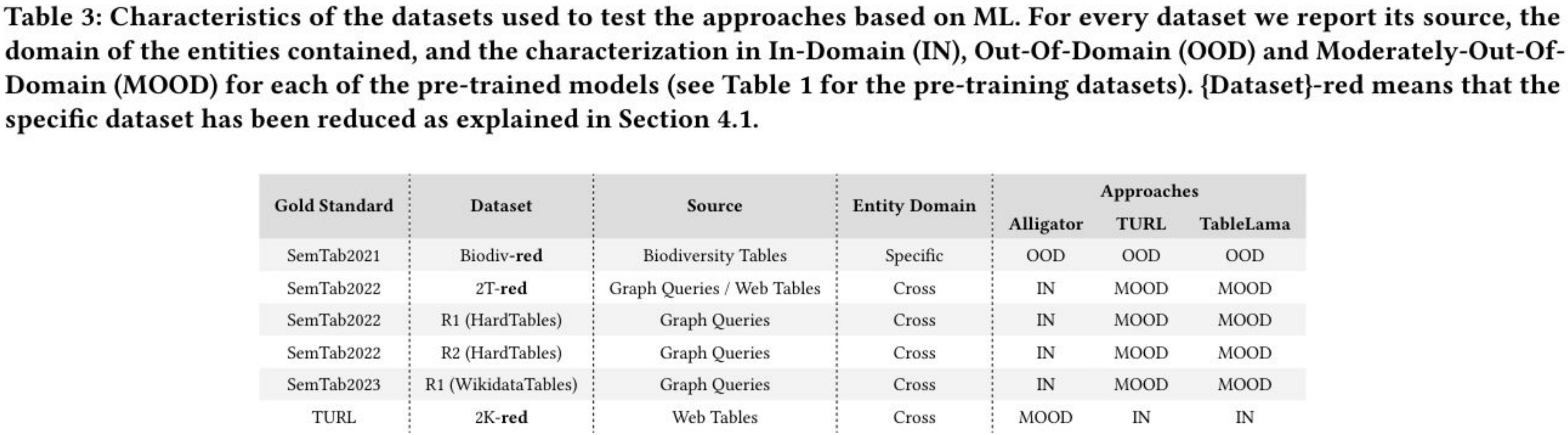
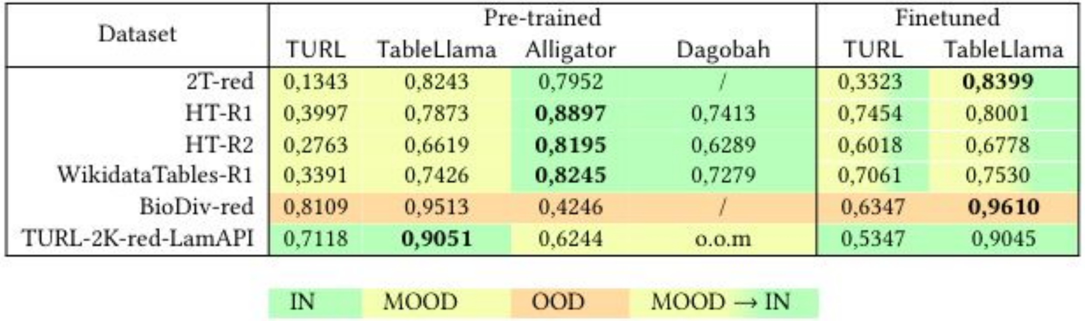


Table 3: Characteristics of the datasets used to test the approaches based on ML. For every dataset we report its source, the domain of the entities contained, and the characterization in In-Domain (IN), Out-Of-Domain (OOD) and Moderately-Out-Of-Domain (MOOD) for each of the pre-trained models (see Table 1 for the pre-training datasets). {Dataset}-red means that the specific dataset has been reduced as explained in Section 4.1.

Gold Standard	Dataset	Source	Entity Domain	Approaches		
				Alligator	TURL	TableLama
SemTab2021	Biodiv-**red**	Biodiversity Tables	Specific	OOD	OOD	OOD
SemTab2022	2T-**red**	Graph Queries / Web Tables	Cross	IN	MOOD	MOOD
SemTab2022	R1 (HardTables)	Graph Queries	Cross	IN	MOOD	MOOD
SemTab2022	R2 (HardTables)	Graph Queries	Cross	IN	MOOD	MOOD
SemTab2023	R1 (WikidataTables)	Graph Queries	Cross	IN	MOOD	MOOD
TURL	2K-**red**	Web Tables	Cross	MOOD	IN	IN

Results



Dataset	Pre-trained				Finetuned	
	TURL	TableLlama	Alligator	Dagobah	TURL	TableLlama
2T-red	0,1343	0,8243	0,7952	/	0,3323	**0,8399**
HT-R1	0,3997	0,7873	**0,8897**	0,7413	0,7454	0,8001
HT-R2	0,2763	0,6619	**0,8195**	0,6289	0,6018	0,6778
WikidataTables-R1	0,3391	0,7426	**0,8245**	0,7279	0,7061	0,7530
BioDiv-red	0,8109	0,9513	0,4246	/	0,6347	**0,9610**
TURL-2K-red-LamAPI	0,7118	**0,9051**	0,6244	o.o.m	0,5347	0,9045

🟩 IN 🟨 MOOD 🟧 OOD 🟩 MOOD → IN

- Alligator
  - excels on HT-R1, HT-R2, WikidataTablesR1 SemTab
  - performs poorly on BioDiv-red and TURL-2K-red-LamAPI while being on par on 2T dataset with TableLlama
- The opposite is observed for TableLlama and TURL
  - TableLlama achieving generally higher accuracy than TURL
- BioDiv and TURL-2K-red-LamAPI contain tables from different specific domains, with misspelt, repeated and abbreviated mentions, whose heterogeneity is difficult to capture with handcrafted features
- Both TURL and TableLlama excel on BioDiv, even though it’s considered OOD for both approaches: we hypothesise that the enormous and general pre-training knowledge retained by these models explains this excellence
- TableLlama and TURL underperforms (w.r.t. original results) on TURL-2K-LamAPI dataset
  - Increased number of candidates we have retrieved with LamAPI impacts on performance

Ablation studies

Figure 4: Accuracies achieved by TableLlama and TURL w.r.t. the number of candidates.

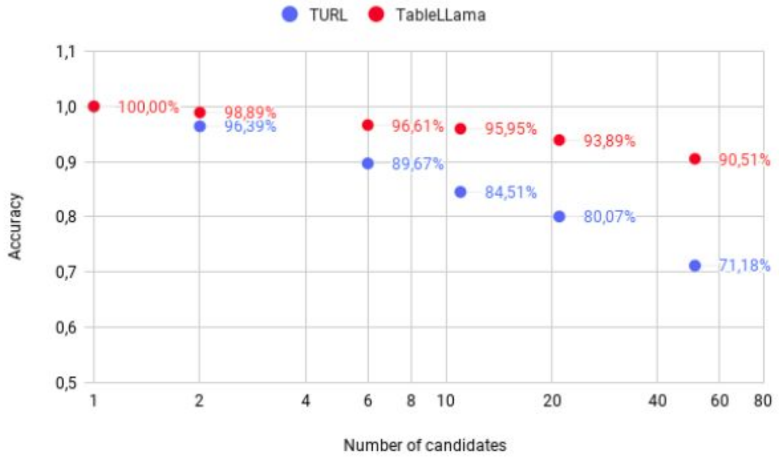
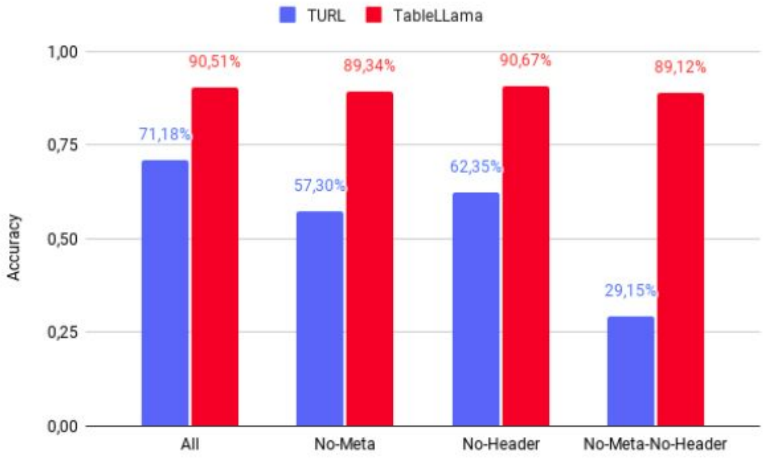


Figure 5: Accuracies achieved by TableLlama and TURL w.r.t. the presence of table's metadata.



Time/Memory consumption

Figure 2: Overall elapsed time per dataset.

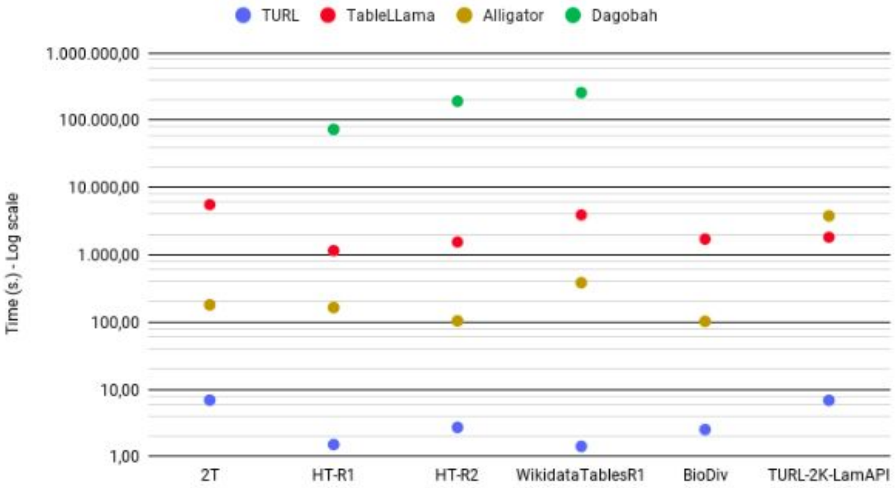
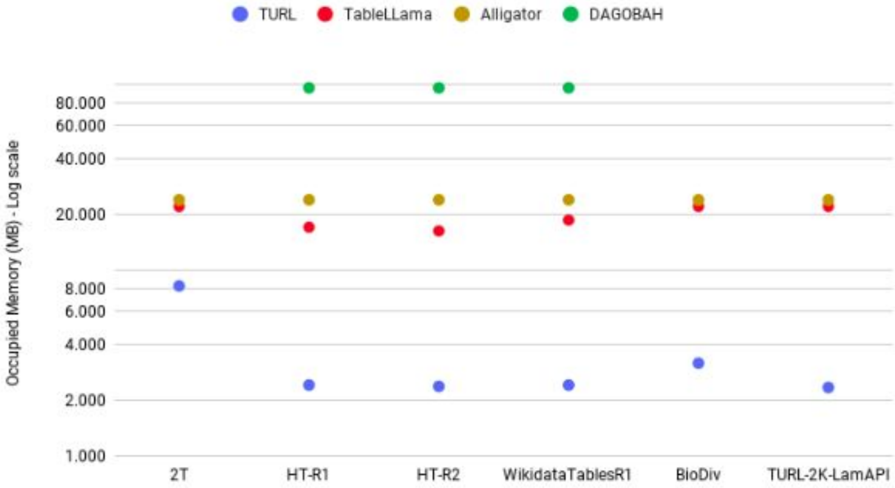


Figure 3: Overall occupied memory per dataset.



2: TableLlama and NIL

Example prompt (from TURL-2K dataset)

input: [TLE] The Wikipedia page is about 2005 NPSL Season. The Wikipedia section is about New Franchises. [TAB] col: | team name | metro area | location | previous affiliation | row 1: | Detroit Arsenal | Detroit area | Berkley, MI | expansion | [SEP] row 2: | Grand Rapids Alliance | Grand Rapids area | Grand Rapids, MI | expansion | [SEP] row 3: | Milwaukee Bavarians | Milwaukee area | Milwaukee, WI | expansion | [SEP] row 4: | Minnesota Blast | Minneapolis area | Burnsville, MN | expansion | [SEP] row 5: | Princeton 56ers | Madison area | Madison, WI | expansion | [SEP] row 6: | San Diego Pumitas | San Diego area | San Diego, CA | expansion | [SEP] row 7: | San Jose Frogs | San Jose area | San Jose, CA | expansion | [SEP] row 8: | St. Paul Twin Stars | St. Paul area | St. Paul, MN | expansion |
question: The selected entity mention in the table cell is: Grand Rapids, MI. The column name for 'Grand Rapids, MI' is location. The referent entity candidates are: <Grand Rapids [DESCRIPTION] city in Itasca County, Minnesota, United States [TYPE] None>, <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>. What is the correct referent entity for the entity mention 'Grand Rapids, MI'?
output: <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>

NIL experiments

Dataset has been created by replacing the correct candidate with **<NIL [DESCRIPTION] Not in list [TYPE] None>**

- TURL-1.7K - NIL: 0.3345
- HTR1 - NIL: 0.0611

Good performance on TURL-1.7K, but …

- 526 candidates lists have length == 1, i.e. 526/1731 ≃ 30.39%
- 837 candidates lists have length <= 5, i.e. 837/1731 ≃ 48.35%

The performances on HTR1-2023 are possibly driven by the fact that:

- 70 candidates lists have length == 1, i.e. 70/1375 ≃ 5.09%
- 157 candidates lists have length <= 5, i.e. 157/1375 ≃ 11.42%

NIL experiments

Experiment with a further downsampled dataset where prompts with mentions with less than 10 candidates were removed:

TURL-1.7K - NIL-reduced:	0.0064
HTR1 - NIL-reduced:	0.0067

Which demonstrates that TableLlama is not able to understand NIL entities

NIL examples

"input": "[TLE] The Wikipedia page is about USA Today All-USA high school football team (2000\u201309). The Wikipedia section is about 2002 team. [TAB] col: | player | position | school | hometown | college | row 1: | Chris Leak | Quarterback | Independence High School | Charlotte, North Carolina | Florida | [SEP] row 2: | Reggie Bush | Running back | Helix High School | La Mesa, California | USC | [SEP] row 3: | Demetris Summers | Running back | Lexington High School | Lexington, South Carolina | South Carolina | [SEP] row 4: | Jorrie Adams | Offensive line | Jasper High School | Jasper, Texas | Texas A&M | [SEP] row 5: | Ofa Mohetau | Offensive line | Trinity High School | Euless, Texas | BYU | [SEP] row 6: | Ryan Harris | Offensive line | Cretin-Derham Hall High School | Saint Paul, Minnesota | Notre Dame | [SEP] row 7: | Mike Jones | Offensive line | Richards High School | Oak Lawn, Illinois | Iowa | [SEP] row 8: | Greg Olsen | Tight end | Wayne Hills High School | Wayne, New Jersey | Miami (FL) | [SEP] row 9: | Sean Bailey | Wide receiver | Milton High School | Alpharetta, Georgia | Georgia | [SEP] row 10: | Steve Smith | Wide receiver | Taft High School | Woodland Hills, California | USC | [SEP] row 11: | Tom Zbikowski | Athlete | Buffalo Grove High School | Buffalo Grove, Illinois | Notre Dame | [SEP] row 12: | Chase Goggans | Placekicker | Coffee High School | Douglas, Georgia | Florida State |"

"question": "The selected entity mention in the table cell is: South Carolina. The column name for 'South Carolina' is college.

"gt": "<nil [desc] not in list [type] none>",

"output": "<south carolina state university [desc] public university in south carolina, usa [type] university>"

NIL examples

"input": "[TLE] The Wikipedia page is about Basingstoke (UK Parliament constituency). The Wikipedia section is about Members of Parliament. [TAB] col: | election | election | member | party | row 1: | | 1885 | George Sclater-Booth | Conservative | [SEP] row 2: | | 1887 by-election | Arthur Frederick Jeffreys | Conservative | [SEP] row 3: | | 1906 | Arthur Salter | Conservative | [SEP] row 4: | | 1917 by-election | Auckland Campbell Geddes | Unionist | [SEP] row 5: | | 1920 by-election | Arthur Richard Holbrook | Coalition Conservative | [SEP] row 6: | | 1923 | Reginald Fletcher | Liberal | [SEP] row 7: | | 1924 | Arthur Richard Holbrook | Conservative | [SEP] row 8: | | 1929 | Viscount Lymington | Conservative | [SEP] row 9: | | 1934 by-election | Henry Maxence Cavendish Drummond Wolff | Conservative | [SEP] row 10: | | 1935 | Patrick Donner | Conservative | [SEP] row 11: | | 1955 | Denzil Freeth | Conservative | [SEP] row 12: | | 1964 | David Mitchell | Conservative | [SEP] row 13: | | 1983 | Andrew Hunter | Conservative | [SEP] row 14: | | 2002 | Andrew Hunter | Independent Conservative | [SEP] row 15: | | 2004 | Andrew Hunter | Democratic Unionist | [SEP] row 16: | | 2005 | Maria Miller | Conservative |"

"question": "The selected entity mention in the table cell is: Conservative. The column name for 'Conservative' is party.

"gt": "<nil [desc] not in list [type] none>",

"output": "<conservative party [desc] political party in romania founded in 1991 [type] politicalparty>"

NIL examples

"input": "[TLE] The Wikipedia page is about J/80. The Wikipedia section is about World championships. [TAB] col: | year | location | entries | winning boat | country | skipper | row 1: | 2013 | Marseille , France | 117 | New territoriess | POR | Hugo Rocha | [SEP] row 2: | 2012 | Dartmouth , UK | 76 | Nilfisk | ESP | Jos\u00e9 van der Ploeg | [SEP] row 3: | 2011 | Copenhagen , Denmark | 67 | Nextel Engineering | ESP | Ignacio Camino | [SEP] row 4: | 2010 | Newport , USA | 61 | ECC Viviendas | ESP | Jose Maria Torcida | [SEP] row 5: | 2009 | Santander , Spain | 134 | Princesa Yaiza | ESP | Rayco Tabares | [SEP] row 6: | 2008 | Kiel , Baltic Sea | 64 | Nextel Engineering | ESP | Ignacio Camino | [SEP] row 7: | 2007 | La Trinit\u00e9 sur Mer , France | 125 | ECC Viviendas | ESP | Jose Maria Torcida | [SEP] row 8: | 2006 | Corpus Christi, TX , USA | 34 | L'Glide | USA | Glenn Darden | [SEP] row 9: | 2005 | Falmouth , UK | 52 | Volvo | GBR | Ruairidh Scott | [SEP] row 10: | 2004 | Royal Swedish Yacht Club, Stockholm | 56 | Out of Space | SWE | Peder Arvefors | [SEP] row 11: | 2003 | Ft. Worth , Texas , USA | 49 | Synergy | USA | Jay Lutz | [SEP] row 12: | 2002 | La Rochelle , France | 46 | TENDRISSE | FRA | Pascal Abignoli |"

"question": "The selected entity mention in the table cell is: USA. The column name for 'USA' is country

"gt": "<nil [desc] not in list [type] none>"

"output": "<usa [desc] wikimedia disambiguation page [type] none>"

NIL examples

"input": "[TLE] The Wikipedia page is about Roger Stern. The Wikipedia section is about External links. [TAB] col: | precededby len wein | the incredible hulk writer 1978\u20131980 | succeededby steven grant | row 1: | Precededby Roger McKenzie | Captain America writer 1980\u20131981 (with John Byrne in early 1980) | Succeededby J.M. DeMatteis | [SEP] row 2: | Precededby Dennis O'Neil | The Amazing Spider-Man writer 1982\u20131984 | Succeededby Tom DeFalco | [SEP] row 3: | Precededby Steven Grant | The Avengers writer 1983\u20131987 | Succeededby Ralph Macchio | [SEP] row 4: | Precededby John Byrne | Fantastic Four writer 1986\u20131987 | Succeededby Steve Englehart | [SEP] row 5: | Precededby John Byrne | Superman writer 1988\u20131989 | Succeededby Jerry Ordway | [SEP] row 6: | Precededby John Byrne | Action Comics writer 1988\u20131994 | Succeededby David Michelinie | [SEP] row 7: | Precededby Tom McCraw and Tom Peyer | Legionnaires writer with Tom McCraw 1996\u20131999 | Succeededby Dan Abnett and Andy Lanning | [SEP] row 8: | Precededby Kurt Busiek | Iron Man writer 1998\u20132000 (with Kurt Busiek) | Succeededby Joe Quesada |",

"question": "The selected entity mention in the table cell is: Fantastic Four. The column name for 'Fantastic Four' is the incredible hulk writer 1978\u20131980

"gt": "<nil [desc] not in list [type] none>",

"output": "<fantastic four [desc] marvel comics featuring the super heros titled the fantastic four [type] comic>"

Why TableLlama sucks on SemTab data?

A possible answer can be found here

Need for a curated benchmark dataset

- From the examples before it emerges the need for a curated benchmark dataset
- This dataset should contain*
    - Task-specific tables, e.g.:
        - Tables where info for linking is not directly contained in the table
        - Tables where info is contained in the table, but spread into different cols
        - Tables where info is directly contained in the mention
        - Tables where misspelled mentions could lead to wrong linking (Paris vs Pariss)
        - Tables with acronyms (CR7 → Cristiano Ronaldo)
        - Tables with homonyms (Paris [France] vs Paris [Texas])
        - Tables where no correct candidate is present
        - Tables with no header
- Mention-based vs Table-based?

*What the data should contain depends on what we want to measure

Why is it useful?

If we have a curated and high-quality benchmark containing tables with specific and orthogonal characteristics, then:

- We can leverage meaningful and shared benchmark
- We could precisely identify where and why algorithms make mistakes
- We can apply mechanistic interpretability [1][2]

To summarize…

1. From initial results TableLlama performs well on IN and OOD data
    a. Fine-tuning helps improving performances
    b. Still underperforms Alligator on MOOD data
    c. Memory and time consuming
2. Alligator performs well on IN (SemTab) data
    a. We hypothesize that hand-crafted features capture signals from Wikidata, but no Wikipedia or Web Tables
        i. To be further demonstrated with additional experiments
    b. 1 order of magnitude faster than TableLlama
3. TableLlama is not able to understand NIL entities
4. LLMs sometimes have difficulties to understand context
    a. Sequential approach to linking on tables
    b. Matching literals (dates, strings, numbers, …)
5. We need a curated benchmark dataset
    a. Task-specific tables
    b. Enables interpretability (e.g. mechanistic)

Future directions

- Understand why Alligator excels on SemTab data but fails on Web tables
- Single-task curated benchmark creation
  - Mechanistic interpretability
  - Further tests to compare different approaches
- Improvements over TableLlama
  - Can we improve prompt? Can we use Chain-of-Thought and/or In-Context-Learning? TextGrad?
  - Fine-tune Phi-3-mini (3.8B) or Gemma-2 (2B)
    - Introduce NIL entities (whether they are not in the candidate list or not in the KB)
    - Reduce table contex (5 rows vs full table)
    - Masking headers and/or mentions to favor generalization
- View linking on tables as a sequential problem
  - RL can be useful?
- Human-In-The-Loop through imitation
  - Inspired by RLHF literature

Future directions

- Assessing the uncertainty of a decision in matching problem is necessary for the utilization of matching algorithms in real world settings
  - Checking accuracy
  - Revising results
- Can we estimate uncertainty by looking at inner measures from LLMS?
  - E.g., perplexity
  - E.g., more complex approaches:
    - https://arxiv.org/pdf/2406.02543v1,
    - https://arxiv.org/pdf/2401.03426
    - See also results for “estimating uncertainty in LLm based matching”
- Topic of interest for both entity matching and entity linking in tables; potential collaboration?

Thank you for your

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Additional TableLlama experiments

Problem

Loosely inspired by [Oren23], we want to assess the ability of TableLlama to correctly answer CEA questions when:

- Rows and columns are shuffled
- The header is missing (replaced with “col0”, “col1”, …, “colN”)
- The mention is masked (mention references inside the prompt are replaced with “[MASK]”)
  - Inspired by Fill-the-gap elementary/middle school tests
- The table context is reduced (1 row, 3 rows, 5 rows and 11 rows + header)
  - If C is the context-size, then we take the (C-1)/2 rows above and below the row containing the mention to be linked, wrapping around if the selected row is the first or the last one
- Table structure is removed ([TAB] col: | x | y | z | row: 1 | a | b | c | → [TAB] x y z a b c)

Example prompt - No header

input: [TLE] The Wikipedia page is about 2005 NPSL Season. The Wikipedia section is about New Franchises. [TAB] col: | **col0** | **col1** | **col2** | **col3** | row 1: | Detroit Arsenal | Detroit area | Berkley, MI | expansion | [SEP] row 2: | Grand Rapids Alliance | Grand Rapids area | Grand Rapids, MI | expansion | [SEP] row 3: | Milwaukee Bavarians | Milwaukee area | Milwaukee, WI | expansion | [SEP] row 4: | Minnesota Blast | Minneapolis area | Burnsville, MN | expansion | [SEP] row 5: | Princeton 56ers | Madison area | Madison, WI | expansion | [SEP] row 6: | San Diego Pumitas | San Diego area | San Diego, CA | expansion | [SEP] row 7: | San Jose Frogs | San Jose area | San Jose, CA | expansion | [SEP] row 8: | St. Paul Twin Stars | St. Paul area | St. Paul, MN | expansion |
question: The selected entity mention in the table cell is: Grand Rapids, MI. The column name for 'Grand Rapids, MI' is **col2**. The referent entity candidates are: <Grand Rapids [DESCRIPTION] city in Itasca County, Minnesota, United States [TYPE] None>, <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>. What is the correct referent entity for the entity mention 'Grand Rapids, MI'?
output: <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>

Example prompt - No header + Masked

input: [TLE] The Wikipedia page is about 2005 NPSL Season. The Wikipedia section is about New Franchises. [TAB] col: | **col0** | **col1** | **col2** | **col3** | row 1: | Detroit Arsenal | Detroit area | Berkley, MI | expansion | [SEP] row 2: | Grand Rapids Alliance | Grand Rapids area | **[MASK]** | expansion | [SEP] row 3: | Milwaukee Bavarians | Milwaukee area | Milwaukee, WI | expansion | [SEP] row 4: | Minnesota Blast | Minneapolis area | Burnsville, MN | expansion | [SEP] row 5: | Princeton 56ers | Madison area | Madison, WI | expansion | [SEP] row 6: | San Diego Pumitas | San Diego area | San Diego, CA | expansion | [SEP] row 7: | San Jose Frogs | San Jose area | San Jose, CA | expansion | [SEP] row 8: | St. Paul Twin Stars | St. Paul area | St. Paul, MN | expansion |
question: The selected entity mention in the table cell is: **[MASK]**. The column name for '**[MASK]**' is **col2**. The referent entity candidates are: <Grand Rapids [DESCRIPTION] city in Itasca County, Minnesota, United States [TYPE] None>, <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>. What is the correct referent entity for the entity mention '**[MASK]**'?
output: <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>

Example prompt - No header + Masked + Flatten

input: [TLE] The Wikipedia page is about 2005 NPSL Season. The Wikipedia section is about New Franchises. [TAB] **col0 col1 col2 col3** Detroit Arsenal Detroit area Berkley, MI expansion Grand Rapids Alliance Grand Rapids area **[MASK]** expansion Milwaukee Bavarians Milwaukee area Milwaukee, WI expansion Minnesota Blast Minneapolis area Burnsville, MN expansion Princeton 56ers Madison area Madison, WI expansion San Diego Pumitas San Diego area San Diego, CA expansion San Jose Frogs San Jose area San Jose, CA expansion St. Paul Twin Stars St. Paul area St. Paul, MN expansion

question: The selected entity mention in the table cell is: **[MASK]**. The column name for '**[MASK]**' is **col2**. The referent entity candidates are: <Grand Rapids [DESCRIPTION] city in Itasca County, Minnesota, United States [TYPE] None>, <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>. What is the correct referent entity for the entity mention '**[MASK]**'?

output: <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>

Example prompt - No header + Masked + Reduced ctx

input: [TLE] The Wikipedia page is about 2005 NPSL Season. The Wikipedia section is about New Franchises. [TAB] col: | **col0** | **col1** | **col2** | **col3** | row 1: | St. Paul Twin Stars | St. Paul area | St. Paul, MN | expansion | [SEP] row 2: | Detroit Arsenal | Detroit area | Berkley, MI | expansion | [SEP] row 3: | Grand Rapids Alliance | Grand Rapids area | **[MASK]** | expansion | [SEP] row 4: | Milwaukee Bavarians | Milwaukee area | Milwaukee, WI | expansion | [SEP] row 5: | Minnesota Blast | Minneapolis area | Burnsville, MN | expansion |

question: The selected entity mention in the table cell is: **[MASK]**. The column name for '**[MASK]**' is **col2**. The referent entity candidates are: <Grand Rapids [DESCRIPTION] city in Itasca County, Minnesota, United States [TYPE] None>, <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>. What is the correct referent entity for the entity mention '**[MASK]**'?

output: <Grand Rapids [DESCRIPTION] city in and county seat of Kent County, Michigan, United States [TYPE] None>

Datasets

- **TURL-1.7K**: starting with the TURL-2K dataset we
  - Downsampled it to contain only prompts with a number of tokens <= 8192 (**subsampled**)
    - 1788 examples
  - Have created all the datasets with the reduced ctx (**ctx-reduced**) from **subsampled**
    - 1731 examples (due to some table-parsing error)
  - Further downsampled **subsampled** to have the same cardinality of **ctx-reduced**
    - 1731 examples
- **HTR1-2023**: manually created a dataset from the HTR1-Valid-2023
  - Candidate retrieval with LamAPI (50 candidates)
  - Downsampled it to contain only prompts with a number of tokens <= 8192 (**subsampled**)
    - 1375 examples (out of 1405)

Dataset	Shuffled	Masked	Header	Context	Accuracy
HTR1	❌	❌	❌	1	0,7425
HTR1	❌	✅	❌	1	0,3193
HTR1	❌	❌	❌	3	0,7818
HTR1	❌	✅	❌	3	0,3927
HTR1	❌	❌	❌	5	0,7803
HTR1	❌	✅	❌	5	0,4124
HTR1	❌	❌	❌	11	0,7789
HTR1	❌	✅	❌	11	0,4022
HTR1	❌	❌	❌	Full	0,784
HTR1	❌	✅	❌	Full	0,4109
HTR1	✅	❌	❌	Full	0,7803
HTR1	✅	✅	❌	Full	0,392

Underlined are the results with the default downsampled dataset: no-mask, no-shuffle and w/-header

Dataset	Shuffled	Masked	Header	Context	Accuracy
TURL-1.7K	❌	❌	✅	1	0,9468
TURL-1.7K	❌	✅	✅	1	0,7926
TURL-1.7K	❌	✅	❌	1	0,6765
TURL-1.7K	❌	❌	❌	1	0,9076
TURL-1.7K	❌	❌	✅	3	0,9434
TURL-1.7K	❌	✅	✅	3	0,7787
TURL-1.7K	❌	✅	❌	3	0,6799
TURL-1.7K	❌	❌	❌	3	0,926
TURL-1.7K	❌	❌	✅	5	0,9463
TURL-1.7K	❌	✅	✅	5	0,7545
TURL-1.7K	❌	✅	❌	5	0,6649
TURL-1.7K	❌	❌	❌	5	0,9249
TURL-1.7K	❌	❌	✅	11	0,9428
TURL-1.7K	❌	✅	✅	11	0,7429
TURL-1.7K	❌	✅	❌	11	0,6534
TURL-1.7K	❌	❌	❌	11	0,9214
TURL-1.7K	❌	❌	✅	Full	0,9405
TURL-1.7K	❌	✅	✅	Full	0,7348
TURL-1.7K	✅	❌	✅	Full	0,9099
TURL-1.7K	✅	✅	✅	Full	0,647

No-flatten

Dataset	Shuffled	Masked	Header	Context	Accuracy
HTR1	❌	❌	❌	1	0,7425
HTR1	❌	✅	❌	1	0,3193
HTR1	❌	❌	❌	3	0,7818
HTR1	❌	✅	❌	3	0,3927
HTR1	❌	❌	❌	5	0,7803
HTR1	❌	✅	❌	5	0,4124
HTR1	❌	❌	❌	11	0,7789
HTR1	❌	✅	❌	11	0,4022
HTR1	❌	❌	❌	Full	0,7840
HTR1	❌	✅	❌	Full	0,4109
HTR1	✅	❌	❌	Full	0,7803
HTR1	✅	✅	❌	Full	0,3920

Flatten

Dataset	Shuffled	Masked	Header	Context	Accuracy
HTR1	❌	❌	❌	1	0,7447
HTR1	❌	✅	❌	1	0,2843
HTR1	❌	❌	❌	3	0,7629
HTR1	❌	✅	❌	3	0,3476
HTR1	❌	❌	❌	5	0,7731
HTR1	❌	✅	❌	5	0,3396
HTR1	❌	❌	❌	11	0,7658
HTR1	❌	✅	❌	11	0,3491
HTR1	❌	❌	❌	Full	0,7614
HTR1	❌	✅	❌	Full	0,3491
HTR1	✅	❌	❌	Full	0,7673
HTR1	✅	✅	❌	Full	0,3614

No-flatten

Dataset	Shuffled	Masked	Header	Context	Accuracy
TURL-1.7K	❌	❌	✅	1	0,9468
TURL-1.7K	❌	✅	✅	1	0,7926
TURL-1.7K	❌	✅	❌	1	0,6765
TURL-1.7K	❌	❌	❌	1	0,9076
TURL-1.7K	❌	❌	✅	3	0,9434
TURL-1.7K	❌	✅	✅	3	0,7787
TURL-1.7K	❌	✅	❌	3	0,6799
TURL-1.7K	❌	❌	❌	3	0,9260
TURL-1.7K	❌	❌	✅	5	0,9463
TURL-1.7K	❌	✅	✅	5	0,7545
TURL-1.7K	❌	✅	❌	5	0,6649
TURL-1.7K	❌	❌	❌	5	0,9249
TURL-1.7K	❌	❌	✅	11	0,9428
TURL-1.7K	❌	✅	✅	11	0,7429
TURL-1.7K	❌	✅	❌	11	0,6534
TURL-1.7K	❌	❌	❌	11	0,9214
TURL-1.7K	❌	❌	✅	Full	0,9405
TURL-1.7K	❌	✅	✅	Full	0,7348
TURL-1.7K	✅	❌	✅	Full	0,9099
TURL-1.7K	✅	✅	✅	Full	0,6470

Flatten

Dataset	Shuffled	Masked	Header	Context	Accuracy
TURL-1.7K	❌	❌	✅	1	0,9411
TURL-1.7K	❌	✅	✅	1	0,7516
TURL-1.7K	❌	✅	❌	1	0,6672
TURL-1.7K	❌	❌	❌	1	0,9110
TURL-1.7K	❌	❌	✅	3	0,9388
TURL-1.7K	❌	✅	✅	3	0,7412
TURL-1.7K	❌	✅	❌	3	0,6464
TURL-1.7K	❌	❌	❌	3	0,9157
TURL-1.7K	❌	❌	✅	5	0,9422
TURL-1.7K	❌	✅	✅	5	0,7337
TURL-1.7K	❌	✅	❌	5	0,6297
TURL-1.7K	❌	❌	❌	5	0,9081
TURL-1.7K	❌	❌	✅	11	0,9382
TURL-1.7K	❌	✅	✅	11	0,7099
TURL-1.7K	❌	✅	❌	11	0,6280
TURL-1.7K	❌	❌	❌	11	0,9110
TURL-1.7K	❌	❌	✅	Full	0,9330
TURL-1.7K	❌	✅	✅	Full	0,6967
TURL-1.7K	✅	❌	✅	Full	0,9122
TURL-1.7K	✅	✅	✅	Full	0,6031

Discussion

- A 5-context table is enough to maintain high performance (even without header)
  - TURL-1.7K: 0.9463 vs 0.9405 (5-ctx vs full-ctx resp.)
  - HTR1-2023: 0.7803 vs 0.7840 (5-ctx vs full-ctx resp.)
  - Is this due to the fact that the model has seen the full table during the fine-tuning?
- Masking drops performances (especially when combined with no-header)
  - TURL-1.7K: 0.7348 vs 0.9405 (masked vs not-masked resp.)
  - HTR1-2023: 0.4109 vs 0.7840 (masked vs not-masked resp.)
- Shuffling rows and cols affects more TURL-1.7K than HTR1-2023
  - TURL-1.7K: 0.9099 vs 0.9405 (shuffled vs not-shuffled resp.)
  - HTR1-2023: 0.7803 vs 0.7840 (shuffled vs not-shuffled resp.)
  - Is this a sign of data overfitting?
- Decreasing the table-ctx increases the performance on TURL-1.7K
  - Is this a sign of overfitting?

Discussion

- Table structure influences the most when the mention is masked
  - TURL-1.7K: 0.9330 vs 0.9405 (full-ctx-flatten vs full-ctx resp.)
  - TURL-1.7K: 0.6967 vs 0.7348 (full-ctx-masked flatten vs full-ctx masked resp.)
  - HTR1-2023: 0.7614 vs 0.7840 (full-ctx-flatten vs full-ctx resp.)
  - HTR1-2323: 0.3491 vs 0.4109 (full-ctx-masked flatten vs full-ctx masked resp.)
- NIL cannot be recognized
  - If the semantic of a candidate resembles the mention's label, then TableLlama chooses that candidate
- Has already observed, header influences the performances, especially when the table-ctx is reduced

Next steps

- ~~Add experiments where table structure is missing~~
  - ~~[TAB] col: | x | y | z | [SEP] row 1: | a| b | c | → [TAB] x y z a b c~~
  - ~~Does the table structure have an impact on the overall performance?~~
- Add experiments for TURL-1.7K with candidates retrieved with LamAPI
- Check for seen/unseen candidates and mentions
- Efficiency
  - Is there a way to re-use the kv-cache for the same table?
    - Yes, everything except the question can be reused
- Ask TableLlama if it has some knowledge regarding the table structure as TableQA
  - Give me the row/col/row-col index of the mention X?
  - What is the col name of the mention X?
- Multi-Modality: text + image
  - Can multi-modality improve the performances?

Next steps

- ~~Test NIL on TableLlama~~
  - ~~Replace correct candidate with NIL candidate~~

Thank you for your

$$\text{Attention}(Q,K,V) \;=\; \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$