A Journey of Effective Data Curation: from Data Annotation to Data Integration and Organization.

Yushi Sun

Last updated 2025/1/12

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

My publications

- **Cross-Domain-Aware Worker Selection with Training for Crowdsourced Annotation**, ICDE 2024.
- **RECA: Related Tables Enhanced Column Semantic Type Annotation Framework**, VLDB 2023.
- **Are Large Language Models a Good Replacement of Taxonomies?**, VLDB 2024.
- ……

Data Annotation

Data Integration

Data Organization

Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Integration: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- Data Organization: Are Large Language Models a Good Replacement of Taxonomies?
- Future Vision and Opportunities

Background: Data Curation

- The process of data curation involves all essential processes for systematic and regulated data annotation, integration, and organization, along with the ability to enhance the value of that data [1, 2].
    - Data Annotation: annotating raw data to provide standardized context and meaning.

What kind of flower is shown?

petunia	?
morning glory	?
desert rose	?

- The necessity of domain knowledge and the inherent difficulties of the annotation tasks call for a novel cross-domain annotator training and selection scheme.

[1] A. Freitas and E. Curry, “Big data curation,” New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe, pp. 87–118, 2016.
[2] R. J. Miller et al., “Big data curation.” in COMAD, 2014, p. 4.

Background: Data Curation

- The process of data curation involves all essential processes for systematic and regulated data annotation, integration, and organization, along with the ability to enhance the value of that data [1, 2].
  - Data Integration: combining data from different sources to provide a unified view or dataset.

Actor Name

Name	Date
Tom Cruise	03/07/1962
Tom Hanks	09/07/1956
Morgan Freeman	01/06/1937

Product Name

Name	Date
MacBook Pro 14"	01/01/2025
MacBook Air 13"	24/12/2024
Mac Studio	01/09/2024

  - Need for a deeper understanding of table context to clarify the subtle differences in column semantic -> accurate column semantic type annotation

[1] A. Freitas and E. Curry, "Big data curation," New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe, pp. 87–118, 2016.
[2] R. J. Miller et al., "Big data curation." in COMAD, 2014, p. 4.

Background: Data Curation

- The process of data curation involves all essential processes for systematic and regulated data annotation, integration, and organization, along with the ability to enhance the value of that data [1, 2].
  - Data Organization: involves categorizing, storing, and maintaining data in a way that makes it easy to use.

  - Further exploration of novel data organization paradigm in the era of LLMs.

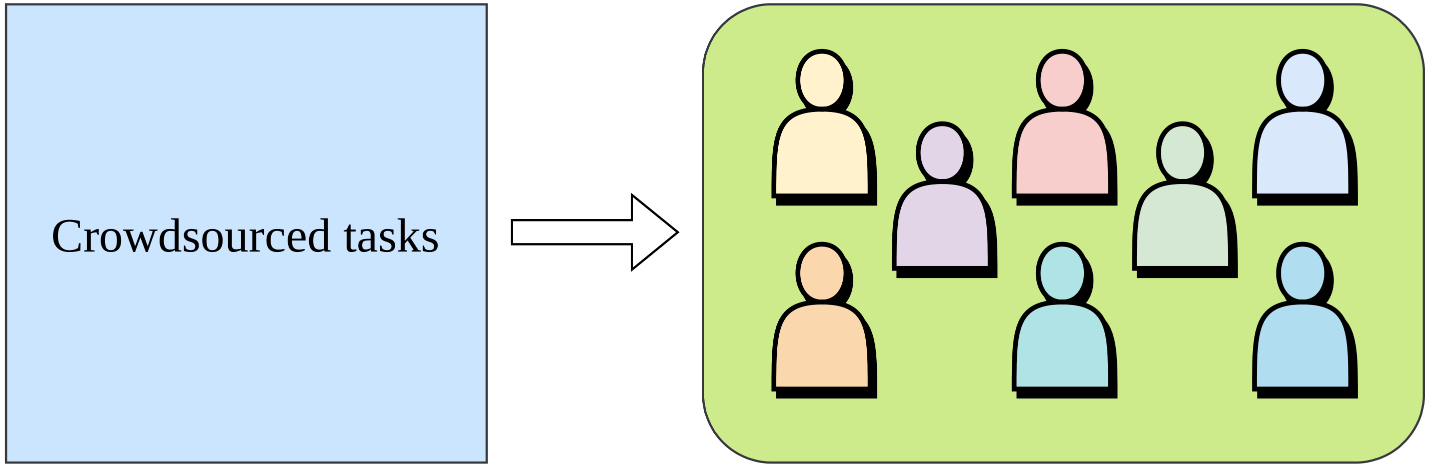
[1] A. Freitas and E. Curry, "Big data curation," New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe, pp. 87–118, 2016.
[2] R. J. Miller et al., "Big data curation." in COMAD, 2014, p. 4.
[3] Andreas, "Taxonomy: Tracing Its Greek Roots to Modern Biological Classification - U speak Greek," U speak Greek, Dec. 25, 2023. https://uspeakgreek.com/science/biology/taxonomy-tracing-its-greek-roots-to-modern-biological-classification/ (accessed Aug. 18, 2024).

Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Integration: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- Data Organization: Are Large Language Models a Good Replacement of Taxonomies?
- Future Vision and Opportunities

Overview

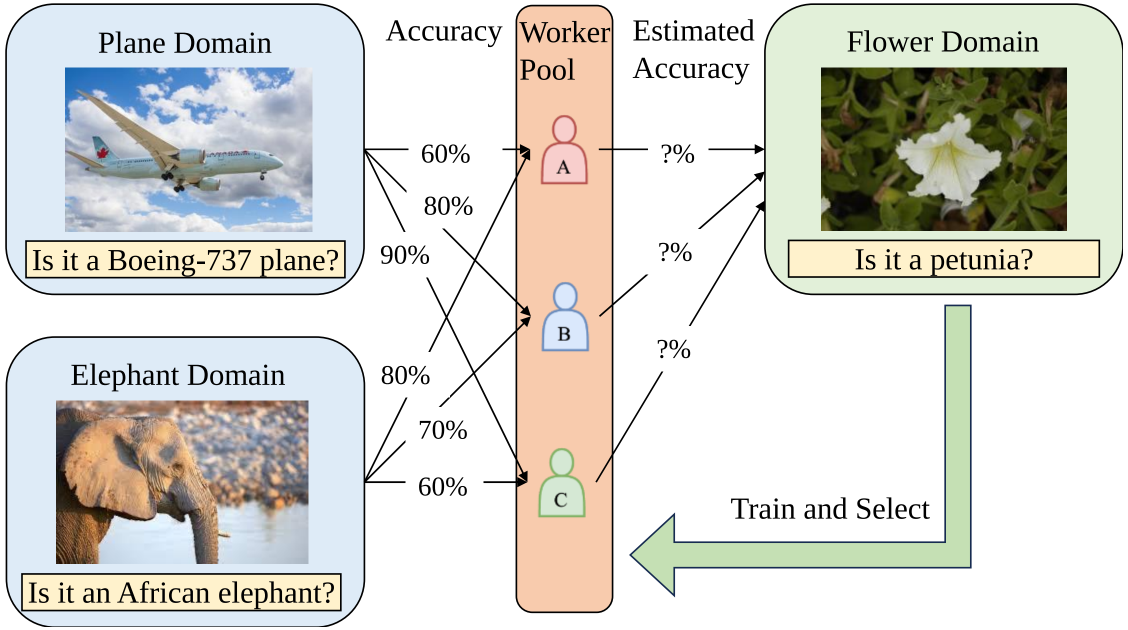
- **Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation (ICDE 2024)**
  - Crowdsourcing is preferable for obtaining high-quality data labels for large-scale datasets.
  - Worker Selection is important in Crowdsourcing.
  - How to design an allocation scheme for golden questions (questions with ground truth answers that are used for worker training/selection) to **train and select** high-performance crowd workers for the incoming crowdsourced tasks remains a challenge.



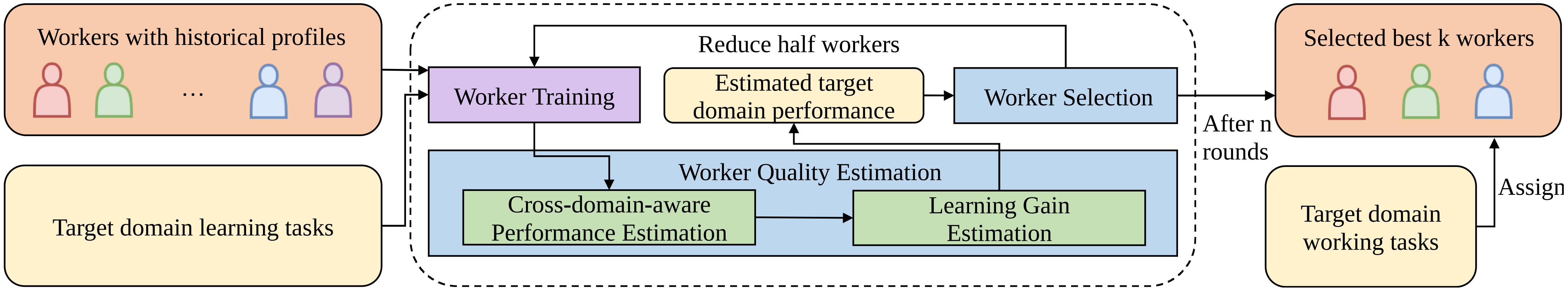
[4] Y. Sun, et al., "Cross-Domain-Aware Worker Selection with Training for Crowdsourced Annotation," in 2024 IEEE 40th International Conference on Data Engineering (ICDE), Utrecht, Netherlands, 2024 pp. 249-262. doi: 10.1109/ICDE60146.2024.00026

Background

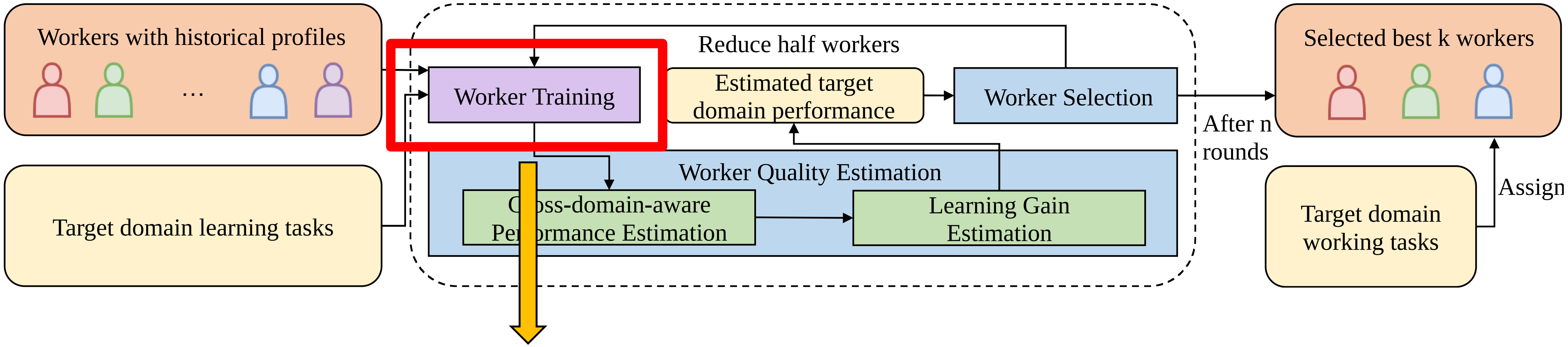
- Many companies such as JD, Alibaba, and Baidu have their commercial crowdsourcing platforms with worker pools, which record the answering history of workers.
- The answering history of workers (prior domain knowledge) can help select high-quality workers when annotating a new domain (target domain task).

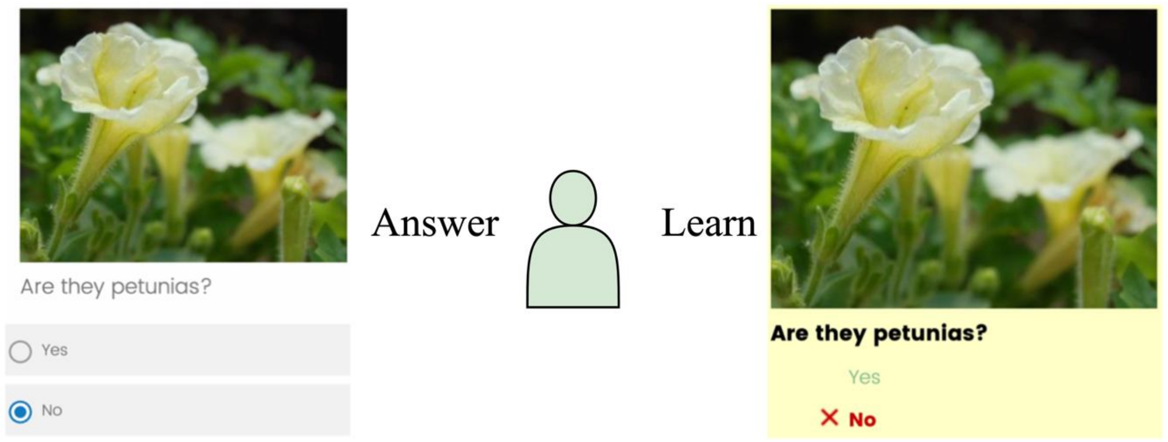


Methodology

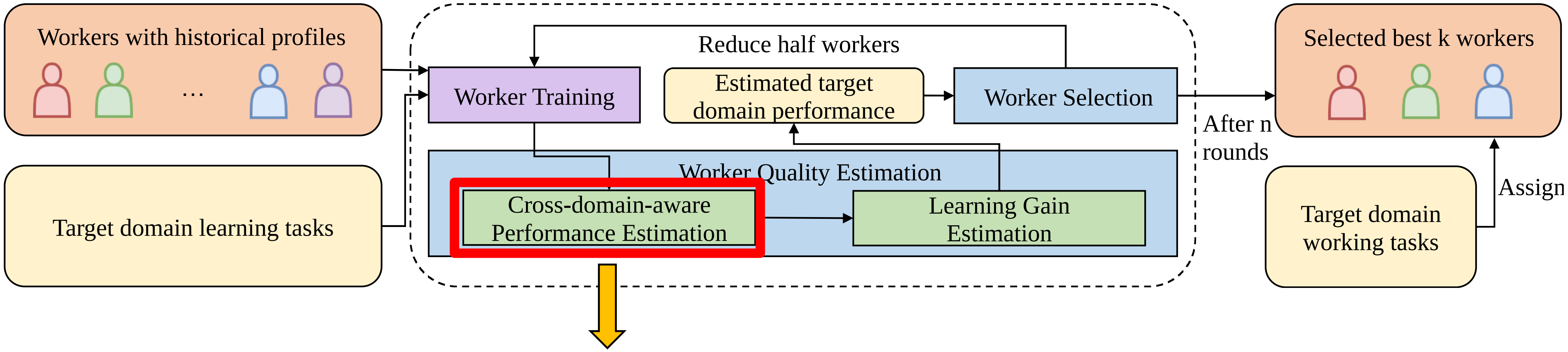


Methodology





Methodology



- Multi-variate normal distribution to model the correlation of the crowd-worker as a group over different domains.
- Maximum Likelihood Estimation to estimate the parameters in the distribution based on the worker training results.

Methodology

- Maximum likelihood estimation:

$$
\begin{aligned}
\log L &= \sum_{i=1}^{|W_c|} \log P(h_{i,T}|h_i) \\
&= \sum_{i=1}^{|W_c|} \log \int_0^1 h_{i,T}^{C_{i,c}} (1-h_{i,T})^{X_{i,c}} \frac{e^{-\Psi}}{\sqrt{2\pi|\bar{\Sigma}|}} \mathrm{d}h_{i,T} \\
&= \sum_{i=1}^{|W_c|} \Big[\log \int_0^1 h_{i,T}^{C_{i,c}} (1-h_{i,T})^{X_{i,c}} e^{-\Psi} \mathrm{d}h_{i,T} \\
&\quad + \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2} \log |\bar{\Sigma}| \Big],
\end{aligned}
$$

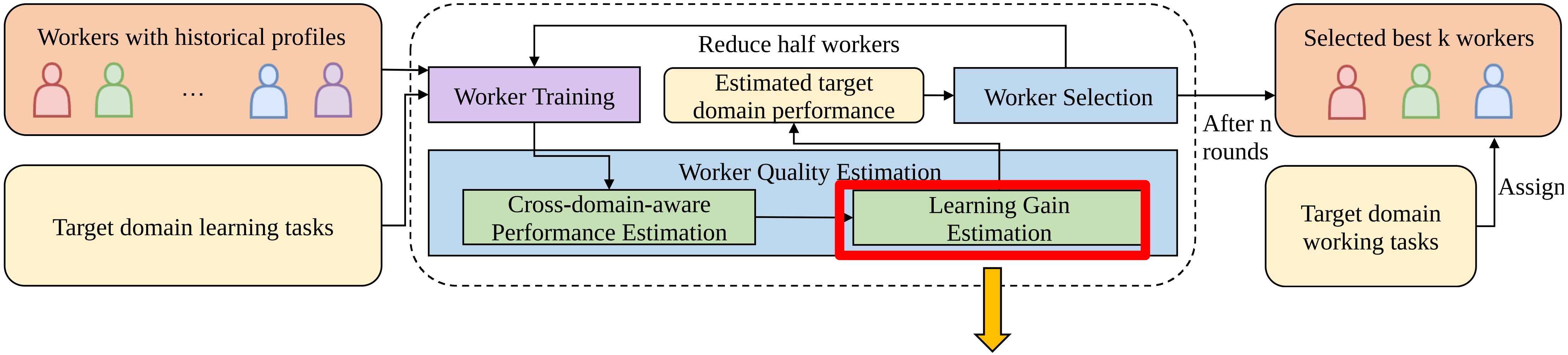
$$
\begin{aligned}
\bar{\mu} &= \mu_T + \Sigma_{1\times D} \Sigma_{D\times D}^{-1} (h_i - \mu_{1\sim D}), \\
\bar{\Sigma} &= \Sigma_{1\times 1} - \Sigma_{1\times D} \Sigma_{D\times D}^{-1} \Sigma_{D\times 1},
\end{aligned}
$$

and $\Psi = \frac{(h_{i,T}-\bar{\mu})^{\mathsf{T}}(h_{i,T}-\bar{\mu})}{2\bar{\Sigma}}$.

- Updated annotation accuracy:

$$
\begin{aligned}
p_{c,i} &= E[h_{i,T}|h_i] \\
&= \int_0^1 h_{i,T} P(h_{i,T}|h_i) \mathrm{d}h_{i,T} \\
&= \int_0^1 h_{i,T} \frac{P(h_i, h_{i,T})}{P(h_i)} \mathrm{d}h_{i,T},
\end{aligned}
$$

Methodology



- Item Response Theory (IRT) to model the dynamic worker knowledge change during the training process for each individual worker.

Methodology

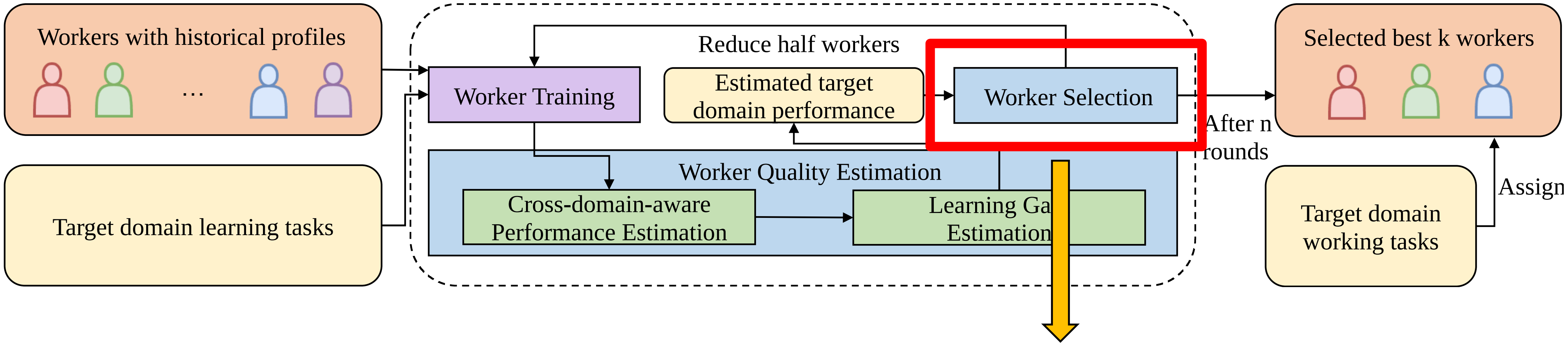
- IRT score:

$$\hat{p}_{j,i,d} = g(\alpha_i, \beta_d, K_j) = \frac{1}{1 + e^{-(\alpha_i \ln(K_j+1) - \beta_d)}}.$$

- Update the learning parameter α_i:

$$\alpha_i = \arg\min_{\alpha_i} \left[\sum_{d=1}^{D} (\hat{p}_{1,i,d} - h_{i,d})^2 + \sum_{j=1}^{c} (\hat{p}_{j-1,i,t} - p_{j,i})^2 \right]$$

Methodology



- Medium Elimination, preserve the better half of the workers in the current round and enter the next round.
- Error bound: $O(\sqrt{\frac{nk}{B}\ln\frac{1}{\delta_c}})$.

Datasets

- Datasets:

TABLE II
DATASET STATISTICS

Datasets	|W|	Q	k	total # of batches	B
RW-1	27	10	7	3	540
RW-2	35	10	9	3	700
S-1	40	20	5	7	2400
S-2	50	20	5	7	3000
S-3	80	20	5	15	6400
S-4	160	20	5	31	16000

|W|: number of crowdsourced workers

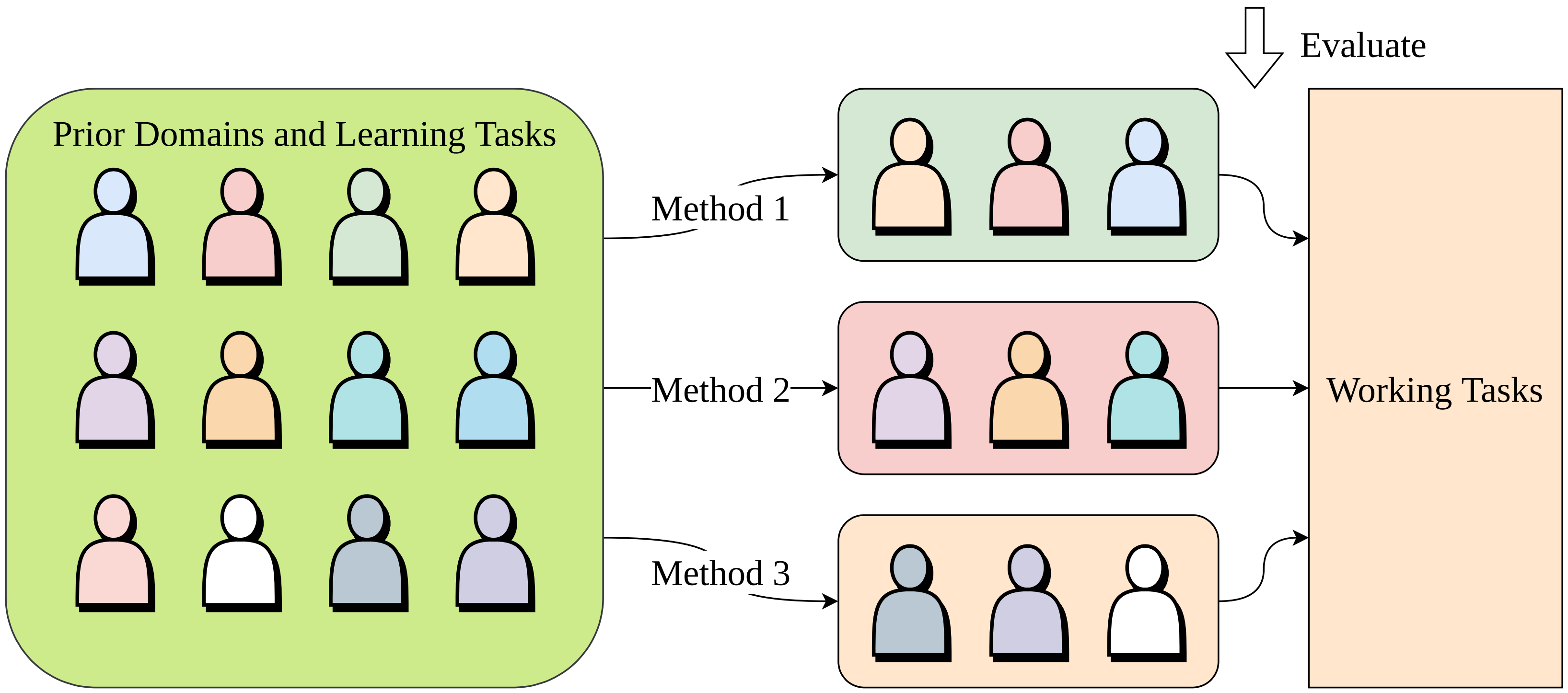
Q: number of learning tasks per batch

k: number of top-k desired workers

B: total worker selection budget

Metrics

- Metric: averaged annotation accuracy of the selected top-k workers on the target domain working task.



Baselines

- Baselines: We considered three baselines, Universal Sampling (US), Medium Elimination (ME), and Li et al.
  - US: use the budget for all the workers equally and select the top k workers
  - ME: allocates the budget in rounds and eliminates the workers by half in each round based on the accuracy of the learning tasks
  - Li et al.: compute the correlation between the prior domain historical results with the target domain performance

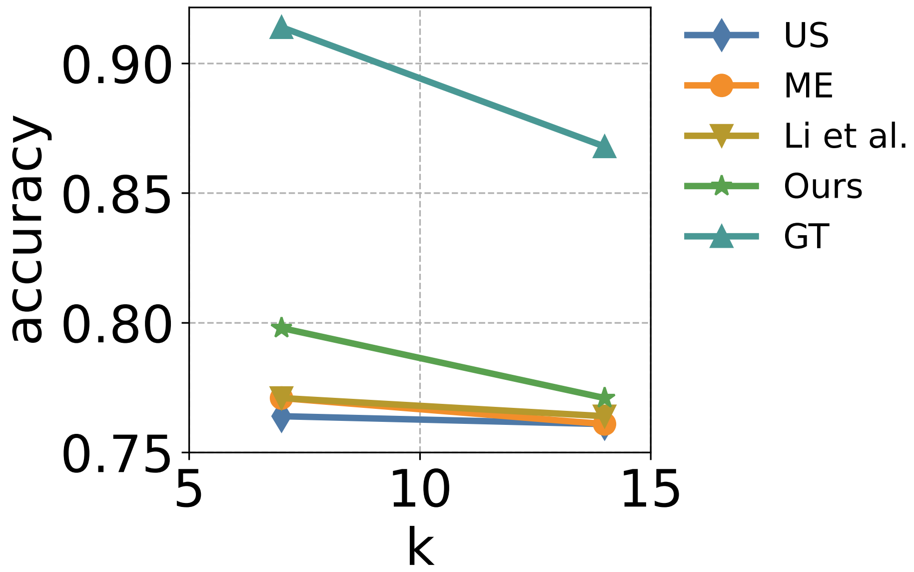
Experiments

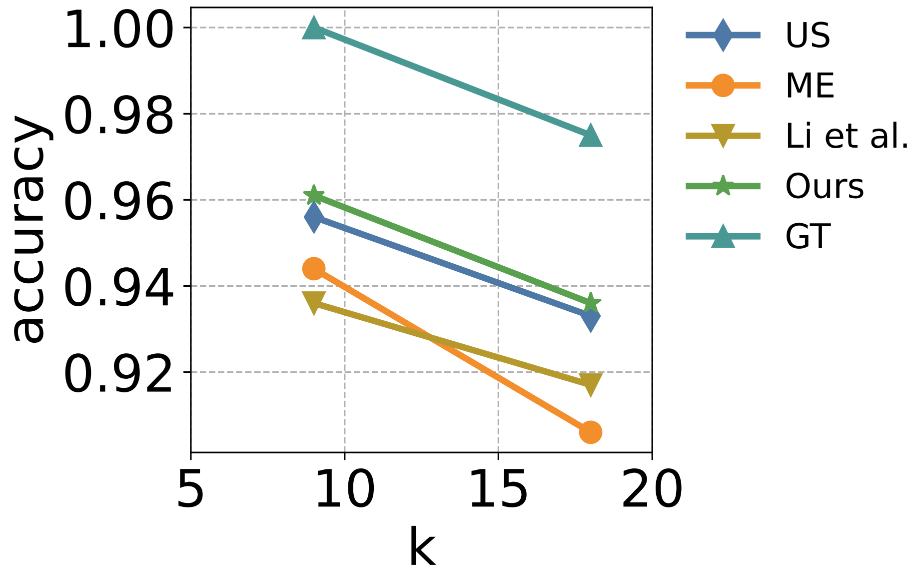
TABLE V
EXPERIMENT RESULTS

	RW-1	RW-2	S-1	S-2	S-3	S-4
US [11], [19]	0.764 (4.5% ↑)	0.956 (0.5% ↑)	0.765 (8.5% ↑)	0.775 (6.8% ↑)	0.815 (4.3% ↑)	0.865 (2.4% ↑)
ME [11], [19]	0.771 (3.5% ↑)	0.944 (1.8% ↑)	0.720 (15.3% ↑)	0.785 (5.5% ↑)	0.795 (6.9% ↑)	0.880 (0.7% ↑)
Li et al. [31]	0.771 (3.5% ↑)	0.936 (2.7% ↑)	0.780 (6.4% ↑)	0.805 (2.9% ↑)	0.845 (0.6% ↑)	0.870 (1.8% ↑)
Ours	**0.798**	**0.961**	**0.830**	**0.828**	**0.850**	**0.886**
Ground Truth	0.914	1.000	0.885	0.875	0.915	0.975

Experiments

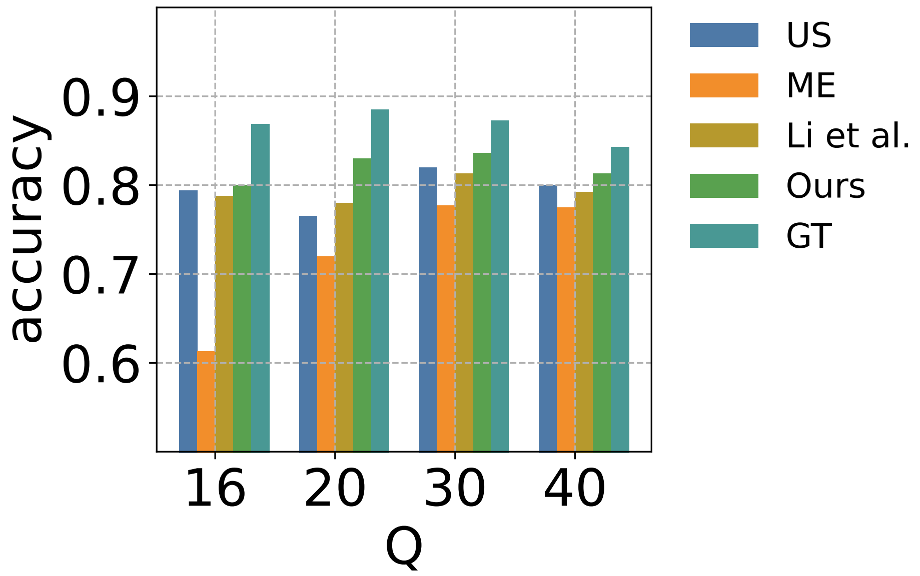
- Stability over the parameter k (number of desired workers)

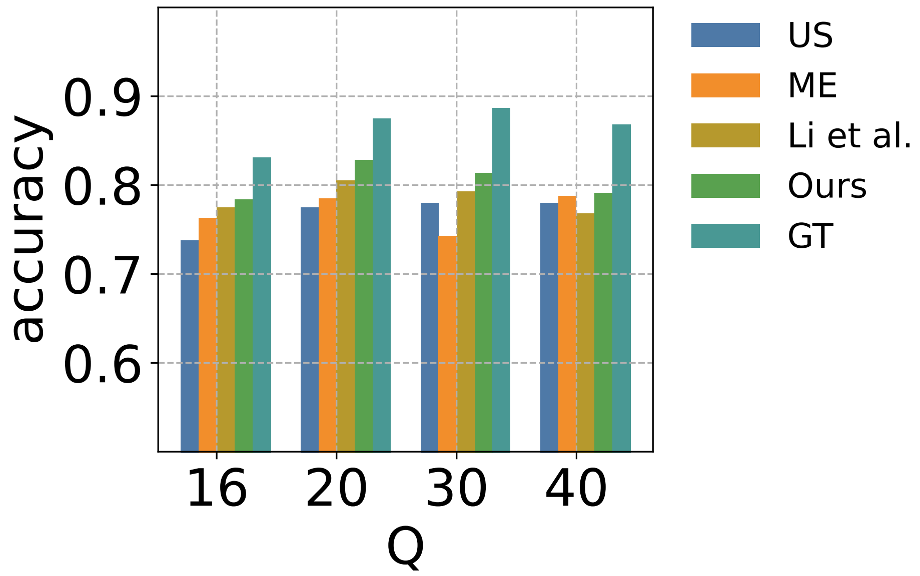




Experiments

- Stability over the parameter Q (number of learning tasks per batch)





Summary

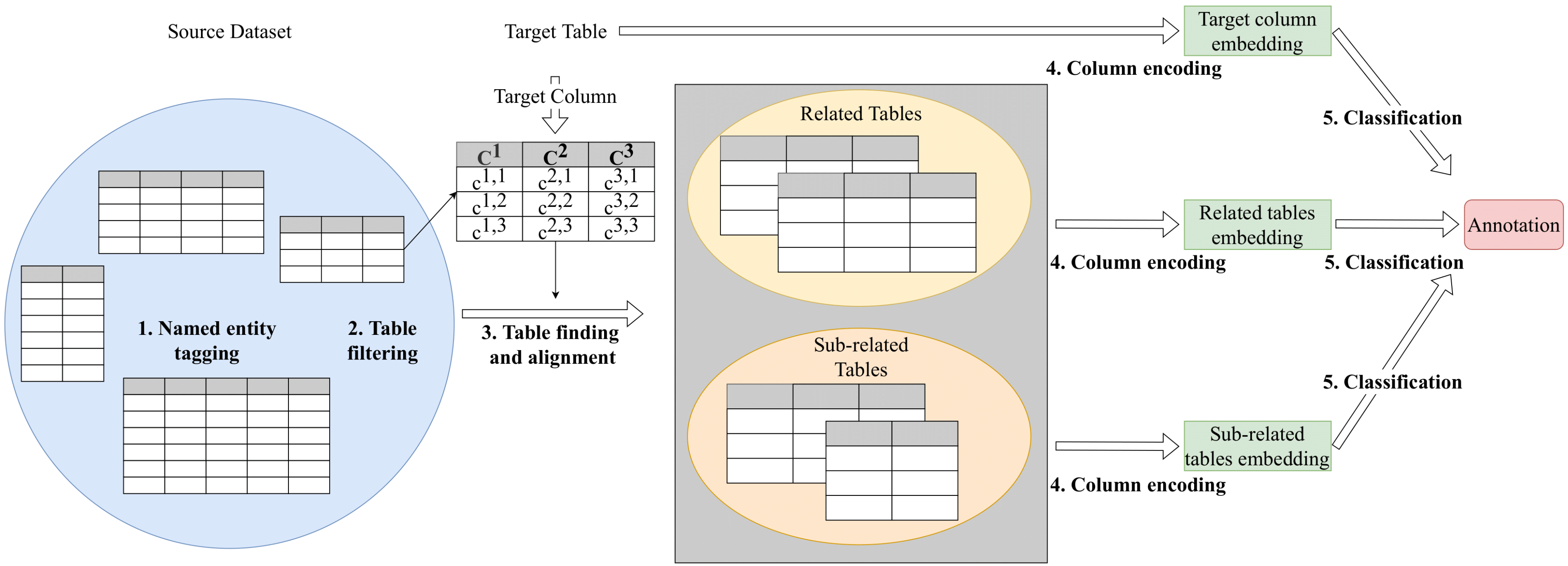
- We incorporate the **cross-domain knowledge** information and propose a novel **Median Elimination-based** worker selection with training algorithm to find high-quality workers for data annotation.
- We comprehensively consider the **learning gain** of workers during the learning task worker training process over the new domain to get a better estimate of the **dynamic change** in worker quality.
- We collect **two novel cross-domain worker selection datasets** for the community to study the problem of cross-domain worker selection with training.
- We conduct **extensive experiments** on real-world and synthesized datasets to evaluate the performance of our proposed method comprehensively.

Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Integration: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- Data Organization: Are Large Language Models a Good Replacement of Taxonomies?
- Future Vision and Opportunities

Overview

- **RECA: Related Tables Enhanced Column Semantic Type Annotation Framework (VLDB 2023)**
- Focus on enhancing **table column semantic type annotation** with **inter-table** context information.



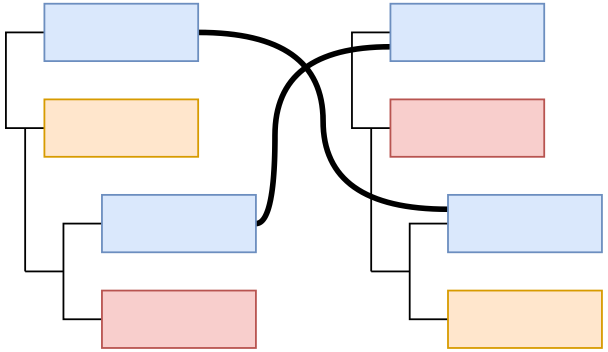
[5] Y. Sun, H. Xin, and L. Chen, "RECA: Related Tables Enhanced Column Semantic Type Annotation Framework," *Proceedings of the VLDB Endowment*, vol. 16, no. 6, pp. 1319–1331, Feb. 2023, doi: https://doi.org/10.14778/3583140.3583149.

Definition

- (Column semantic type annotation): Given a table T from the data lake D, denote the target column as C_t in T. The column semantic type annotation model W annotates C_t with a semantic type $\bar{y}_t = W\ (C_t, T, D)$, such that $\bar{y}_t$ best fits the semantics of C_t.

Background

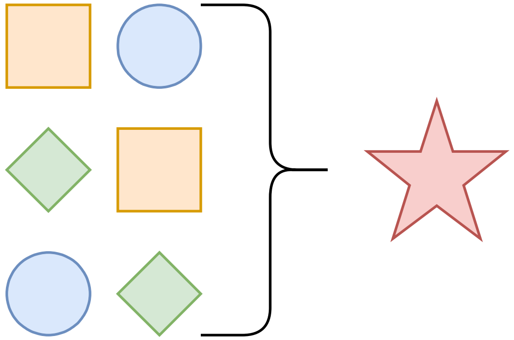
- Accurate column semantic type labeling is important for various applications:
  - schema matching, data cleaning, data integration, etc.



schema matching

Title 1	Title 2	Title 3
Value 1	Value 2	Value 3
Value 4	? ? ?	Value 6
Value 7	Value 8	Value 9
Value 10	Value 11	Value 12

data cleaning



data integration

Challenges

- Existing works (Sherlock, Sato, DODUO, TABBIE, etc.) focus on incorporating the inner-table context.
- Our work focuses on the utilization of inter-table context, which is challenging.

?	?	?	?
Amorcito corazón	L. Suárez	D. Olivera	2012-06-10
A Nero Wolfe Mystery	S. M. Kaminsky	M. Chaykin	2002-08-18

?	?	?	?
Chōriki Sentai Ohranger	T. Inoue	T. Satō	1996-02-23
Chōjin Sentai Jetman	T. Inoue	T. Wakamatsu	1992-02-14
Brewster Place	M. Angelou	O. Winfrey	1990-05-30
Anne of Green Gables: The Continuing Story	K. Sullivan	J. Crombie	2000-07-30
Angry Boys	C. Lilley	C. Lilley	2011-07-27
Alex Haley's Queen	A. Haley	Ann-Margret	1993-02-18
...	...	...	...

WPPD

WPPD

Motivation

- Named Entity Schema: table schema generated based on the most frequent named entity type extracted from each column.
- Tables with the same/similar named entity schemata tend to be from the same/similar data source and thus tend to have the same/similar column semantic types.

?	?	?	?
Amorcito corazón	L. Suárez	D. Olivera	2012-06-10
A Nero Wolfe Mystery	S. M. Kaminsky	M. Chaykin	2002-08-18

WPPD

?	?	?	?
Chōriki Sentai Ohranger	T. Inoue	T. Satō	1996-02-23
Chōjin Sentai Jetman	T. Inoue	T. Wakamatsu	1992-02-14
Brewster Place	M. Angelou	O. Winfrey	1990-05-30
Anne of Green Gables: The Continuing Story	K. Sullivan	J. Crombie	2000-07-30
Angry Boys	C. Lilley	C. Lilley	2011-07-27
Alex Haley's Queen	A. Haley	Ann-Margret	1993-02-18
...	...	...	...

WPPD

?	?	?	?
Donkey Kong Country	Nintendo	2006-12-08	2006
F-Zero	Nintendo	2006-12-08	2006
SimCity	Nintendo	2006-12-29	2006
Super Castlevania IV	Konami	2006-12-29	2006
Street Fighter II: The World Warrior	Capcom	2007-01-19	2007
...	...	...	...

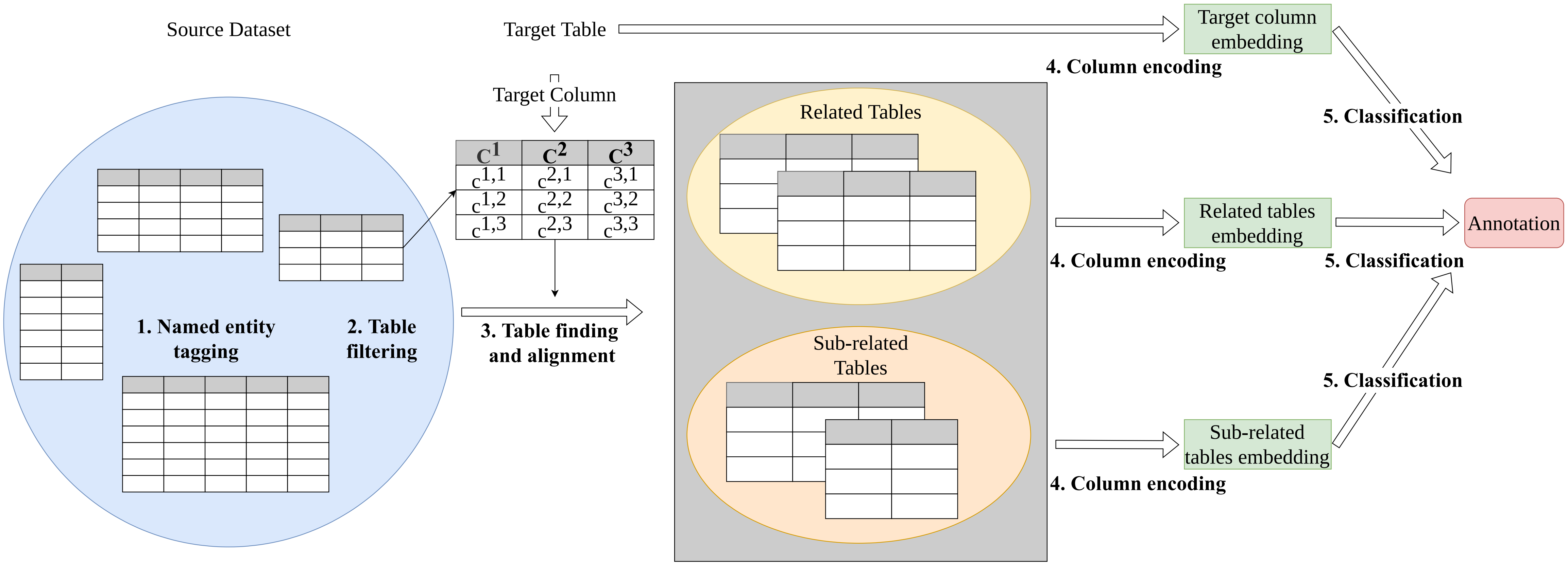
WODD

- W: Work of art; P: Person; D: Date; O: Organization

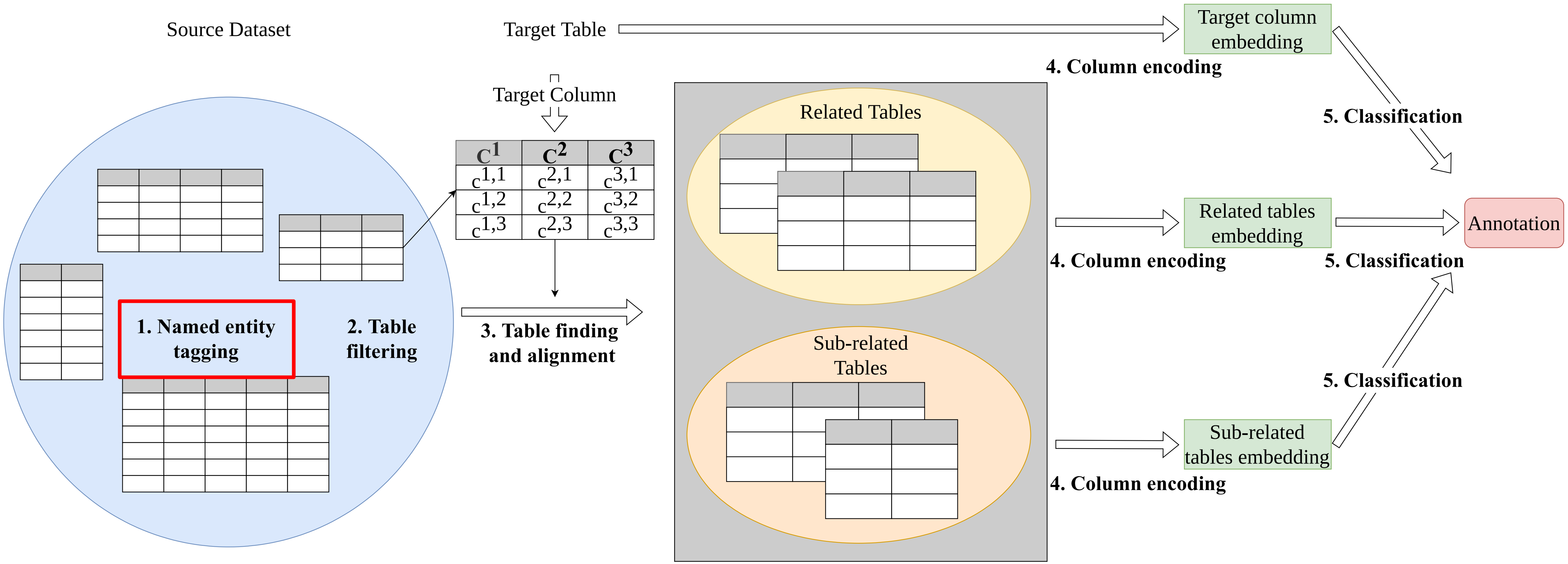
Concepts

- Related Tables: The tables that share the same named entity schema and are similar in content (Jaccard Similarity > δ) with the original table.
- Sub-related Tables: The tables that share a similar named entity schema (the edit distance between their named entity schemata is less than a threshold) and are similar in content (Jaccard Similarity > δ) with the original table.

Methodology



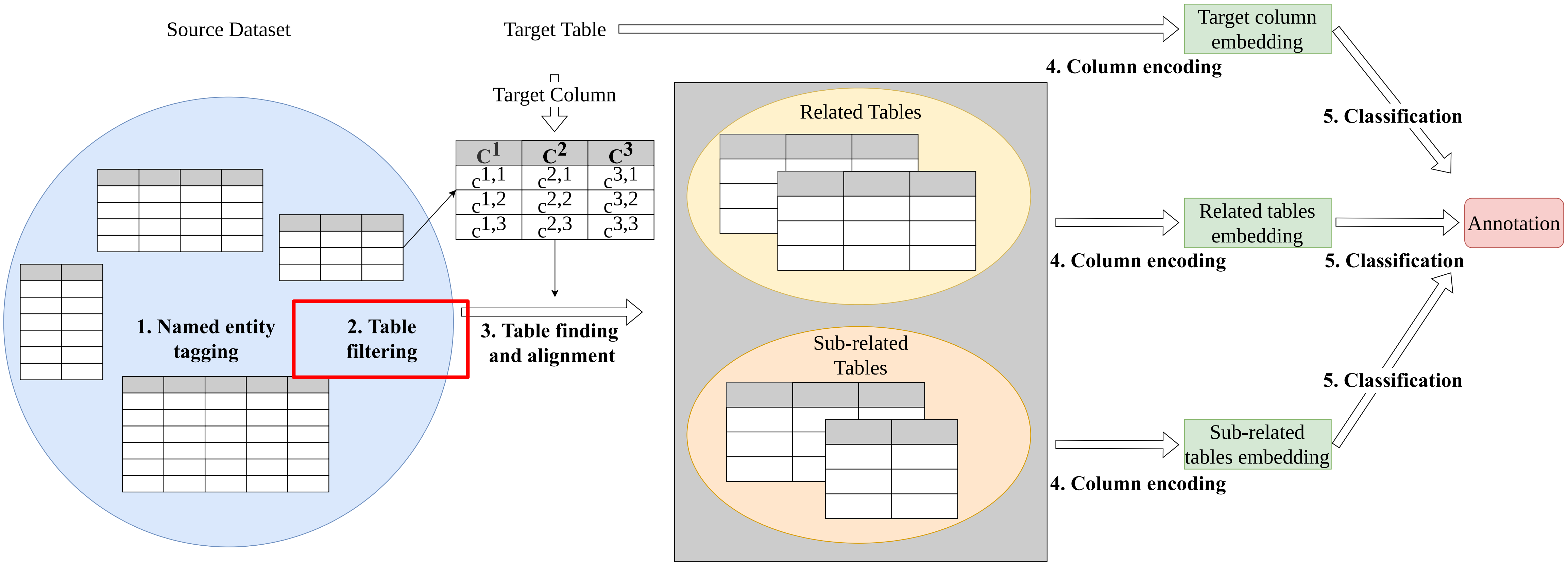
Methodology



Methodology

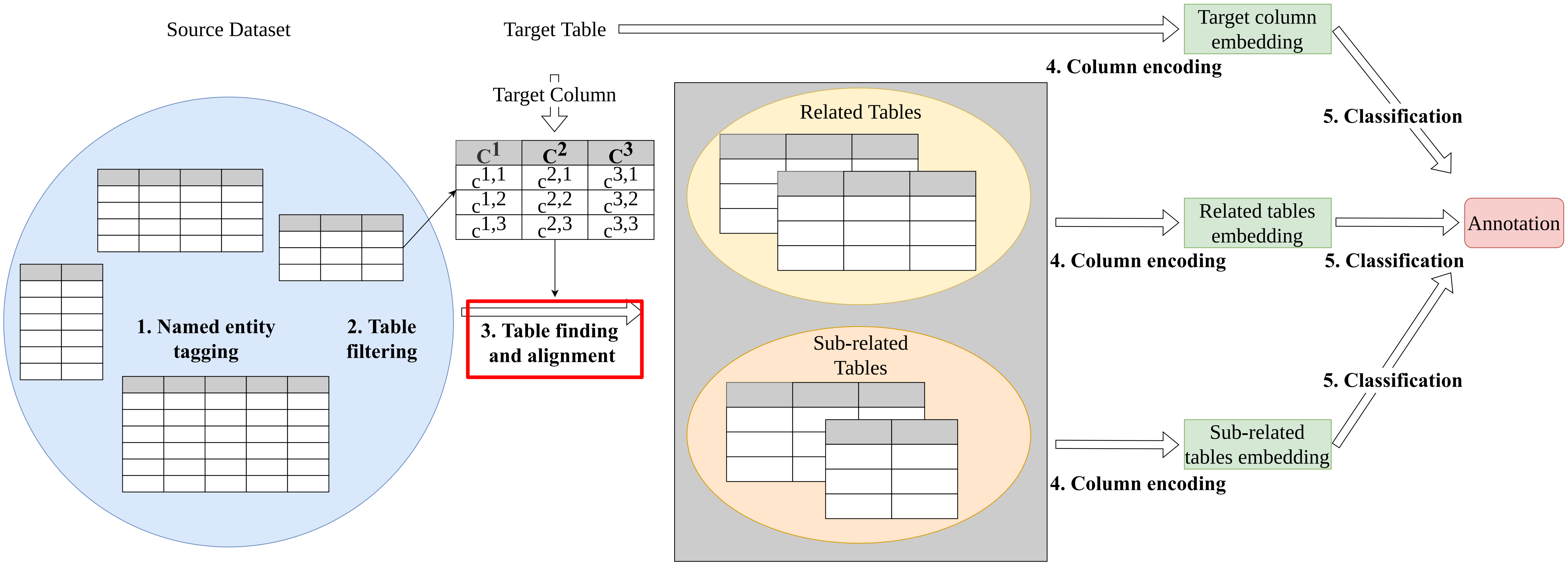
- Given a table T with M columns and N rows, we use the spaCy tagging tool to identify the named entities in each column and tag them.
- We further classify the DATE and PERSON types based on the data format.
  - E.g. DD-MM-YYYY; YYYY; January 16^{th} 2022; 2023
  - E.g. J. K. Rowling; Anna
- We include an additional EMPTY type.
- The most frequent named entity type in each column forms the named entity schema.

Methodology



$$\text{Jaccard}(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$$

Methodology

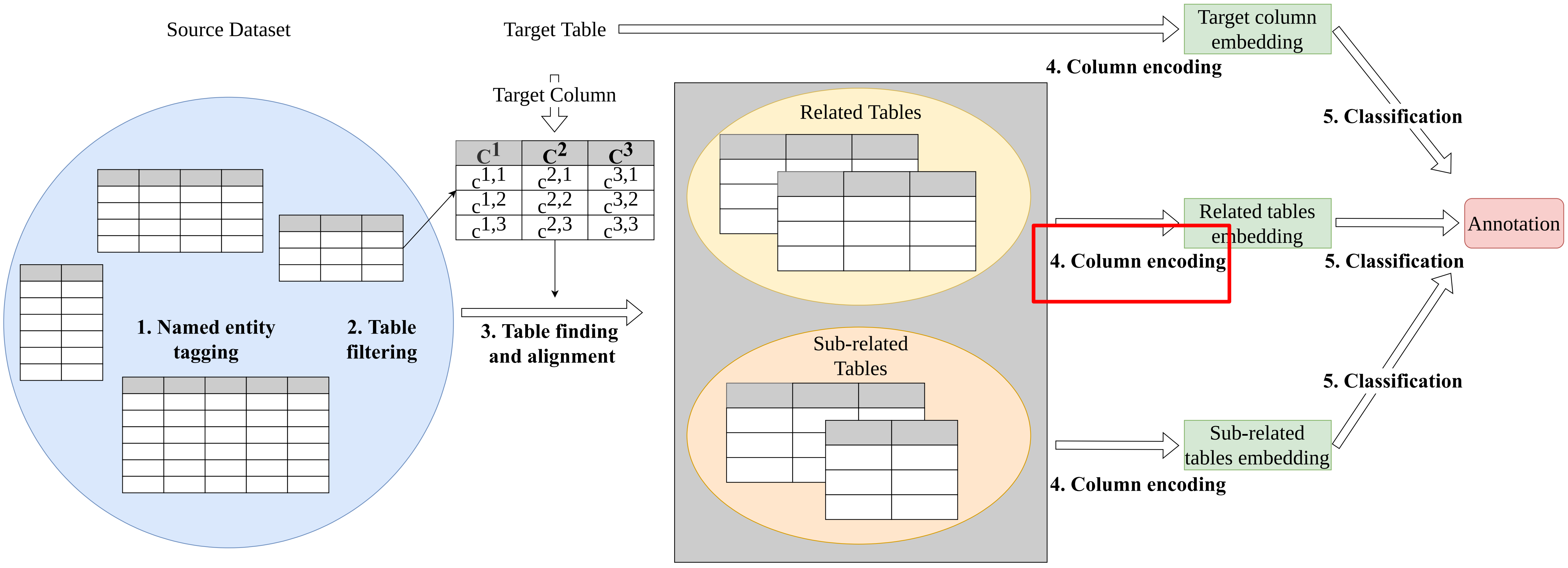


Named Entity Schema & Jaccard Similarity

Methodology

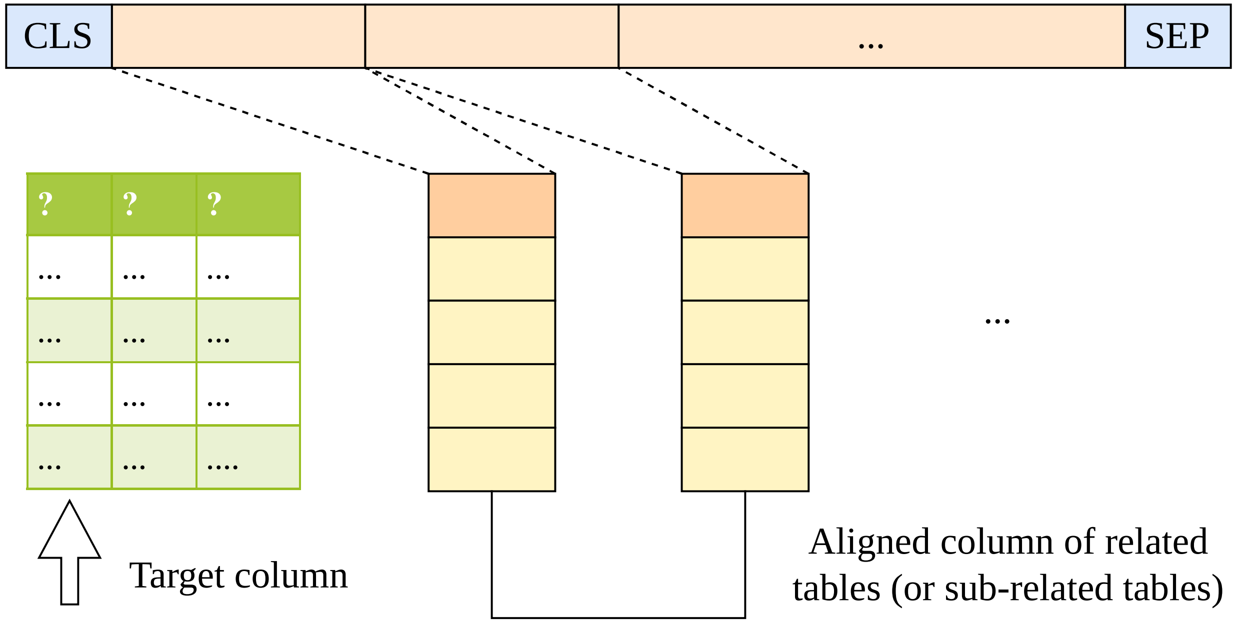
- Related tables: candidate tables T_j that share the same named entity schema as T_i.
- Sub-related tables: we consider the following two requirements:
  - Schema similarity: the named entity schemata should not be very different (edit distance less than a threshold).
  - Column location alignment: The named entity type of the target column matches with that of the column at the identical location in the sub-related table.

Methodology

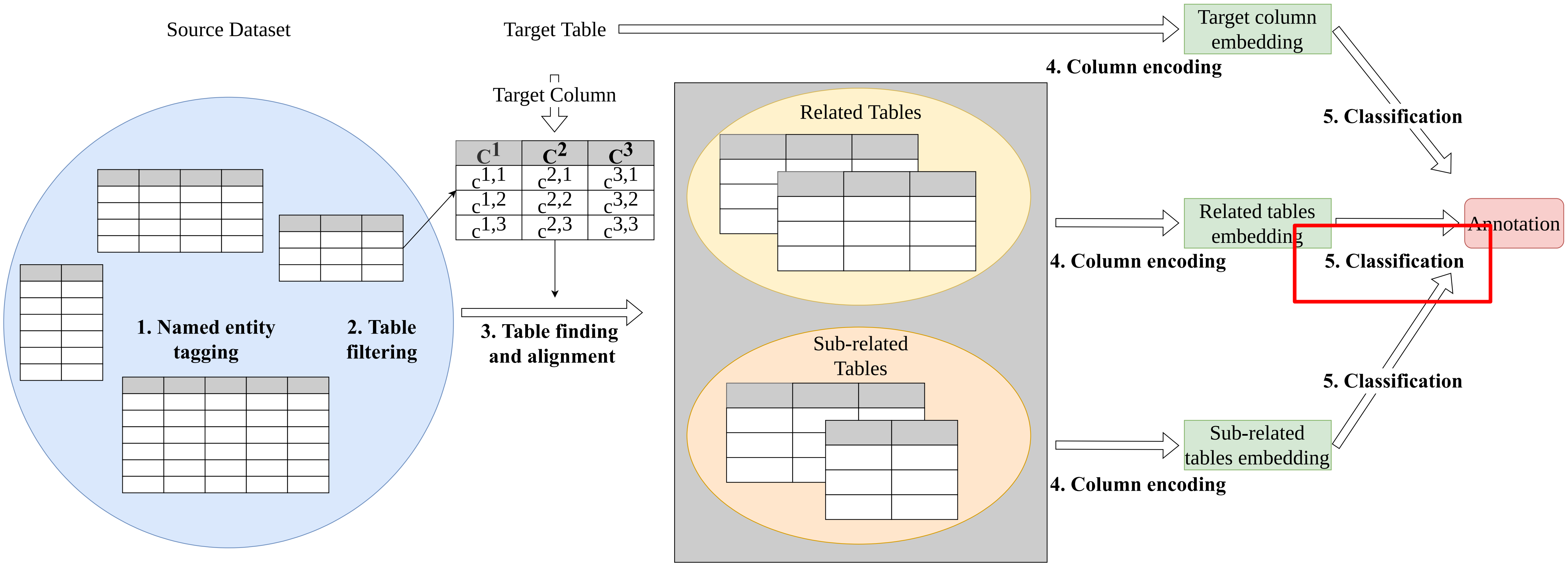


Methodology

- The target column is encoded with BERT solely.
- The aligned columns in related tables and sub-related tables are encoded separately with BERT.
- The tokens are allocated fairly to each related table (or sub-related table).



Methodology



$$a_i^t = \alpha * \hat{v}_i^t + \beta * \hat{r}_i^t + \gamma * \hat{x}_i^t$$

Methodology

- The embeddings of the target column, related tables, and sub-related tables are passed to three corresponding classification modules.
- Each classification module contains two layers: dropout and linear layers.
- The generated output embeddings are combined with learnable weights:

$$a_i^t = \alpha * \hat{v}_i^t + \beta * \hat{r}_i^t + \gamma * \hat{x}_i^t$$

- We use the cross-entropy loss as the loss function.

Experiments

- Datasets:

	WebTables	Semtab2019
# semantic types	78	275
# tables	32262	3045
# annotated columns	74141	7603
Avg. # rows	20.0	69.0
Avg. # columns	2.3	4.5
Avg. # annotated columns	2.3	2.5

- Metrics:
  - Support-weighted F1: weighted support of per type F1 scores
  - Macro average F1: average of per type F1 scores (emphasize on long-tail types)

Experiments

- RECA outperforms all the state-of-the-arts in terms of the F1 scores.

	Semtab2019 dataset		WebTables dataset	
Model names	Support-weighted F1	Macro average F1	Support-weighted F1	Macro average F1
Sherlock [15]	0.646 ± 0.006	0.440 ± 0.009	0.844 ± 0.001	0.670 ± 0.010
TaBERT [35]	0.768 ± 0.011	0.413 ± 0.019	0.896 ± 0.005	0.650 ± 0.011
TABBIE [16]	0.799 ± 0.013	0.607 ± 0.011	0.929 ± 0.003	0.734 ± 0.019
DODUO [30]	0.820 ± 0.009	0.630 ± 0.015	0.928 ± 0.001	0.742 ± 0.012
RECA	**0.853** ± 0.005	**0.674** ± 0.007	**0.937** ± 0.002	**0.783** ± 0.014

Experiments

- We conducted ablation study on RECA:
  - RECA target only: only encode the target column
  - RECA w/o re: encode both target column and aligned columns in sub-related tables
  - RECA w/o sub: encode both target column and aligned columns in related tables
- Performance drops on macro average F1 scores are greater than that on support-weighted F1 scores – incorporating inter-table context can improve the annotation quality on less-populated semantic types.

	Semtab2019 dataset		WebTables dataset	
Model names	Support-weighted F1	Macro average F1	Support-weighted F1	Macro average F1
RECA *target only*	0.808 ± 0.017	0.586 ± 0.039	0.911 ± 0.001	0.688 ± 0.014
RECA *w/o re*	0.836 ± 0.012	0.641 ± 0.037	0.927 ± 0.001	0.748 ± 0.024
RECA *w/o sub*	0.848 ± 0.009	0.650 ± 0.019	0.936 ± 0.002	0.774 ± 0.011
RECA	**0.853** ± 0.005	**0.674** ± 0.007	**0.937** ± 0.002	**0.783** ± 0.014

Experiments

- RECA is efficient in utilizing the learning data and the input data.

Learning data utilization

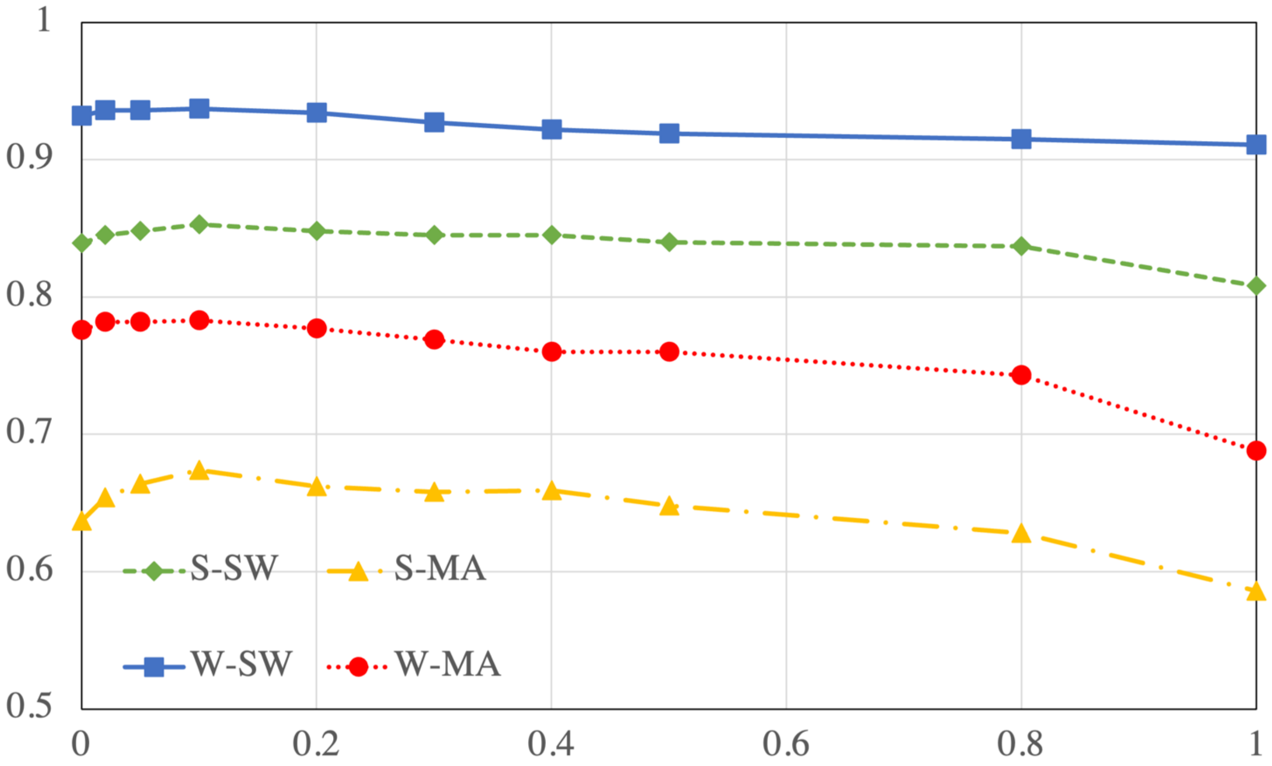
Datasets	[%]	Support-weighted F1	Macro average F1
Semtab2019	25	0.697 ± 0.041	0.442 ± 0.074
Semtab2019	50	0.792 ± 0.020	0.566 ± 0.045
Semtab2019	75	0.820 ± 0.021	0.631 ± 0.047
Semtab2019	100	**0.853** ± 0.005	**0.674** ± 0.007
WebTables	25	0.909 ± 0.002	0.680 ± 0.008
WebTables	50	0.924 ± 0.004	0.738 ± 0.019
WebTables	75	0.930 ± 0.002	0.772 ± 0.013
WebTables	100	**0.937** ± 0.002	**0.783** ± 0.014

Input data utilization

Datasets	Max	Support-weighted F1	Macro average F1
Semtab2019	8	0.540 ± 0.009	0.319 ± 0.010
Semtab2019	16	0.654 ± 0.013	0.436 ± 0.006
Semtab2019	32	0.728 ± 0.010	0.507 ± 0.020
Semtab2019	128	0.816 ± 0.017	0.620 ± 0.033
Semtab2019	256	0.851 ± 0.011	0.662 ± 0.024
Semtab2019	512	**0.853** ± 0.005	**0.674** ± 0.007
WebTables	8	0.907 ± 0.004	0.737 ± 0.011
WebTables	16	0.923 ± 0.002	0.762 ± 0.011
WebTables	32	0.931 ± 0.002	0.780 ± 0.010
WebTables	128	**0.937** ± 0.002	**0.783** ± 0.014
WebTables	256	0.936 ± 0.003	**0.783** ± 0.020
WebTables	512	0.936 ± 0.001	0.780 ± 0.011

Experiments

- RECA achieves stable performance when the Jaccard threshold is in the range of [0, 0.3].



- S-SW and S-MA stand for the support-weighted and macro average F1 scores on the Semtab2019 dataset; W-SW and W-MA stand for the support-weighted and macro average F1 scores on the WebTables dataset.

Summary

- We propose RECA for column semantic type annotation. RECA extracts and leverages inter-table context to enhance the annotation quality of the target column.
- We define a novel named entity schema for RECA to efficiently align related and sub-related tables, which resolves the difficulty of incorporating inter-table context.
- We conduct extensive experiments on two real-world web table datasets to show that RECA outperforms all the state-of-the-art methods. The result demonstrates the effectiveness of utilizing the inter-table context to annotate column semantic types accurately.
- We show that RECA is data efficient and learning efficient, since it requires shorter input token sequences and fewer training data to achieve high annotation performance.

Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Integration: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- Data Organization: Are Large Language Models a Good Replacement of Taxonomies?
- Future Vision and Opportunities

Overview

- **Are Large Language Models a Good Replacement of Taxonomies? (VLDB 2024)**
- Taxonomies provide a structured way to organize and categorize knowledge, which is indeed a kind of ``knowledge about knowledge'' (meta-knowledge).
- Typically, nodes in taxonomies follow a tree-like structure and the relationships between nodes are depicted as hypernymy (Is-A) links (e.g., HKUST is a type of University).
- Recently, we have witnessed the rapid advancements of large language models (LLMs) such as GPTs and Llamas. These LLMs have demonstrated impressive abilities in internalizing knowledge
- Can LLMs internalize the taxonomy structures?

[4] Andreas, "Taxonomy: Tracing Its Greek Roots to Modern Biological Classification - U speak Greek," U speak Greek, Dec. 25, 2023. https://uspeakgreek.com/science/biology/taxonomy-tracing-its-greek-roots-to-modern-biological-classification/ (accessed Aug. 18, 2024).
[6] Y. Sun, et al., "Are Large Language Models a Good Replacement of Taxonomies?," *Proceedings of the VLDB Endowment*, vol. 17, no. 11, pp. 2919–2932, Aug. 2024, doi:https://doi.org/10.14778/3681954.3681973.

Background

- Why this study is important?
  - If internalizing taxonomy data in LLMs is feasible, we can save a large amount of labor work for the construction and maintenance of taxonomies, which is a core asset for data organization.
  - If internalizing taxonomy data in LLMs is feasible, we may witness a change in the data management paradigm, with much of the explicitly stored data (such as tree structure in taxonomies) potentially transformed or partially transformed to exist in an implicit form of model internalized knowledge (neural-symbolic form).

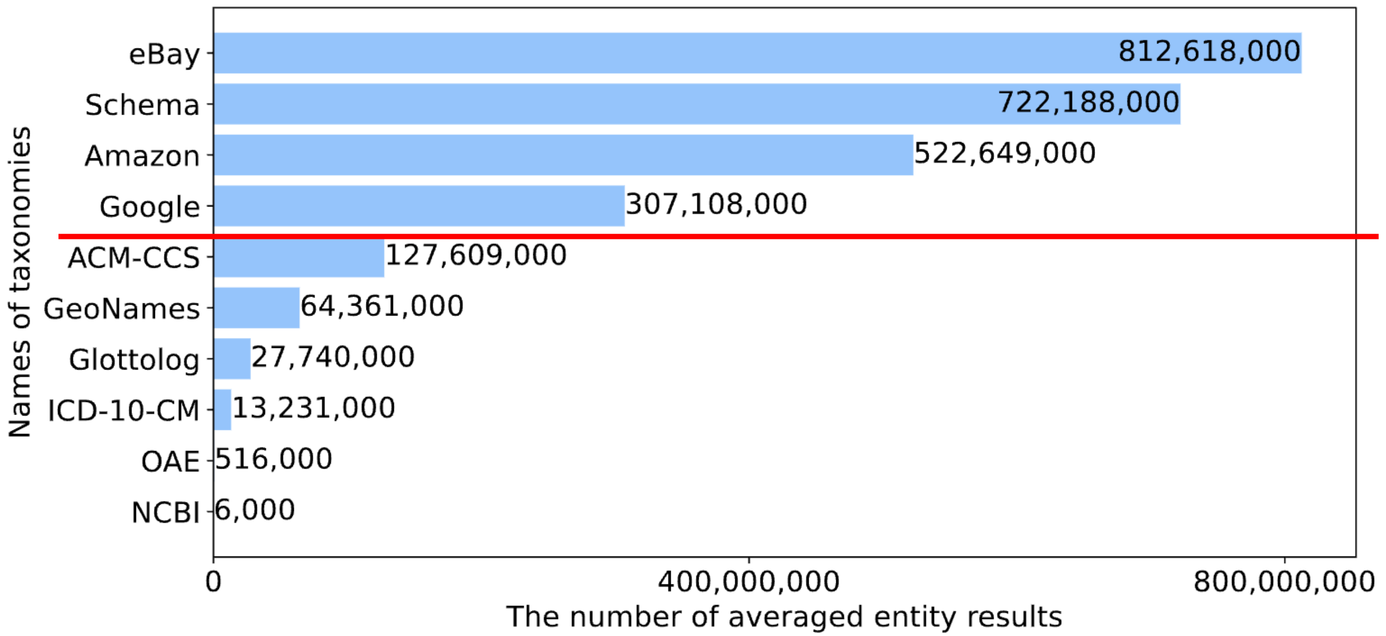
Background

Common to Specialized Domains

Domain	Taxonomy	Root to Leaf Levels						
Shopping	(a) eBay	Books, Movies & Music	Music	Music CDs				
General	(b) Schema	Thing	Intangible	Enumeration	StatusEnumeration	PaymentStatusType	PaymentComplete	
Shopping	(c) Amazon	Electronics	Computers & Accessories	Networking Products	Network Adapters	Bluetooth Network Adapters		
Shopping	(d) Google	Home & Garden	Kitchen & Dining	Cookware & Bakeware	Cookware Accessories	Pot & Pan Lids		
Computer Science	(e) ACM-CCS	Computing methodologies	Machine learning	Machine learning algorithms	Ensemble methods	Boosting		
Geography	(f) GeoNames	Stream, lake, ...	tidal creek (s)					
Language	(g) Glottolog	Sino-Tibetan	Sinitic	Classical-Middle-Modern-Sinitic	Middle-Modern-Sinitic	Hakka-Chinese	Hailu	
Health	(h) ICD-10-CM	Certain infectious & parasitic diseases	Intestinal infectious diseases	Cholera	Cholera due to Vibrio cholerae 01, biovar cholerae			
Medical	(i) OAE	investigation result abnormal AE	electrolyte lab test abnormal AE	electrolyte blood test abnormal AE	blood phosphorus abnormal AE	blood phosphate abnormal AE		
Biology	(j) NCBI	Eukaryota	Streptophyta	Magnoliopsida	Lamiales	Scrophulariaceae	Verbascum	Verbascum chaixii

Data Collection

- Taxonomies: 10 taxonomies on 8 domains:
- Common taxonomies:
  - Shopping domain: eBay, Amazon, Google
  - General domain: Schema.org
- Specialized taxonomies:
  - CS domain: ACM-CCS
  - Geography domain: GeoNames
  - Language domain: Glottolog
  - Health domain: ICD-10-CM
  - Medical domain: OAE
  - Biology domain: NCBI



Question Templates

- Design of questions: adopt simple True/False question

Domains	Question Templates
Shopping	Are <child-type> products a type of <parent-type> products? answer with (Yes/No/I don't know)
General	Is <child-type> entity type a type of <parent-type> entity type? answer with (Yes/No/I don't know)
Computer Science	Is <child-type> computer science research concept a type of <parent-type> computer science research concept? answer with (Yes/No/I don't know)
Geography	Is <child-type> geographical concept a type of <parent-type> geographical concept? answer with (Yes/No/I don't know)
Language	Is <child-type> language a type of <parent-type> language? answer with (Yes/No/I don't know)
Health / Biology	Is <child-type> a type of <parent-type>? answer with (Yes/No/I don't know)
Medical	Is <child-type> Adverse Events concept a type of <parent-type> Adverse Events concept? answer with (Yes/No/I don't know)

Question Sets

- Generation of question set

	eBay	Amazon	Google	Schema	ACM-CCS	GeoNames	Glottolog	ICD-10-CM	OAE	NCBI
Level 1-root	176	438	258	34	138	492	500	222	638	344
Level 2-1	430	700	597	276	450	n/a	564	550	700	439
Level 3-2	n/a	748	653	394	567	n/a	584	690	670	636
Level 4-3	n/a	758	626	410	370	n/a	600	n/a	572	741
Level 5-4	n/a	n/a	n/a	320	n/a	n/a	732	n/a	n/a	766
Level 6-5	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	770
Total	606	2644	2134	1434	1525	492	2980	1462	2580	3696

LLMs

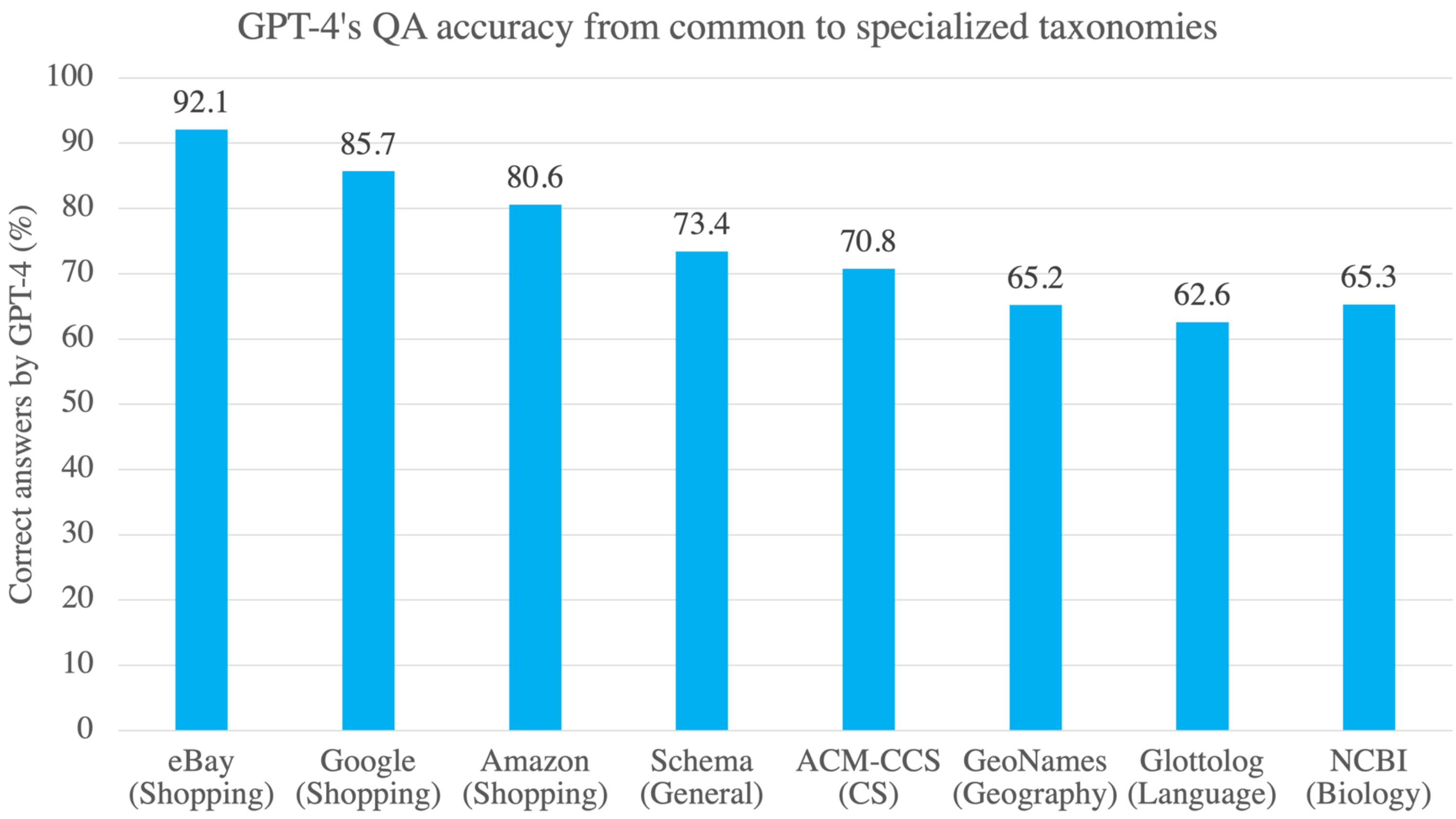
- LLMs considered:
  - Open-source:
    - Llama-2s: 7B, 13B, 70B
    - Llama-3s: 8B, 70B
    - Flan-T5s: 3B, 11B
    - Falcons: 7B, 40B
    - Vicunas: 7B, 13B, 33B
    - Mistrals: 7B, 8*7B
- Closed-source:
  - GPTs: GPT 3.5, GPT 4
  - Claude-3-Opus
- Fine-tuned:
  - LLMs4OL

Experiment Overview

- We experimented with 18 SOTA LLMs on different taxonomies from common to specialized domains and root-to-leaf levels to see whether the existing LLMs internalize the taxonomy knowledge (zero-shot annotation on taxonomy data).
- Specifically, we ask each LLM about whether a child entity is a type of its parent entity.
- Record the QA accuracy for each LLM on each level of different taxonomies.

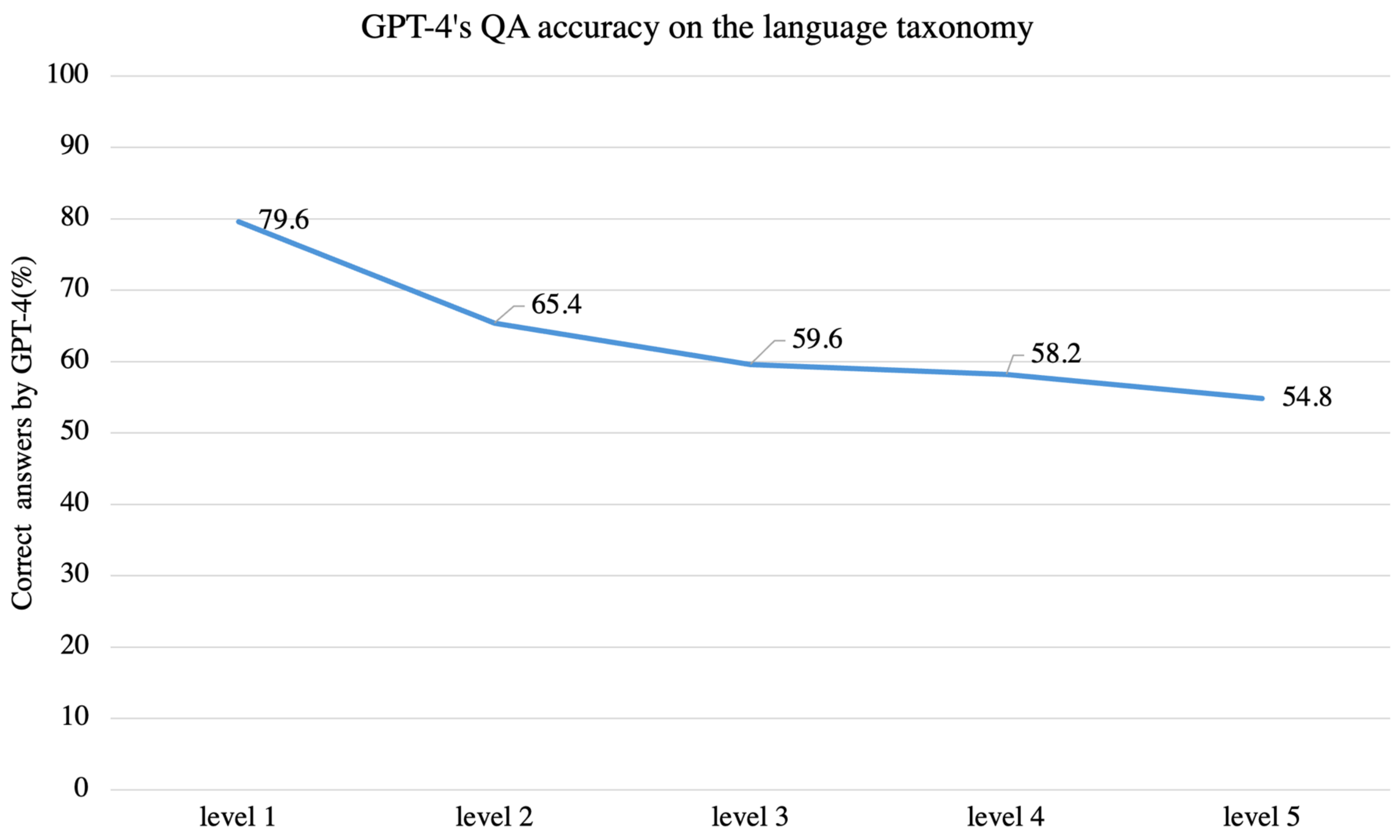
Experiments

- RQ1: How reliable are LLMs for discovering hierarchical structures in different taxonomies?
- The best LLMs perform well on common taxonomies (e.g., eBay, with over 90% accuracy); however, the performance downgrades on specialized taxonomies to around 60%.



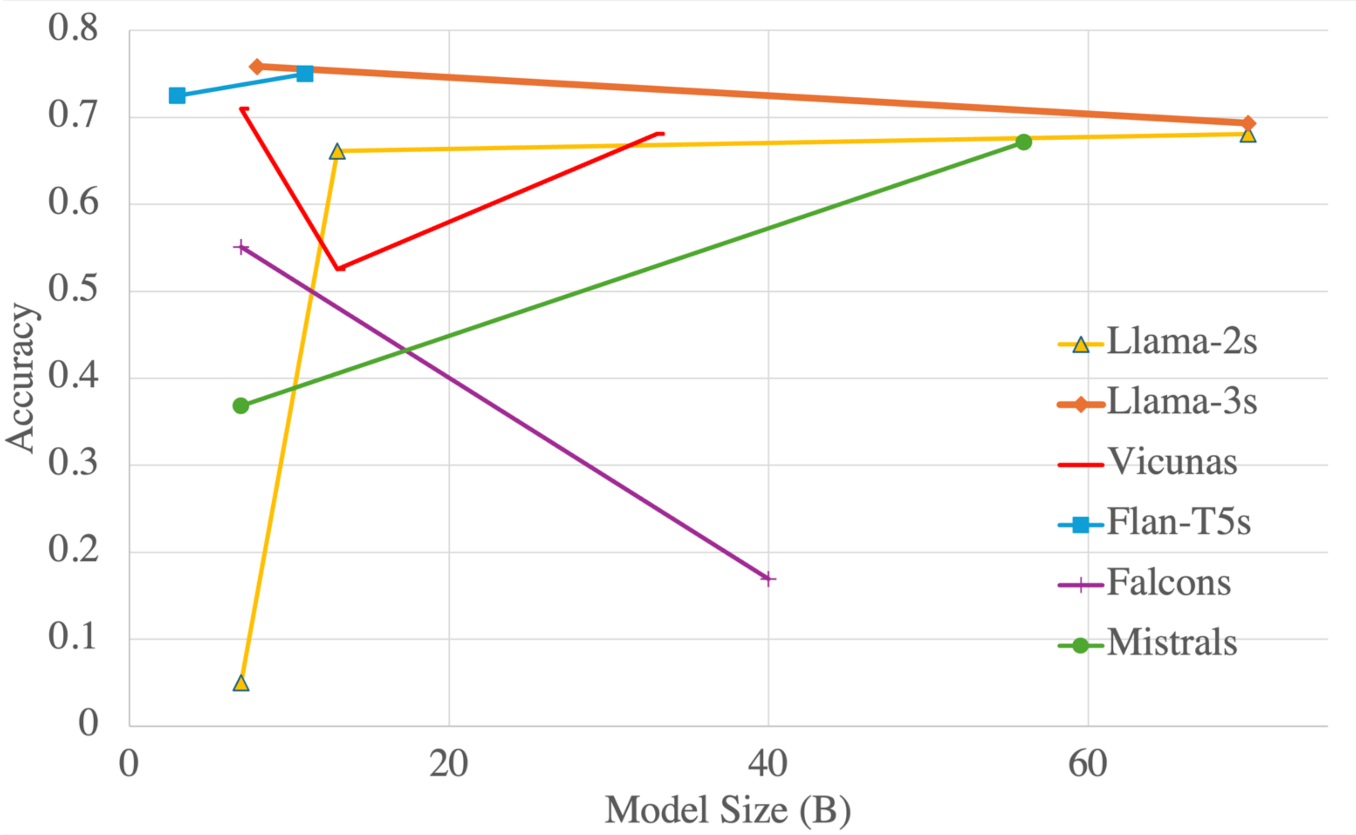
Experiments

- RQ2: Do LLMs perform equally well among different levels of taxonomies?
- LLMs roughly achieve progressively worse performance from root to leaf in most taxonomies (e.g., drops by relatively over 30% on Language taxonomy).



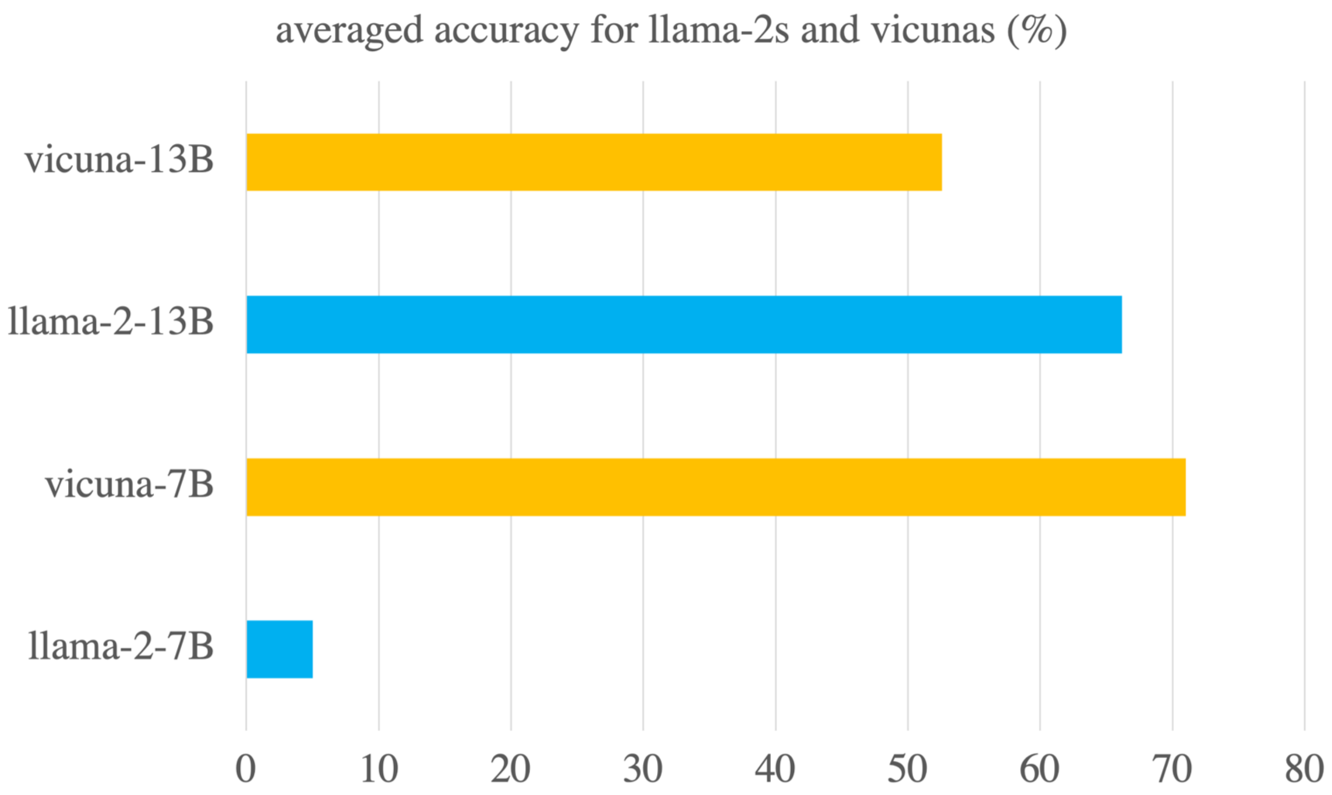
Experiments

- RQ3: Do normal methods that improve LLMs increase the accuracy?
  - RD3.1: Can we improve LLMs' performance by increasing the sizes of the LLMs used?
  - The increase in sizes of LLMs may not lead to an increase in performance.



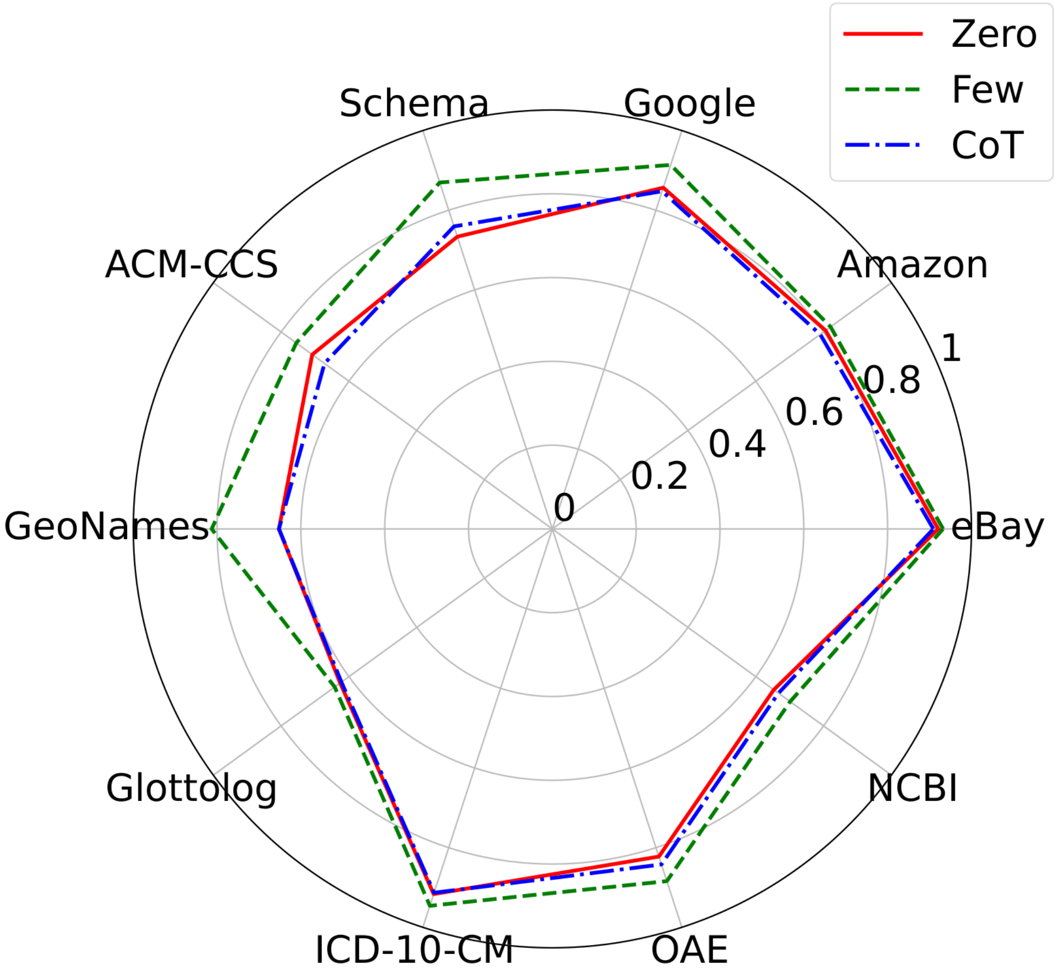
Experiments

- RQ3: Do normal methods that improve LLMs increase the accuracy?
  - RD3.2: Can we improve LLMs' performance by adopting domain-agnostic fine-tuning?
  - The adoption of domain-agnostic fine-tuning of LLMs may not lead to an increase in performance.



Experiments

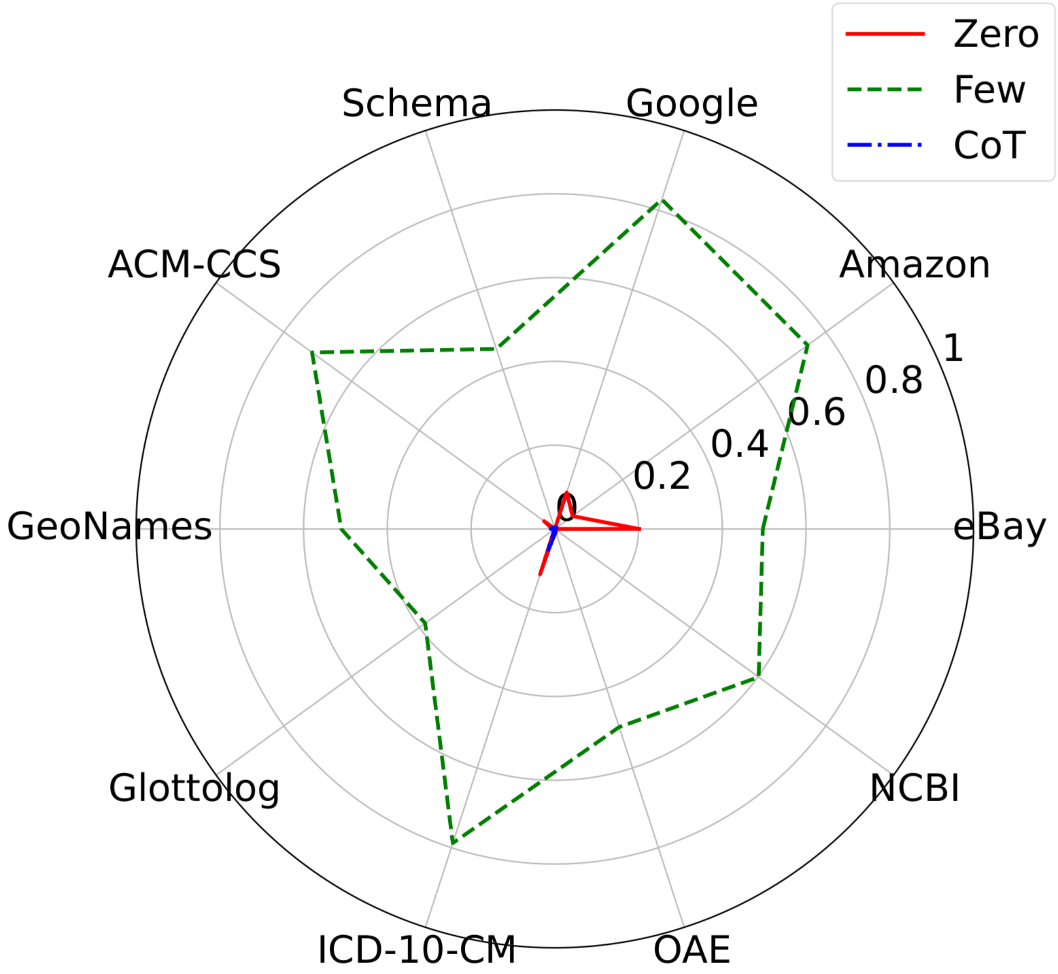
- RQ4: Do different prompting settings influence the performance?
- The performance changes of best LLMs brought by few-shot and Chain-of-Thoughts prompting settings are minimal. The main effect of prompting settings is to influence the miss rates instead of the accuracy of LLMs.



GPT-4 accuracy

Experiments

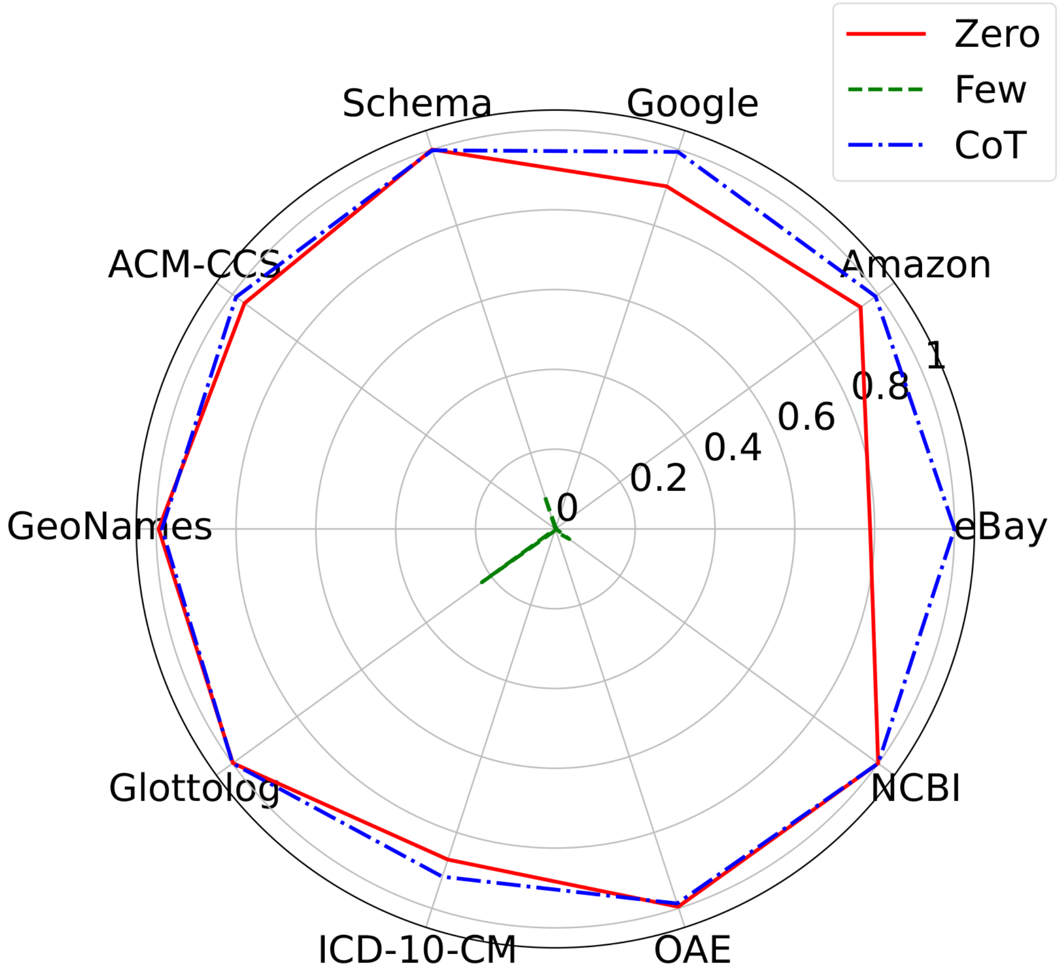
- RQ4: Do different prompting settings influence the performance?
- The performance changes of best LLMs brought by few-shot and Chain-of-Thoughts prompting settings are minimal. The main effect of prompting settings is to influence the miss rates instead of the accuracy of LLMs.



Llama-2-7B accuracy

Experiments

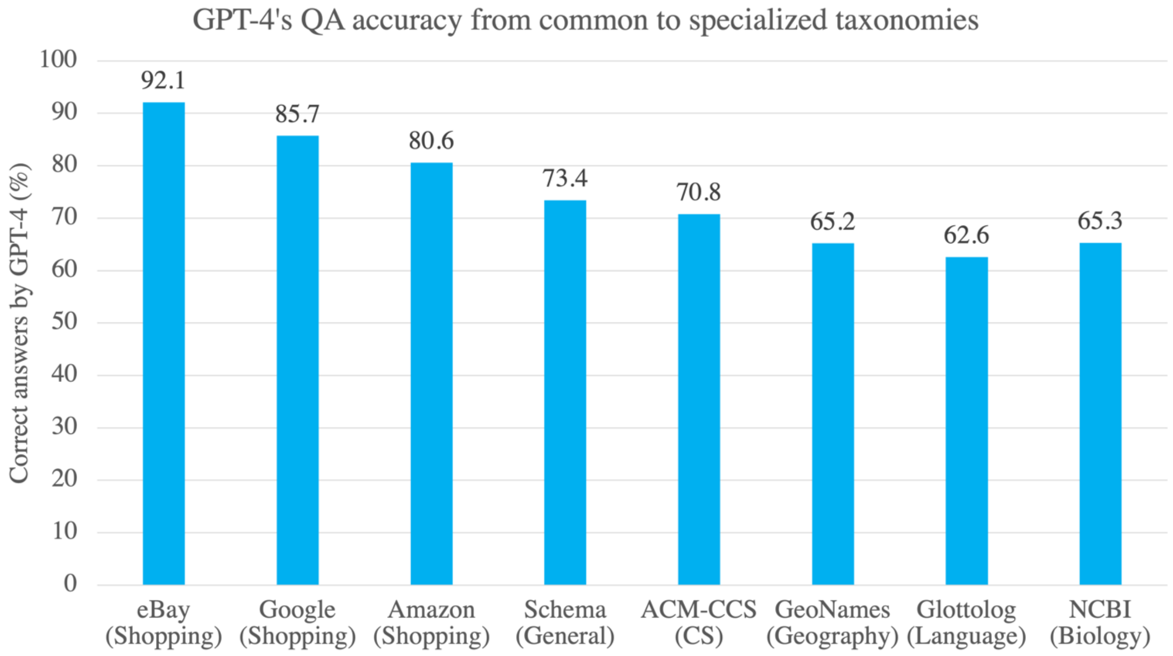
- RQ4: Do different prompting settings influence the performance?
- The performance changes of best LLMs brought by few-shot and Chain-of-Thoughts prompting settings are minimal. The main effect of prompting settings is to influence the miss rates instead of the accuracy of LLMs.

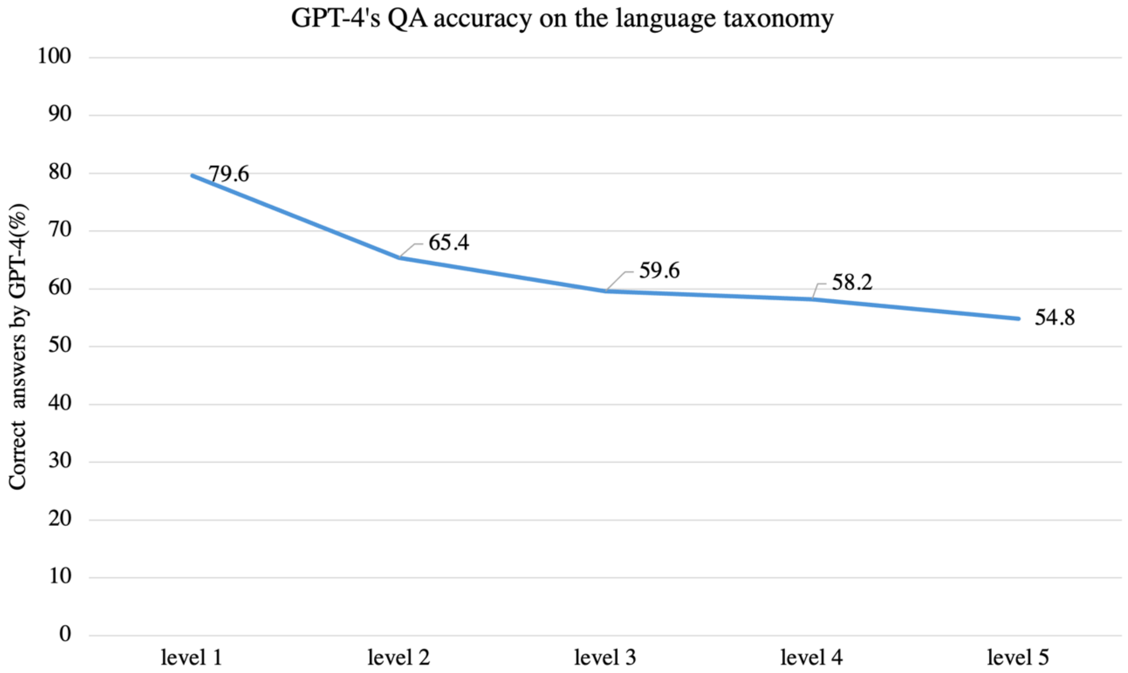


Llama-2-7B miss rate

Experiment Summary

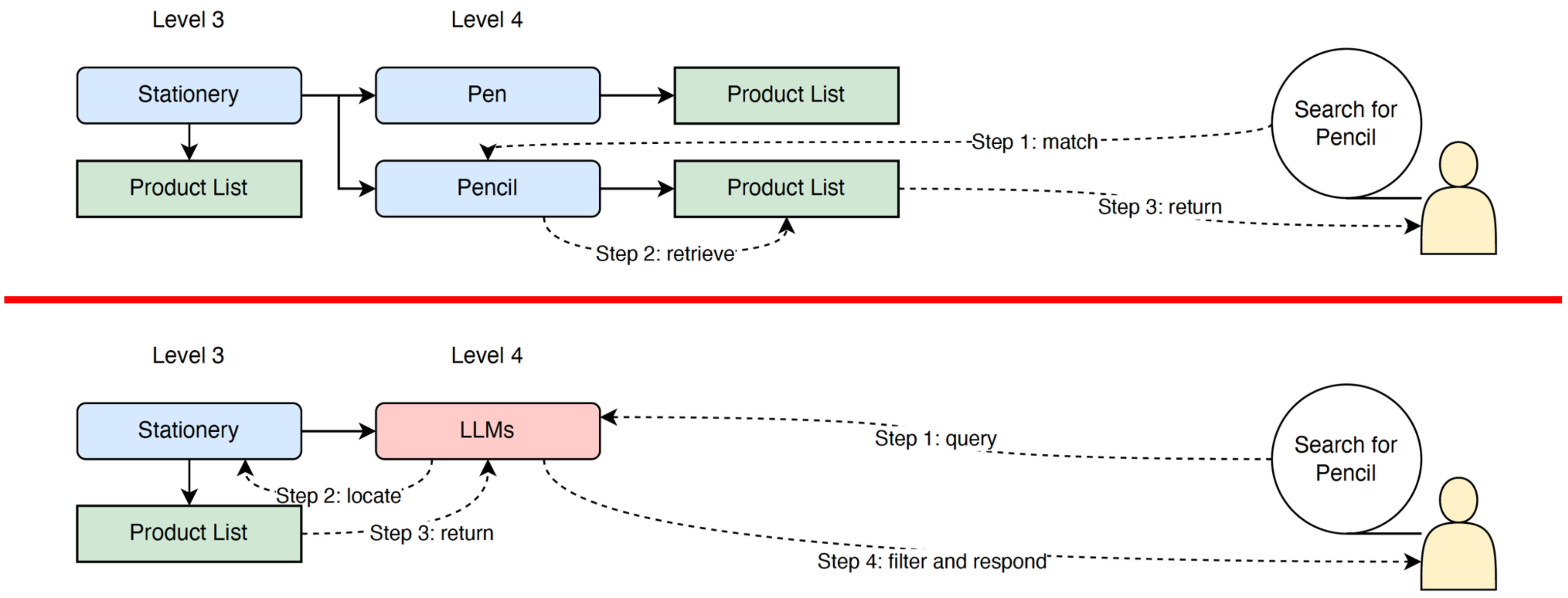
- Insights: LLMs are good at common domains and head (root-level) entities. But less reliable on specialized domains and tail (leaf-level) entities.
- Still cannot be zero-shot, all-rounded, and perfect on domain-specific tasks.





Case Study

- A concrete example of the integration of traditional taxonomy structure and LLMs:
  - Replaced the nodes in level 4 or lower of the Amazon Product Category with the Llama-2-70B model while preserving the nodes in root to level 3.



  - We report the precision and recall of the returned product list.

Case Study

- By performing the LLM replacement on Amazon Product Taxonomy, we reduce 59% of taxonomy construction and maintenance costs. ☺
  - (Number of nodes in each level of Amazon Product Taxonomy: 41-507-3910-13579-25777; cost saved: 25777/43814 = 59%)
- The precision and recall of the integrated solution are 0.713 and 0.792 respectively. ☺
- The cost can be further reduced if we replace more levels of taxonomy.
- The precision and recall are expected to be improved along with the advancements of LLMs.

Summary

- In this paper, we introduced TaxoGlimpse, a novel taxonomy hierarchical structure benchmark that comprehensively evaluates the data annotation performance of LLMs over different taxonomies from common to specialized domains, from root to leaf levels.
- Four highly concerned research questions were proposed and resolved and we provided valuable insights into future research.
- Our comprehensive evaluation shows that LLMs present unsatisfactory annotation performances at specialized taxonomies and for entities near the leaf levels. In response, we suggest future research directions to combine the LLMs with traditional taxonomies to create novel neural-symbolic taxonomies that have the best of both worlds.

Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Integration: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- Data Organization: Are Large Language Models a Good Replacement of Taxonomies?
- Future Vision and Opportunities

Research Opportunities: Advanced Designs in Column Type Annotation Support

- Properly design fine-tuning mechanisms that help the large-language-model-based/pre-trained-model-based approaches generalize well on new data lakes (requires research in training data selection and augmentation).

	Generalizability	Accuracy
human-in-the-loop-based	low, need training	high
pre-trained-model-based	medium, require finetuning data	high with domain-specific finetuning
large-language-model-based	high, only need few-shot examples	low, without domain-specific finetuning
large-language-model-based*	relatively low, require finetuning data	high, with domain-specific finetuning

* means finetuning

Bad Good

Research Opportunities: RAG and Data Curation

- We conduct a preliminary study that evaluates the performance of LLMs accessing different modalities and sources of data (Our CRAG benchmark paper, NeurIPS 2024)
- We identify that the existing LLM-based methods fail to provide correct responses when the annotations are fast-changing or require complex access to external databases (range query, set query, etc.).
- How to make database content more accessible to LLM and thus help QA solutions better in the RAG settings remains a challenge and an interesting topic to explore.

[7] Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen tau Yih, and Xin Luna Dong. 2024. CRAG – Comprehensive RAG Benchmark. arXiv preprint arXiv:2406.04744 (2024). https://arxiv.org/abs/2406.04744