When Data Quality Meets Language Models: Past, Status-quo, and Future.

Yushi Sun

Last updated 22/09/2024

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

About me ☺

- Yushi Sun (Steve)
- 4th and final year PhD student at HKUST.
- Supervised by Prof. Lei Chen.
- Research interest in data quality (data labeling and preparation), LLMs, and RAG.
- Fortunate to collaborate with experts in these fields: Prof. Nan Tang and Dr. Xin Luna Dong.

Outline

- Background
- LM4DQ
  - Past: Crowd-sourced / Human-in-the-loop
  - Status-quo: Pre-train+fine-tune LMs
  - Status-quo: Low-resource LMs
  - Future: Zero-shot LMs
- Future Vision and Opportunities
  - Preliminary study on DQ4LM
  - LM4DQ and DQ4LM

Background: DQ and LM

- Data quality defines fitness for the use of data [1]:
  - Accuracy
  - Completeness

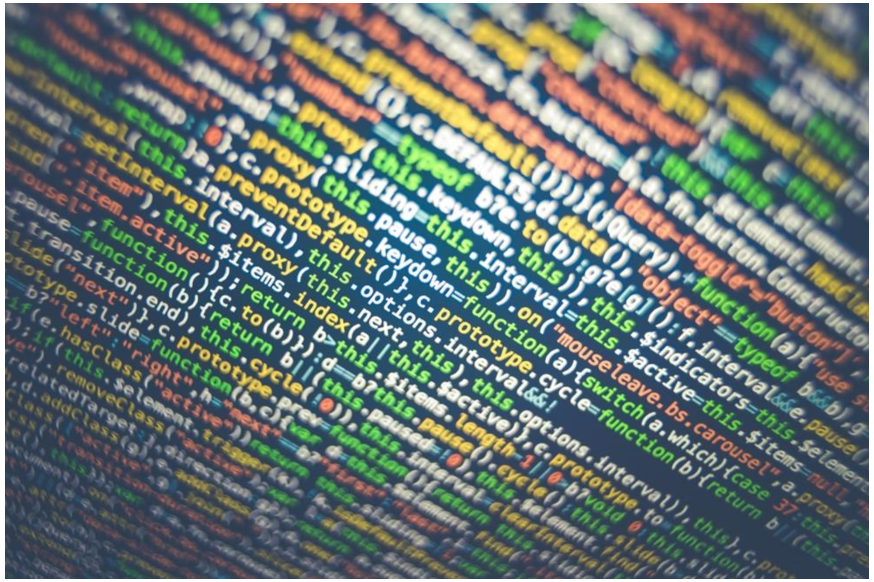
  High-quality and efficient data labeling / preparation
  - Consistency
  - Timeliness
  - …
- Language Models:
  - A language model is a probabilistic model of a natural language.
  - LMs predict or generate natural language text by capturing text patterns.
  - Good at processing textual data.

[1] S. Mohammed, "A Data Quality Glossary". Zenodo, Jan. 09, 2024. doi: 10.5281/zenodo.10474880.

Outline

- Background
- LM4DQ
    - Past: Crowd-sourced / Human-in-the-loop
    - Status-quo: Pre-train+fine-tune LMs
    - Status-quo: Low-resource LMs
    - Future: Zero-shot LMs
- Future Vision and Opportunities
    - Preliminary study on DQ4LM
    - LM4DQ and DQ4LM

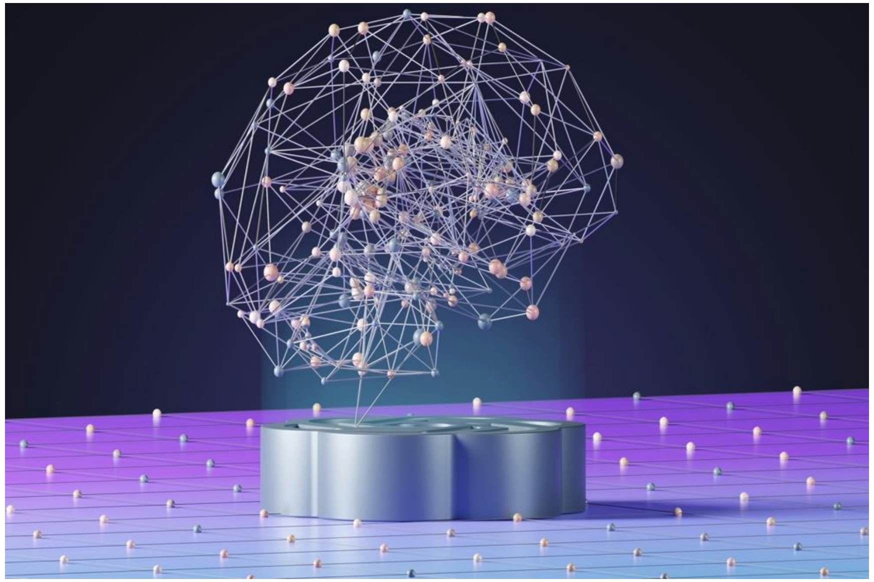
LMs-powered Data Quality



Data Quality (focus on data labeling)

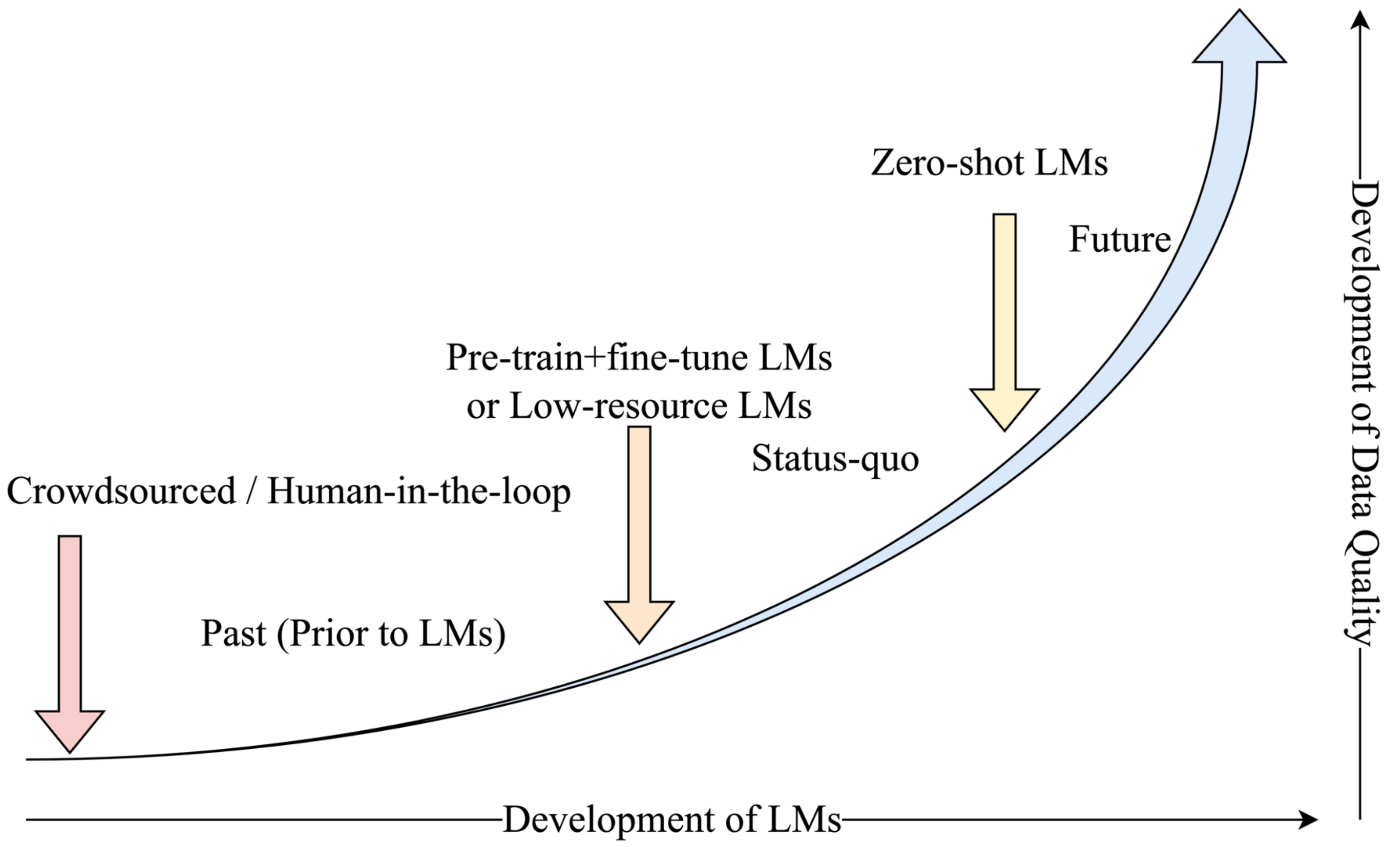
LM4DQ

Reduced labeling cost
Improved data labeling performance



Language Models (BERT, GPT, Llama, …)

Data Quality: Past, Status-quo, and Future

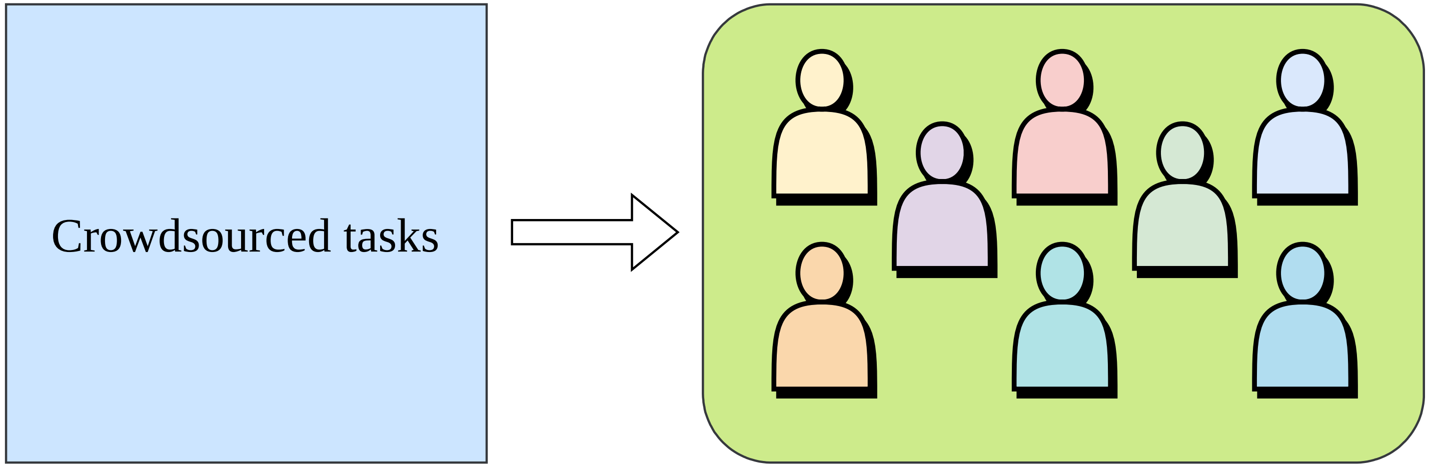


Outline

- Background
- LM4DQ
  - Past: Crowd-sourced / Human-in-the-loop
  - Status-quo: Pre-train+fine-tune LMs
  - Status-quo: Low-resource LMs
  - Future: Zero-shot LMs
- Future Vision and Opportunities
  - Preliminary study on DQ4LM
  - LM4DQ and DQ4LM

Crowd-sourced / Human-in-the-loop - overview

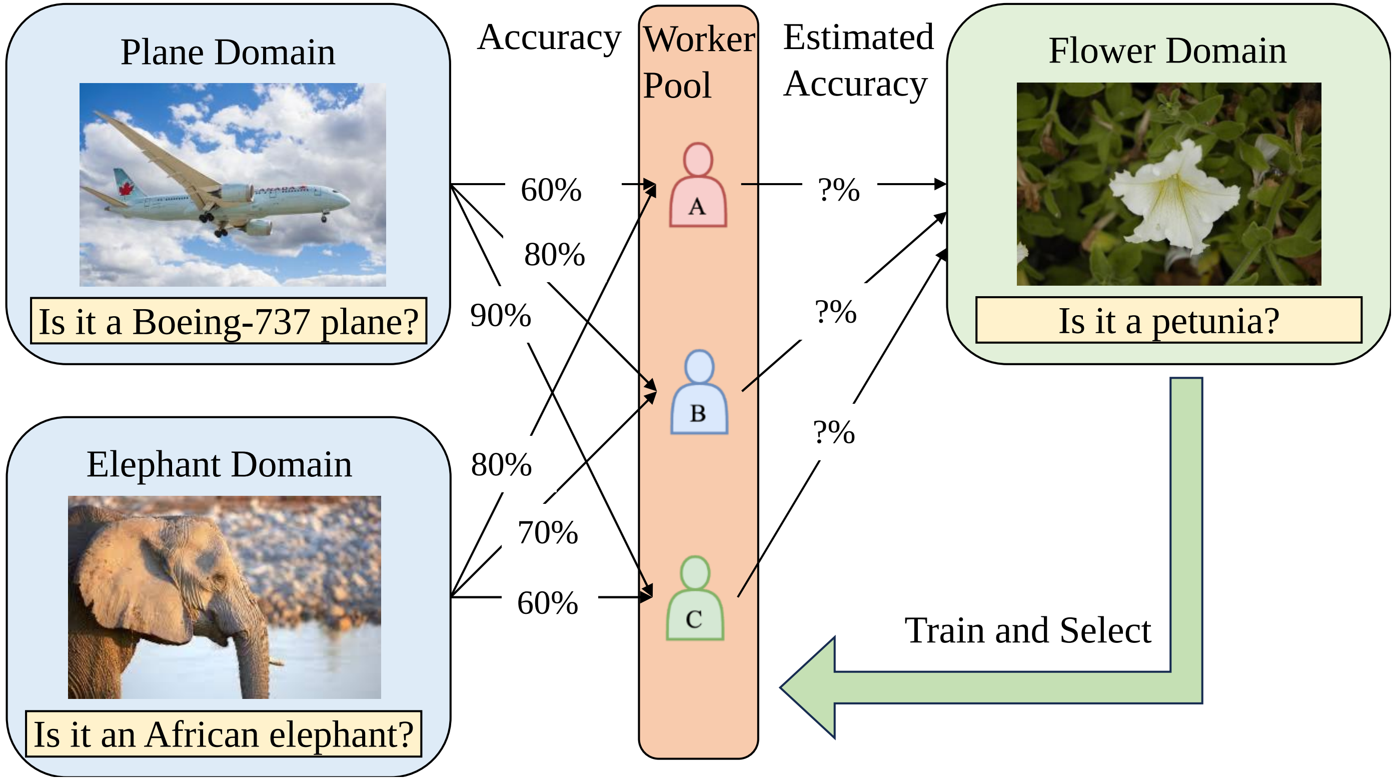
- **Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation (ICDE 2024)**
  - Crowdsourcing is preferable for obtaining high-quality data labeling for large-scale datasets.
  - Worker Selection is important in Crowdsourcing.
  - How to design an allocation scheme for golden questions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



[2] 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

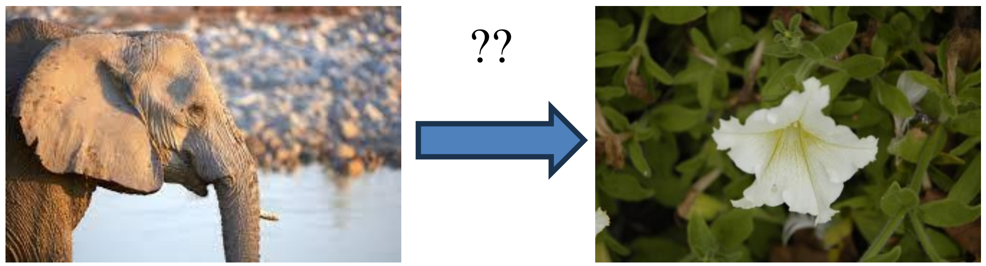
Crowd-sourced / Human-in-the-loop - background

- The **answering history of workers** (prior domain knowledge) can help select high-quality workers when **annotating a new domain** (target domain task).



Crowd-sourced / Human-in-the-loop - challenges

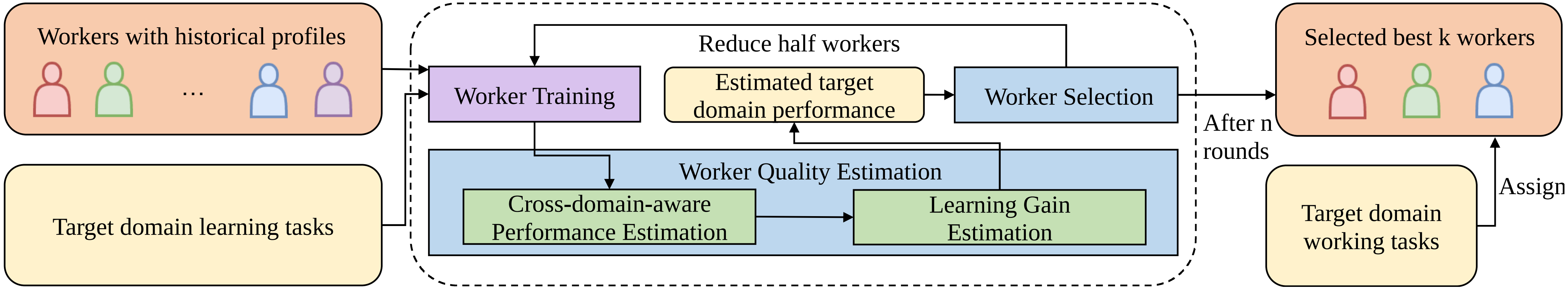
- Difficulty in accurately estimating the correlation between domains with a limited budget.
- Difficulty in estimating the workers' dynamic knowledge change during the question-answering worker training process.



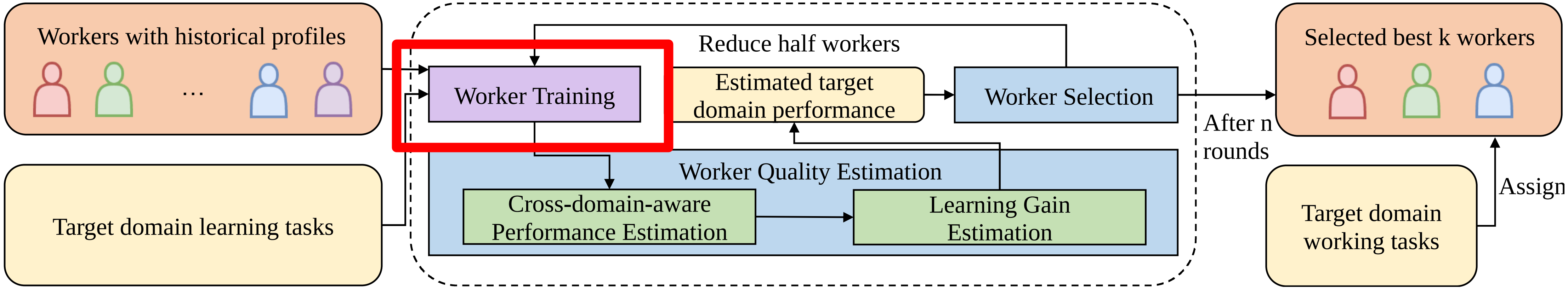
Crowd-sourced / Human-in-the-loop - definition

- Cross-domain-aware worker selection with training:
  - Given target domain tasks $T = \{T_l, T_w\}$, the total budget B, and worker pool W with each worker w_i's historical profile h_i.
  - Cross-domain-aware worker selection with training problem is to 1) assign no more than B tasks to |W| workers **for training** and 2) **select top k workers** with the highest possible annotation accuracy on working tasks T_w.

Crowd-sourced / Human-in-the-loop - methodology

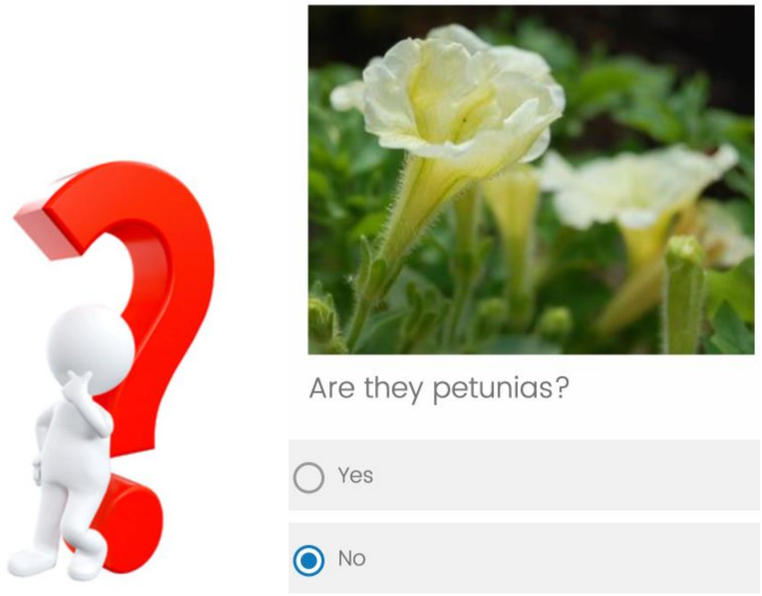


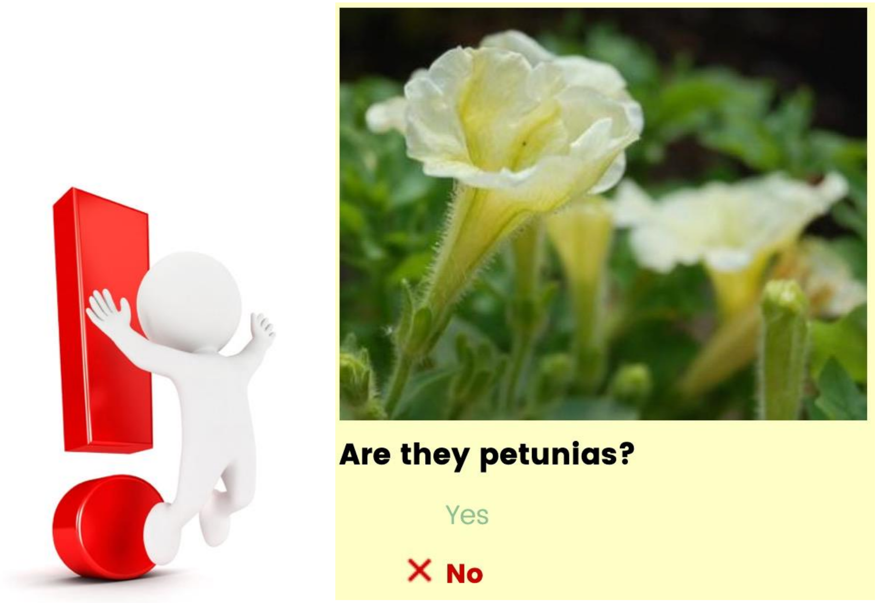
Crowd-sourced / Human-in-the-loop - methodology



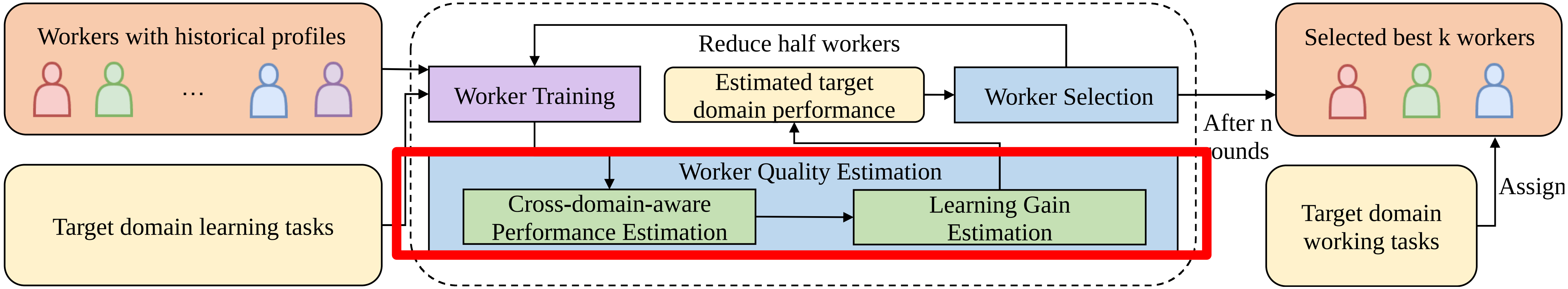
Crowd-sourced / Human-in-the-loop - methodology

- Worker training is treated as an **"Answer and learn"** process for workers.





Crowd-sourced / Human-in-the-loop - methodology

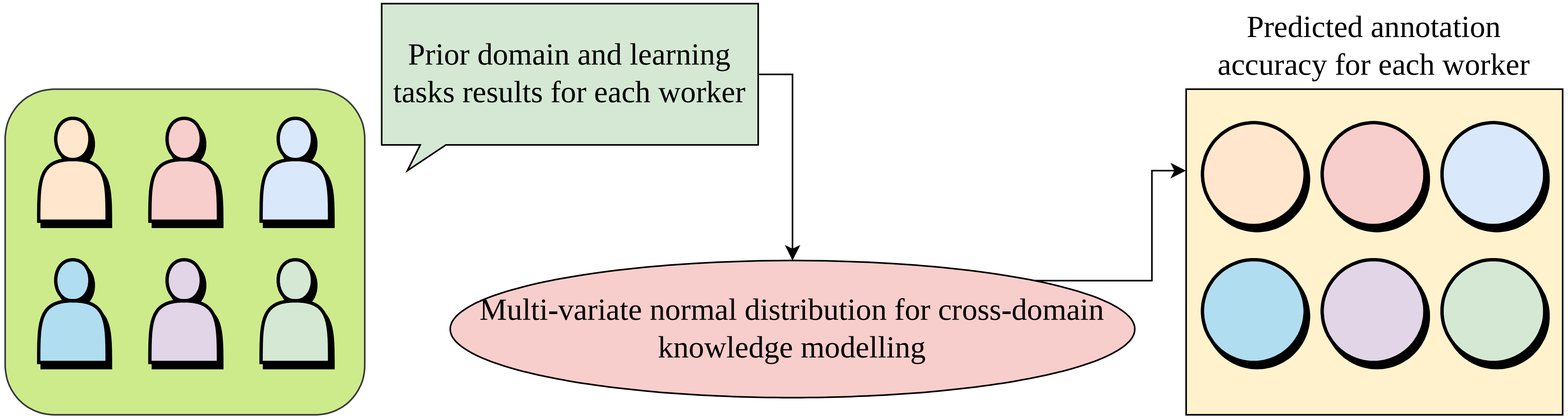


Crowd-sourced / Human-in-the-loop - methodology

- We consider two factors in estimating workers' quality:
- **Cross-domain correlation** – Cross-domain-aware Performance Estimation (CPE)
- **Worker learning gain** – Learning Gain Estimation (LGE)

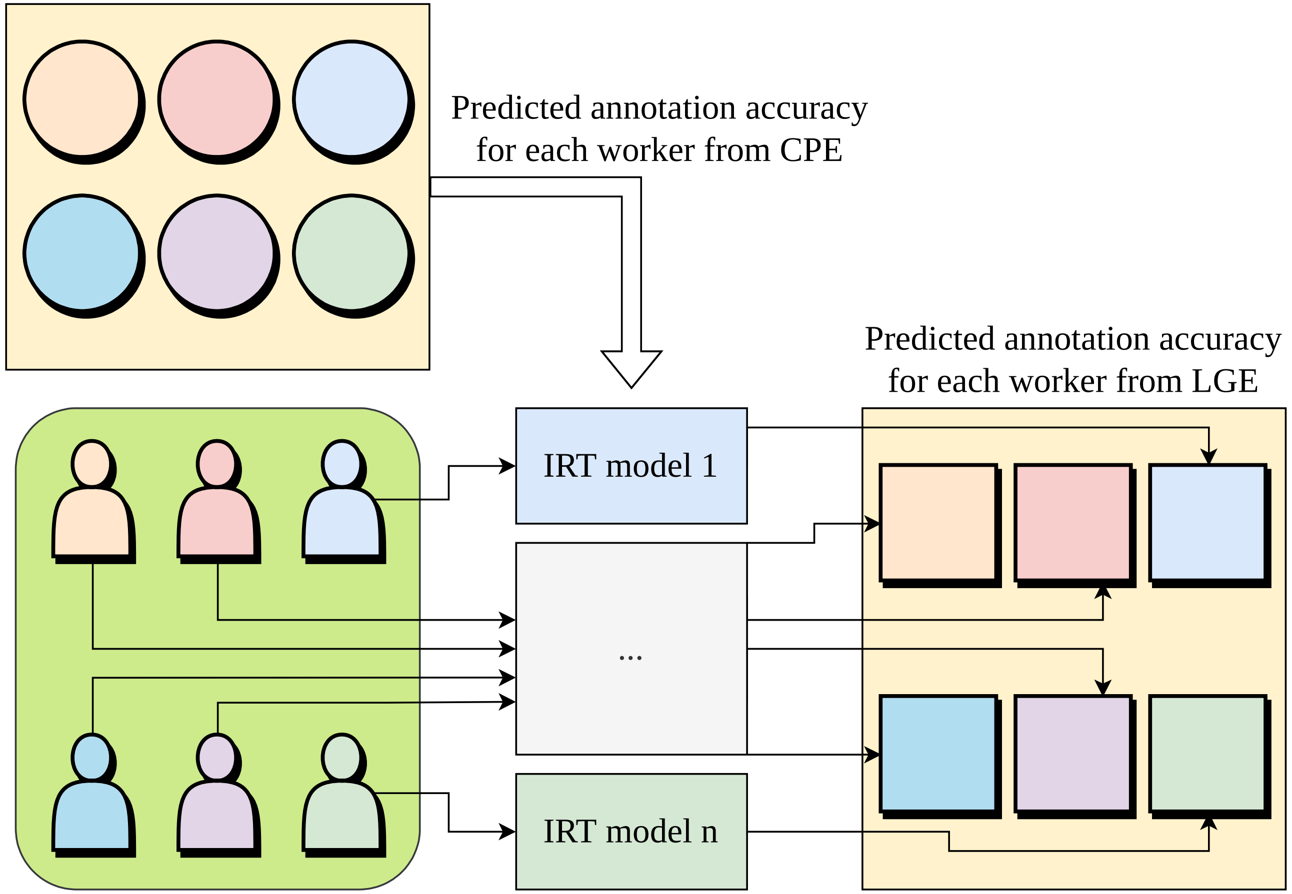
Crowd-sourced / Human-in-the-loop - methodology

- Model the **correlation** between workers' **prior knowledge and the target domain** knowledge as a **multivariate normal distribution**.
- Record **the correct and wrong number** of learning tasks for each worker.
- Update the distribution with **maximum likelihood estimation**.
- **Predict the annotation accuracy** of each worker.

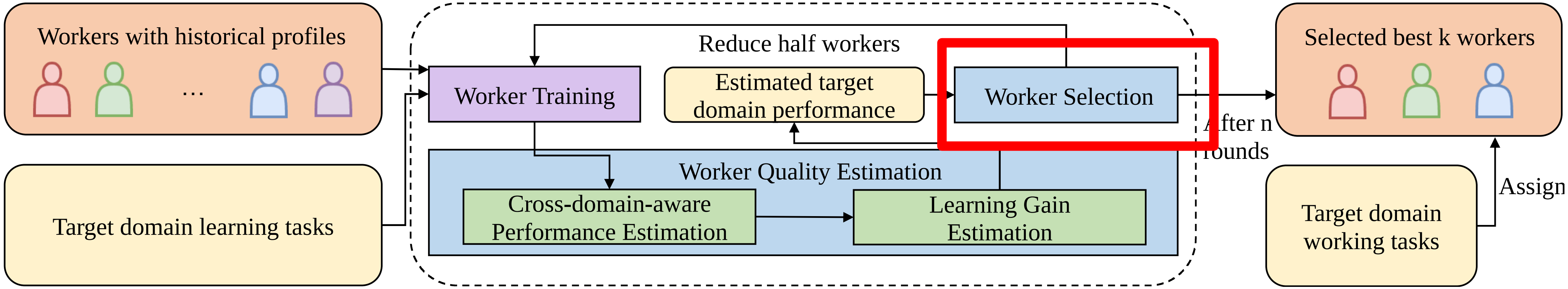


Crowd-sourced / Human-in-the-loop - methodology

- Adapt the **Item Response Theory (IRT)** model to estimate the learning gain.
    - Compute the IRT scores on the **prior domains**.
    - Compute the IRT scores on the **target domain learning tasks**.
    - Update the learning parameter α_i for each worker based on the **CPE scores and answering history**.
- Predict the estimated scores in the current round.

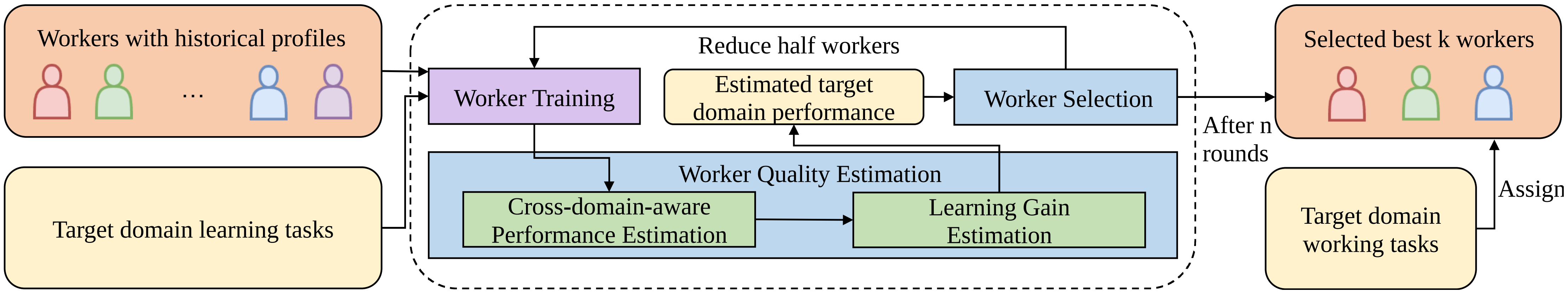


Crowd-sourced / Human-in-the-loop - methodology



Crowd-sourced / Human-in-the-loop - methodology

- Adapt the ME algorithm to select the **top half of the workers** in the current round.
- Error bound: $O(\sqrt{\frac{nk}{B}\ln\frac{1}{\delta_c}})$.



Crowd-sourced / Human-in-the-loop - 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

Crowd-sourced / Human-in-the-loop - takeaways

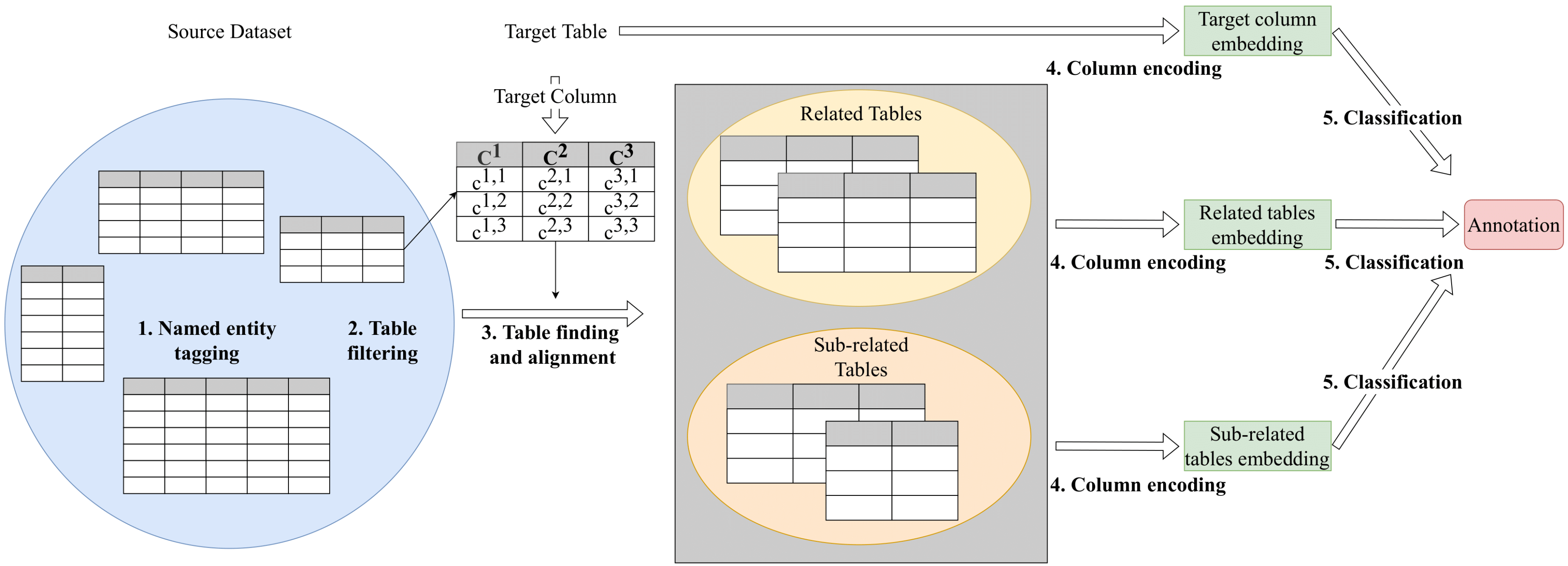
- Before the emergence of LM in data labeling, crowd-sourced / human-in-the-loop approaches were the main approaches that we can count on.
  - Pros:
    - Compared to black-box LM, easy debugging on the data labeling results (You can ask the crowd-workers about their choices).
    - Quality control and guarantee (You can monitor the results given by the crowd-workers and replace workers when the quality becomes low).
    - Accurate.
  - Cons:
    - Human labeling costs are high.
    - Human labeling is relatively slow.
  - Research Opportunities:
    - How to combine human labeling and LM-based labeling to reduce costs, improve speed, and guarantee quality.

Outline

- Background
- LM4DQ
  - Past: Crowd-sourced / Human-in-the-loop
  - Status-quo: Pre-train+fine-tune LMs
  - Status-quo: Low-resource LMs
  - Future: Zero-shot LMs
- Future Vision and Opportunities
  - Preliminary study on DQ4LM
  - LM4DQ and DQ4LM

Pre-train+fine-tune LMs - overview

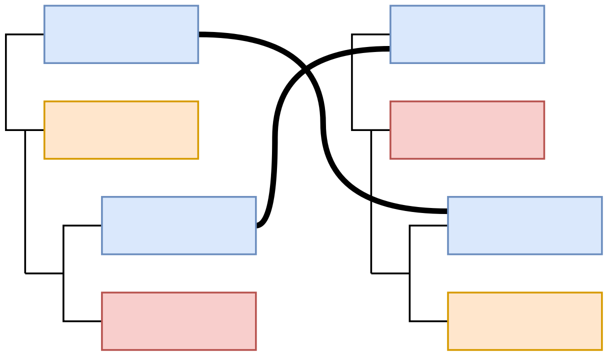
- **RECA: Related Tables Enhanced Column Semantic Type Annotation Framework (VLDB 2023)**
- Focus on enhancing tabular data labeling with inter-table context information.



[3] 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.

Pre-train+fine-tune LMs - background

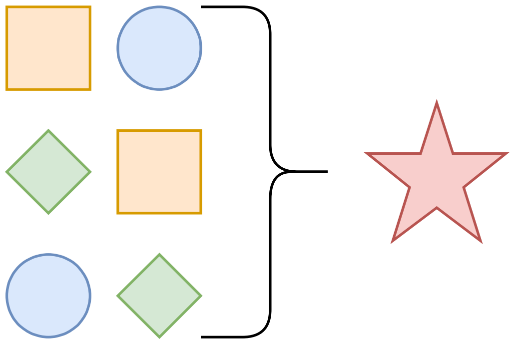
- Accurate column semantic type labeling is important for various applications:
  - schema matching, data cleaning, data integration, etc.



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

Pre-train+fine-tune LMs - challenges

- The utilization of inter-table context

?	?	?	?
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

Pre-train+fine-tune LMs - motivation

- 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

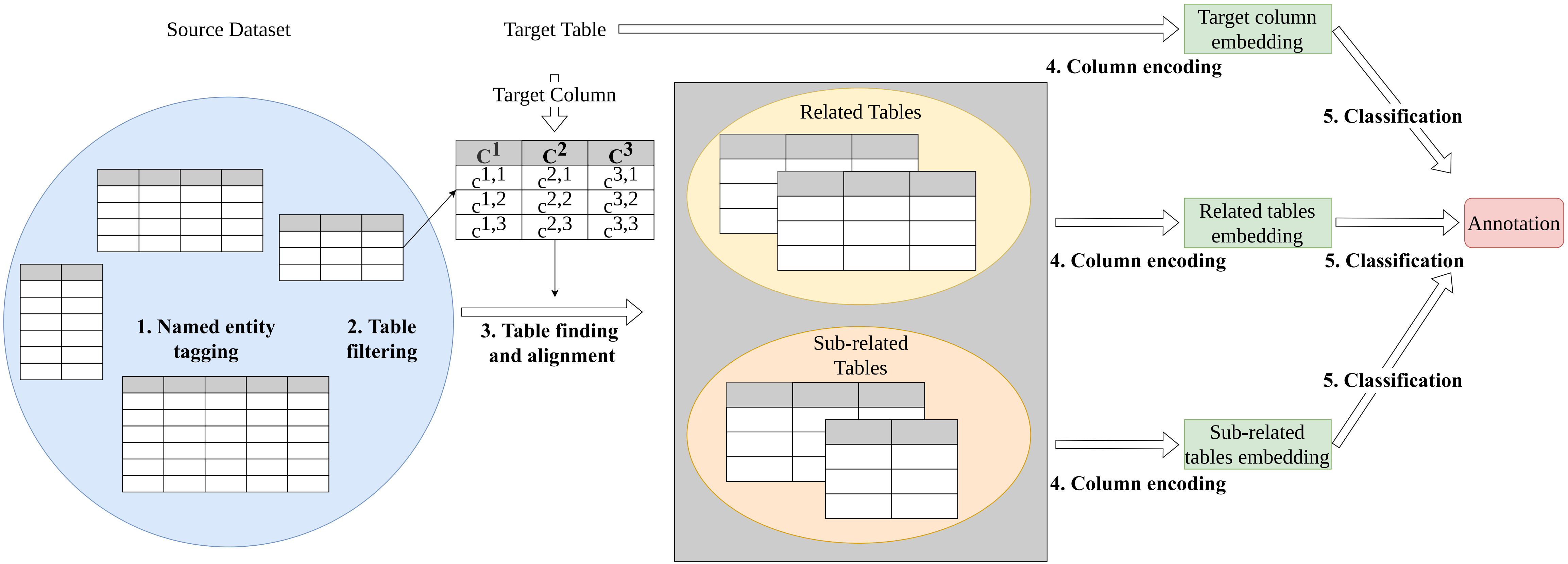
Pre-train+fine-tune LMs - definition

- Named Entity Schema: Named Entity Schema is the table schema generated based on the most frequent named entity type extracted from each column.
- 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.

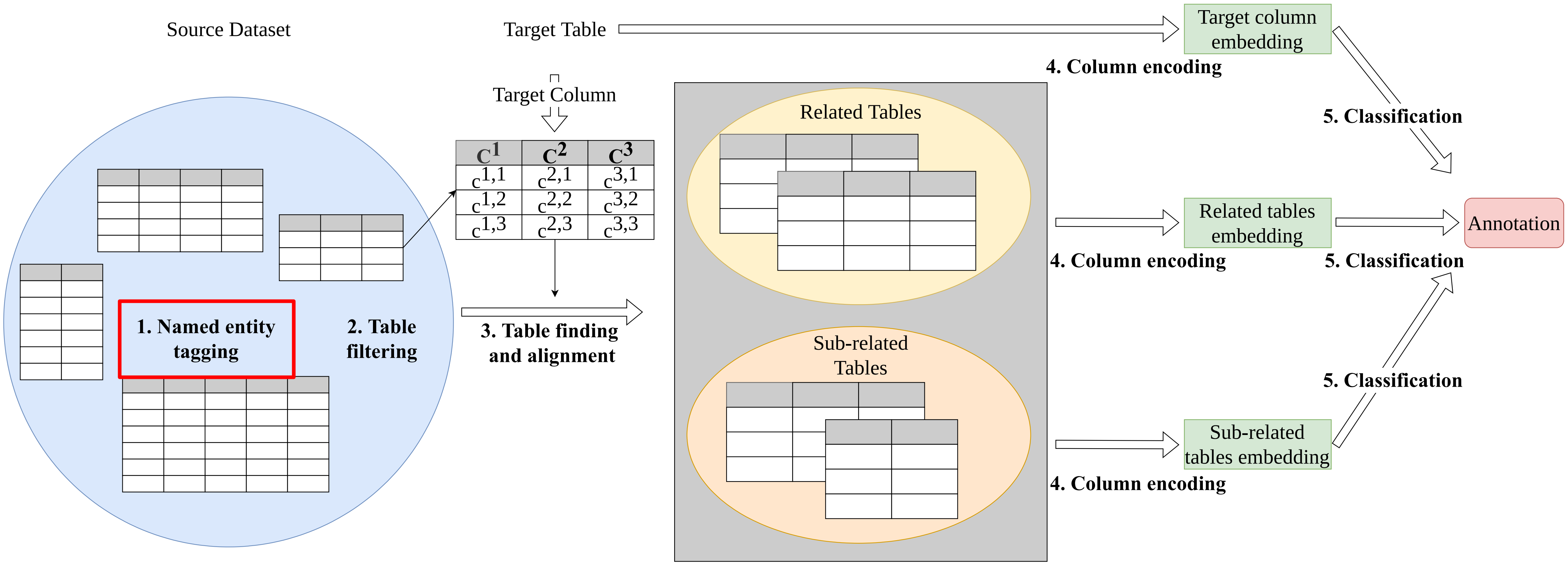
Pre-train+fine-tune LMs - definition

- (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.

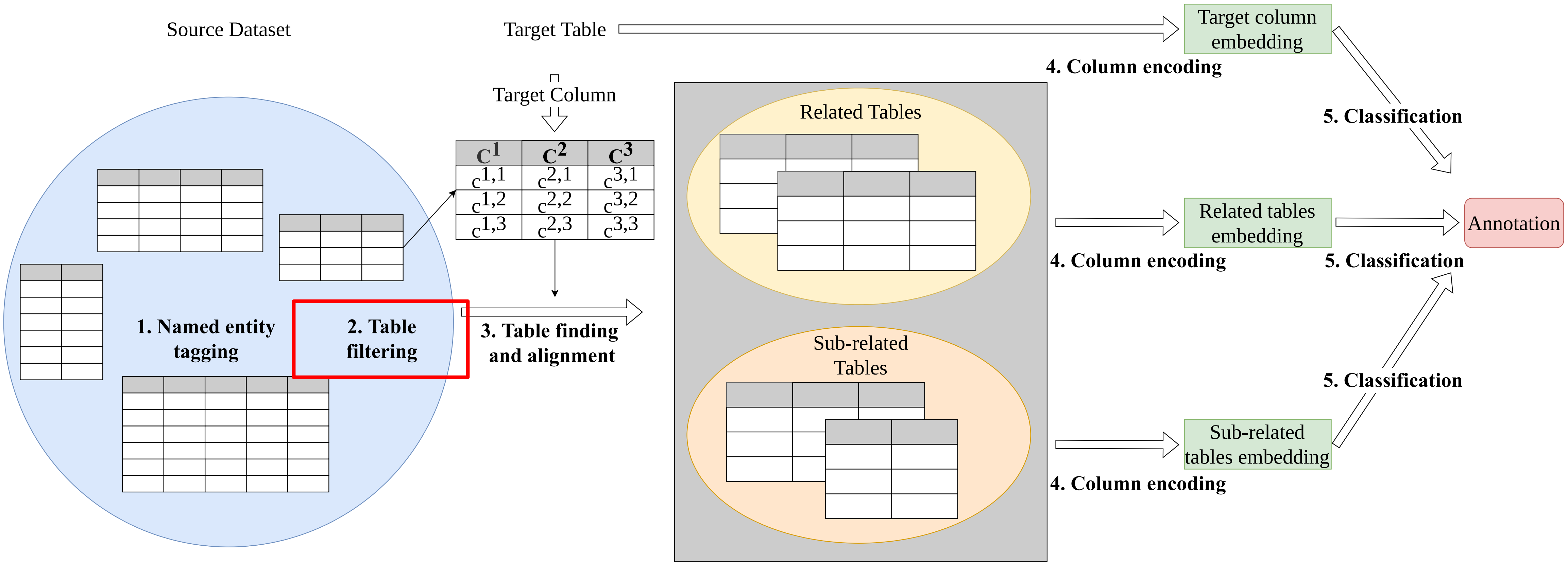
Pre-train+fine-tune LMs - methodology



Pre-train+fine-tune LMs - methodology

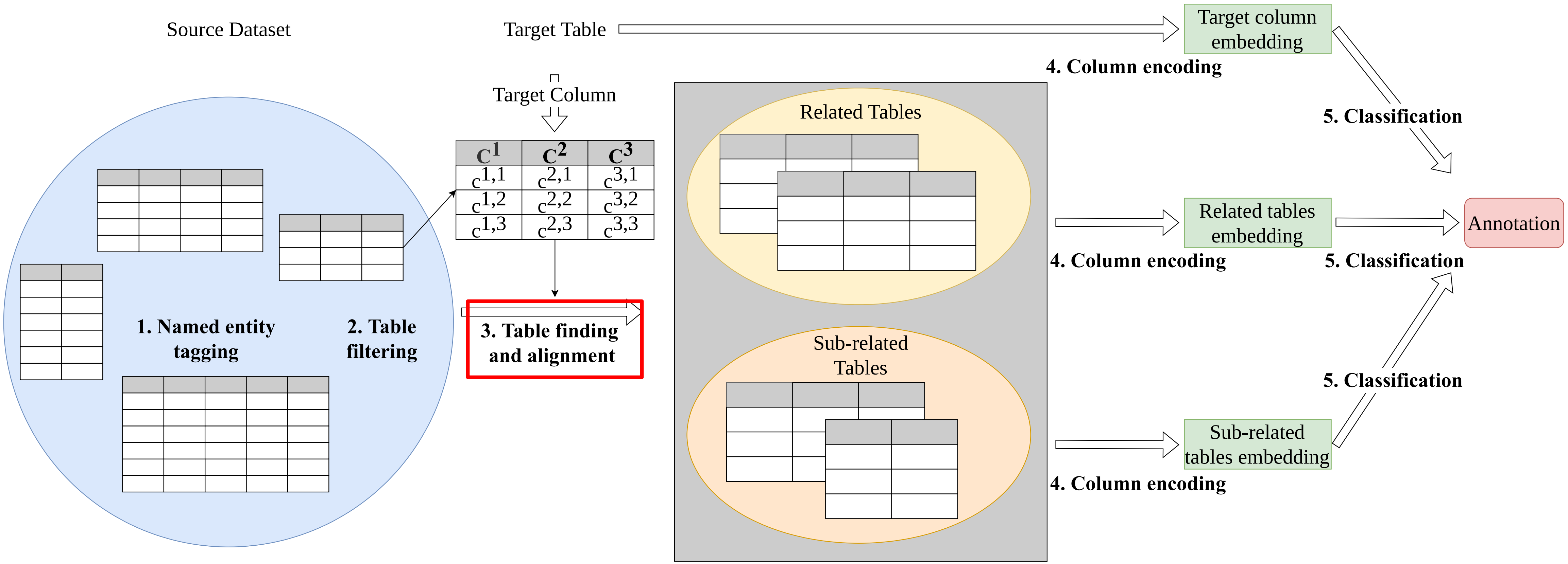


Pre-train+fine-tune LMs - methodology



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

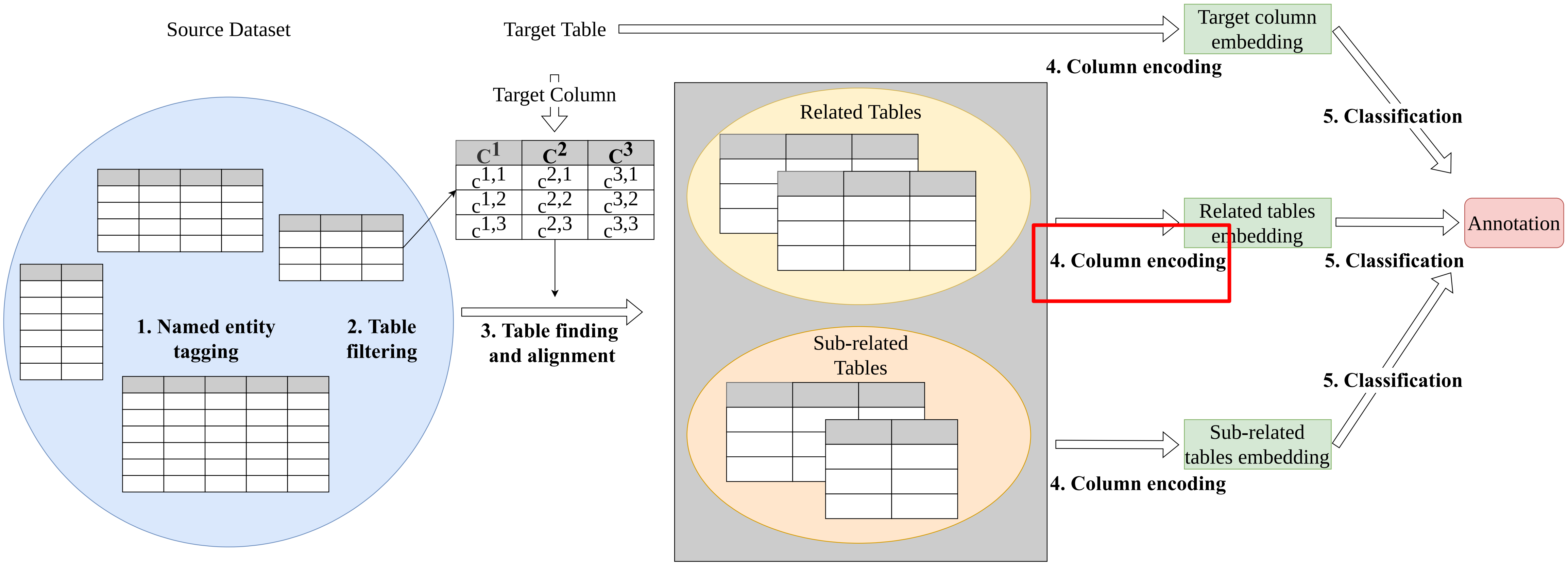
Pre-train+fine-tune LMs - methodology



Pre-train+fine-tune LMs - methodology

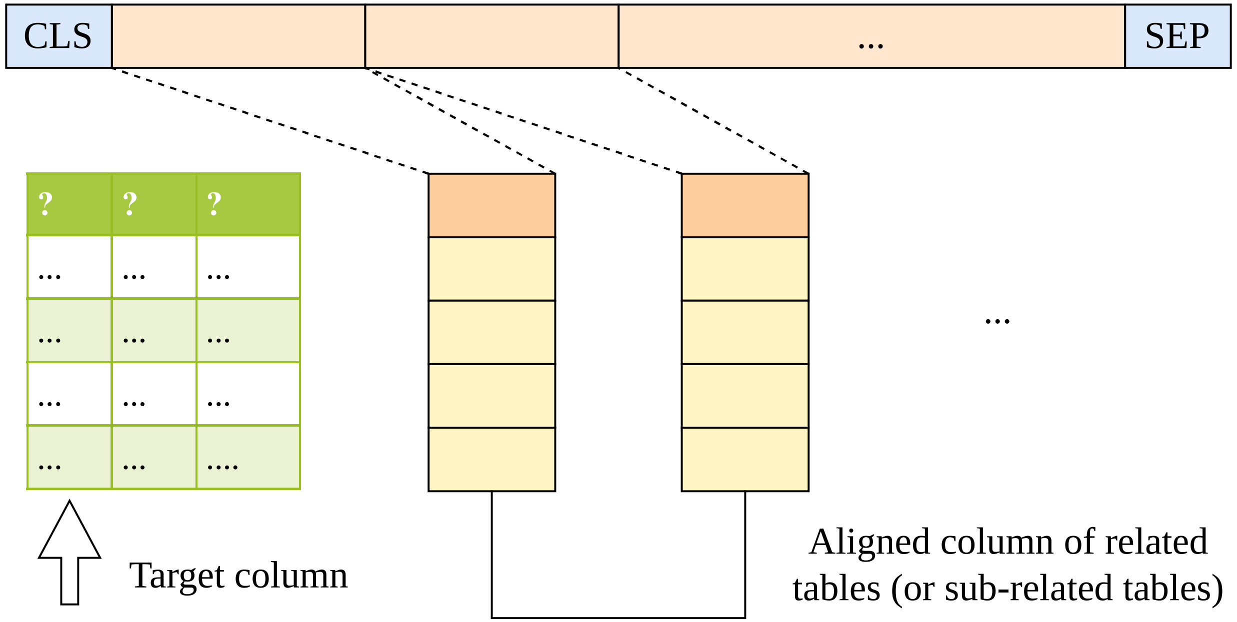
- Related tables: candidate tables T_j that share the same named entity schema as T_i.
- Sub-related tables: we consider the following two requirements:
  - Schema similarity: the named entity schemata should not be very different (edit distance less than a threshold).
  - Column location alignment: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

Pre-train+fine-tune LMs - methodology

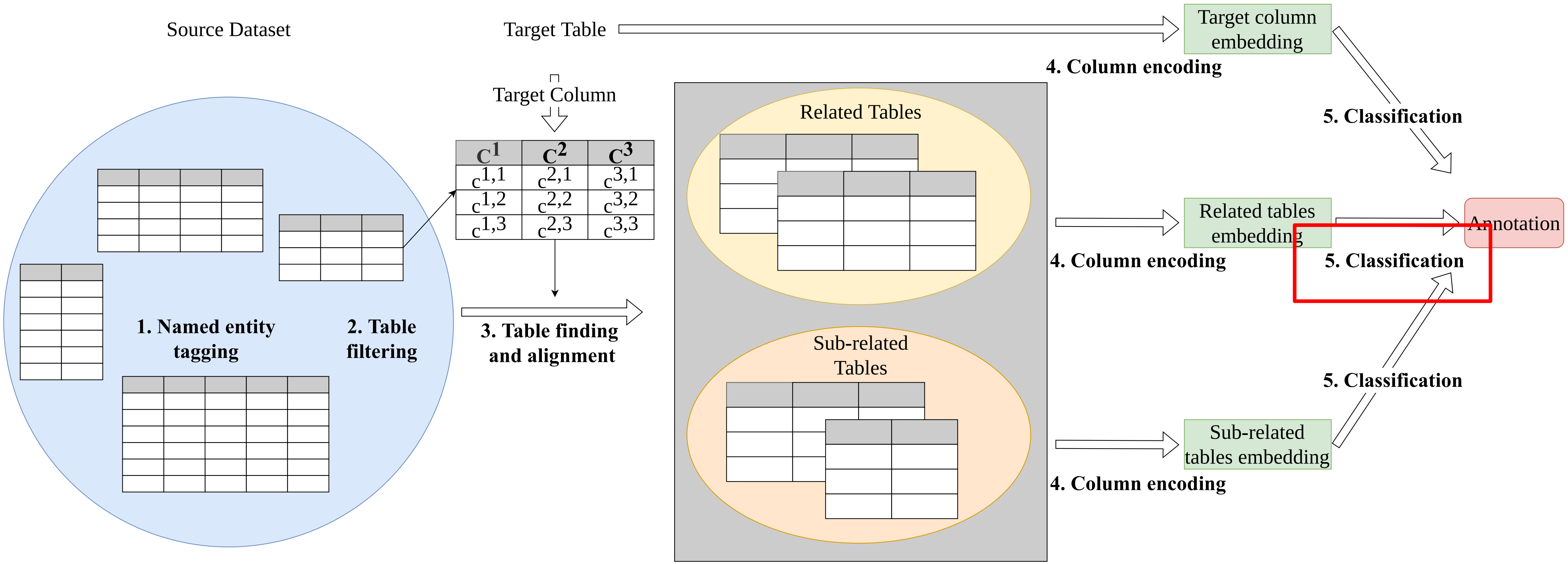


Pre-train+fine-tune LMs - methodology

- 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).



Pre-train+fine-tune LMs - methodology



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

Pre-train+fine-tune LMs - 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

Pre-train+fine-tune LMs - takeaways

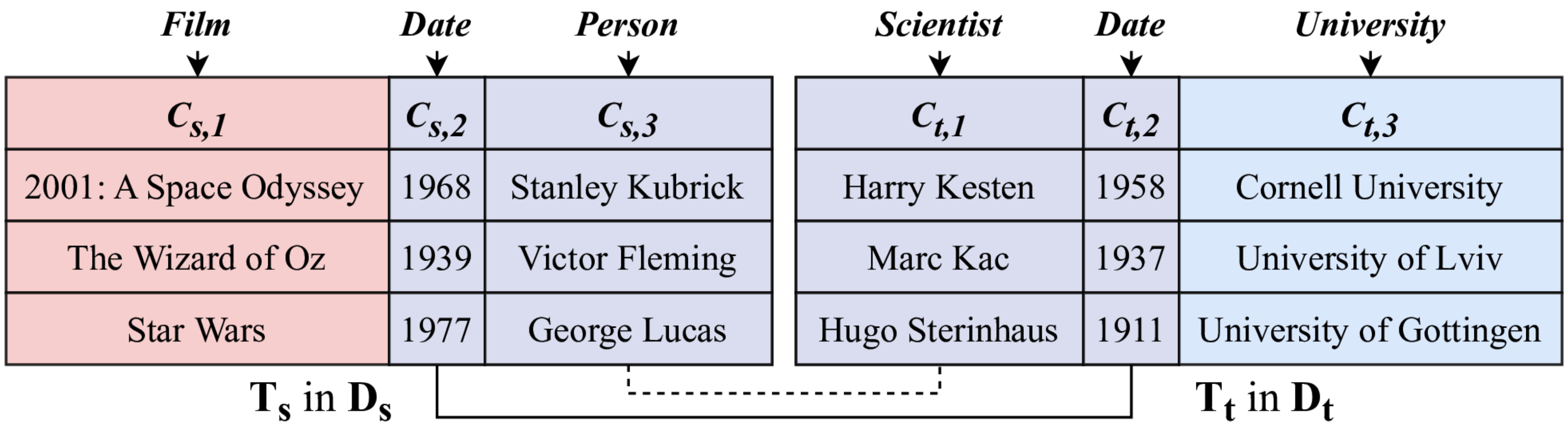
- The emergence of LM in data labeling opens up opportunities for utilizing LMs for DQ.
  - Pros:
    - Low annotation cost.
  - Cons:
    - Require annotated fine-tuning data for LMs (upon new data lakes).
  - Research Opportunities:
    - How to reduce the labeled training data required for LMs on performing DQ tasks / generalizing to new data lakes.

Outline

- Background
- LM4DQ
  - Past: Crowd-sourced / Human-in-the-loop
  - Status-quo: Pre-train+fine-tune LMs
  - Status-quo: Low-resource LMs
  - Future: Zero-shot LMs
- Future Vision and Opportunities
  - Preliminary study on DQ4LM
  - LM4DQ and DQ4LM

Low-resource LMs - overview

- **LakeHopper: Cross Data Lakes Column Type Annotation through Model Adaptation (submitted to ICDE 2025)**
- Focus on enhancing cross-domain tabular data labeling with the interaction of the world model and pre-trained models.



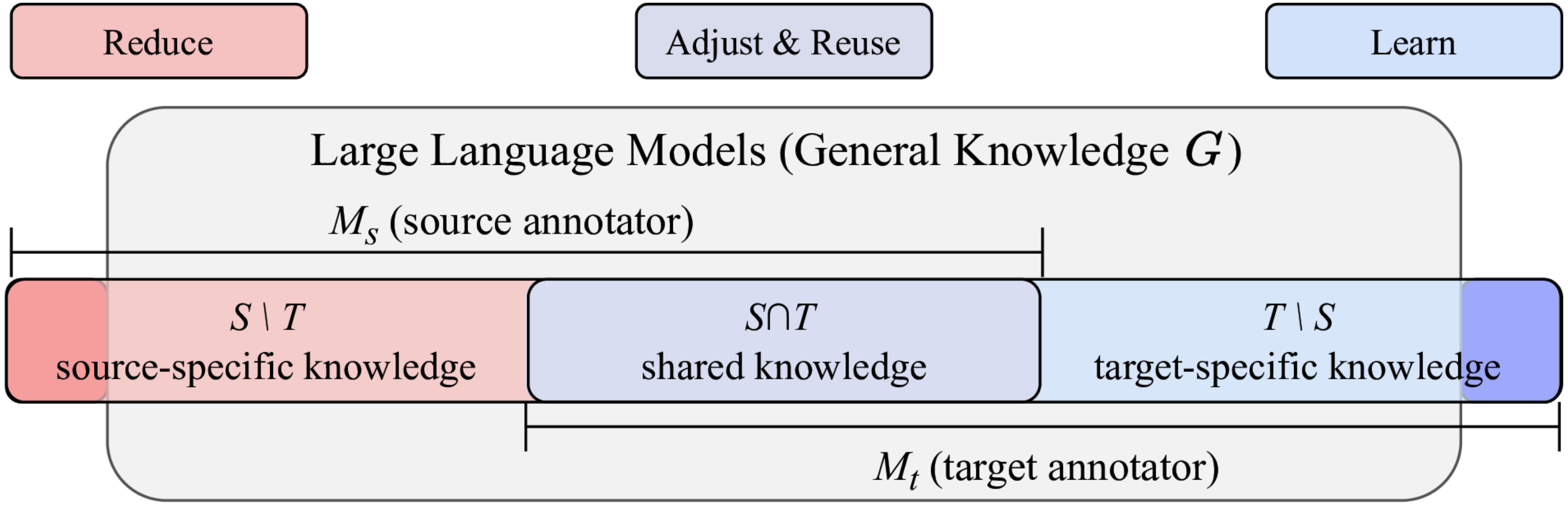
(a) Sample Source and Target Data Lake Tables

 [4] Y. Sun, N. Tang, and L. Chen, "LakeHopper: Cross Data Lakes Column Type Annotation through Model Adaptation," Under Review.

Low-resource LMs - overview

- Transform the source annotator into the target annotator.
- Reduce the source-specific knowledge.
- Adjust and reuse the shared knowledge.
- Learn the target-specific knowledge.

With the help of the general knowledge world model and resource-efficient fine-tuning process



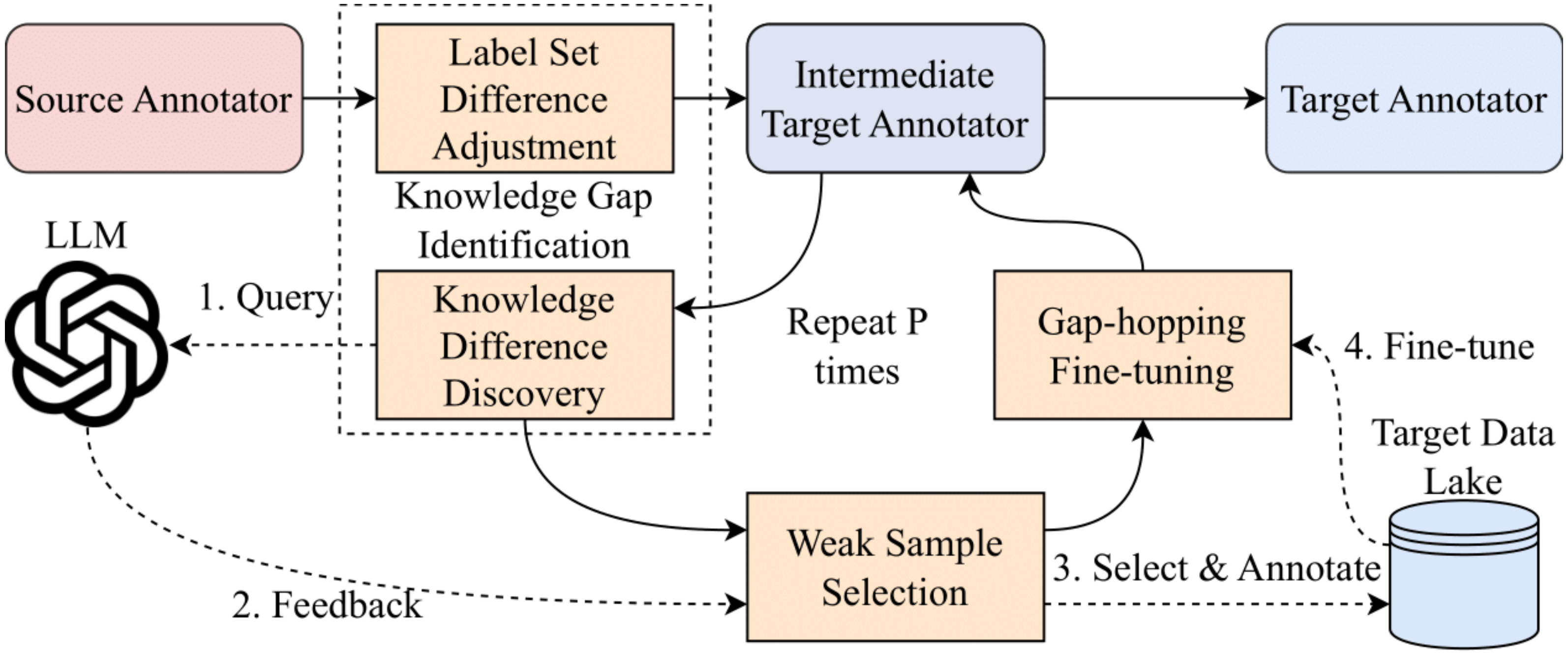
(b) Connections among Source/Target Annotators and LLMs

Low-resource LMs – definition

- (Cross Data Lakes Column Type Annotation): Given a model M_s fine-tuned on a source data lake D_s, a target data lake D_t, and a fixed budget N_t of training samples on the target data lake, the problem of cross data lakes column type annotation is to select at most N_t samples (each sample is a (C_i, y_i) pair) from the target data lake, and then use these training samples to obtain a transformed model M_t for the target data lake, such that M_t achieves the best column type annotation accuracy on the target data lake.

Low-resource LMs – methodology overview

- Knowledge gap identification: label set difference adjustment, knowledge differences found through the interaction with a general knowledge model (such as GPT)
- Weak sample selection: identify the weak samples through clustering
- Gap-hopping fine-tuning: fine-tuning with rehearsal incremental training



Low-resource LMs - experiments

LOW-RESOURCE EXPERIMENTAL RESULTS ON THE PUBLICBI TO VIZNET DATA LAKE TRANSFER.

	low1 1.6% (239 col)		low2 2.5% (364 col)		low3 4.2% (614 col)		low4 5.9% (864 col)		Avg. Gain	
	SW F1	MA F1	SW F1	MA F1	SW F1	MA F1	SW F1	MA F1	SW F1	MA F1
Sherlock [22]	0.344	0.130	0.470	0.238	0.558	0.303	0.591	0.345	-	-
TABBIE [23]	0.505	0.204	0.565	0.268	0.637	0.278	0.709	0.315	-	-
DODUO [51]	0.499	0.190	0.569	0.254	0.644	0.280	0.742	0.416	-	-
Sudowoodo [59]	0.561	0.213	0.601	0.277	0.705	0.374	0.724	0.427	-	-
RECA [53]	0.587	0.206	0.610	0.216	0.716	0.303	0.749	0.312	-	-
LakeHopper(D)	0.612	0.323	0.664	0.343	0.746	0.425	0.783	0.486	15.2% ↑	43.4% ↑
- LLM	0.591	0.256	0.657	0.336	0.714	0.376	0.744	0.440	-	-
LakeHopper(S)	0.609	0.317	0.679	0.384	**0.776**	0.446	**0.814**	**0.558**	11.0% ↑	34.3% ↑
- LLM	0.592	0.269	0.630	0.350	0.706	0.390	0.739	0.455	-	-
LakeHopper(R)	**0.621**	**0.331**	**0.705**	**0.412**	0.749	**0.506**	0.793	0.522	8.0% ↑	71.4% ↑
- LLM	0.555	0.306	0.604	0.334	0.729	0.463	0.767	0.516	-	-

Low-resource LMs - takeaways

- The interactions between domain-specific LMs and general LMs enable the generalization across different domains for DQ tasks.
  - Pros:
    - Low annotation cost.
    - Generalize across domains with relatively low fine-tuning costs.
  - Cons:
    - Still not zero-shot, and requires a small amount of labeled data.
    - Rely on the general knowledge of LMs to generalize across domains.
  - Research Opportunities:
    - How to further improve on the generalizability and reduce the labeling cost.

Outline

- Background
- LM4DQ
  - Past: Crowd-sourced / Human-in-the-loop
  - Status-quo: Pre-train+fine-tune LMs
  - Status-quo: Low-resource LMs
  - Future: Zero-shot LMs
- Future Vision and Opportunities
  - Preliminary study on DQ4LM
  - LM4DQ and DQ4LM

Zero-shot LMs - 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).

[5] 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.
[6] 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).

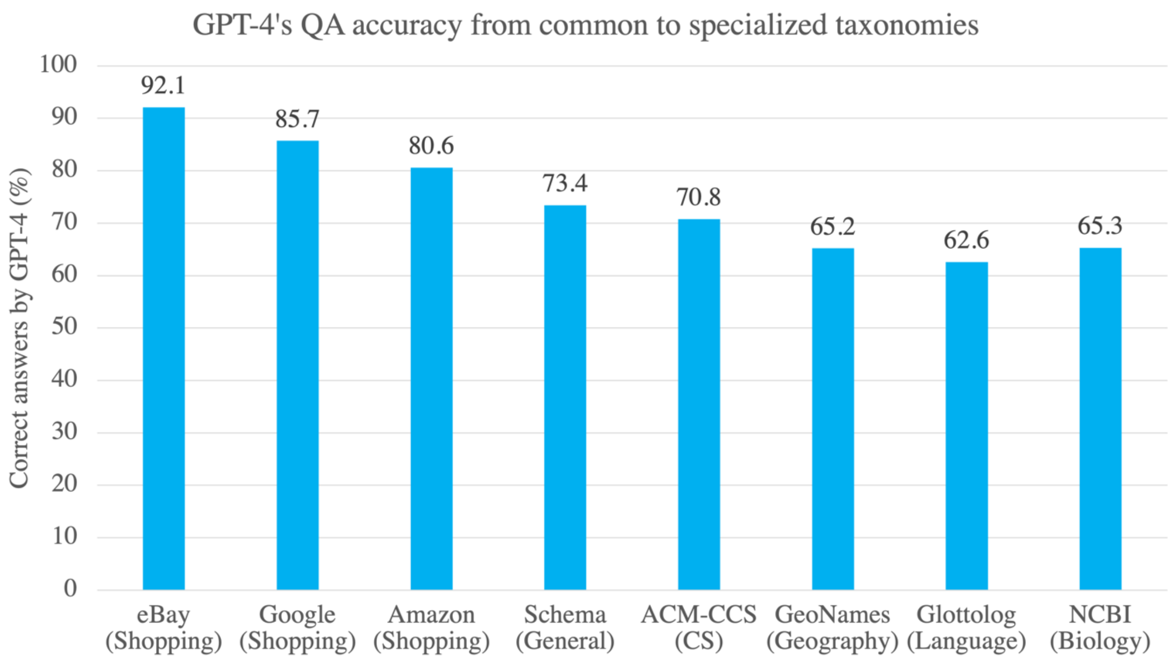
Zero-shot LMs - experiments

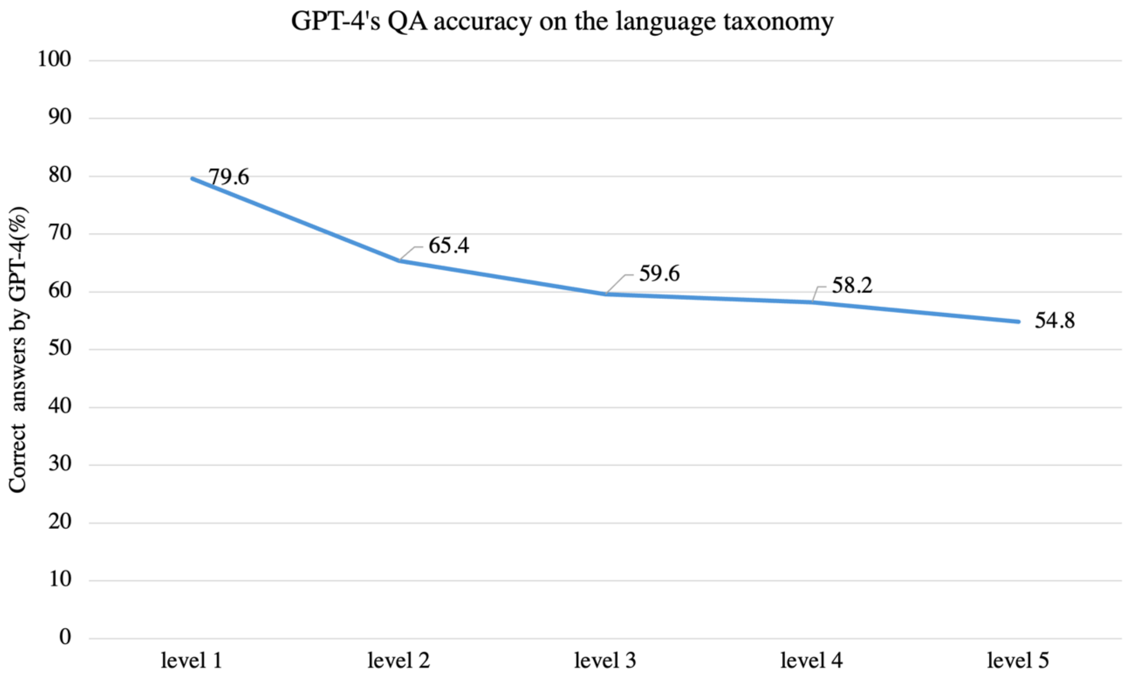
- We experimented with 18 SOTA LLMs on different taxonomies from common to specialized domains and root-to-leaf levels to see whether the existing LLMs internalize the taxonomy knowledge (zero-shot annotation on taxonomy data).
- Specifically, we ask each LLM about whether a child entity is a type of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

Domain	Taxonomy	# of entities	# of levels	# of entities in each level
Shopping	Google	5595	5	21-192-1349-2203-1830
Shopping	Amazon	43814	5	41-507-3910-13579-25777
Shopping	eBay	595	3	13-110-472
General	Schema	1346	6	3-17-215-403-436-272
CS	ACM-CCS	2113	5	13-84-543-1087-386
Geography	GeoNames	689	2	9-680
Language	Glottolog	11969	6	245-712-1048-1205-1366-7393
Biology	NCBI	2190125	7	53-309-514-1859-10215-107615-2069560

Zero-shot LMs - experiments

- 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.





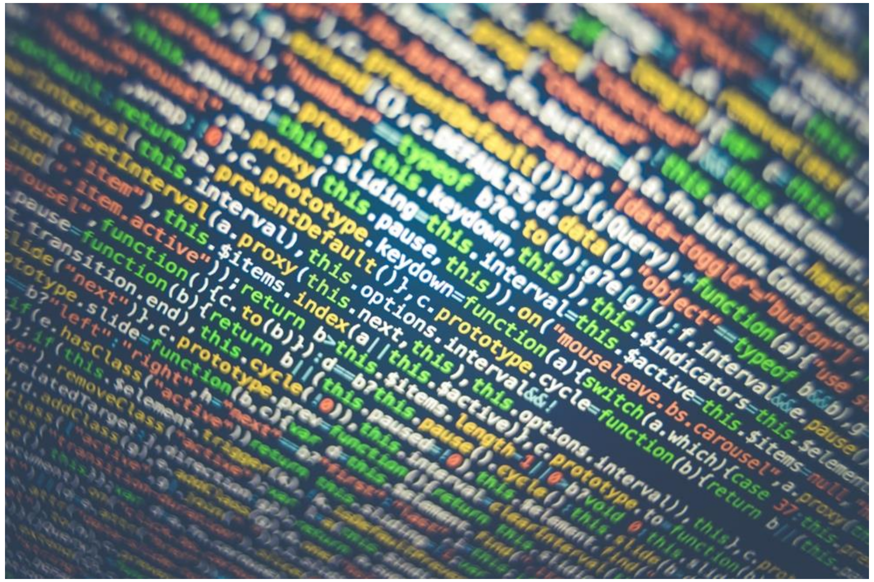
Zero-shot LMs - takeaways

- The advancement of LMs introduces the possibility of zero-shot DQ.
  - Pros:
    - Low annotation cost.
    - Zero generalization cost.
  - Cons:
    - The performance is not stable across different domains and different entities.
  - Research Opportunities:
    - How to achieve zero-shot, all-rounded, stable, unbiased DQ with LM.

Outline

- Background
- LM4DQ
  - Past: Crowd-sourced / Human-in-the-loop
  - Status-quo: Pre-train+fine-tune LMs
  - Status-quo: Low-resource LMs
  - Future: Zero-shot LMs
- Future Vision and Opportunities
  - Preliminary study on DQ4LM
  - LM4DQ and DQ4LM

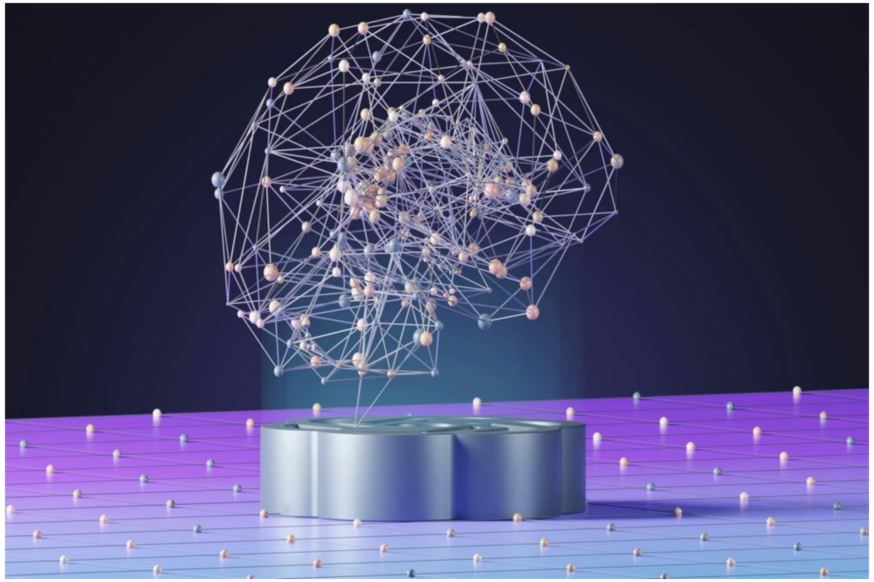
Data-quality-guaranteed LMs



Data Quality (focus on data labeling)

DQ4LM

Improved accuracy,
generalizability, …
Reduced
hallucination, bias, …



Language Models (BERT, GPT, Llama, …)

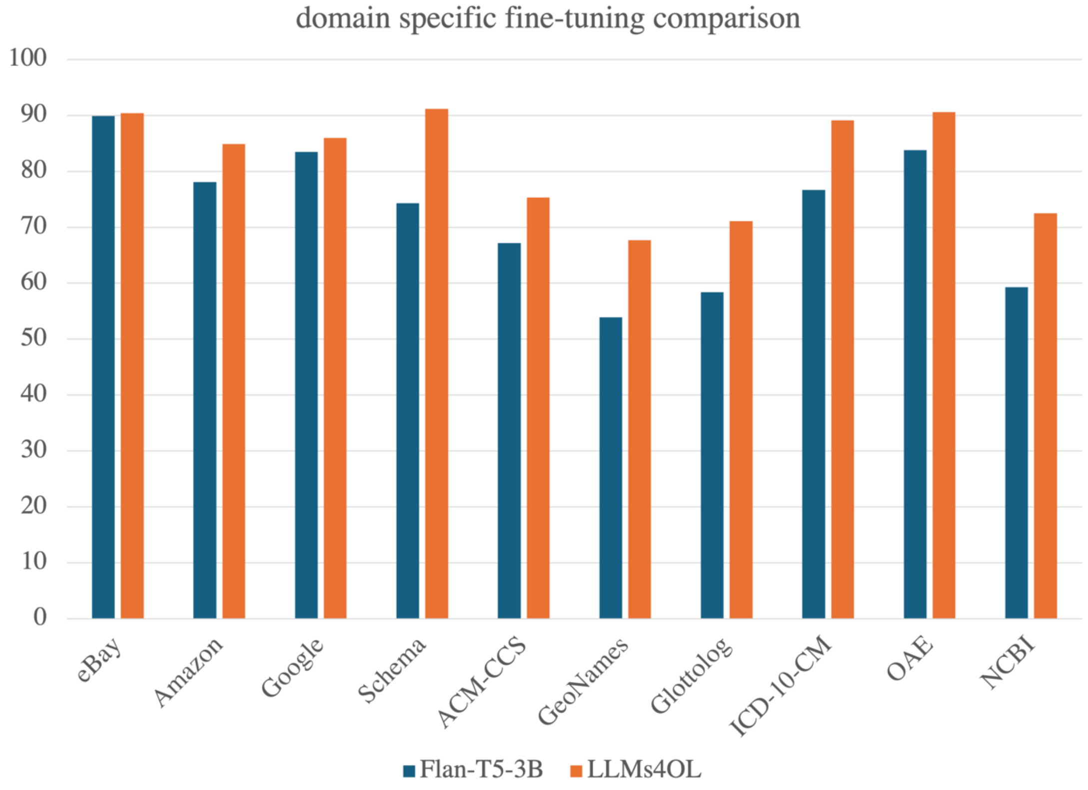
How does DQ influence LMs?

- Training data quality is crucial for LMs
  - Size of data: large-scale data
  - Diversity of data: comprehensive data
  - Fairness of data: unbiased data
  - …
- Garbage in garbage out!
- The quality of training data of LMs is more crucial than the size of the models [5]

[5] 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.

How does DQ influence LMs? Fine-tuning

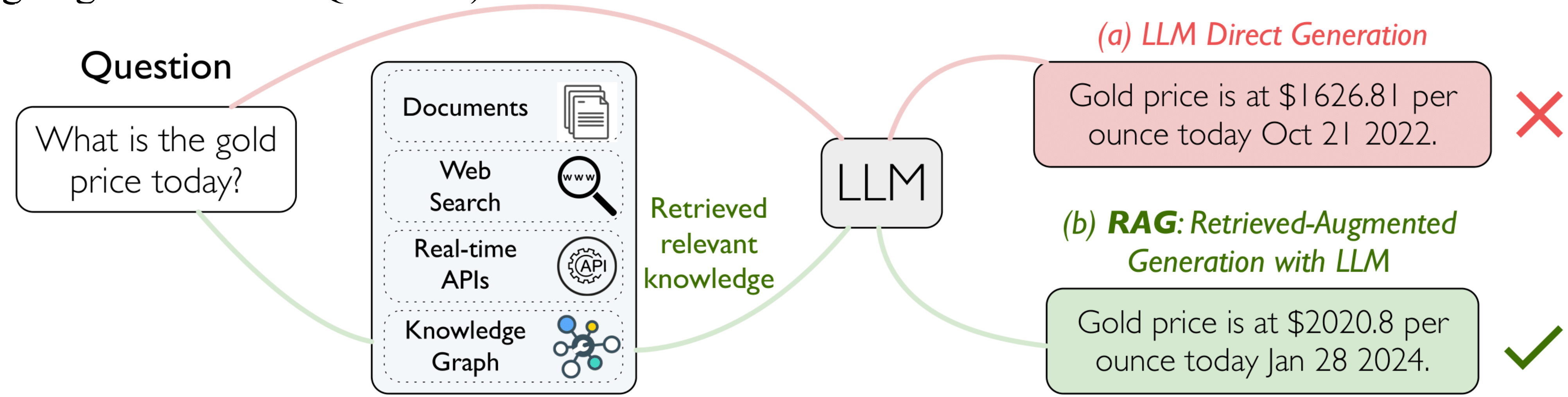
- **Are Large Language Models a Good Replacement of Taxonomies? (VLDB 2024)**
- Insights: High-quality training data can benefit the performance of LMs through fine-tuning.



[5] 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.

How does DQ influence LMs? RAG

- **CRAG – Comprehensive RAG Benchmark (Rebuttal <score: 7,7,7,7>, submitted to NeurIPS 2024)**
- Considered questions based on timeliness and difficulty level.
- Provided both KG and Web data sources.
- Insights: High-quality retrieval data can benefit the performance of LMs through RAG.
- Providing the right and high-quality data is important in the era of LLMs (insight from our other ongoing RAG-based QA work)



[7] Yang X, Sun K, Xin H, Sun Y, Bhalla N, Chen X, et al. CRAG -- Comprehensive RAG Benchmark [Internet]. arXiv.org. 2024 [cited 2024 Sep 5]. Available from: https://arxiv.org/abs/2406.04744

Research Opportunities: LM4DQ and DQ4LM

- My future endeavors: collaboration and fusion of the two fields, towards zero-shot all-rounded DQ and advanced LMs.
- LM4DQ: towards a zero-shot, all-in-one LM-based DQ general method.
- DQ4LM: improving LMs on fairness, timeliness, and domain-specific. Quantifying and optimizing the value/quality (size, diversity, fairness, etc.) of different data (structured, semi-structured, unstructured) for a specific LM (Bert, GPT, Llama) under a specific data usage scenario (fine-tuning, RAG) on different applications (task/domain-dependent).