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

Yushi Sun

Last updated 21/11/2024

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.

My first-authored publications

- **Cross-Domain-Aware Worker Selection with Training for Crowdsourced Annotation**, ICDE 2024.
- **RECA: Related Tables Enhanced Column Semantic Type Annotation Framework**, VLDB 2023.
- **LakeHopper: Cross Data Lakes Column Type Annotation through Model Adaptation**, VLDB 2025 submitted.
- **Are Large Language Models a Good Replacement of Taxonomies?**, VLDB 2024.
- **CRAG -- Comprehensive RAG Benchmark**, NeurIPS 2024 & Hosting KDD Cup 2024.

Crowdsourced DQ

Pretrain+finetune LM-based DQ

Low-resource LM-based DQ

Zero-shot LM-based DQ

Language Model for Data Quality (LM4DQ)

DQ in Fine-tuning

DQ in RAG

Data Quality for Language Model (DQ4LM)

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, LM, and LM4DQ

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

] High-quality and efficient data labeling/preparation

My PhD focus

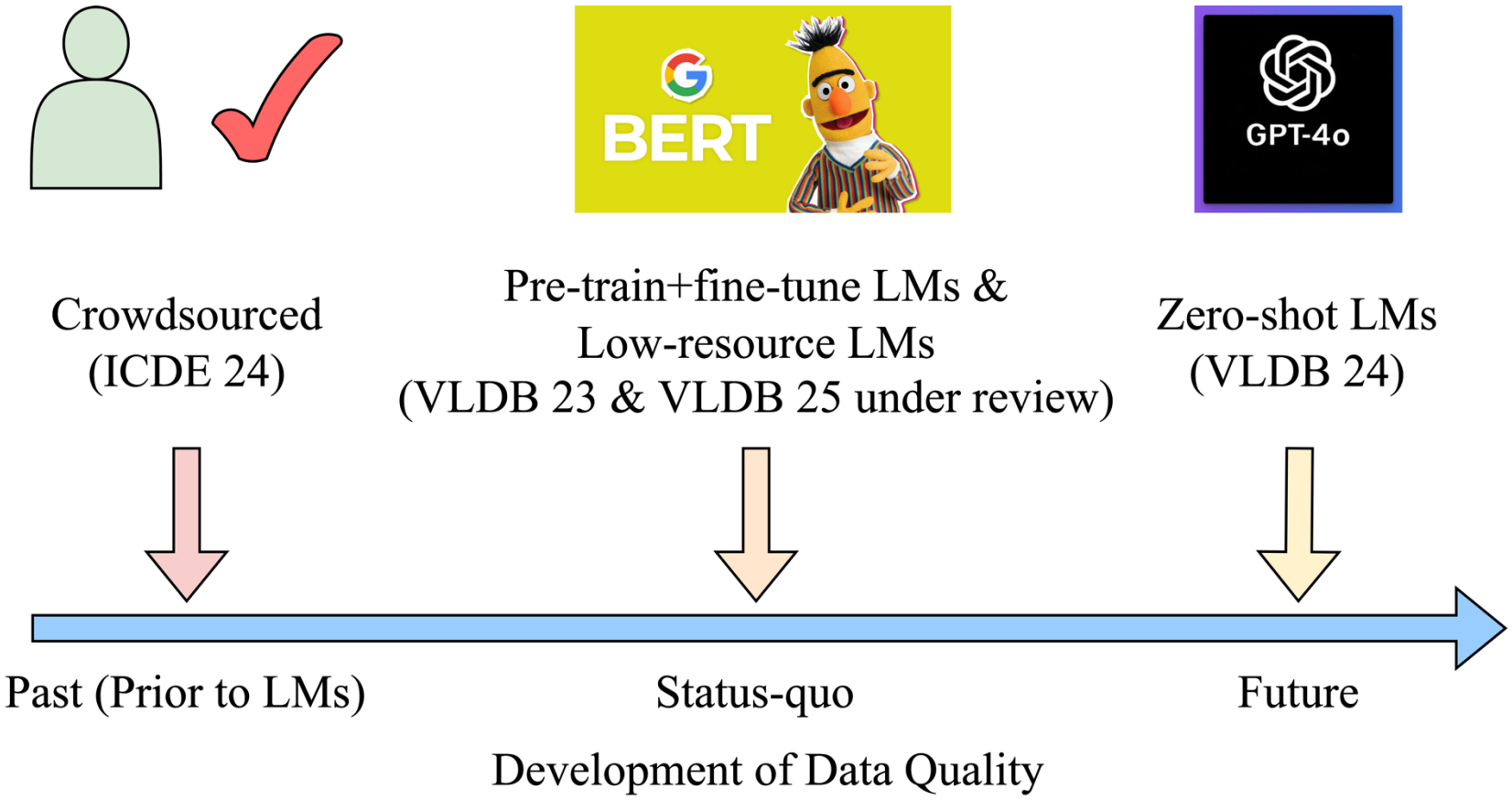
- Language Models:
  - Good at processing textual data.
  - Knowledgeable through pre-training.
- LM4DQ:
  - Apply LMs in completing DQ tasks (e.g., data labeling/preparation).
  - Reduced labeling cost
  - Improved data labeling performance

[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

Data Quality: Past, Status-quo, and Future

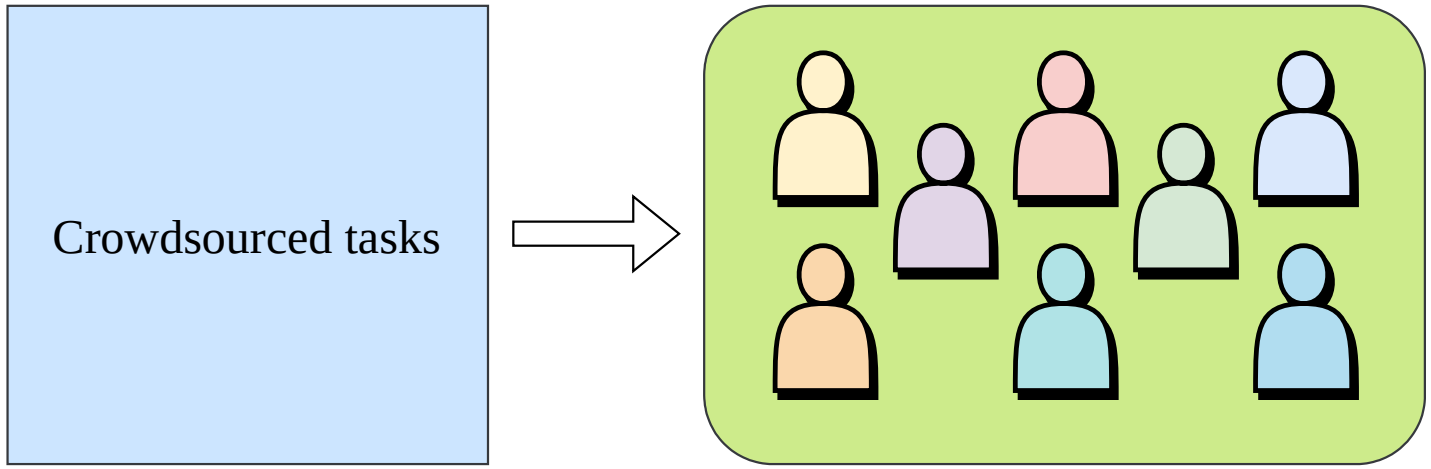


Outline

- Background
- LM4DQ
  - Past: Crowd-sourced
  - 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 - overview

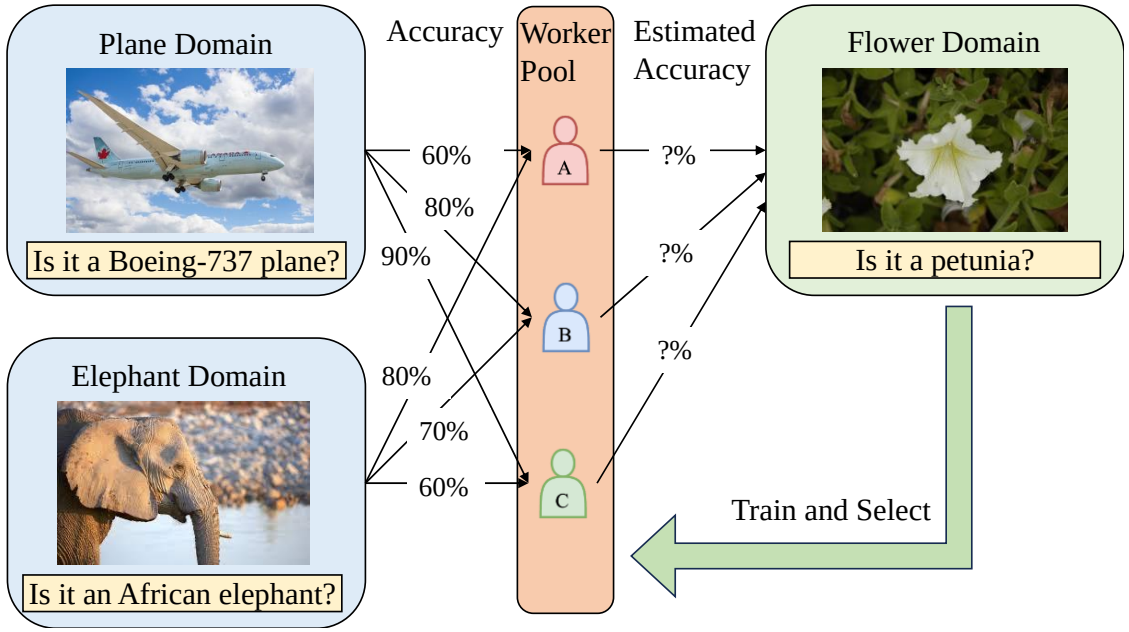
- **Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation (ICDE 2024)**
  - Crowdsourcing is preferable for obtaining high-quality data labels for large-scale datasets.
  - Worker Selection is important in Crowdsourcing.
  - How to design an allocation scheme for golden questions (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.



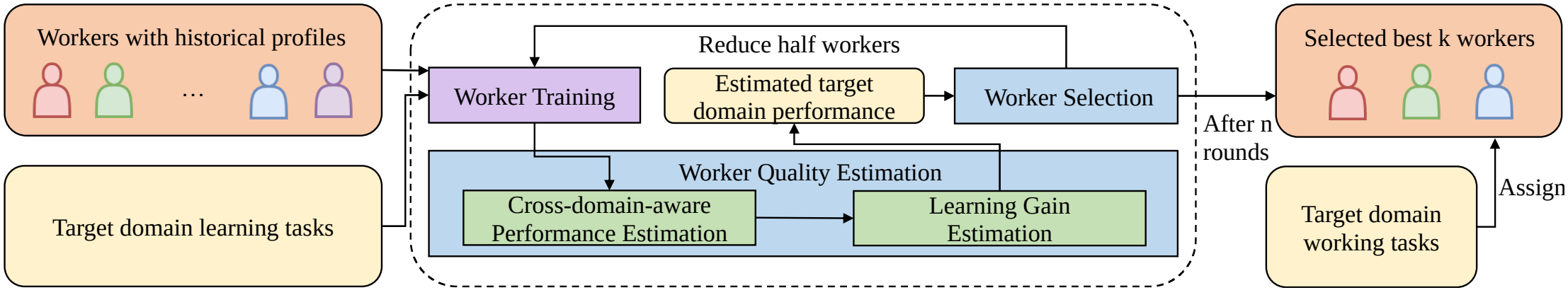
[2] Y. Sun, et al., "Cross-Domain-Aware Worker Selection with Training for Crowdsourced Annotation," in 2024 IEEE 40th International Conference on Data Engineering (ICDE), Utrecht, Netherlands, 2024 pp. 249-262. doi: 10.1109/ICDE60146.2024.00026

Crowd-sourced - background

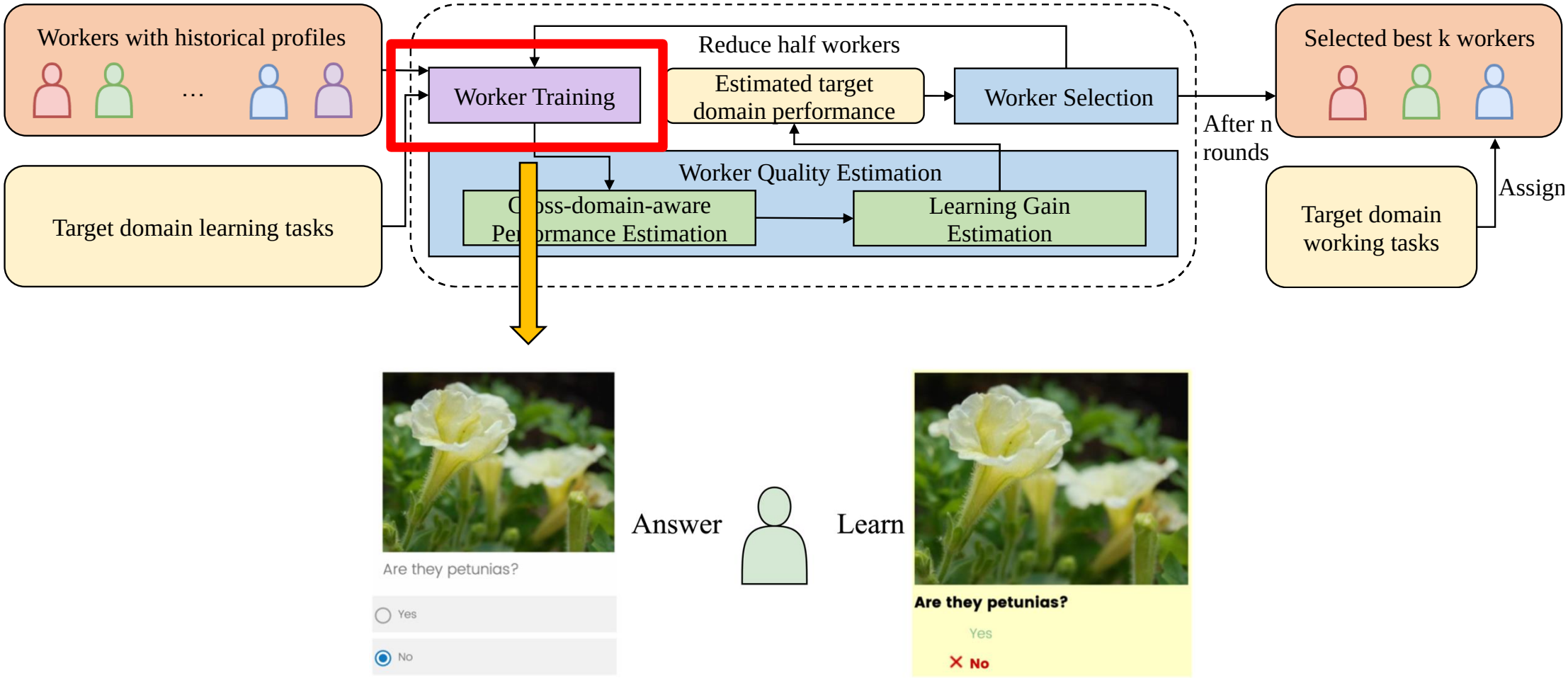
- Many companies such as JD, Alibaba, and Baidu have their commercial crowdsourcing platforms with worker pools, which record the answering history of workers.
- The answering history of workers (prior domain knowledge) can help select high-quality workers when annotating a new domain (target domain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



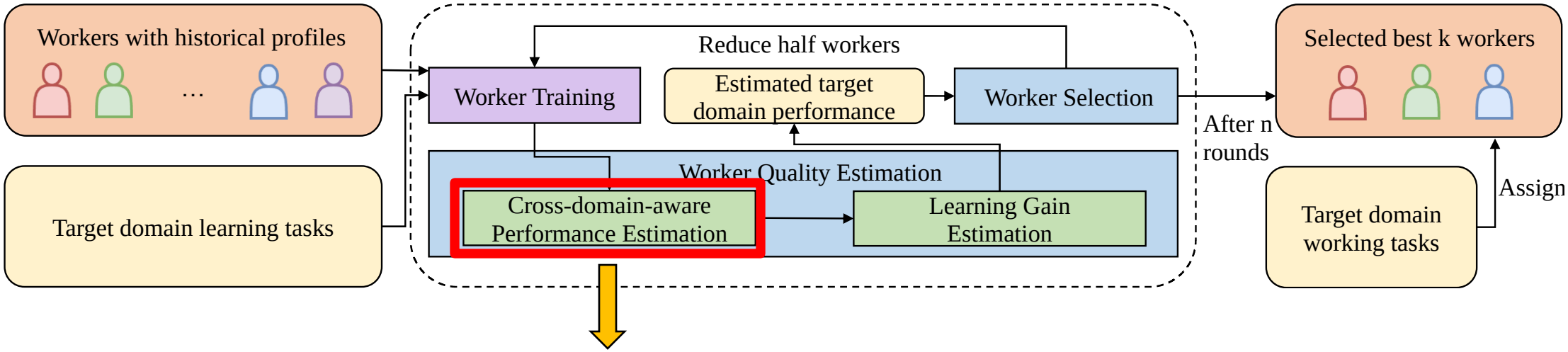
Crowd-sourced - methodology



Crowd-sourced - methodology



Crowd-sourced - 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.

Crowd-sourced - methodology

- Maximum likelihood estimation:

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

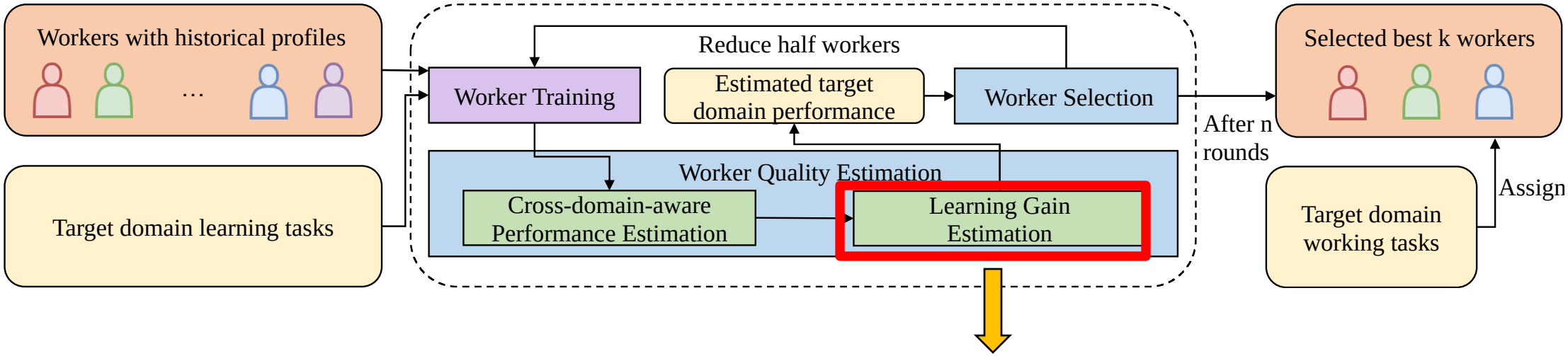
$$\bar{\mu} = \mu_T + \Sigma_{1\times D} \Sigma_{D\times D}^{-1} (h_i - \mu_{1\sim D}),$$
$$\bar{\Sigma} = \Sigma_{1\times 1} - \Sigma_{1\times D} \Sigma_{D\times D}^{-1} \Sigma_{D\times 1},$$

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

- Updated annotation accuracy:

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

Crowd-sourced - methodology



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

Crowd-sourced - methodology

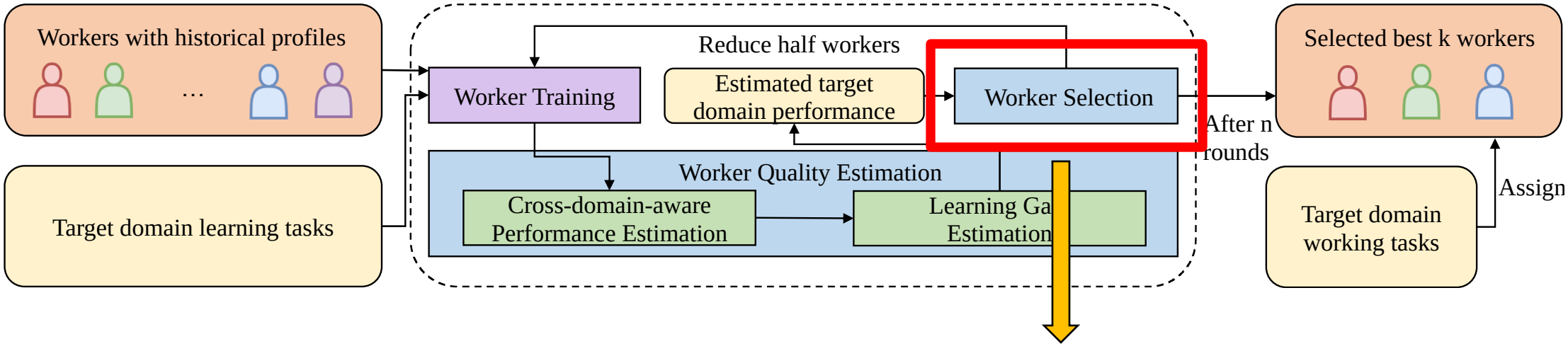
- IRT score:

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

- Update the learning parameter α_i:

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

Crowd-sourced - methodology



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

Crowd-sourced - datasets

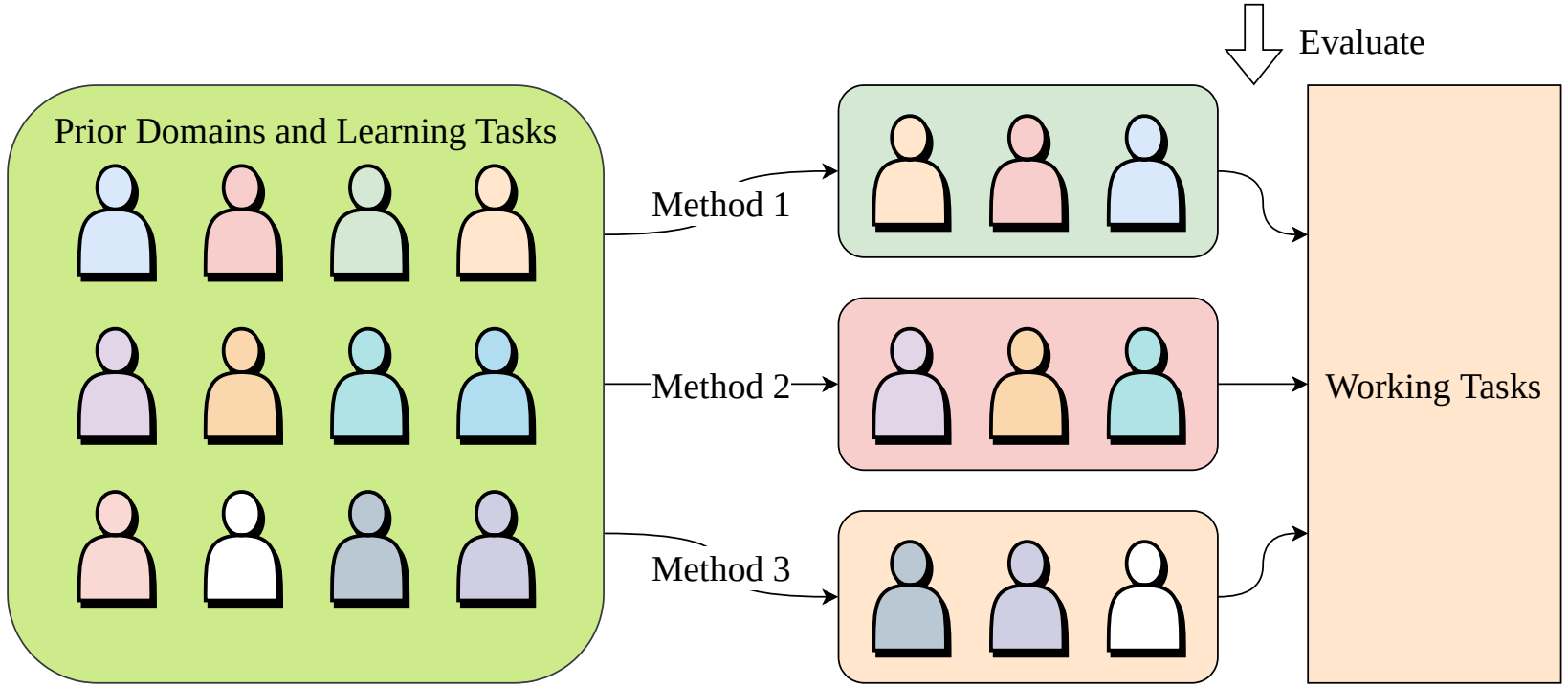
- Datasets:

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

Crowd-sourced - metrics

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



Crowd-sourced - 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

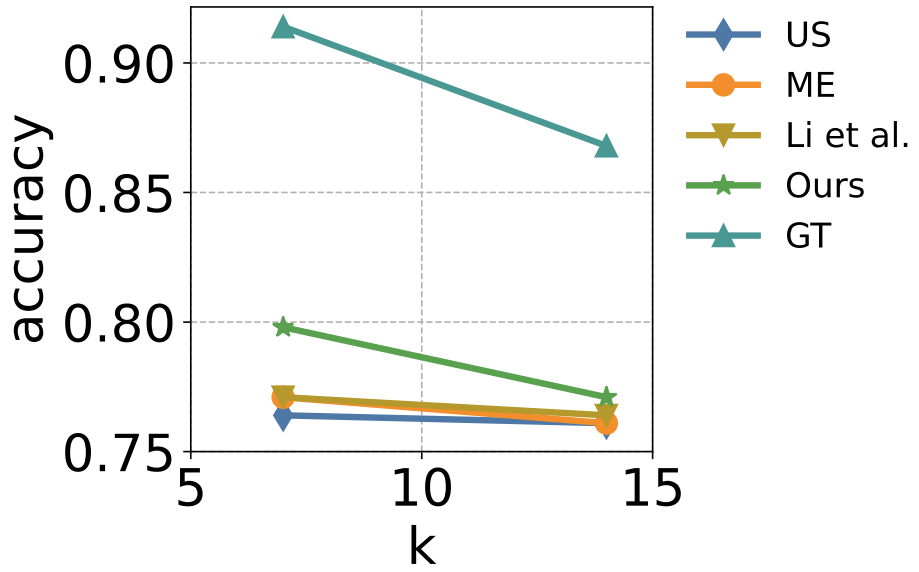
Crowd-sourced - experiments

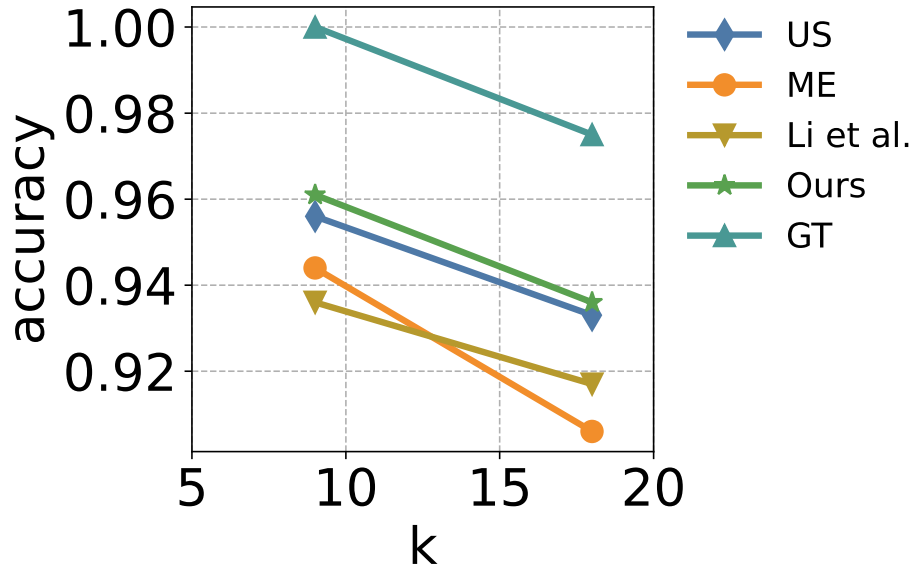
TABLE V
EXPERIMENT RESULTS

	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 - experiments

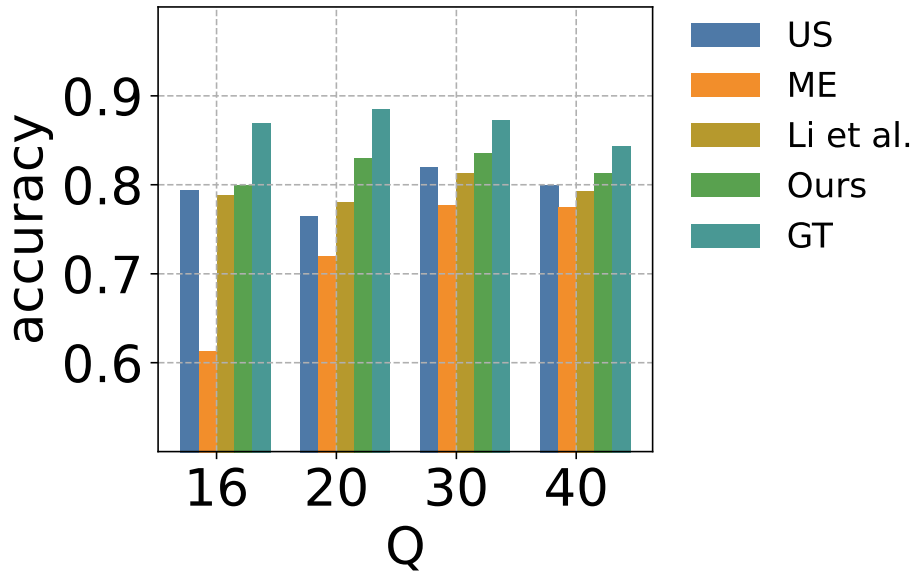
- Stability over the parameter k (number of desired workers)

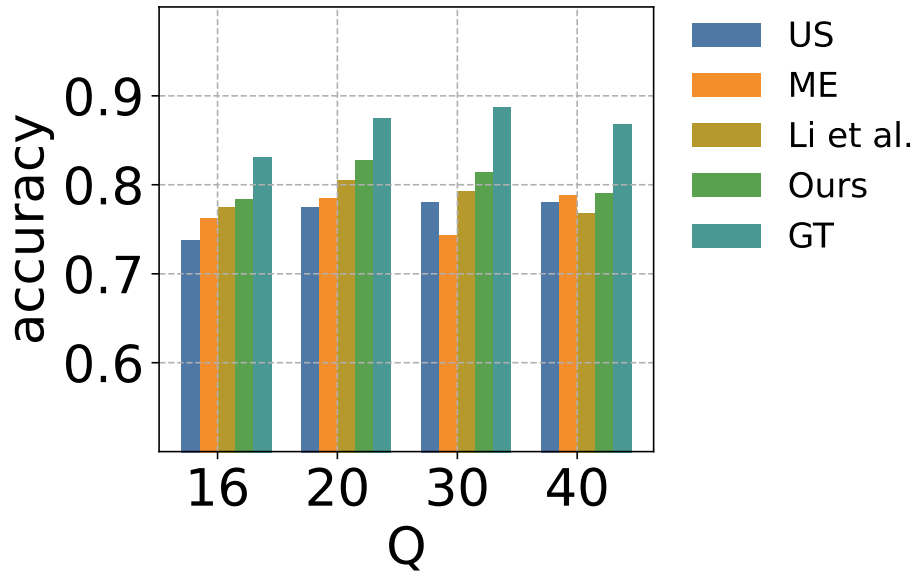




Crowd-sourced - experiments

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





Crowd-sourced - takeaways

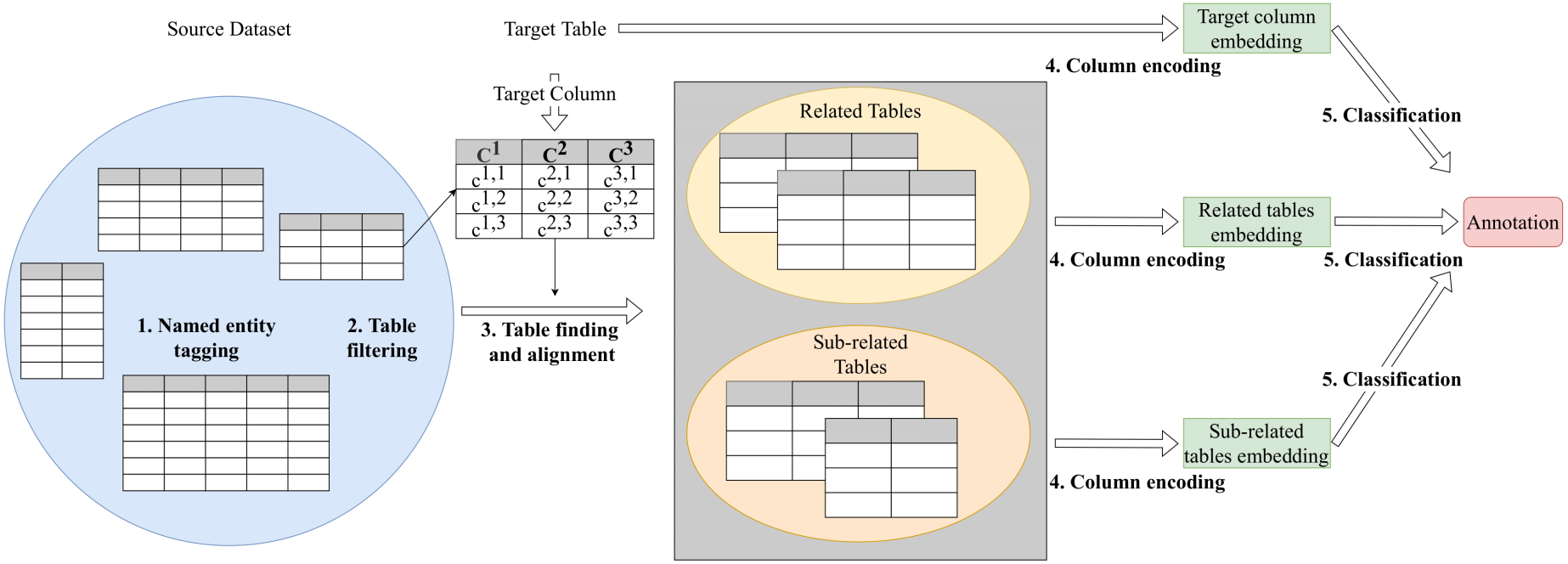
- Before the emergence of LM in data labeling, crowd-sourced approaches were the main approaches that we can count on.
  - Pros:
    - Compared to black-box LM, easy debugging on the data labeling results (You can ask the crowd-workers about their choices).
    - Quality control and guarantee (You can 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

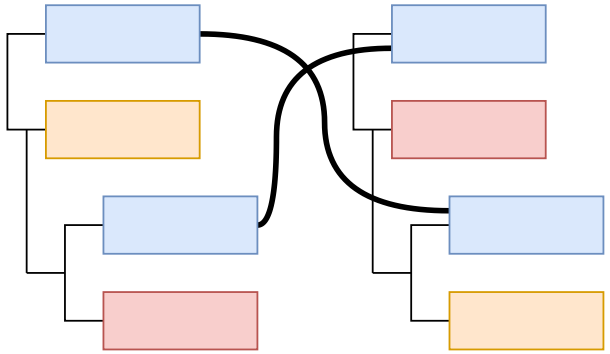
- **RECA: Related Tables Enhanced Column Semantic Type Annotation Framework (VLDB 2023)**
- Focus on enhancing **tabular data labeling** with **inter-table** context information.



[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

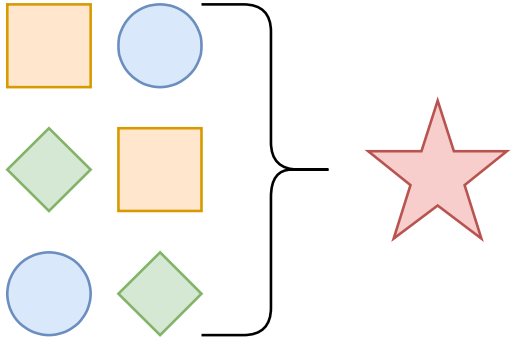
- Accurate column semantic type labeling is important for various applications:
  - schema matching, data cleaning, data integration, etc.



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

- 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

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

- 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

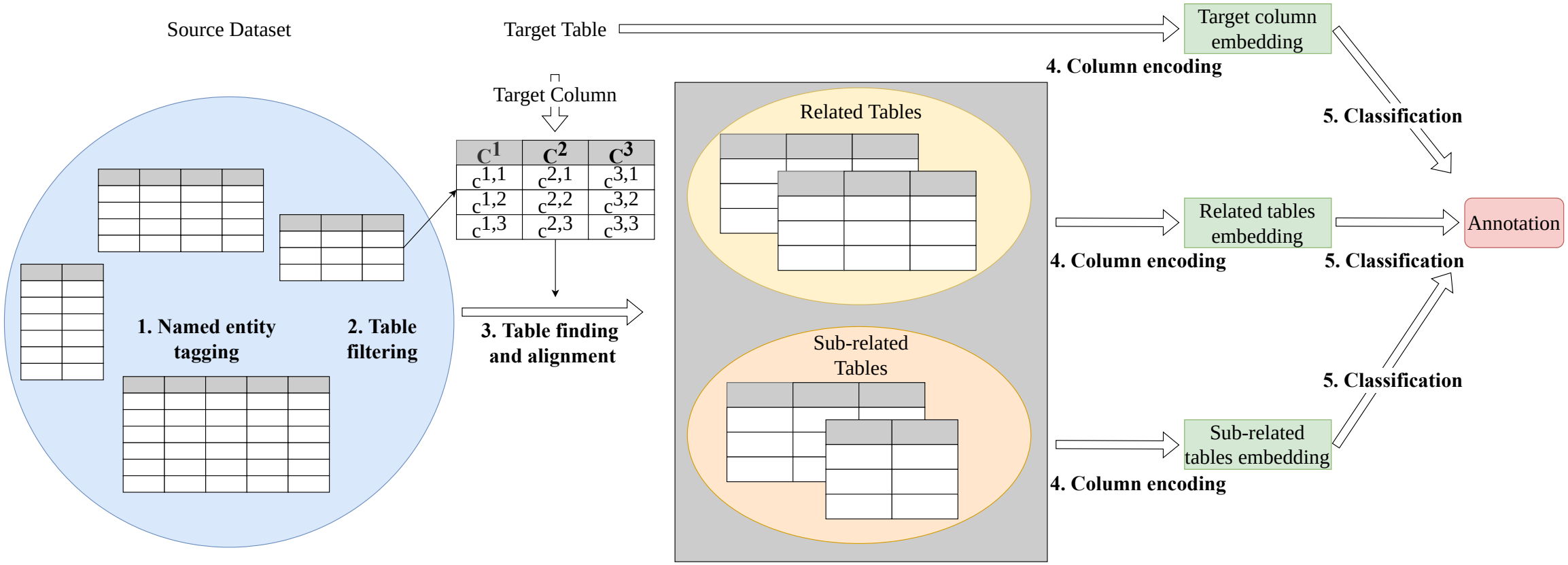
Pre-train+fine-tune LMs - definition

- 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.

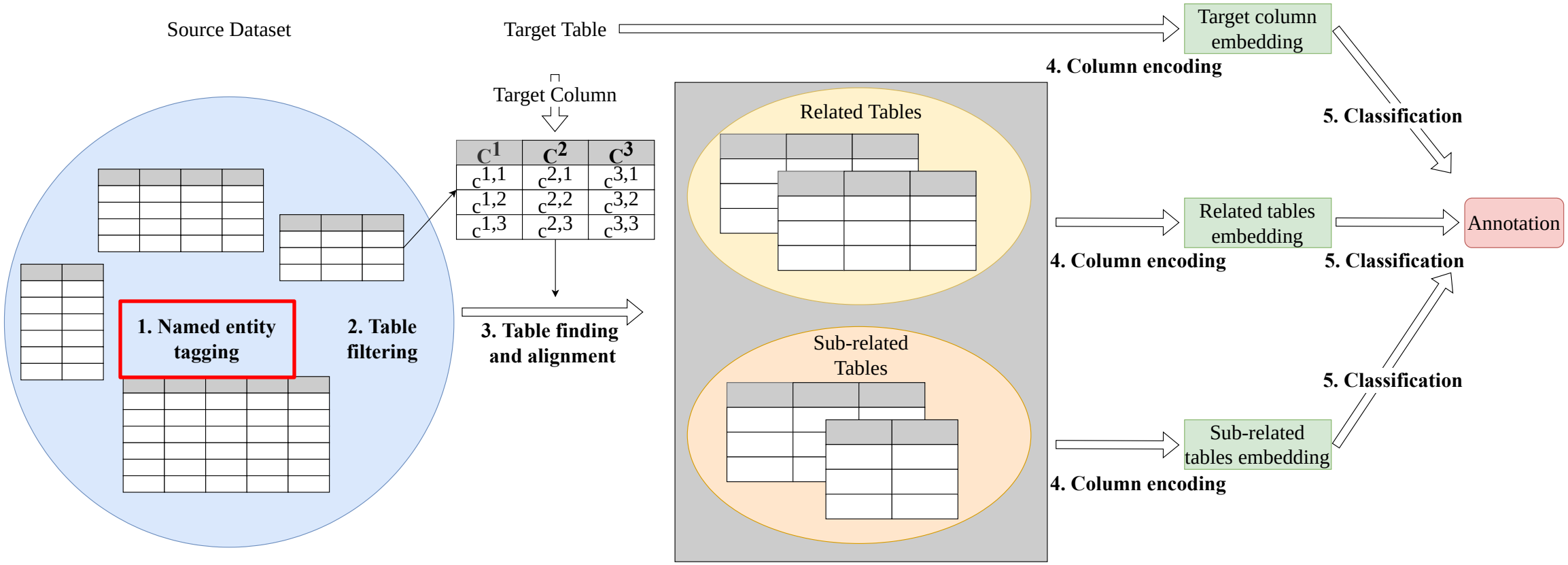
Pre-train+fine-tune LMs - definition

- (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



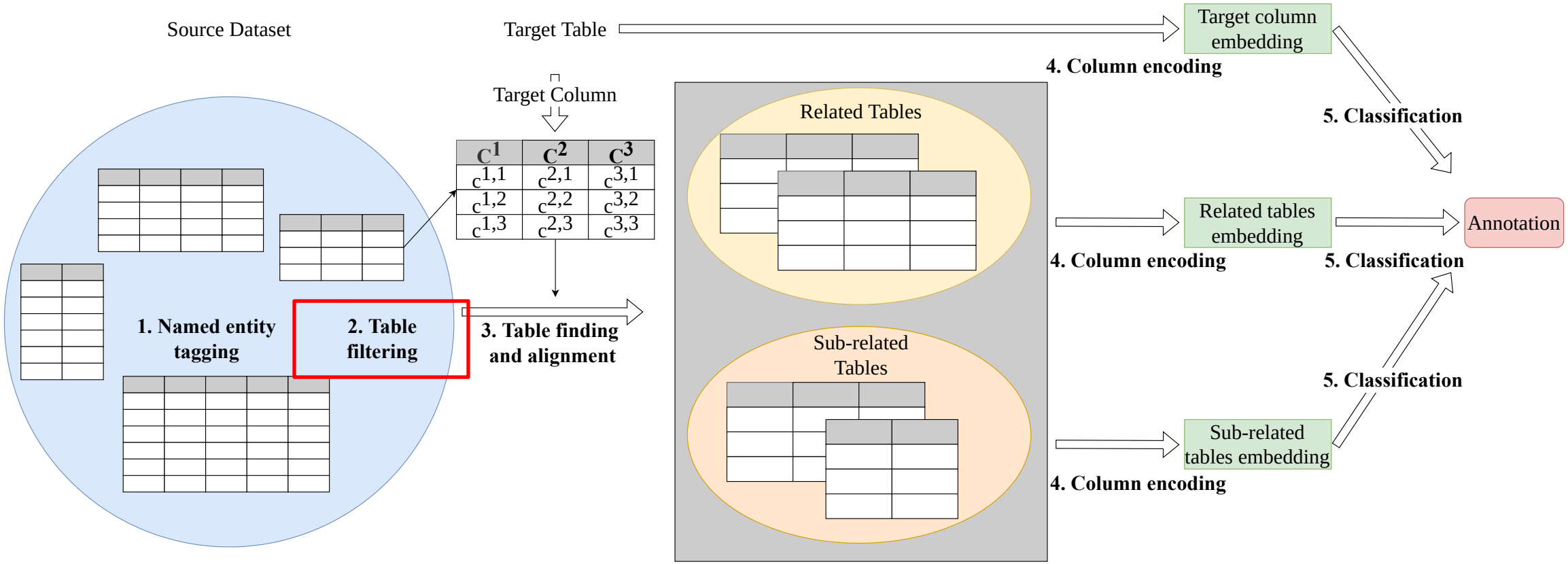
Pre-train+fine-tune LMs - methodology



Pre-train+fine-tune LMs - methodology

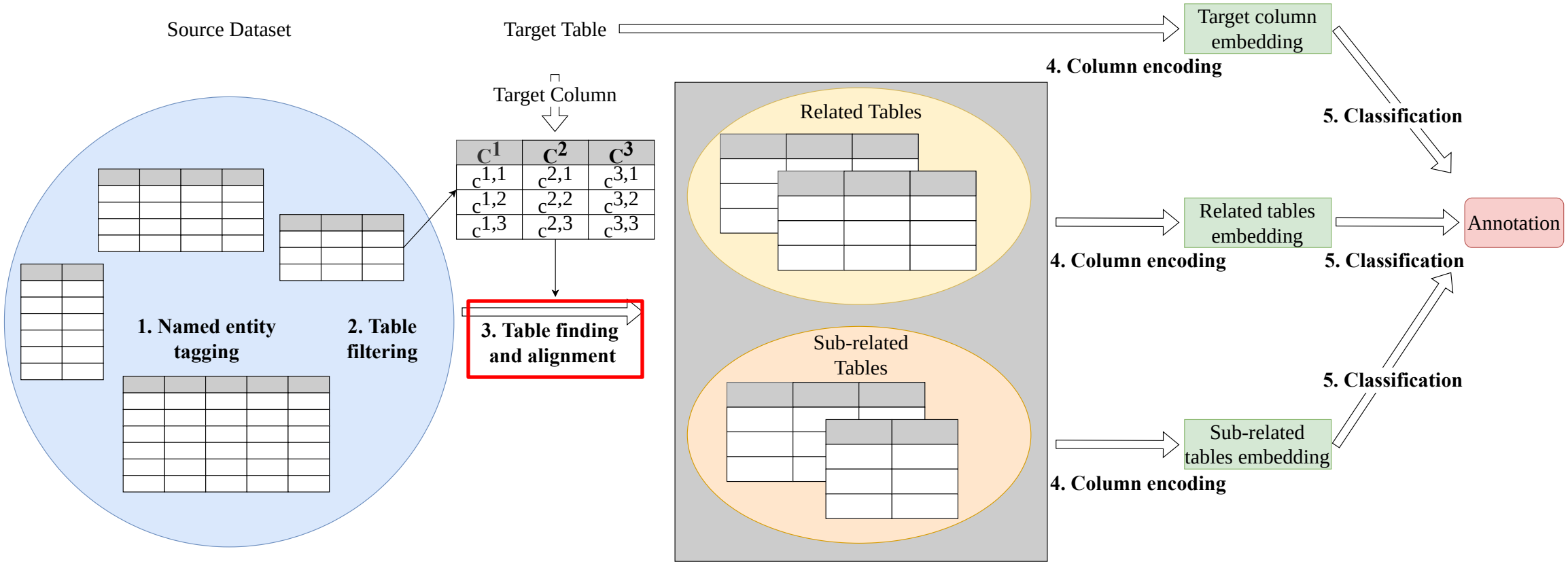
- Given a table T with M columns and N rows, we use the spaCy tagging tool to identify the named entities in each column and tag them.
- We further classify the DATE and PERSON types based on the data format.
  - E.g. DD-MM-YYYY; YYYY; January 16th 2022; 2023
  - E.g. J. K. Rowling; Anna
- We include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

Pre-train+fine-tune LMs - methodology



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

Pre-train+fine-tune LMs - methodology

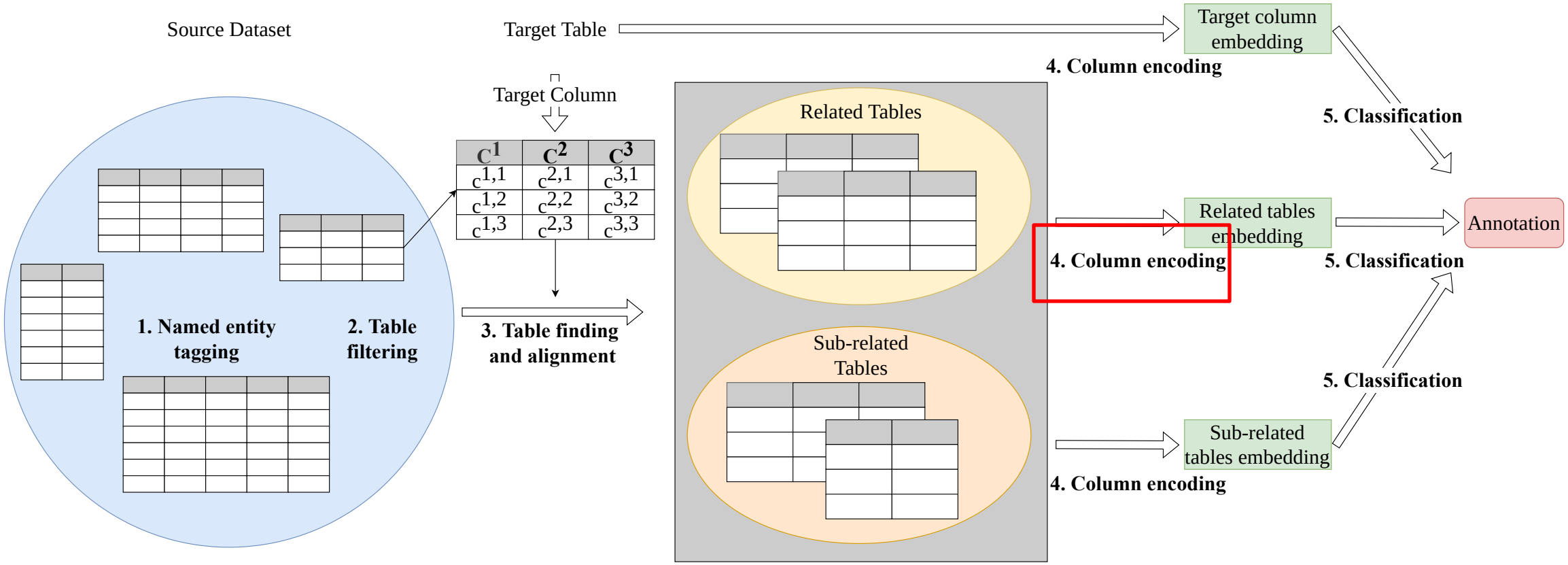


Named Entity Schema & Jaccard Similarity

Pre-train+fine-tune LMs - methodology

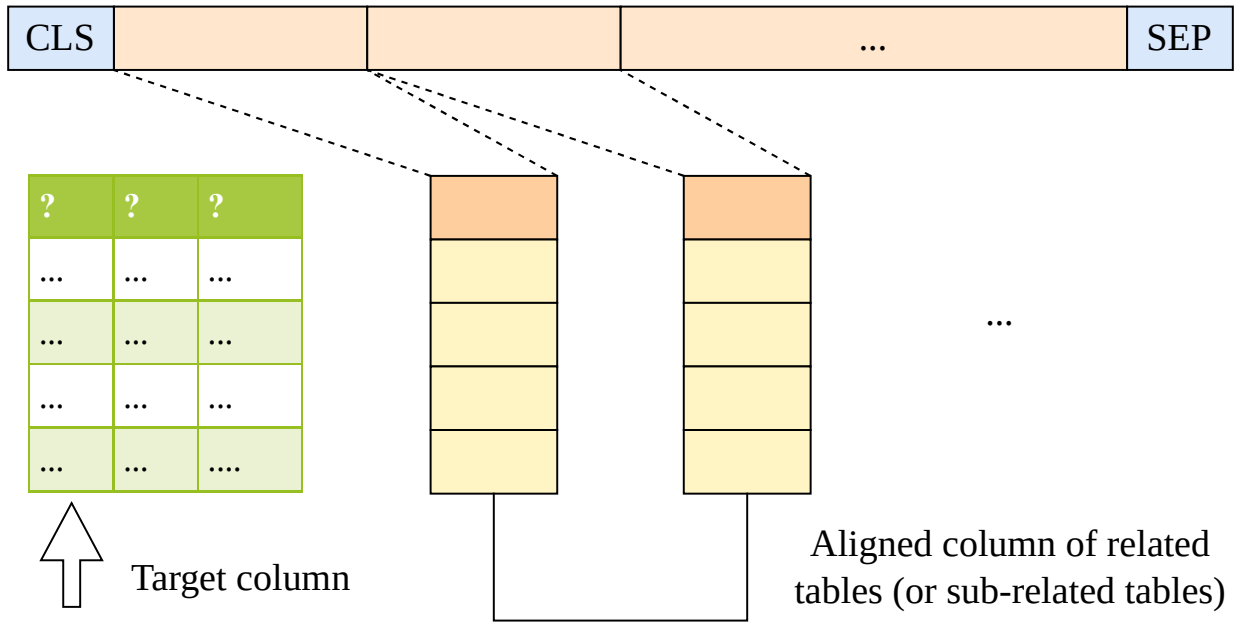
- Related tables: candidate tables T_j that share the same named entity schema as T_i.
- Sub-related tables: we consider the following two requirements:
  - Schema similarity: the named entity schemata should not be very different (edit distance less than a threshold).
  - Column location alignment: 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

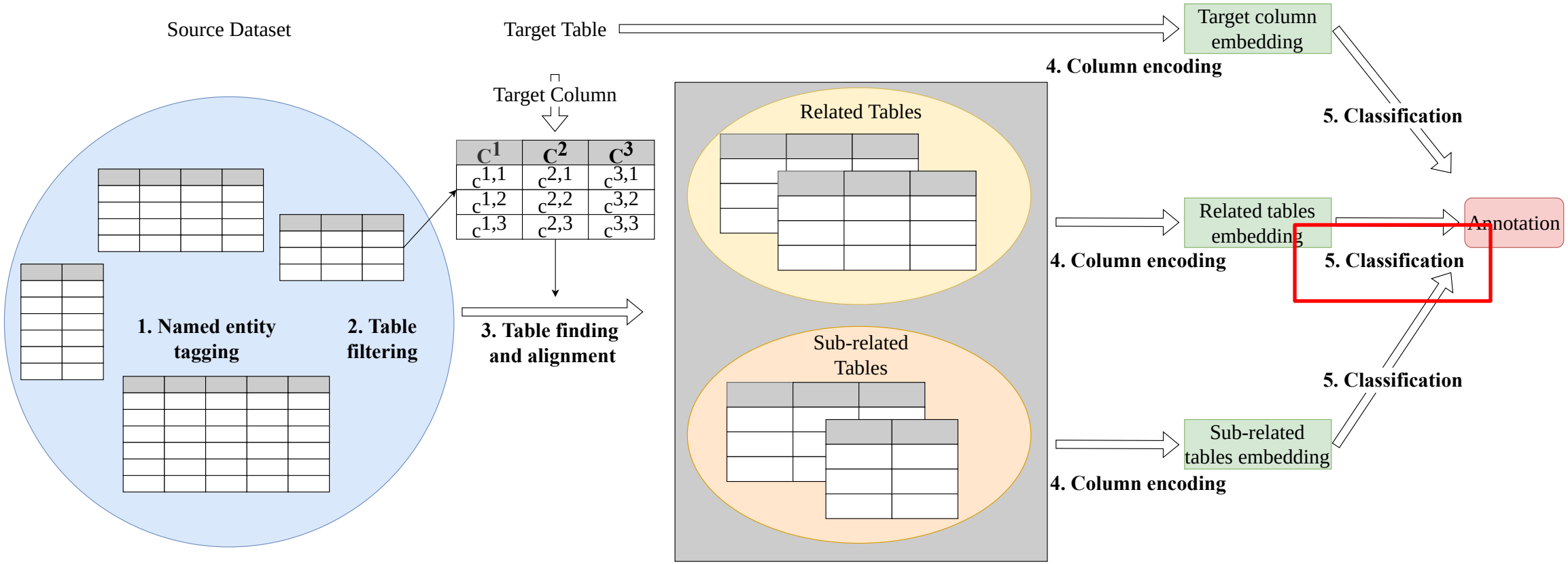


Pre-train+fine-tune LMs - methodology

- 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

- 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)

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

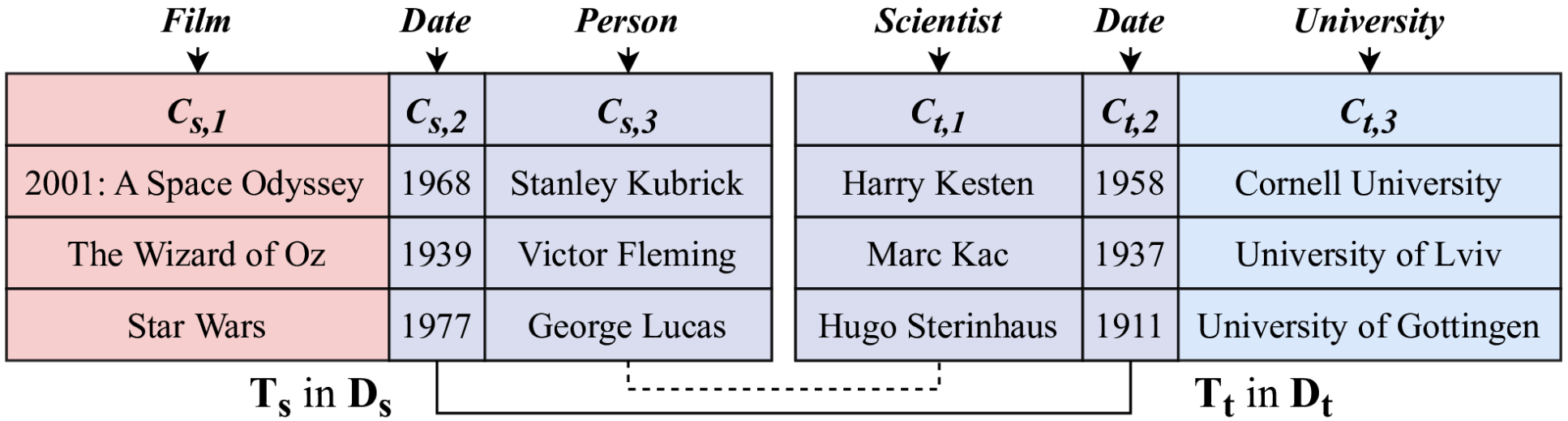
- The emergence of LMs in data labeling opens up opportunities for utilizing LMs for DQ.
  - Pros:
    - Low annotation cost.
  - Cons:
    - Require annotated fine-tuning data for LMs (upon new data lakes).
  - Research Opportunities:
    - How to reduce the labeled training data required for LMs on 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 (VLDB 2025 submitted)**
- Focus on enhancing cross-domain tabular data labeling with the interaction of the world model and pre-trained models.



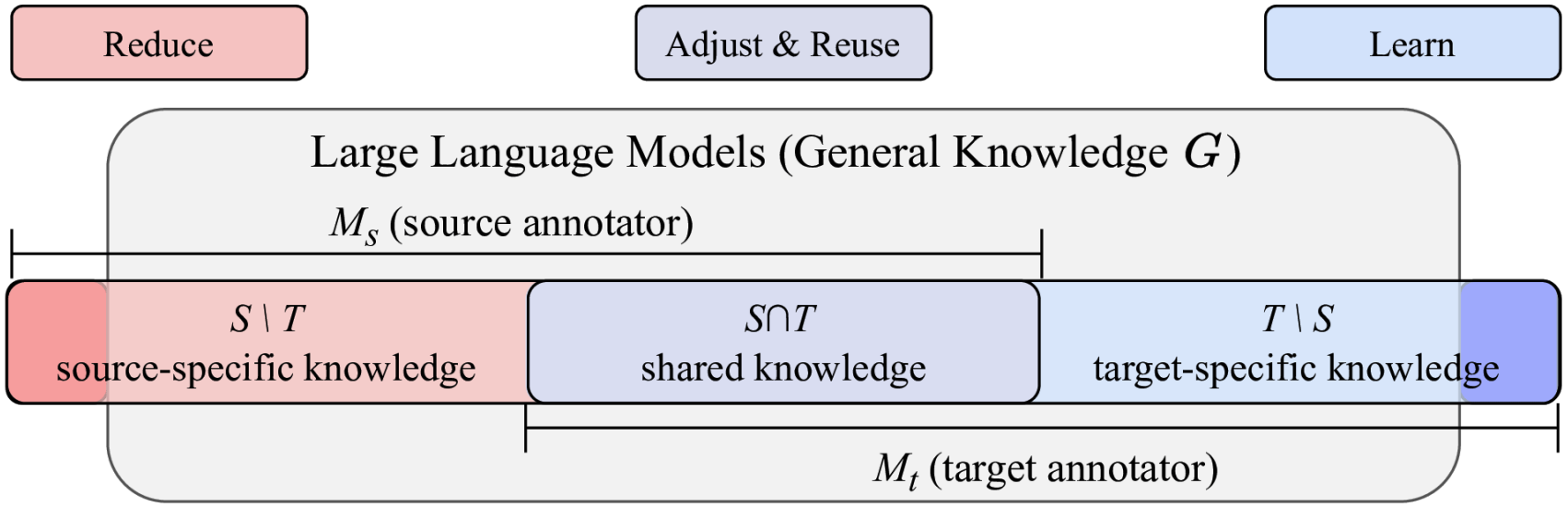
(a) Sample Source and Target Data Lake Tables

[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



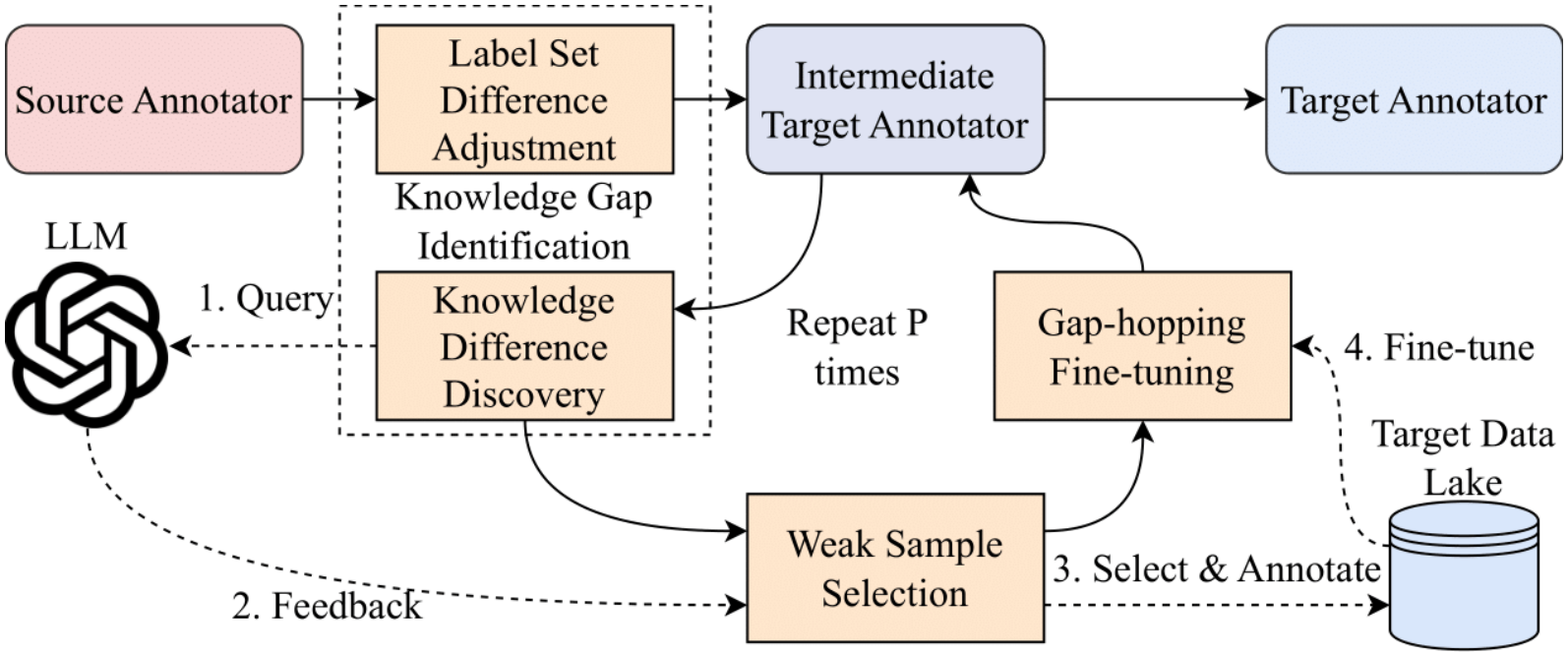
(b) Connections among Source/Target Annotators and LLMs

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.

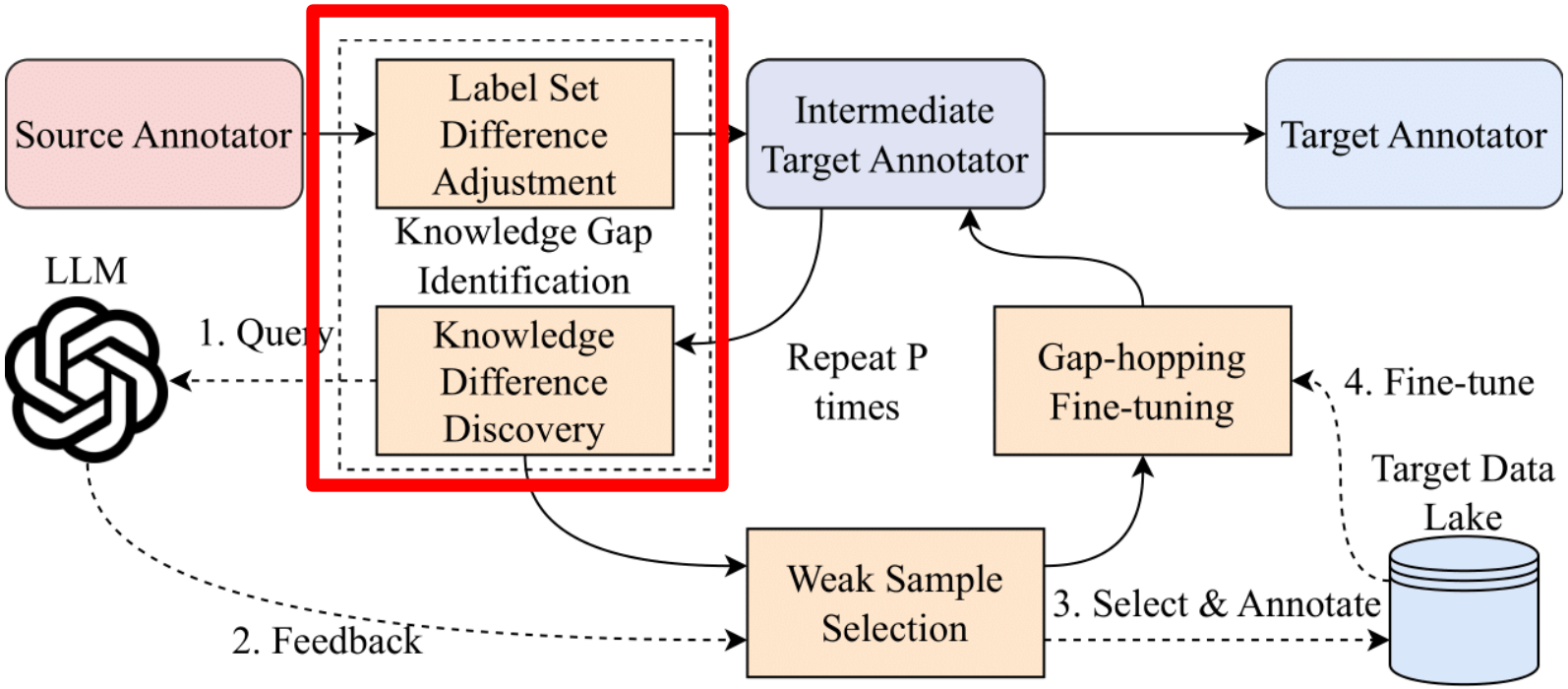
Low-resource LMs – methodology overview

- 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



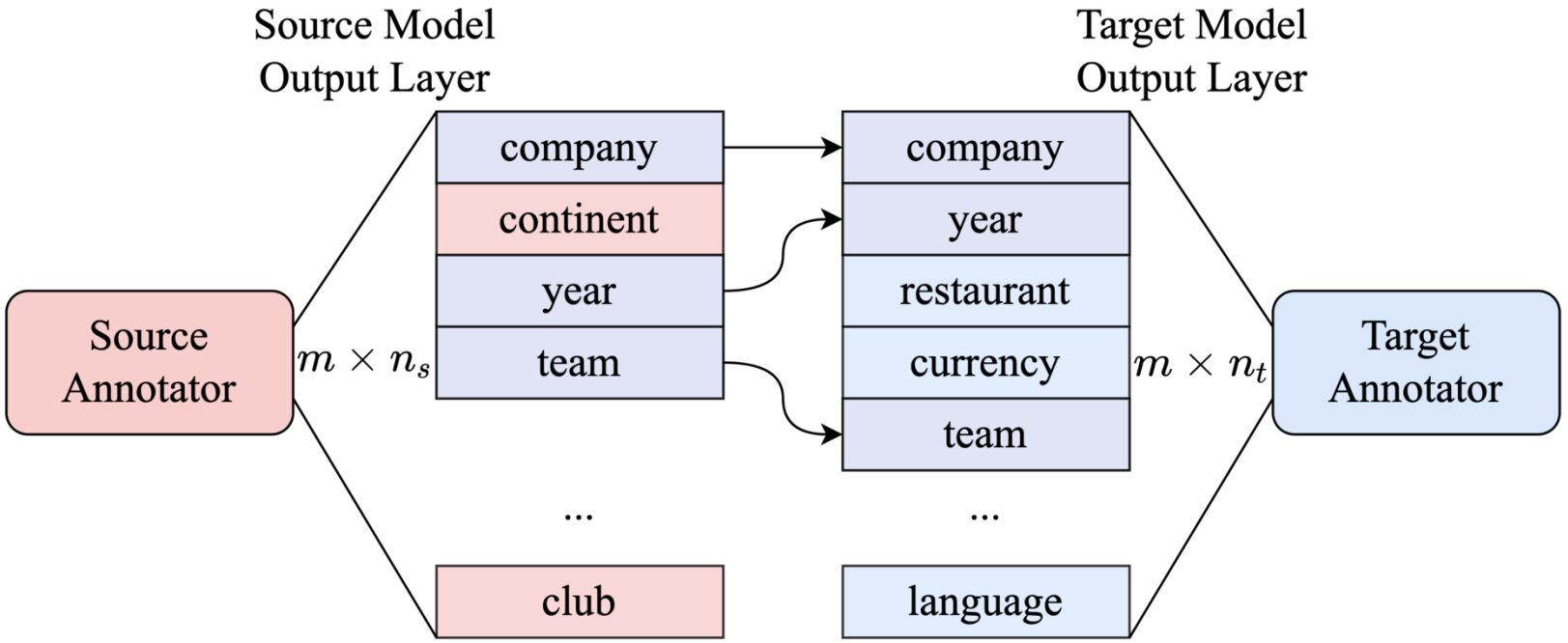
Low-resource LMs – methodology overview

- 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



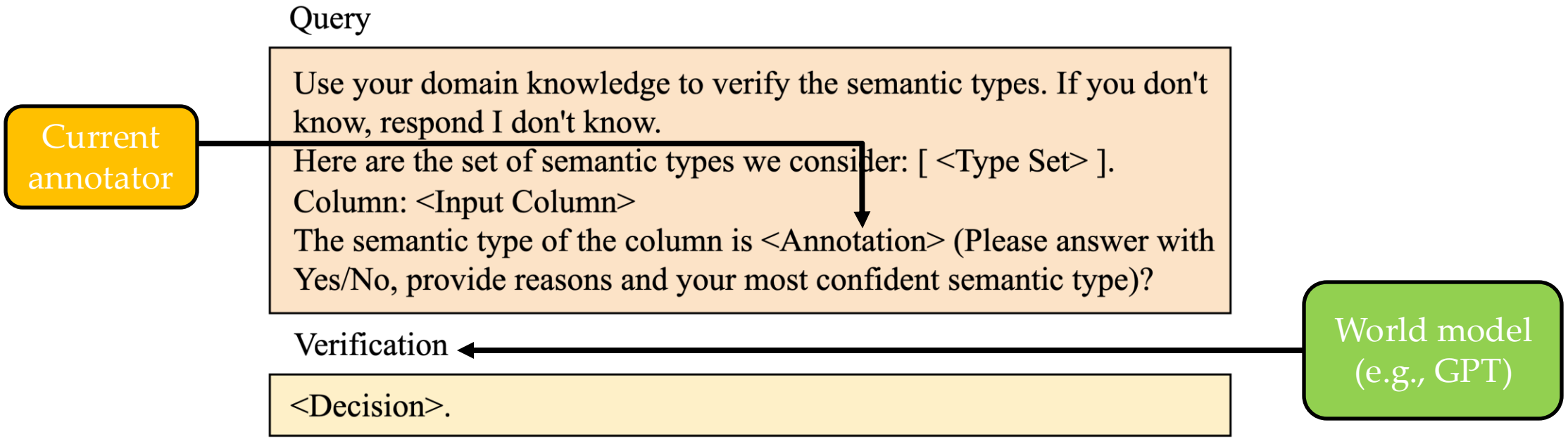
Low-resource LMs – methodology

- Label set difference adjustment



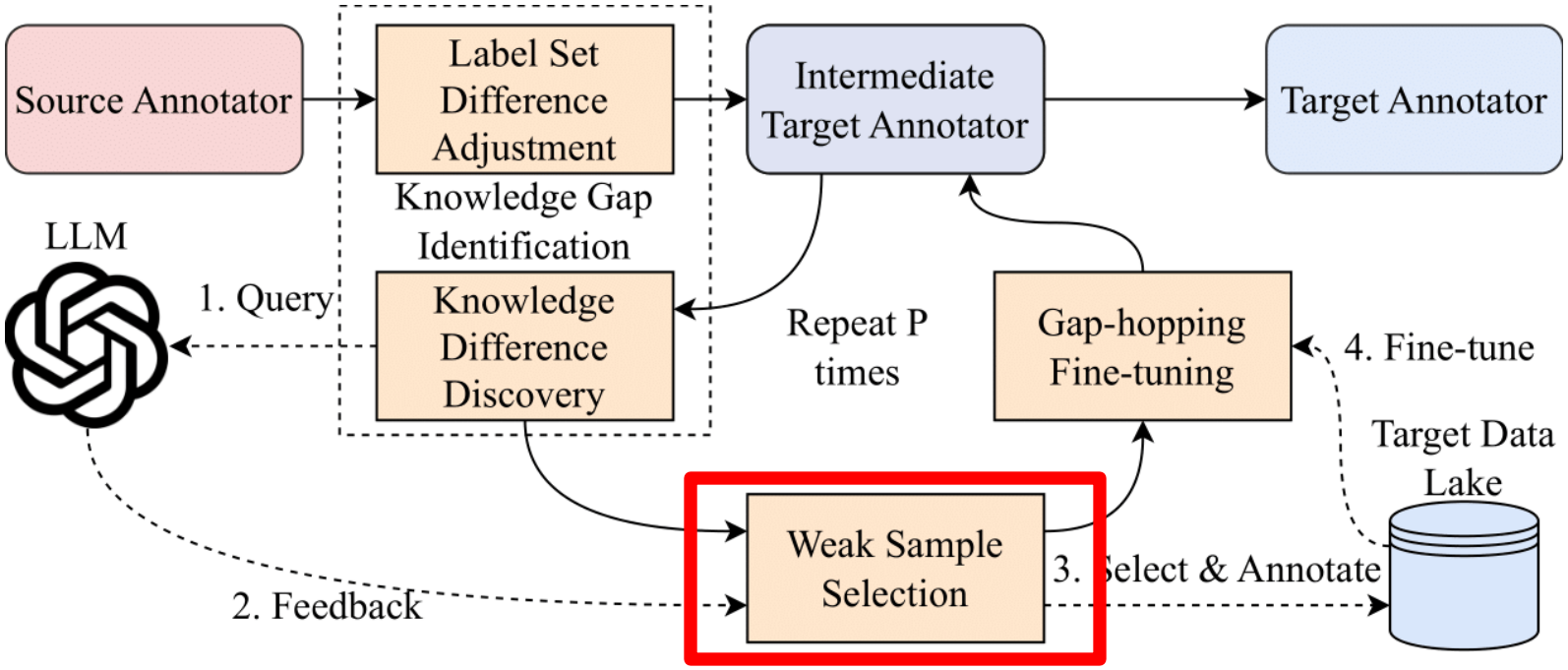
Low-resource LMs – methodology

- Knowledge difference discovery



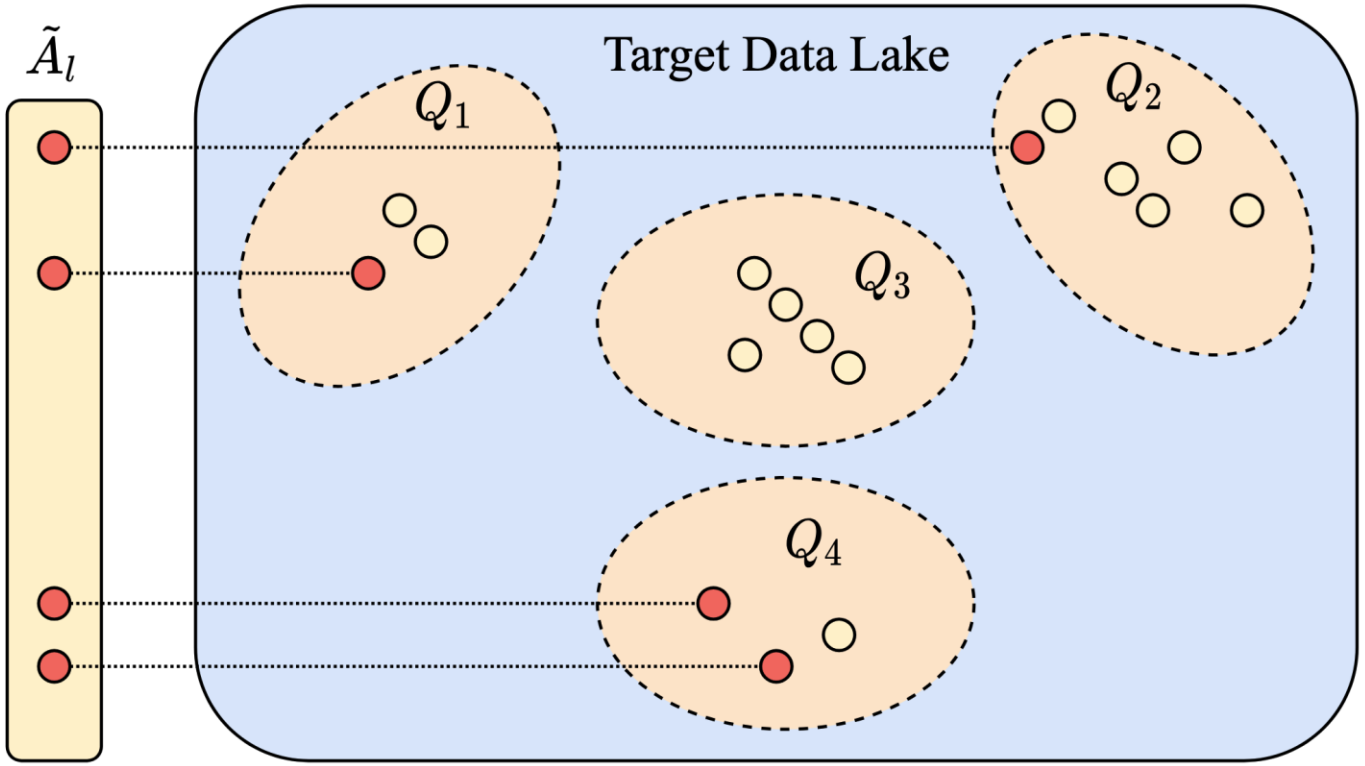
Low-resource LMs – methodology overview

- 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



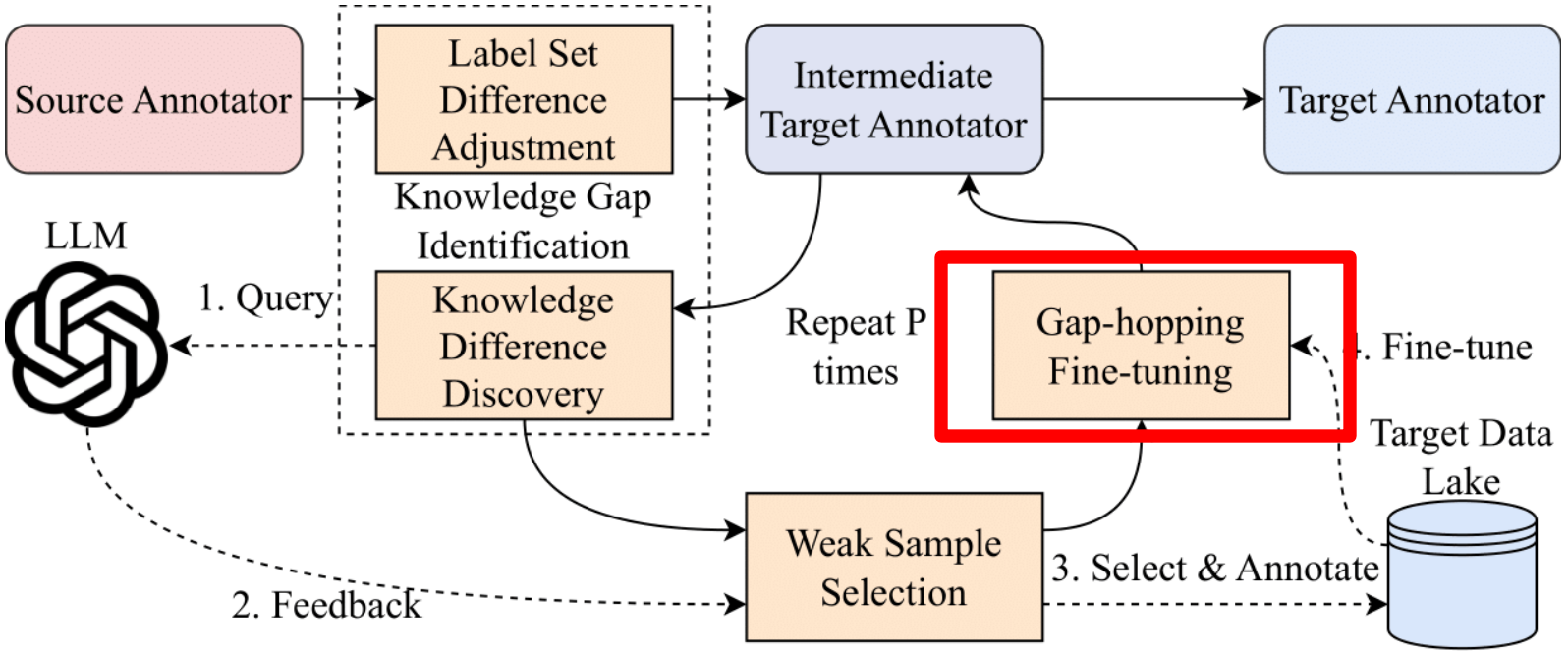
Low-resource LMs – methodology

- Weak Sample Selection



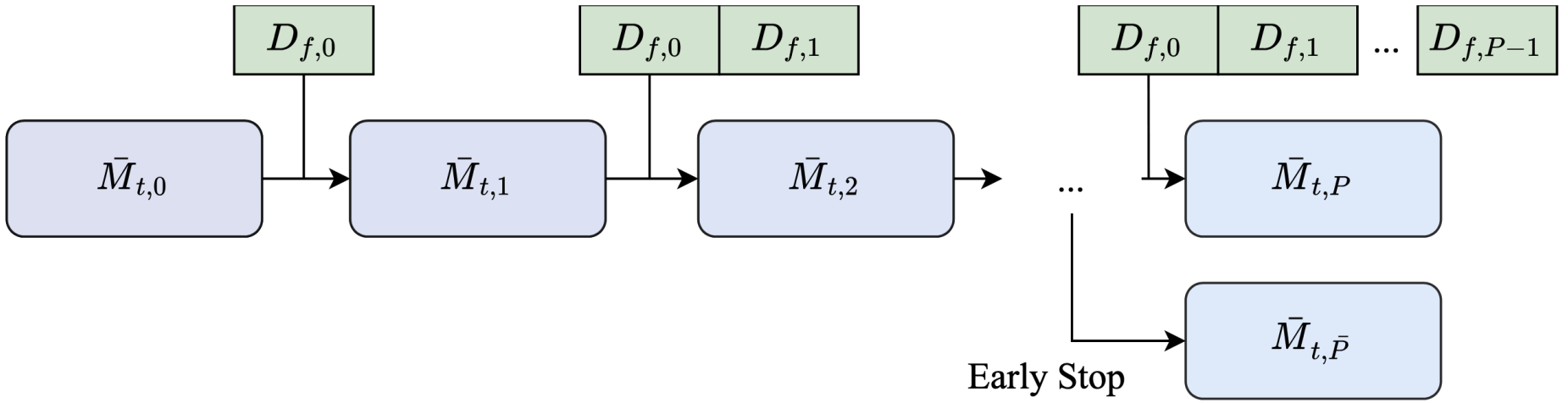
Low-resource LMs – methodology overview

- 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 – methodology

- Incremental Gap-hopping Fine-tuning



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

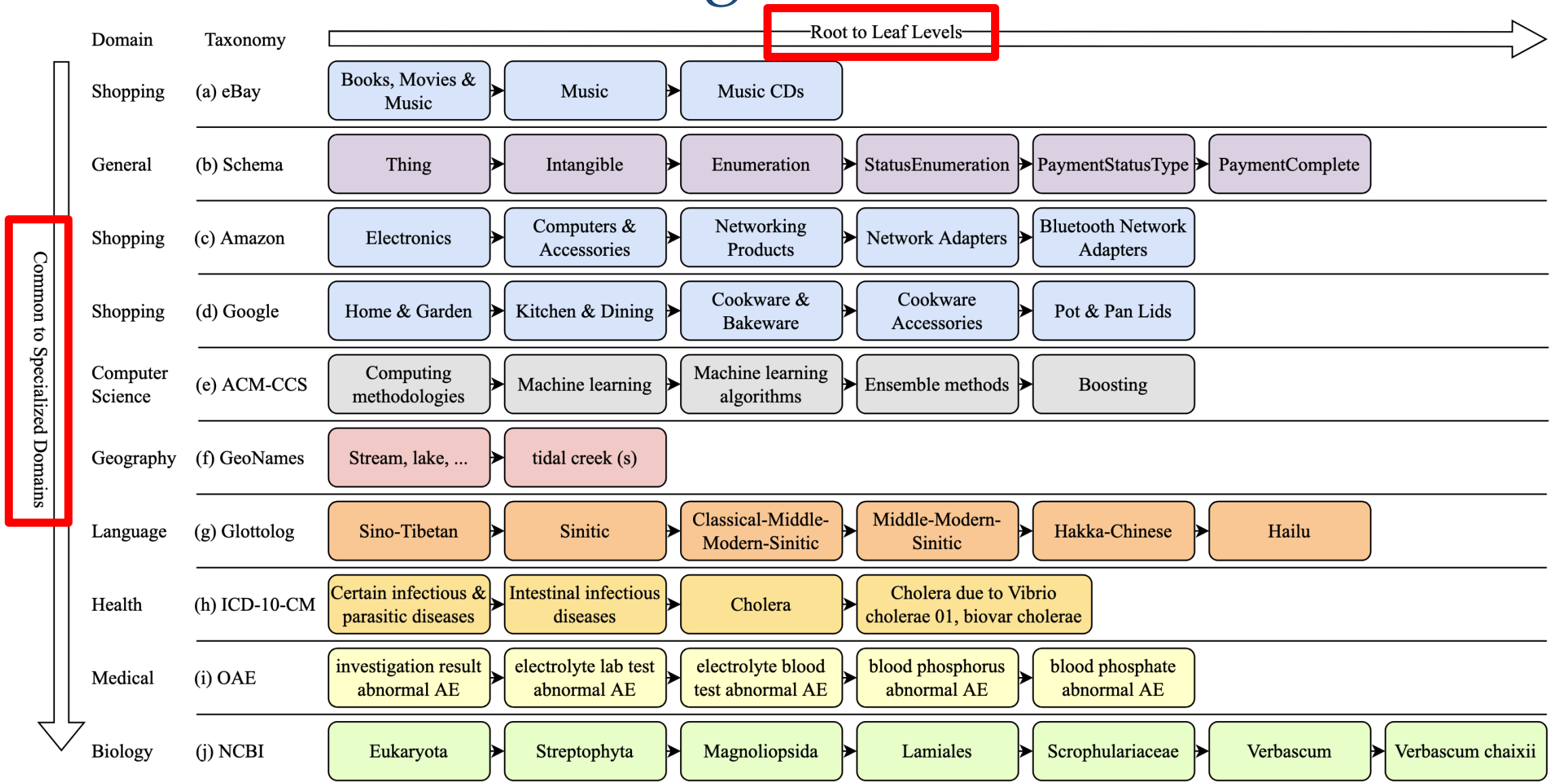
- Background
- LM4DQ
  - Past: Crowd-sourced / Human-in-the-loop
  - Status-quo: Pre-train+fine-tune LMs
  - Status-quo: Low-resource LMs
  - Future: Zero-shot LMs
- Future Vision and Opportunities
  - Preliminary study on DQ4LM
  - LM4DQ and DQ4LM

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).
- 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 LMs perform zero-shot data labeling on taxonomy data?

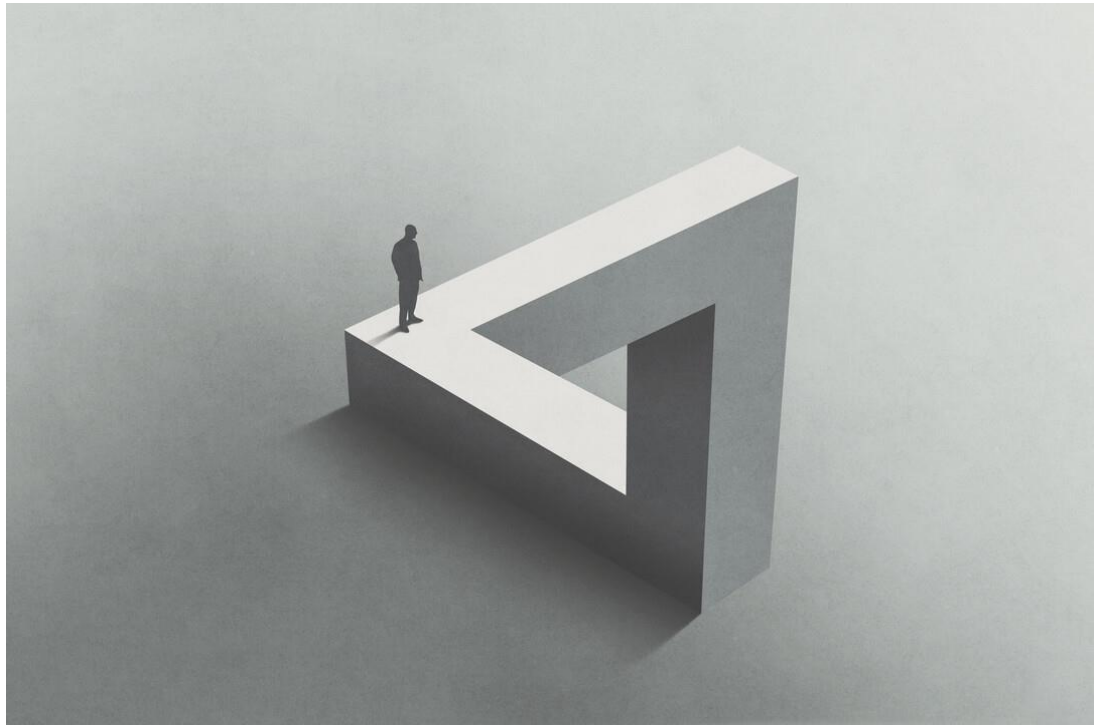
[5] Y. Sun, et al., "Are Large Language Models a Good Replacement of Taxonomies?," *Proceedings of the VLDB Endowment*, vol. 17, no. 11, pp. 2919–2932, Aug. 2024, doi:https://doi.org/10.14778/3681954.3681973.
[6] Andreas, "Taxonomy: Tracing Its Greek Roots to Modern Biological Classification - U speak 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 - background



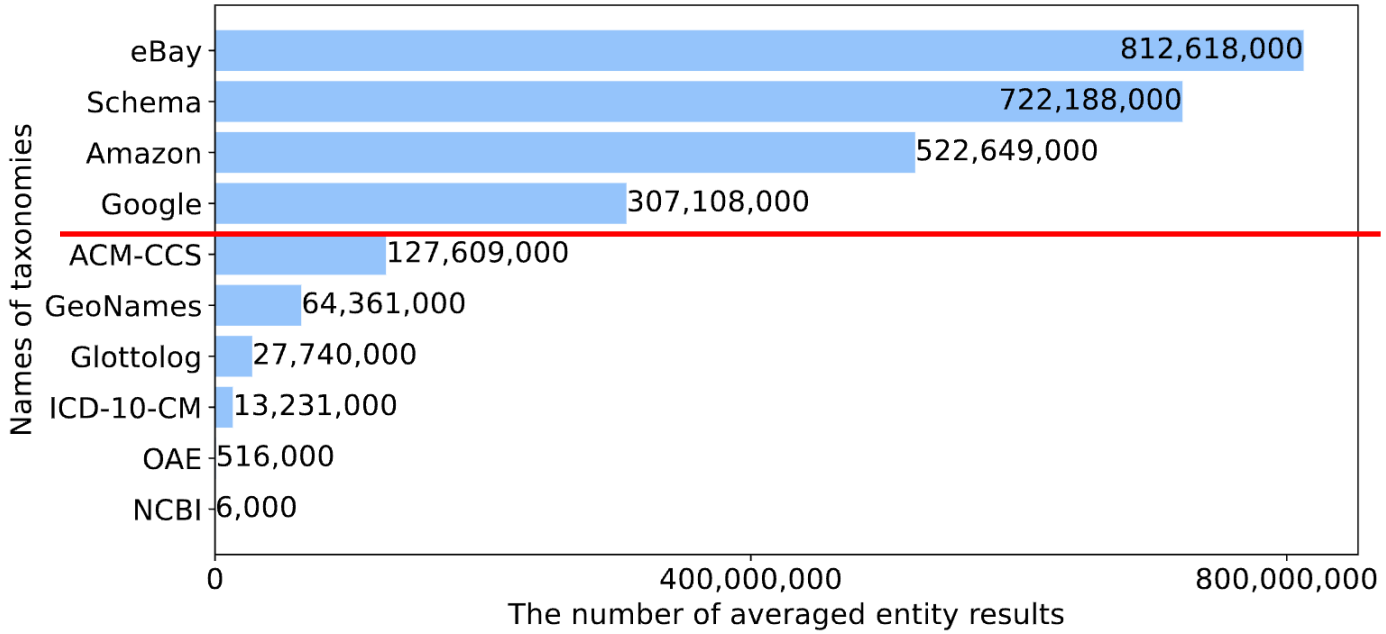
Zero-shot LMs - background

- The importance of the study is three-fold:
- (1) Industrial users can understand if constructing and maintaining traditional taxonomies is worth investing in;
- (2) LLM developers can learn about the pros and cons of their models in taxonomies and improve accordingly to help users better perform taxonomy-related tasks with LLMs; and
- (3) Database researchers can innovate on the novel forms of taxonomy structures, and explore meaningful research problems/application domains that boost the reasoning of LLMs.



Zero-shot LMs – 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



Zero-shot LMs – question template

- 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)

Zero-shot LMs – question set

- 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

Zero-shot LMs – LLMs

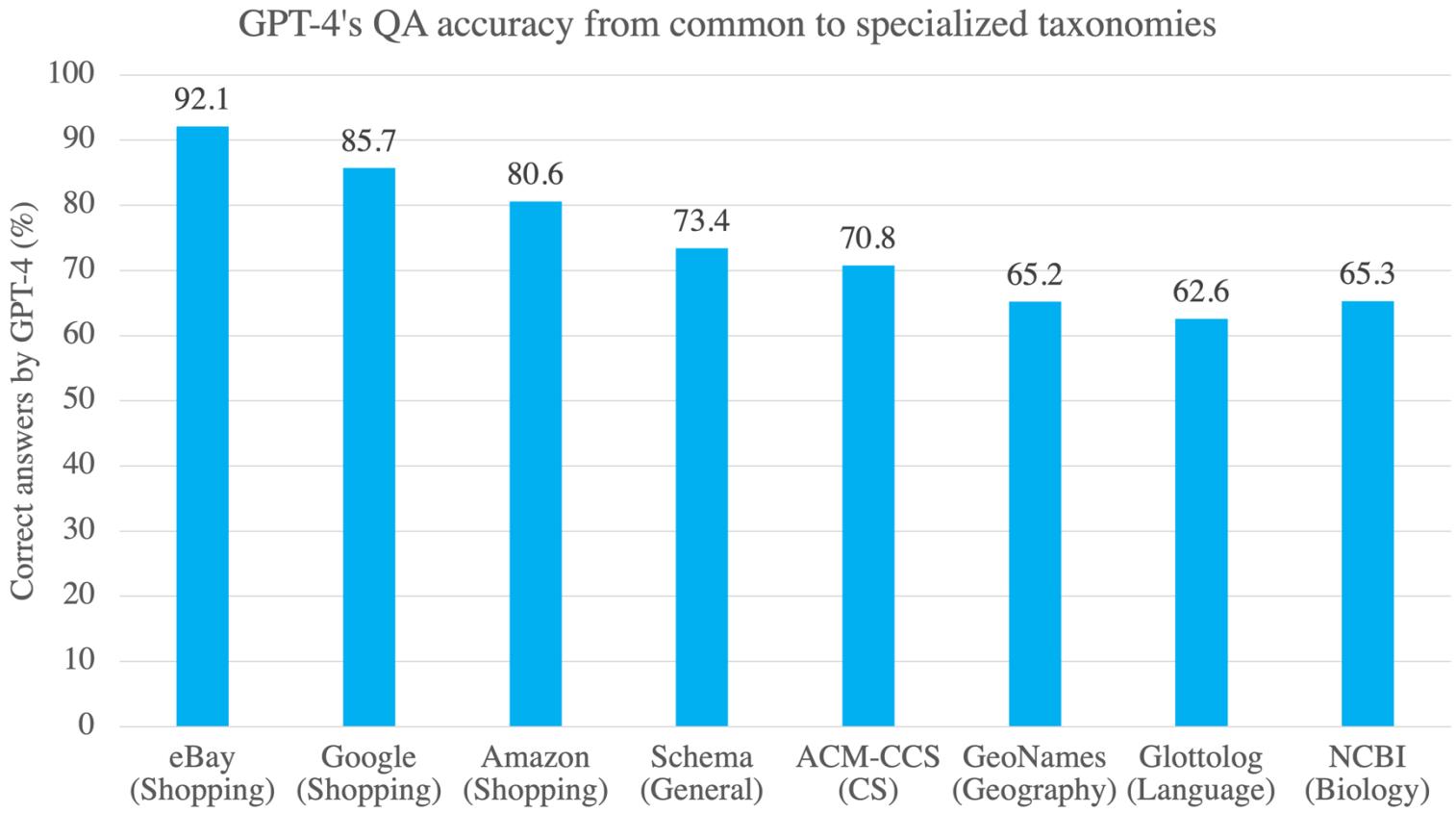
- LLMs considered:
  - Open-source:
    - Llama-2s: 7B, 13B, 70B
    - Llama-3s: 8B, 70B
    - Flan-T5s: 3B, 11B
    - Falcons: 7B, 40B
    - Vicunas: 7B, 13B, 33B
    - Mistrals: 7B, 8*7B
  - Closed-source:
    - GPTs: GPT 3.5, GPT 4
    - Claude-3-Opus
  - Fine-tuned:
    - LLMs4OL

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 of different taxonomies.

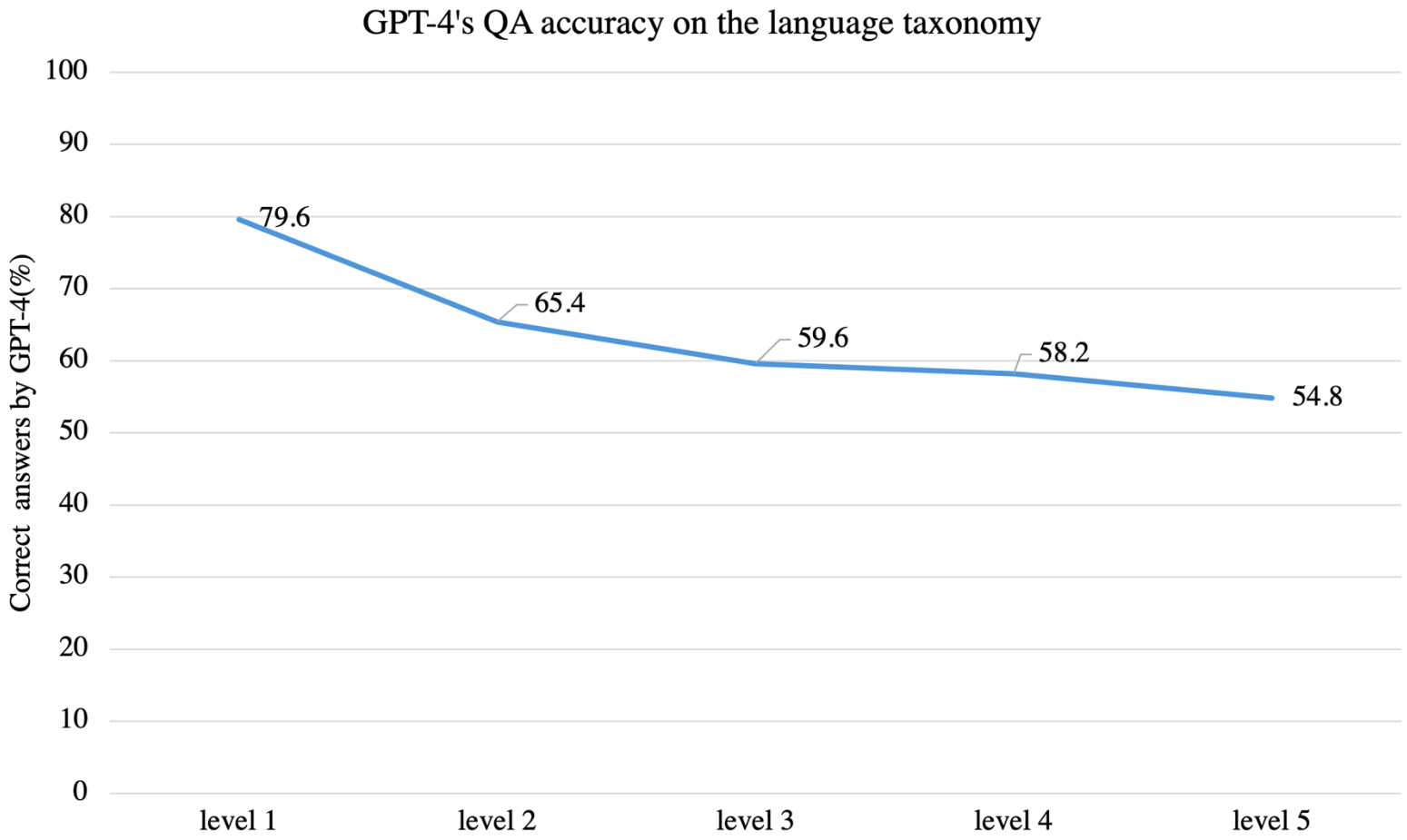
Zero-shot LMs - experiments

- 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%.



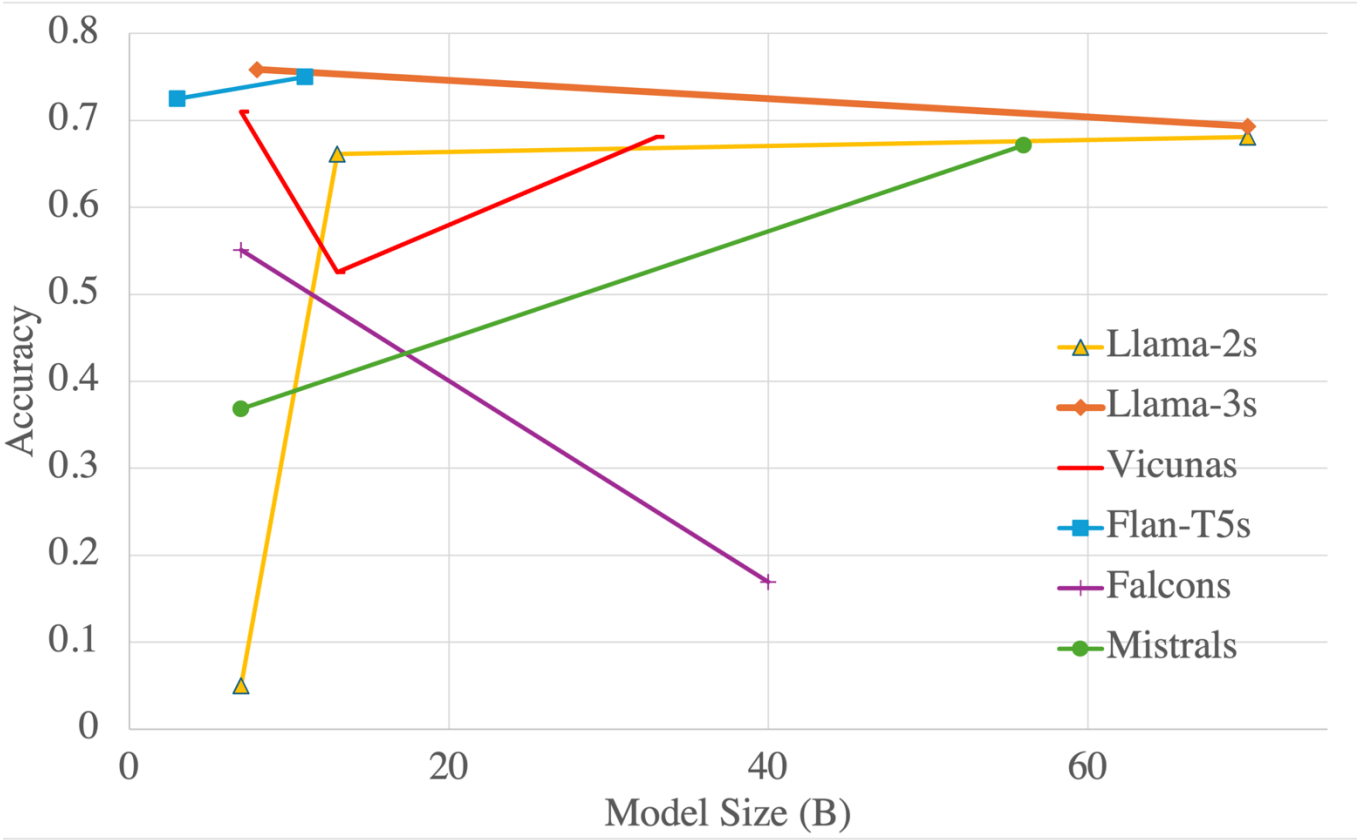
Zero-shot LMs - experiments

- 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).



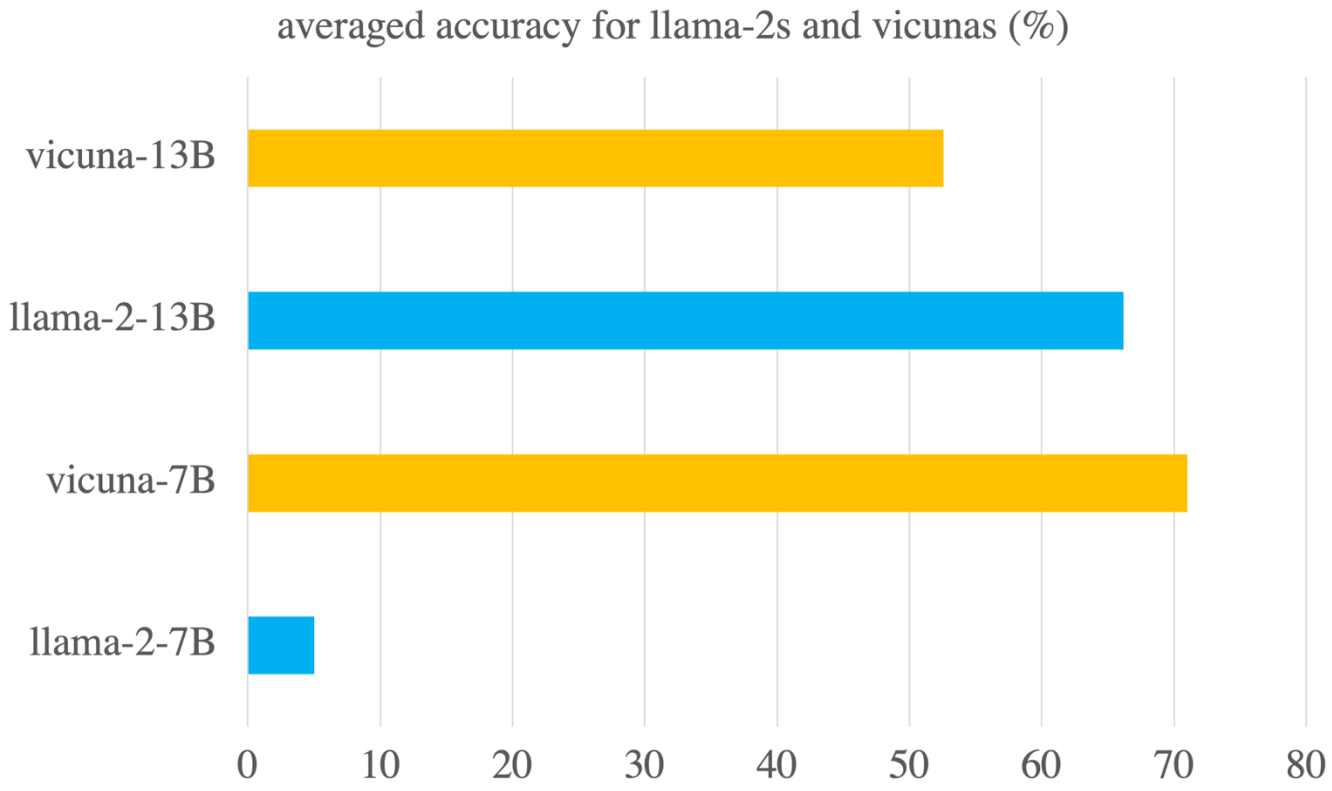
Zero-shot LMs - experiments

- 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.



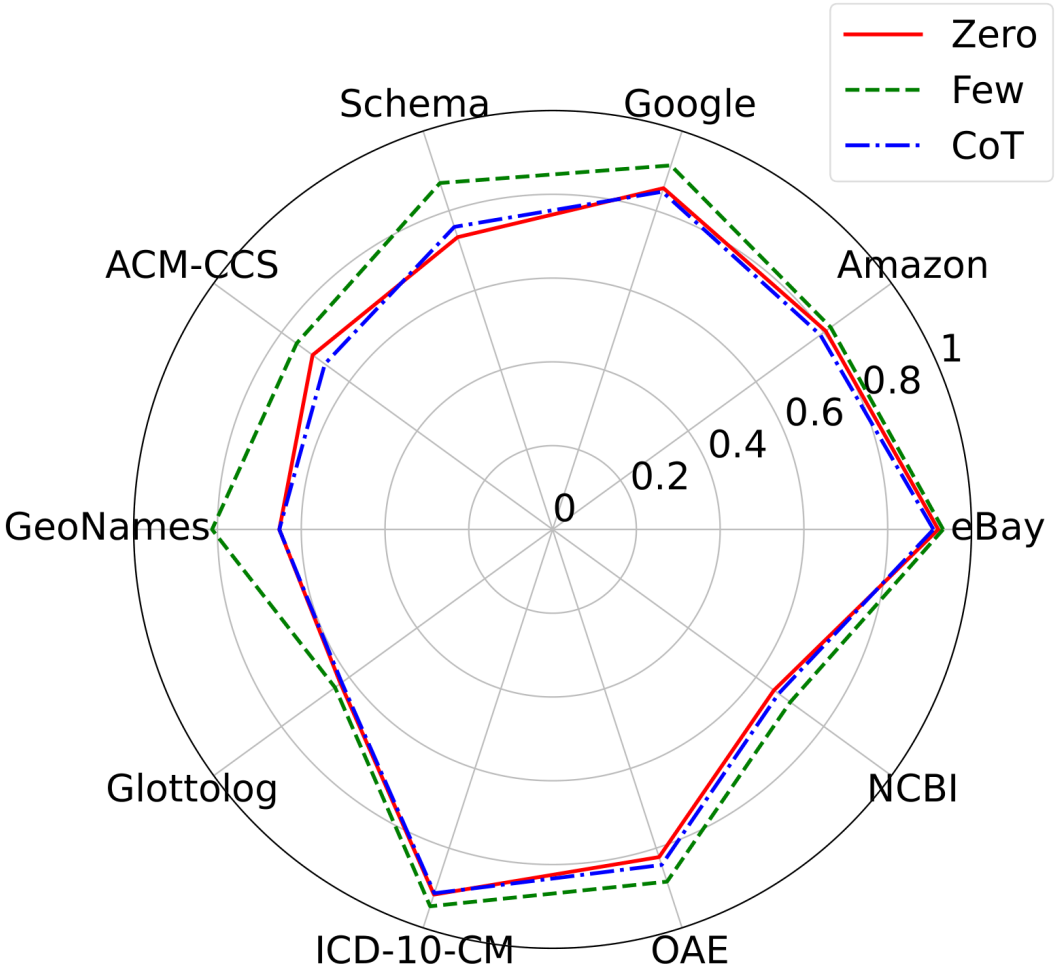
Zero-shot LMs - experiments

- 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.



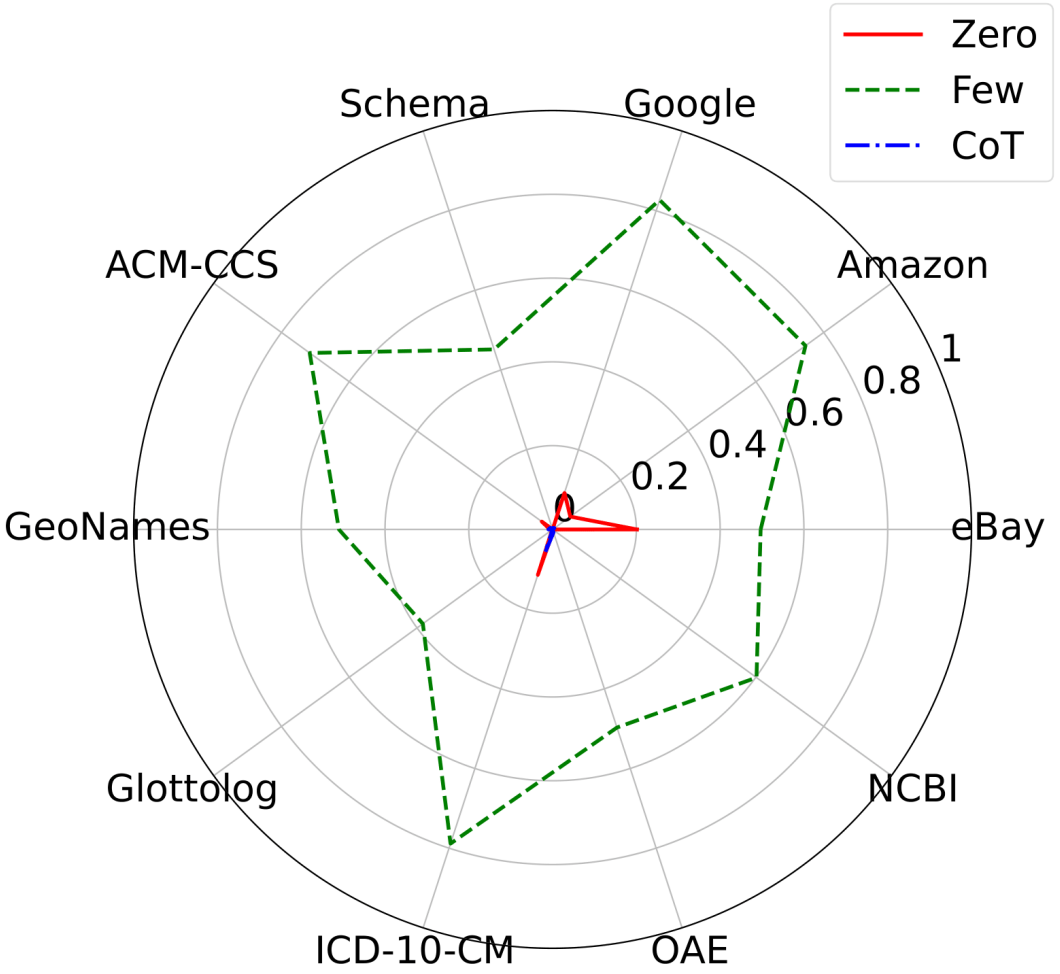
Zero-shot LMs - experiments

- 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.



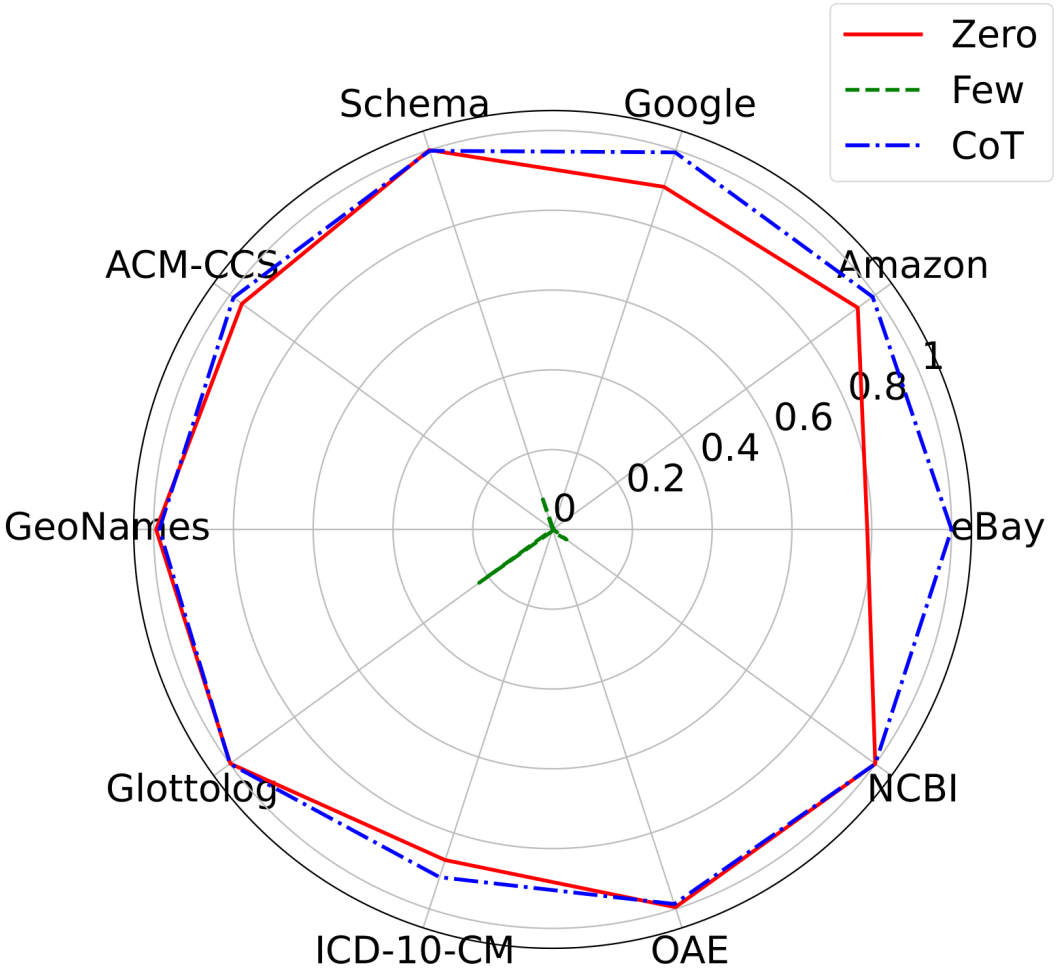
Zero-shot LMs - experiments

- 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.



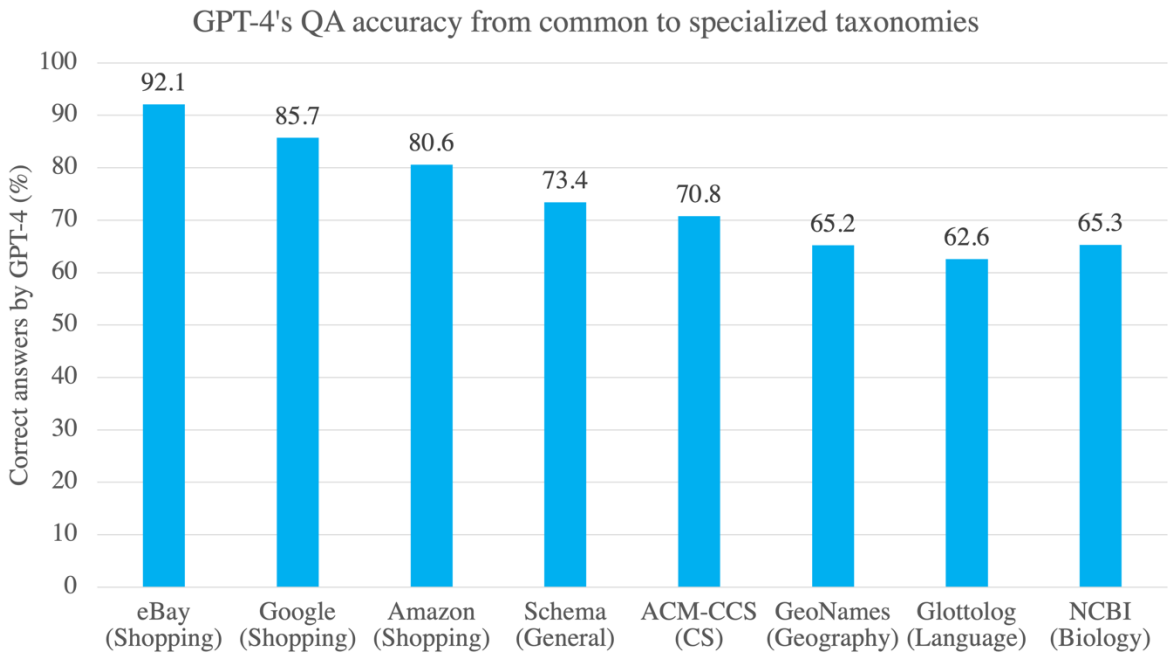
Zero-shot LMs - experiments

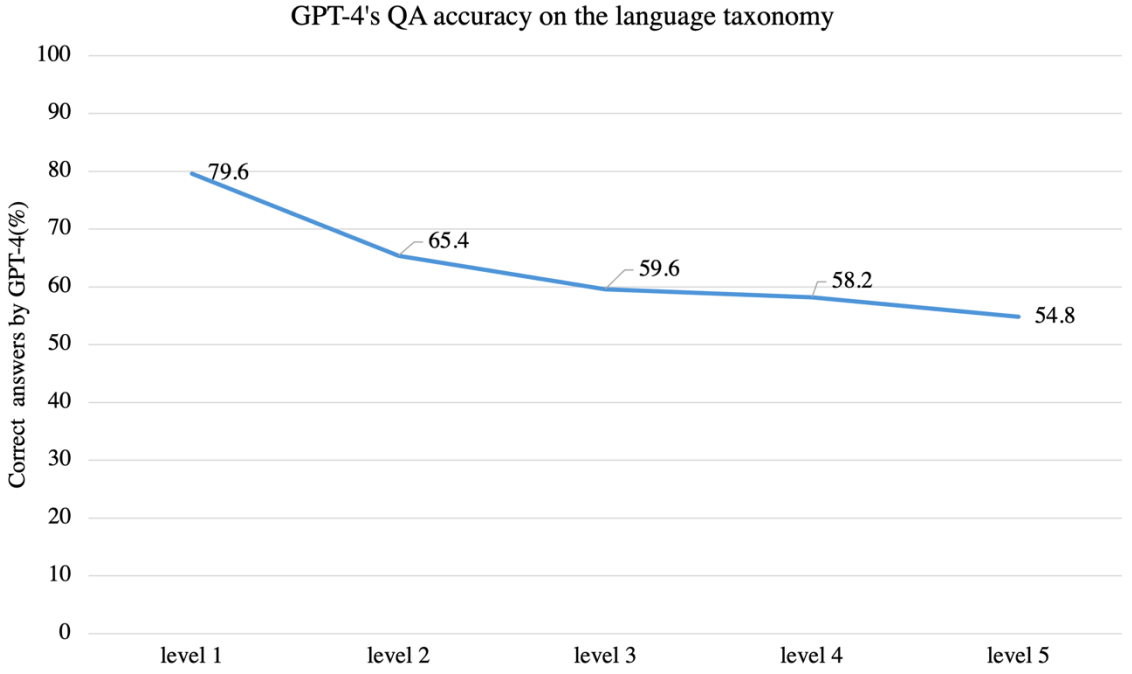
- RQ4: Do different prompting settings influence the performance?
- The performance changes of best LLMs brought by few-shot and Chain-of-Thoughts prompting settings are minimal. The main effect of prompting settings is to influence the miss rates instead of the accuracy of LLMs.



Zero-shot LMs – experiments 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.





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

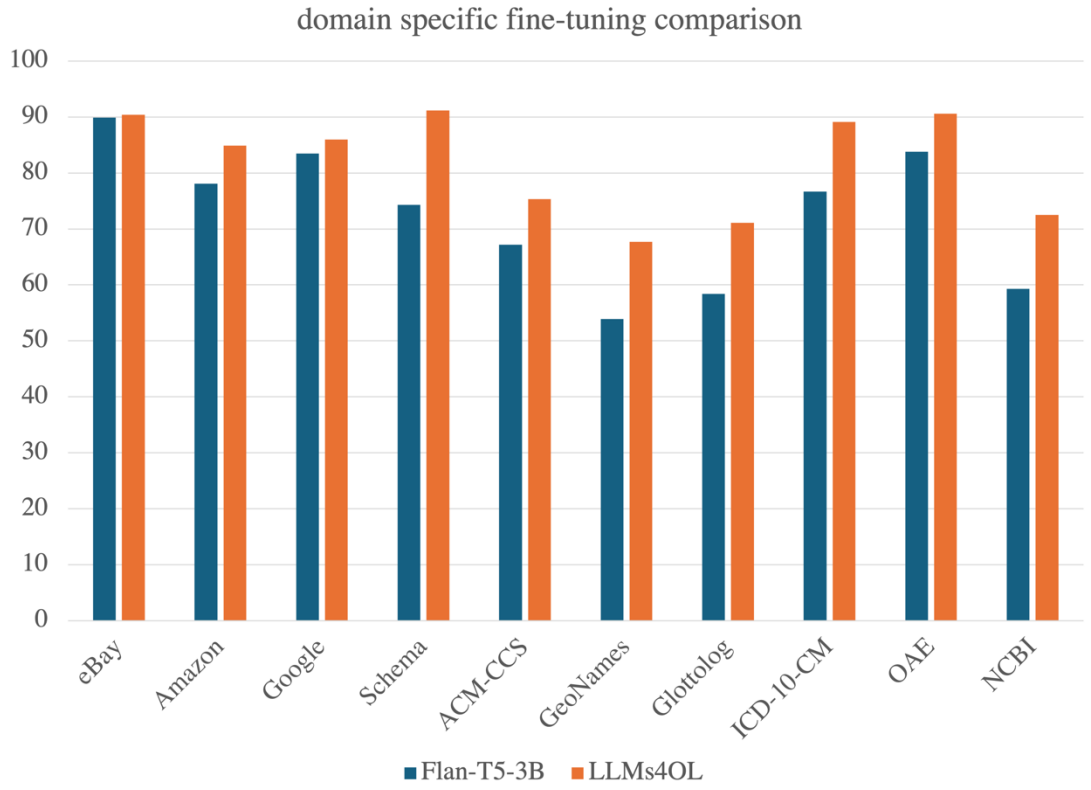
How does DQ influence LMs?

- 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]
- DQ4LM:
  - Improved accuracy, generalizability, …
  - Reduced hallucination, bias, …

[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

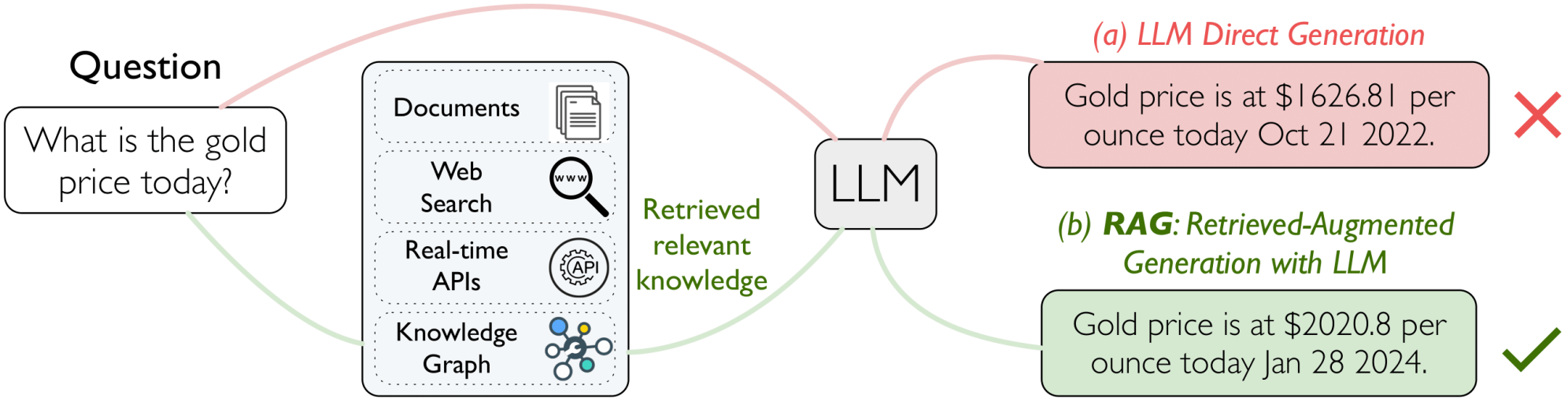
- **Are Large Language Models a Good Replacement of Taxonomies? (VLDB 2024)**
- Insights: High-quality training data can benefit the performance of LMs through fine-tuning.



[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 (accepted by NeurIPS 2024, which is also used for hosting KDD Cup 2024 in Barcelona)**
- Considered questions based on timeliness and difficulty level.
- Provided both KG and Web data sources.
- 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

CRAG – Comprehensive RAG Benchmark

Benchmark	Web retrieval	KG search	Mock API	Dynamic question	Torso and tail facts	Beyond Wikipedia	Question size
QALD-10 [35]	✗	✓	✗	✗	✗	✗	0.8K
MS MARCO [4]	✓	✗	✗	not explicitly	not explicitly	✓	100K
NQ [18]	✓	✗	✗	not explicitly	not explicitly	✗	323K
RGB [6]	✓	✗	✗	✗	✗	✓	1K
FreshLLM [36]	✗	✗	✗	✓	✗	✓	0.6K
CRAG	✓	✓	✓	✓	✓	✓	4.4K

CRAG – Comprehensive RAG Benchmark

Dynamism	Finance	Sports	Music	Movie	Open	Total
Real-time	434 (42)	0 (0)	2 (0)	0 (0)	1 (0)	437 (10)
Fast-changing	204 (20)	275 (33)	40 (6)	17 (2)	28 (4)	564 (13)
Slow-changing	183 (18)	215 (26)	152 (24)	253 (22)	204 (26)	1,007 (23)
Static	218 (21)	343 (41)	430 (69)	855 (76)	555 (70)	2,401 (54)
All	1,039	833	624	1,125	788	4,409

Question type	Finance	Sports	Music	Movie	Open	Total
Simple	466 (45)	23 (3)	112 (18)	519 (46)	85 (11)	1,205 (27)
Simple w. condition	113 (11)	250 (30)	92 (15)	112 (10)	122 (15)	689 (16)
Set	48 (5)	93 (11)	72 (12)	104 (9)	86 (11)	403 (9)
Comparison	146 (14)	85 (10)	102 (16)	105 (9)	98 (12)	536 (12)
Aggregation	69 (7)	137 (16)	96 (15)	71 (6)	116 (15)	489 (11)
Multi-hop	86 (8)	64 (8)	55 (9)	90 (8)	87 (11)	382 (9)
Post-processing heavy	26 (3)	24 (3)	26 (4)	28 (2)	76 (10)	180 (4)
False Premise	85 (8)	157 (19)	69 (11)	96 (9)	118 (15)	525 (12)
All	1,039	833	624	1,125	788	4,409

CRAG – Comprehensive RAG Benchmark

- **Task 1:** We provide up to five web pages for each question. These web pages are likely, but not guaranteed, to be relevant to the question. This task aims to test the answer-generation capability of a RAG system.
- **Task 2:** We provide mock APIs to access information from underlying mock KGs. The mock KGs store structured data relevant to the questions; answers to the questions may or may not exist in the mock KGs. This task tests how well a RAG system 1) queries structured data sources and 2) synthesizes information from different sources.
- **Task 3:** Similar to Task 2, Task 3 also provides both web search results and mock APIs as candidates for retrieval but provides 50 web pages, instead of 5, as candidates. Task 3 tests how a RAG system ranks a larger number of retrieval results.

CRAG – Comprehensive RAG Benchmark

- Metrics: We classify the answers into the following types (four types for human-eval, and three types for auto-eval)

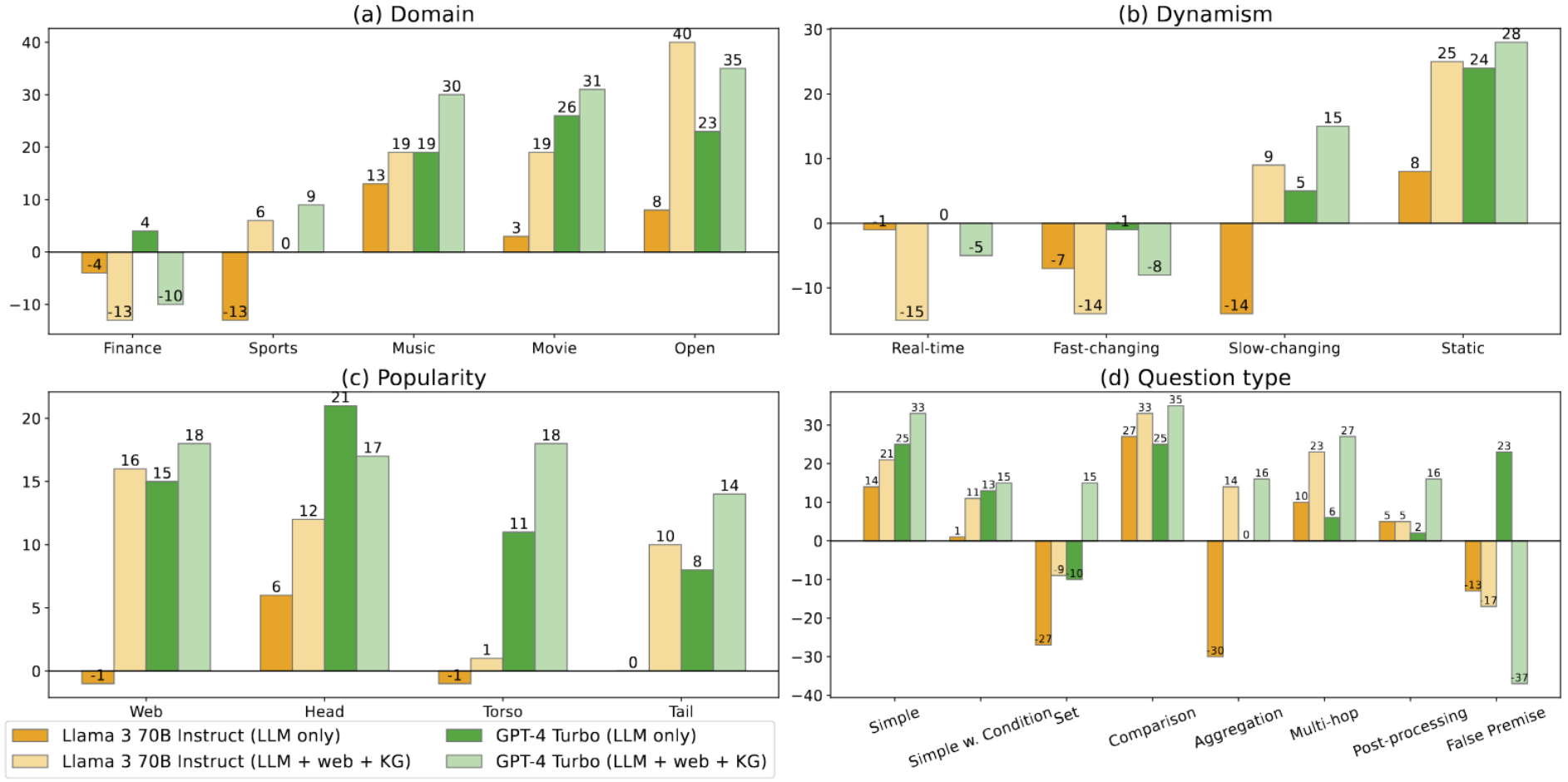
Accurate (1)

- **Perfect (1).** The response correctly answers the user's question and contains no hallucinated content.
- **Acceptable (0.5).** The response provides a useful answer to the user's question but may contain minor errors that do not harm the usefulness of the answer.
- **Missing (0).** The response is "I don't know", "I'm sorry I can't find …", a system error such as an empty response, or a request from the system to clarify the original question.
- **Incorrect (-1).** The response provides wrong or irrelevant information to answer the user's question.

CRAG – Comprehensive RAG Benchmark

	Model	Accuracy	Hallucination	Missing	Truthfulness$_a$
LLM only	Llama 3 70B Instruct	32.3	28.9	38.8	3.4
	GPT-4 Turbo	**33.5**	**13.5**	53.0	**20.0**
Task 1	Llama 3 70B Instruct	35.6	31.1	33.3	4.5
	GPT-4 Turbo	**35.9**	**28.2**	35.9	**7.7**
Task 2	Llama 3 70B Instruct	37.5	29.2	33.3	8.3
	GPT-4 Turbo	**41.3**	**25.1**	33.6	**16.2**
Task 3	Llama 3 70B Instruct	40.6	31.6	27.8	9.1
	GPT-4 Turbo	**43.6**	**30.1**	26.3	**13.4**

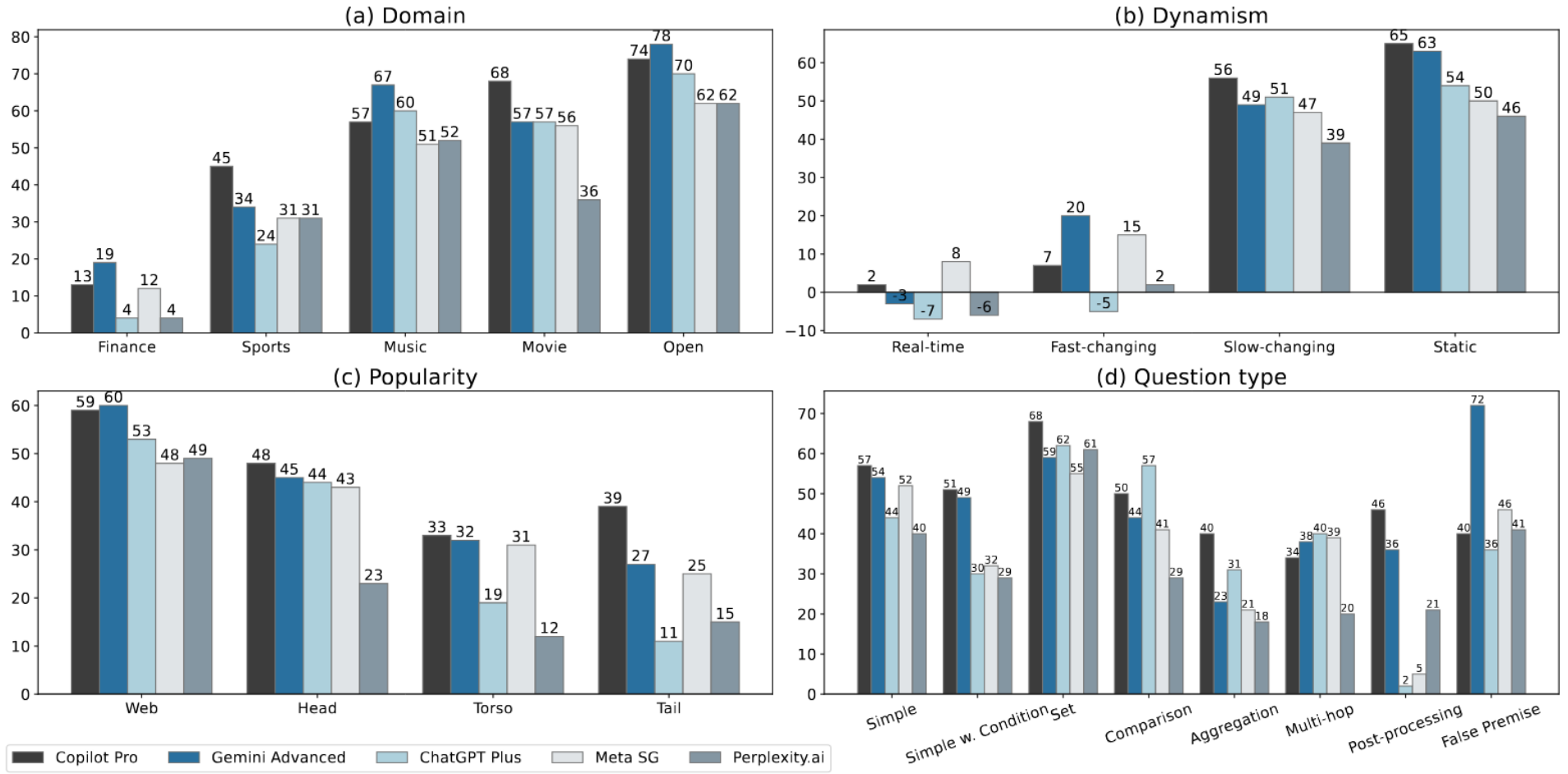
CRAG – Comprehensive RAG Benchmark



CRAG – Comprehensive RAG Benchmark

System	Perfect	Acc.	Hall.	Miss.	Truth$_h$	Latency (ms)
Copilot Pro	**62.6**	11.7	17.9	7.8	**50.6**	11,596
Gemini Advanced	60.8	10.1	16.6	12.5	49.3	5,246
ChatGPT Plus	59.8	**13.3**	25.0	1.9	41.5	6,195
Meta SG	52.5	9.7	**16.0**	21.8	41.4	**3,431**
Perplexity.ai	55.8	8.8	25.3	10.1	34.9	4,634

CRAG – Comprehensive RAG Benchmark



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).