

# Efficient Mixture of Experts based on Large Language Models for Low-Resource Data Preprocessing

Mengyi Yan<sup>1</sup> Yaoshu Wang<sup>2\*</sup> Kehan Pang<sup>1</sup> Min Xie<sup>2</sup> Jianxin Li<sup>1\*</sup>

<sup>1</sup>Beihang University      <sup>2</sup>Shenzhen Institute of Computing Sciences



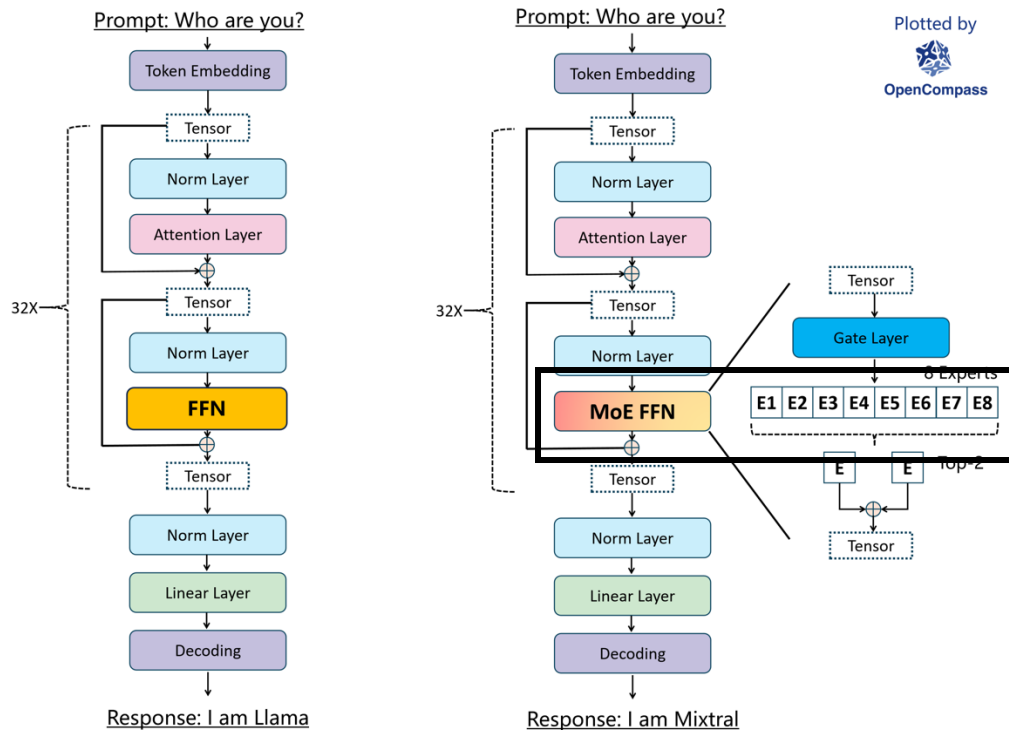
北京航空航天大学  
BEIHANG UNIVERSITY



深圳计算科学研究院  
Shenzhen Institute of Computing Sciences

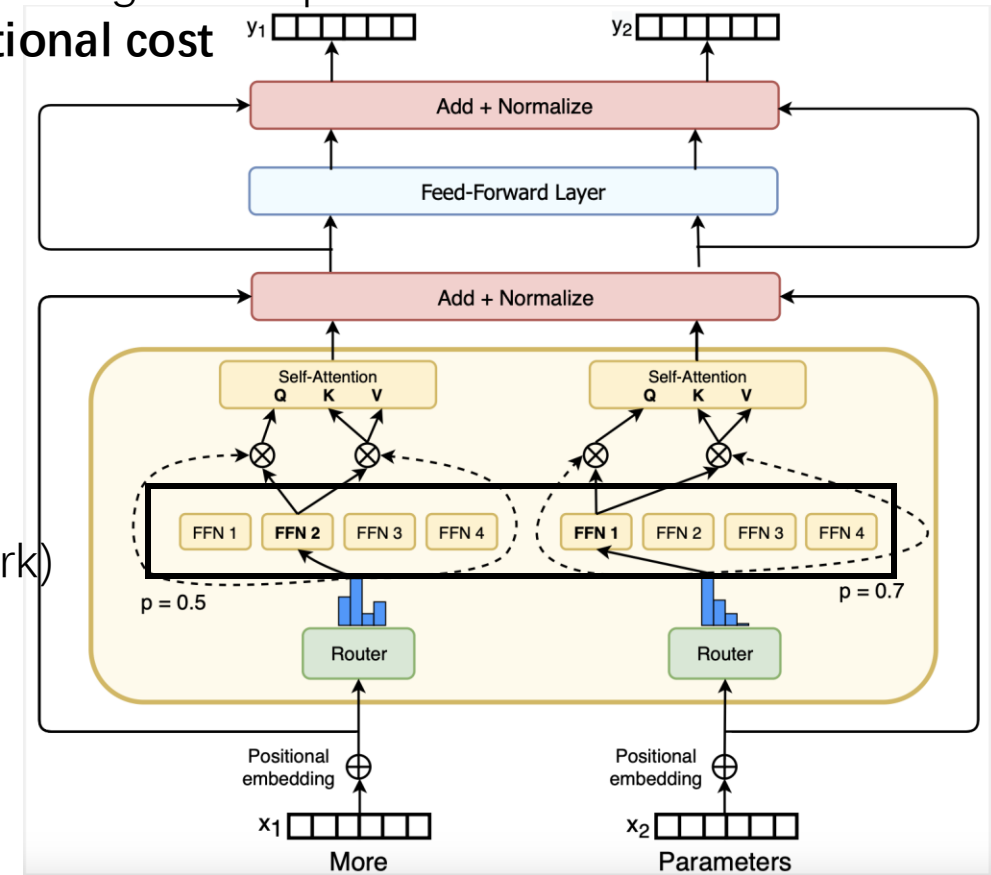
# Preliminary: Mixture-of-Experts(MoE) Architecture

- Contains multiple sub-model(**Multiple Expert Model**)
- Parameter is Sparse Activated (e.g. Top-2 of a total of 8 experts per token for Mixtral)
- Different Expert model can **concentrate on different aspects**, improving overall performance on MTL.
- Scaling up **Model Parameter**, while reducing **inference computational cost**



Mixtral<sup>1</sup>

Gated Network  
(a.k.a Router Network)



Switch Transformer<sup>2</sup>


1. Mixtral of Experts. arXiv preprint arXiv:2401.04088 (2024).

2. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. The Journal of Machine Learning Research 23, 1 (2022), 5232–5270.

# Motivation: Parameter and Data-Efficient Learning in Data Preprocessing for DB

- In Database Area, Data preprocessing(DP) tasks and labeled data are **diverse** and **domain-specific**
  - Data preprocessing tasks vary by domain and require specialized feature engineering.
  - Generalization between task, data and models is challenging.
  - Manual data labeling is expensive and doesn't scale well across multiple domains.
- **Few-shot** and **cross-domain** learning are hard in data preprocessing.
  - The scarcity of labeled data hampers the training of **robust, generalizable** models.
  - **Transferring knowledge** across domains is difficult, affecting model adaptability.

### Entity Matching




**Entity Matching (EM)** Given a pair of tuples  $t_1, t_2$ , our task is to infer whether they refer to the same entity. Formulated as:

$$(Ins^{EM}, D^{EM}, (t_1, t_2)), C^{EM}, C^{EM} = \{match, mismatch\}$$

EM is a binary classification task.

Binary classification, requiring the model to have a complex and clear classification boundary.

### Data Cleaning




**Data Cleaning (DC)** Given a tuple  $t$  and an attribute  $a_i$ , the data cleaning over a relational table is a process that identifies and repairs such cell with the correct values, with a few annotated tuples  $D^{DC}$ . Formulated as:

$$(Ins^{DC}, D^{DC}, (t, a_i), C^{DC})$$

DC is an open-domain generation task, which means the output domain for  $C^{DC}$  has no limits.

Generation task, requiring the model to have the ability to induce and apply rules.

### Relation Extraction



**Relation Extraction (RE)** Given a web table  $T$  and a set of pre-defined knowledge graph (KG) relations  $\mathcal{R}$ , our task is to annotate a column  $h \in T$  with a KG relation type  $r \in \mathcal{R}$ , such that all entities in column  $h$  hold the same relation  $r$ . Formulated as:

$$(Ins^{RE}, D^{RE}, (T, h), C^{RE}), C^{RE} = \mathcal{R}$$

RE is a close-domain ranking task.

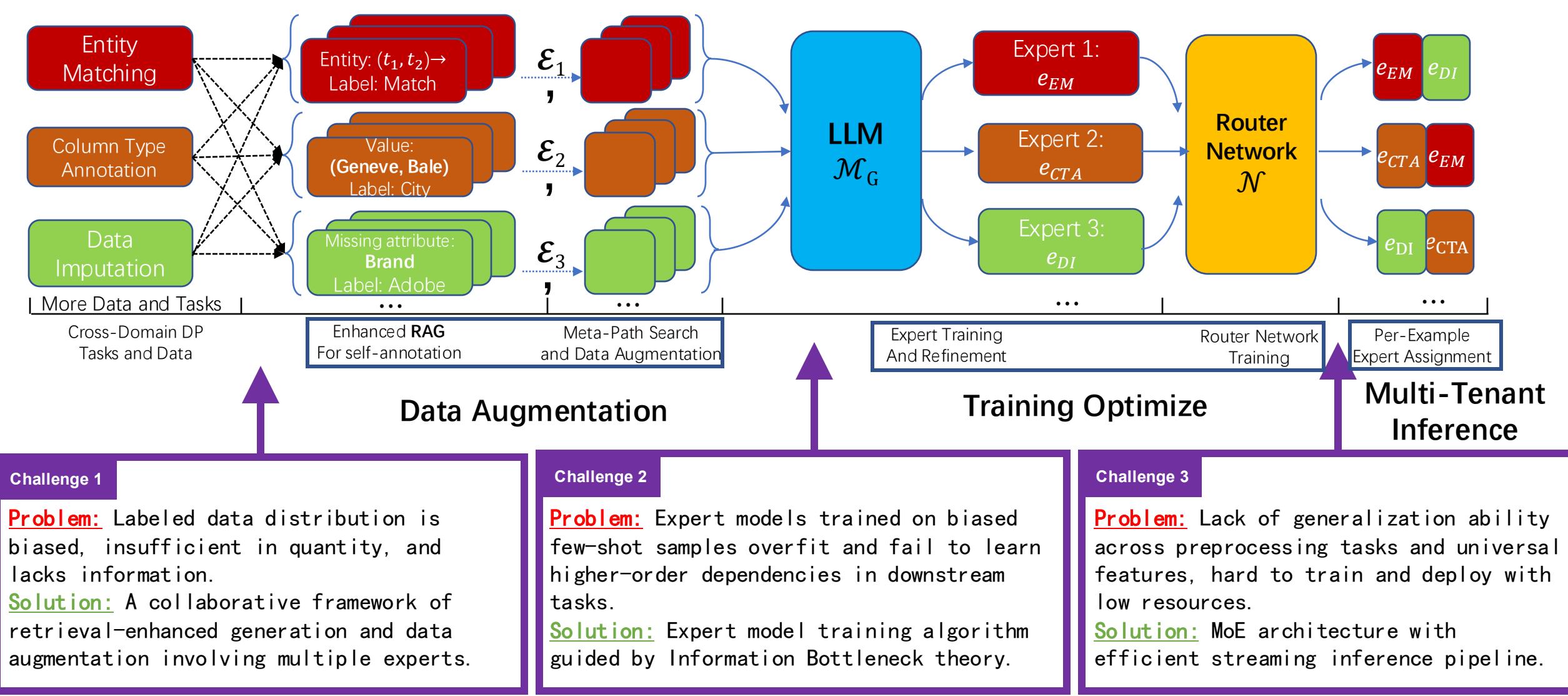
Ranking task, requiring the model to have retrieval-augmentation and classification capabilities.

1. Can different preprocessing tasks be unified into a common framework in **generative manner**? (**Multi-Task**)
2. Can **few-shot** labelling data from different tasks mutually boost each others performance? (**Data-Efficient**)
3. Can **Sparse-Activated Mixture of Expert** models(SMoE) outperform single dense models? (**Computational-Efficient**)

# Limitation: Low-Resource DP in Database

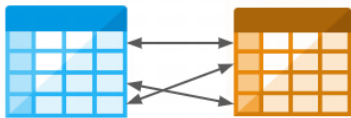
- **Expensive Labelling Cost:**
  - Few-shot and biased labeled data
- **Data Privacy Consideration:**
  - Deployed with local environment with offline model
- **Limitation of Computational Resource:**
  - Deployed in Consumer-level Hardware, e.g. RTX 3090/4090

# MELD: A Few-shot data preprocessing framework based on a mixture of experts



## Unified Various DP Task as Generation Task with LLM

### Entity Matching



**Entity Matching (EM)** Given a pair of tuples  $t_1, t_2$ , our task is to infer whether they refer to the same entity. Formulated as:

$$(Ins^{EM}, D^{EM}, (t_1, t_2), C^{EM}), C^{EM} = \{\text{match, mismatch}\}$$

### Classification Task

### Data Cleaning



**Data Cleaning (DC)** Given a tuple  $t$  and an attribute  $a_i$ , the data cleaning over a relational table is a process that identifies and repairs such cell with the correct values, with a few annotated tuples  $D^{DC}$ . Formulated as:

$$(Ins^{DC}, D^{DC}, (t, a_i), C^{DC})$$

### Generation Task

### Relation Extraction



**Relation Extraction (RE)** Given a web table  $T$  and a set of pre-defined knowledge graph (KG) relations  $\mathcal{R}$ , our task is to annotate a column  $h \in T$  with a KG relation type  $r \in \mathcal{R}$ , such that all entities in column  $h$  hold the same relation  $r$ . Formulated as:

$$(Ins^{RE}, D^{RE}, (T, h), C^{RE}), C^{RE} = \mathcal{R}$$

### Ranking Task

1. Roee et al. 2023. In-context learning creates task vectors. arXiv preprint arXiv:2310.15916 (2023).
2. Fan et al. 2024. Few-shot Adaptation of Multi-modal Foundation Models: A Survey. arXiv preprint arXiv:2401.01736 (2024).
3. Nishanth et al. 2023. On the benefits of learning to route in mixture-of-experts models. EMNLP 9376–9396.

# Theoretical Analysis

## Theorem 1: Task Subspace

- Different Task  $\mathcal{T}_i$  can be **compressed** to low-dimension Task Vector  $\theta_i$  for LLM

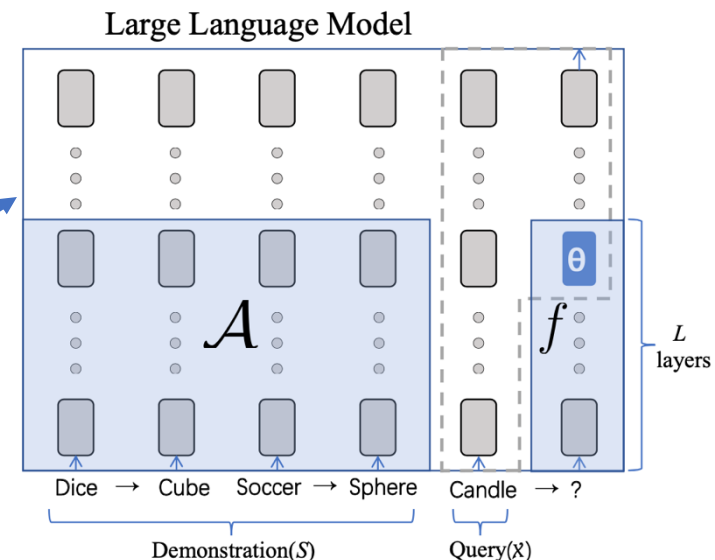
$\theta_i$  is in Low-Dimension  
Intrinsic Task Subspace  $V$

## Theorem 2: Error Bound

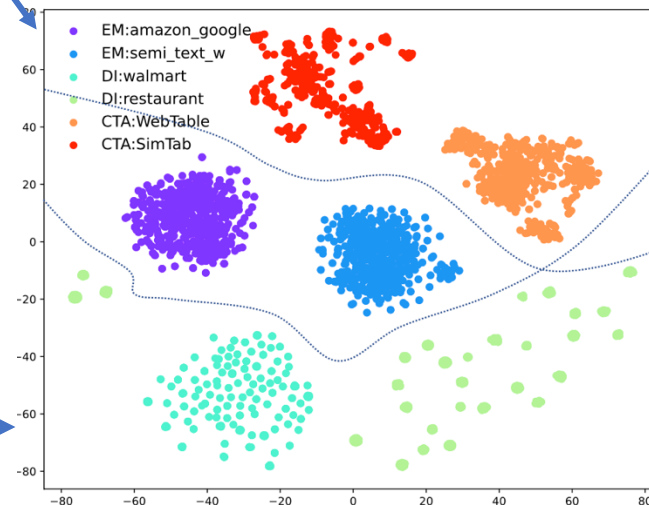
In same Parameter Size,  
**Dense Single Model** falls short in  
Multi-Task Learning  
Than **SMoE Model** in Error Bound

## Theorem 3: Convergence

**Router Network  $\mathcal{N}$  for SMoE Model**  
Dispatch samples to experts by  
Cluster in Latent Space



How LLM learns specific DP Task  $\mathcal{T}_i$



Clusters in Subspace  $V$   
For Task Vector  $\theta_i$

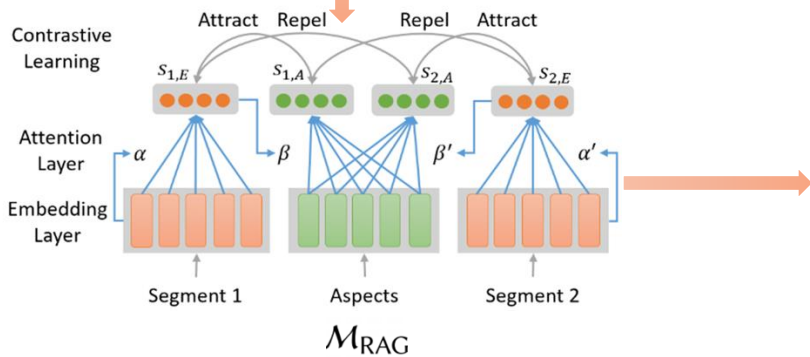


# Cross-task data augmentation based on RAG models

- Retrieve related **examples and contextual information** across tasks and domains to mitigate the issues of insufficient and biased labeled data.
- Generate new samples **using self-supervised labeled data** to expand the training set.
- Train a unified retrieval-enhanced framework using **contrastive learning**.

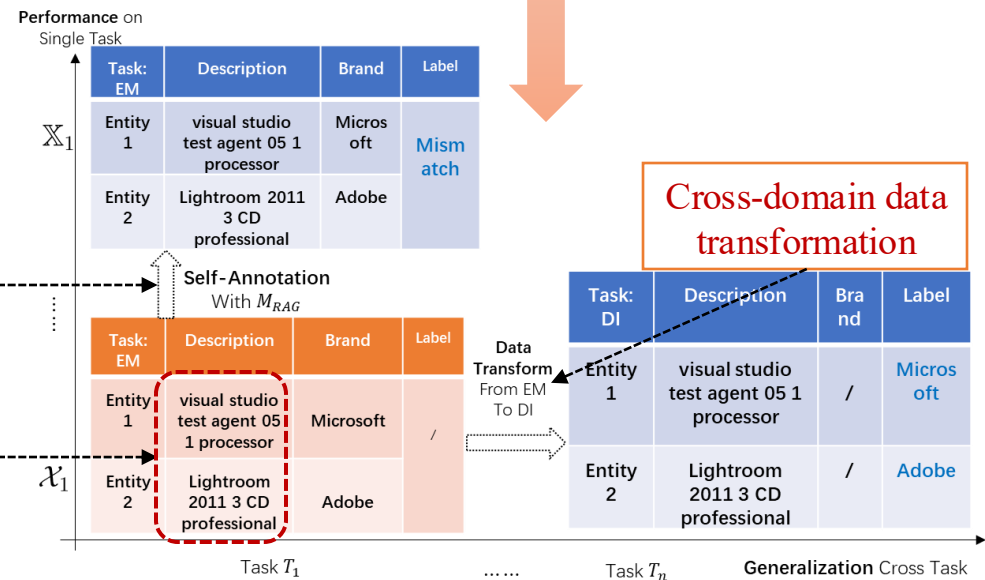
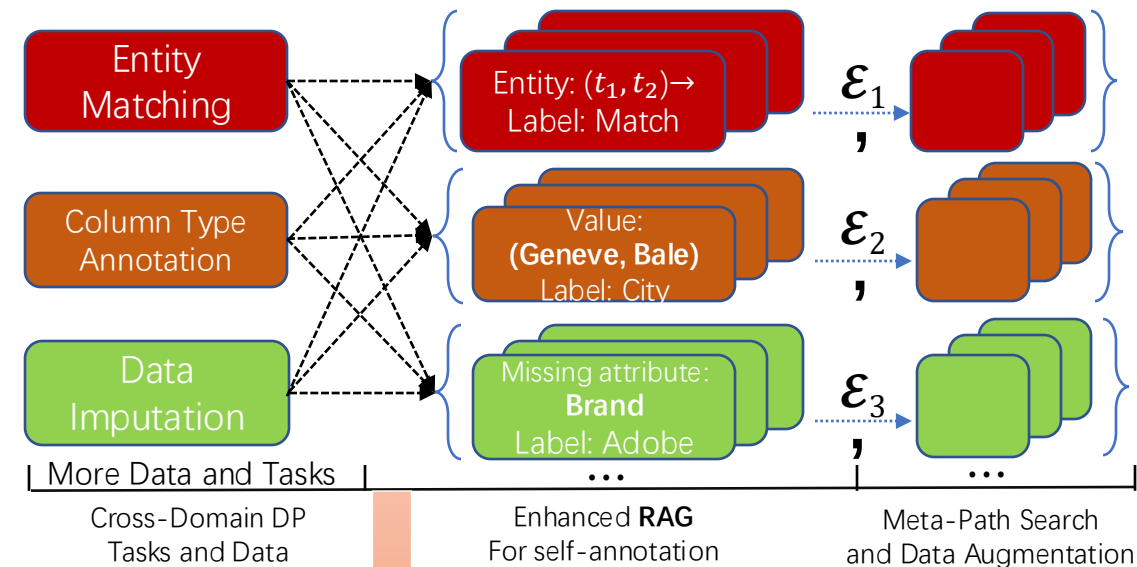
optimization target

$$\min \sum_{p \in \mathcal{P}_q} -\log \frac{\exp(\langle \text{emb}_q, \text{emb}_p \rangle / \tau)}{\sum_{p' \in \mathcal{P}_q \cup \mathcal{N}_q} \exp(\langle \text{emb}_q, \text{emb}_{p'} \rangle / \tau)}$$



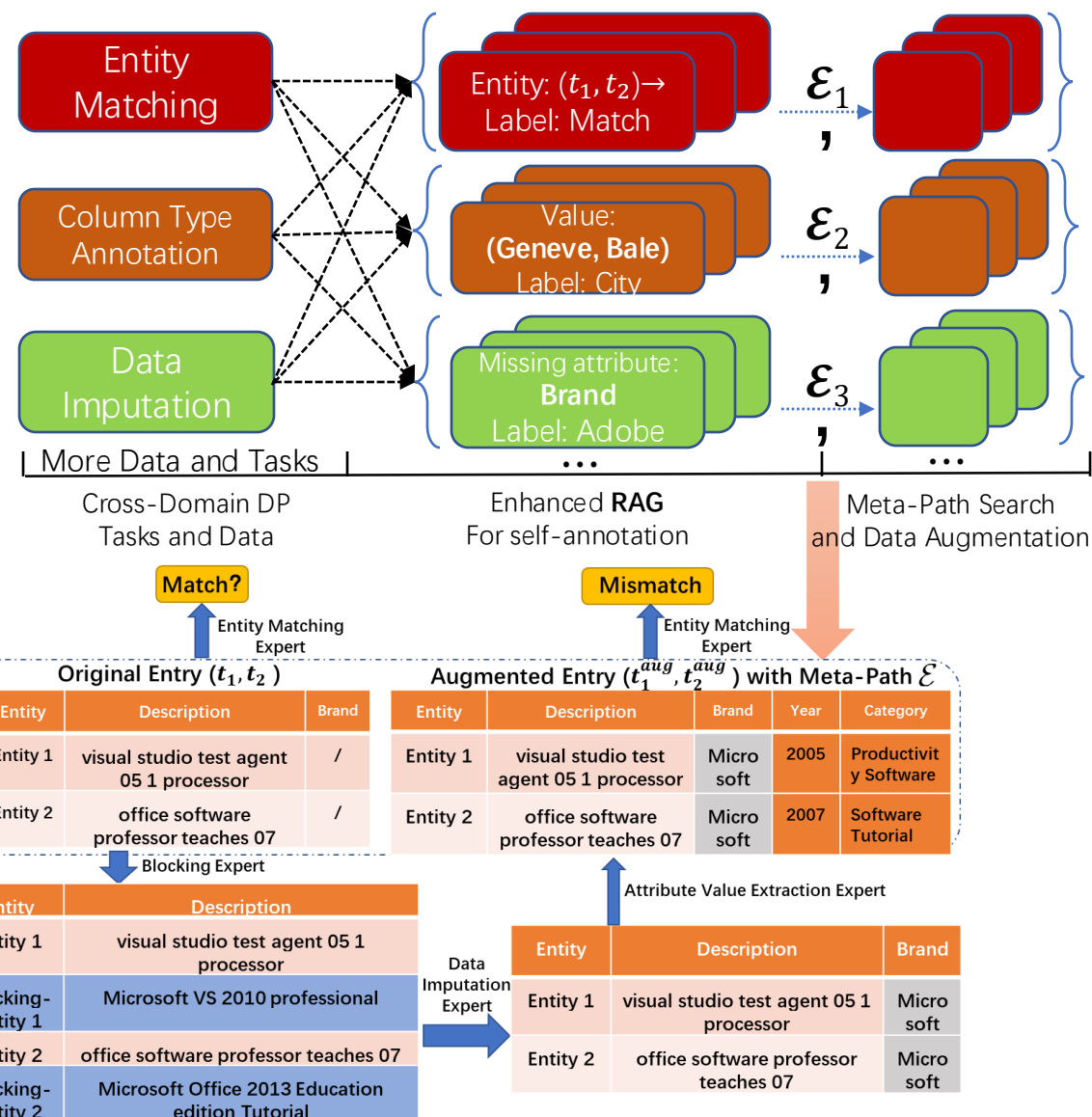
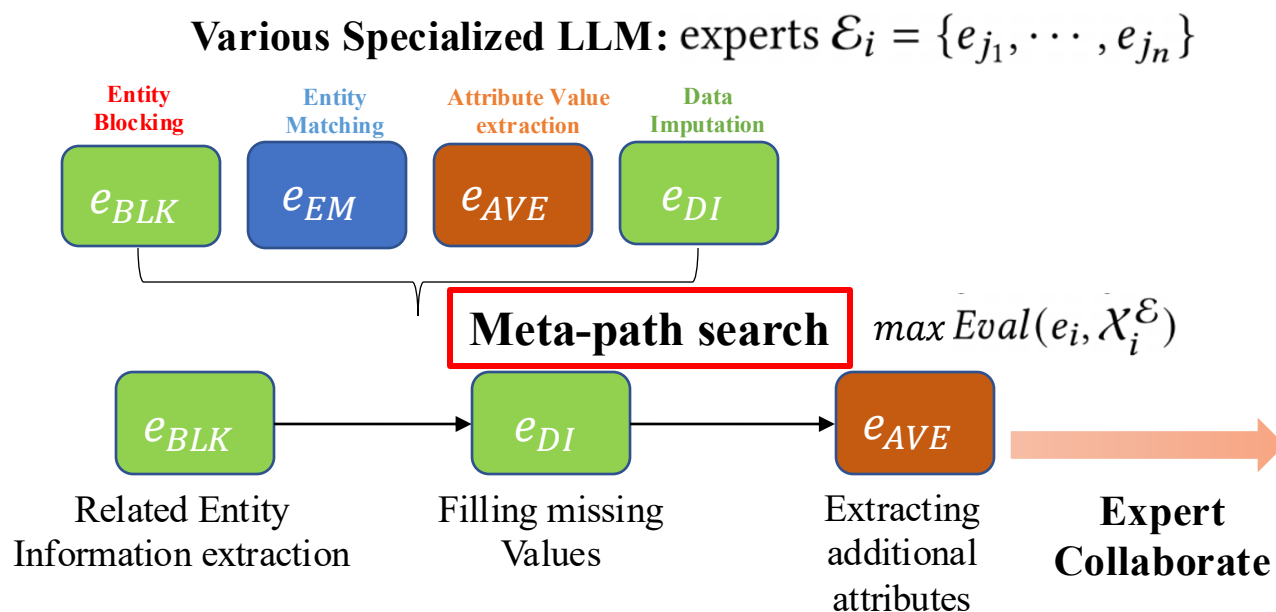
Self-supervised labeling

Contextual RAG



# Multi-expert collaborative enhancing based on meta-path search

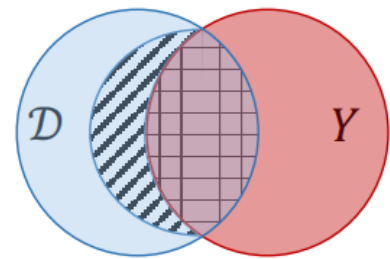
- Search for multi-expert collaborative paths to guide data augmentation.
- Experts focus on various data views, offering complementary advantages.
- Address information loss in **low-quality data**.





# Preliminary: Information Bottleneck(IB)

$$\max_Z I(Z; Y) - \beta I(Z; X)$$



/// irrelevant info.

▦ minimal sufficient info.

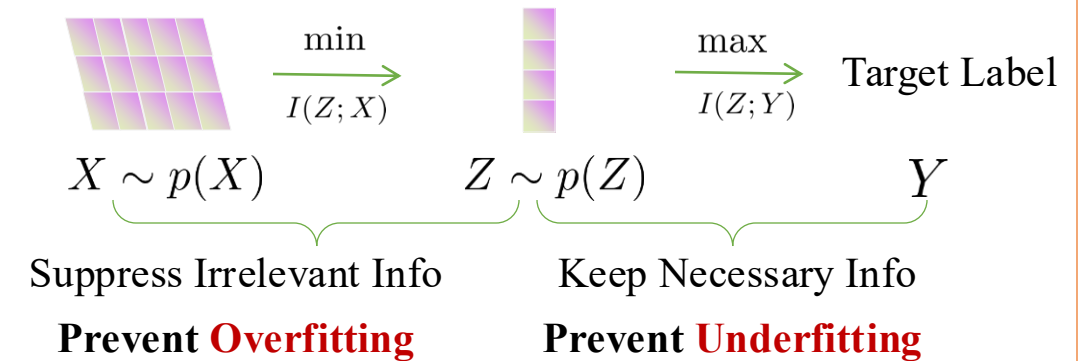


▦ optimal Z

▦ + /// overfitting

- $I(;)$  represents Mutual Information(MI)
- $X$ : Input Data;  $Z$ : Representation  $Y$ : Target/Label

## Concept of Information Bottleneck: Minimal Sufficient



# Expert model training with Information Bottleneck(IB) theory

## Motivations and Observations

- **Diverse** and **Augmented** training data, even from different domain, can activate the generalization ability for LLM, leveraging the overfitting problem, caused by the small size of  $|\mathbf{Q}|$ . (Min of MI)
- A well trained expert  $e_i$  should capture the **intrinsic and high-level** common feature from a diverse of training data  $\mathbf{Q} \cup \mathbf{Q}^*$ , and make the right decision within the constrained domain  $\mathcal{O}$ . (Max of MI)

## Formulation of Min-Max in training LLM-based $e_i$

Training data should be **Diverse and Augmented**

Expert Model should **Learn the Task Sufficient**

$$\min_{\theta_{\text{RAG}}} \max_{\theta_{\text{LLM}} \in \mathcal{M}_g} I(\mathcal{M}_g(\mathbf{Q}); \mathcal{M}_g(\text{RAG}(\mathbf{Q}))) \quad (1)$$

- Maximize: for  $\mathbf{q} = (q_k, l_k) \in \mathbf{Q} \cup \mathbf{Q}^*$ , maximize the mutual information of label  $l_k$  and the model output  $o_k$ .
- Minimize: for  $\mathbf{q} \in \mathbf{Q}$ , minimize the mutual information of  $\frac{1}{|\mathbf{Q}^*|} \sum_{\mathbf{q}' \in \mathbf{Q}^*} I(\mathbf{q}; \mathbf{q}')$

**Input:** Task  $T_i$ , Labeled data  $X_i$   
**Output:** Expert Model  $e_i$

## Findings:

**Diverse** and **augmented** training data address biased distribution and overfitting in few-shot learning

$$\text{Min } I(\mathbf{Z}; \mathbf{X})$$

Training data should match the task's correct distribution, enabling models to capture **inherent, high-level** features and associations

$$\text{Max } I(\mathbf{Z}; \mathbf{Y})$$

Explicitly optimize by fine-tuning large models to perform task  $T_i$

Implicitly optimize by adjusting parameter  $\theta$  of  $\text{RAG}$  to maintain diversity in training data  $X_i$

## Methods:

**Min-Max optimization**

# Router Network Optimization based on IB theory

## Observation

- For a given query  $q_u$ , which represents one or more entities  $ent$ , such raw data can be applied to different tasks in  $\mathcal{T}$  naturally. So the label should be  $l_u^1, \dots, l_u^{|\mathcal{T}|}$
- Top- $\tau$  experts should be diverse enough, that their responding  $o_1, \dots, o_u$  with  $q_u$  should be diverse enough with each other. (Min of MI)
- Top- $\tau$  experts should cover the domain of  $q_u$ , which means experts  $\mathbf{E}'$  should output correct  $o_1, \dots, o_u$  with label  $l_u^1, \dots, l_u^T$  (Max of MI)

- **Input:** Experts Set  $e_1, \dots, e_n$
- **Output:** Gated network  $\mathcal{N}$  that can assign given query  $q$  to top- $k$  relevant experts set

## Formulation of Min-Max in training Gated Network $\mathcal{N}$

Expert selection should be  
task-relevant to  $q_u$

Expert selection should  
ensure diversity

$$\max \sum_{e_i \in \mathcal{N}(q_u)} I(e_i(q_u^i); l_u^i), \quad \min \sum_{\substack{i \neq j \\ e_i, e_j \in \mathcal{N}(q_u)}} I(e_i(q_u^i); e_j(q_u^j)), \quad (2) \text{ where}$$

$$|\mathcal{N}(q_u)| = \tau.$$

Equivalently, if  $q_u$  originally belongs to task  $T_i$ , then any  $(q_u^j, l_u^j), i \neq j$  can be regarded as an augmented output from  $\text{RAG}(q_u)$

Optimization Objective  
For Gated Network  $\mathcal{N}$  :

# Mixture-of-Experts Implementation

- *Divide and Conquer*, Initialize different expert model for different task
- During Inference, mix-up expert weight for cross-domain generalization
- IB-Theory guided training for multi-expert allocation **per query**
- **Multi-Tenant LoRA Serving** for multi-expert inference, support 1 base model and up to 24 experts in single GPU, without merging and quantization

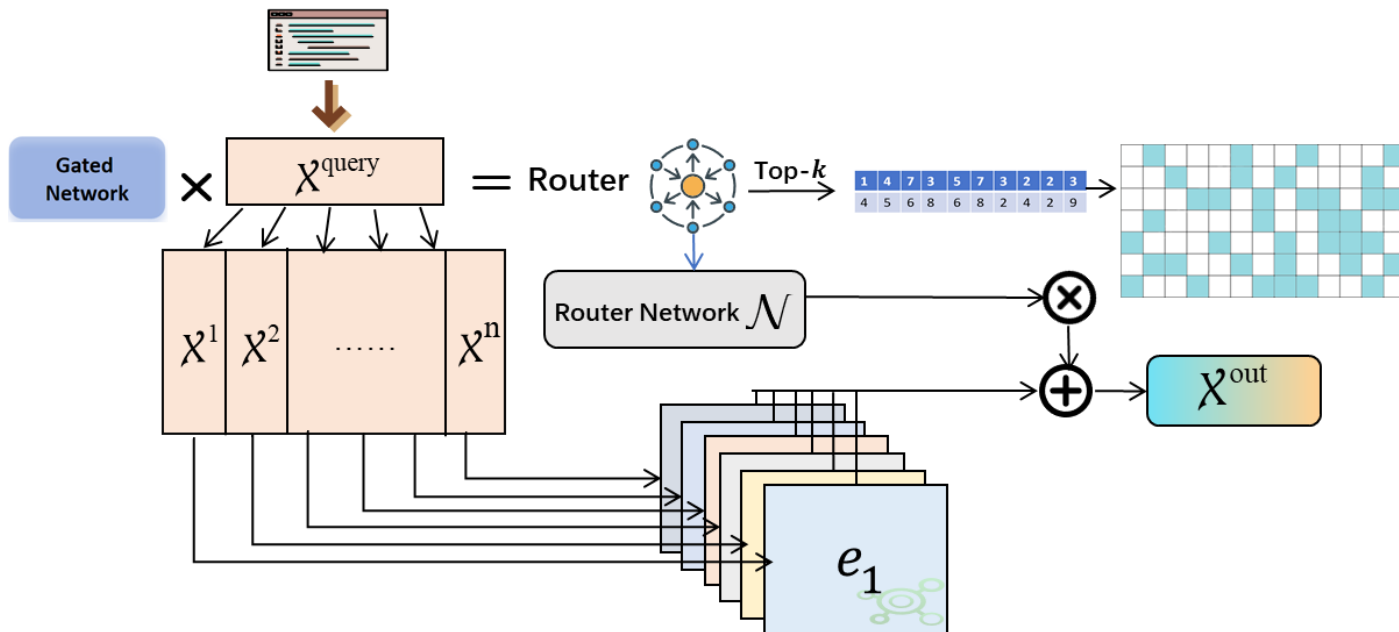


Fig.1 MoE Structure Framework

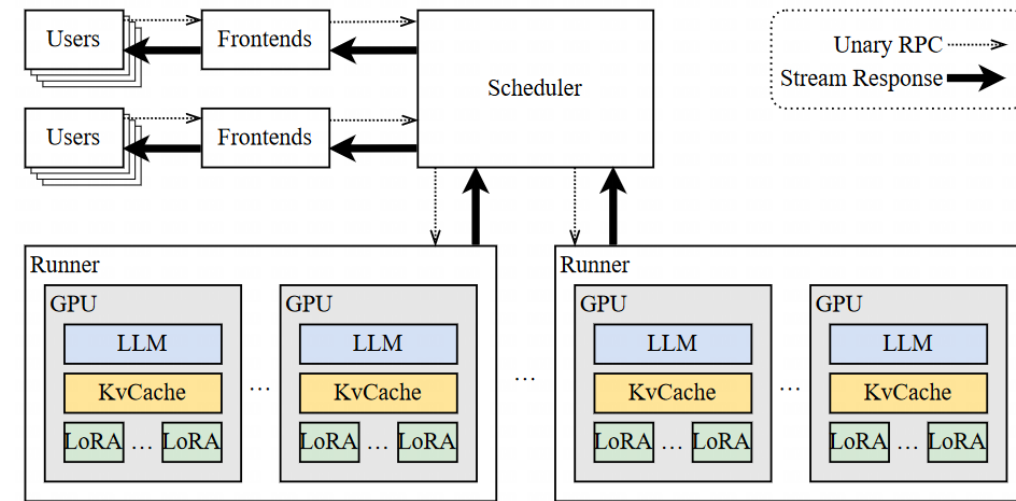


Fig.2 Multi-Tenant LoRA Serving by Punica



# Dataset and Experiment Setting

**Downstream Task: 19 datasets over 10 DP tasks, all with few-shot labelling setting(1%-10%)**

- Entity Matching, EM, F1 score
  - Entity Blocking, BLK, Top-1 Recall
  - Error Detection, ED, F1 score
  - Data Cleaning, DC, F1 score
  - Column Type Annotation, CTA), Micro-F1
  - Relation Extraction, RE, Micro-F1
  - Entity Linking, EL, Top-1 Accuracy
  - Schema Matching, SM, F1 score
  - Data Imputation, DI, Top-1 Accuracy
  - Attribute Value Extraction, AVE, Top-1 Accuracy
- } Entity Resolution
- } Data Cleaning
- } Tabular Interpretation Learning
- } Data Imputation

## Baseline Model:

- 12 non-LLM baselines (Including Feature Engineer/Rule-Discovery/Transformer-Based Deep Learning Method)
- Jellyfish/ExtractGPT (Pre-trained LLM methods in 13B/70B)
- MoE Model(Mixtral 8\*7B)

**Backbone Model for each Expert: Mistral-7B**

**Backbone Model for RAG: Roberta-XL**

*Methods.* We categorized the baselines as follows. (1) Non-LLM methods [86]. (a) ED: Raha[79], (b) DI: IPM[82], (c) Blocking: DeepBlocker[107], (d) EM: Ditto[72] and PromptEM[113], (e) DC: Baran[78] and Garf[92], (f) CTA: RECA[31], (g) RE/EL: TURL[26], (h) SM: CONSchema[117] and SMAT[128], and (k) AVE: MAVe[121].

Task	Dataset	#Instance (few-shot)	#Instance (All)
Entity Matching (EM) & Blocking	Amazon-Google[72]	100	6874
	Walmart-Amazon[72]	100	6144
	WDC-All[72]	100	7229
	Ant-Buy[72]	100	5743
	Semi-Text-Watch[113]	80	5540
	Semi-Text-Computer[113]	80	12538
Error Detection(ED) & Data Cleaning(DC)	Hospital[78]	20	1000
	Rayyan[78]	20	1000
	Beer[78]	20	2410
Column Type Annotation(CTA)	SemTab19[31]	1920	7603
	WebTables[31]	15420	61023
Relation Extraction(RE)	WikiGS-RE[26]	6502	65026
Entity Linking(EL)	WikiGS-EL[26]	5441	54410
Schema Matching(SM)	CMS[128]	20505	20505
	Synthea[128]	23709	23709
Data Imputation(DI)	Walmart[82]	242	2421
	Amazon[82]	2001	20013
	Restaurant[82]	86	864
Attribute Value Extraction(AVE)	OA-mine[5]	286	1452

# Experiment 1 Main Result

Task	Dataset	MELD Few-shot	Non-LLM Baseline Few-shot	LLM Baseline Few-Shot	Mixtral Few-shot
EM & (BLK)	Amazon-Google	<b>83.41(74.12)</b>	61.88(50.47)	65.98(/)	51.28(/)
	Walmart-Amazon	<b>91.42(78.80)</b>	79.09(58.21)	42.03(/)	39.78(/)
	WDC-All	<b>91.97(31.50)</b>	34.35(1.70)	49.80(/)	48.97(/)
	Ant-Buy	<b>91.12(86.20)</b>	84.89(40.66)	71.40(/)	60.42(/)
	Semi-Text-Watch	<b>78.28(59.23)</b>	23.60(2.66)	54.27(/)	40.55(/)
	Semi-Text-Computer	<b>86.46(30.85)</b>	33.90(8.09)	76.80(/)	73.15(/)
DC	Hospital	<b>95.01</b>	67.10	49.30	53.20
	Rayyan	<b>82.15</b>	28.50	9.39	6.68
	Beer	<b>97.30</b>	90.31	51.30	56.27

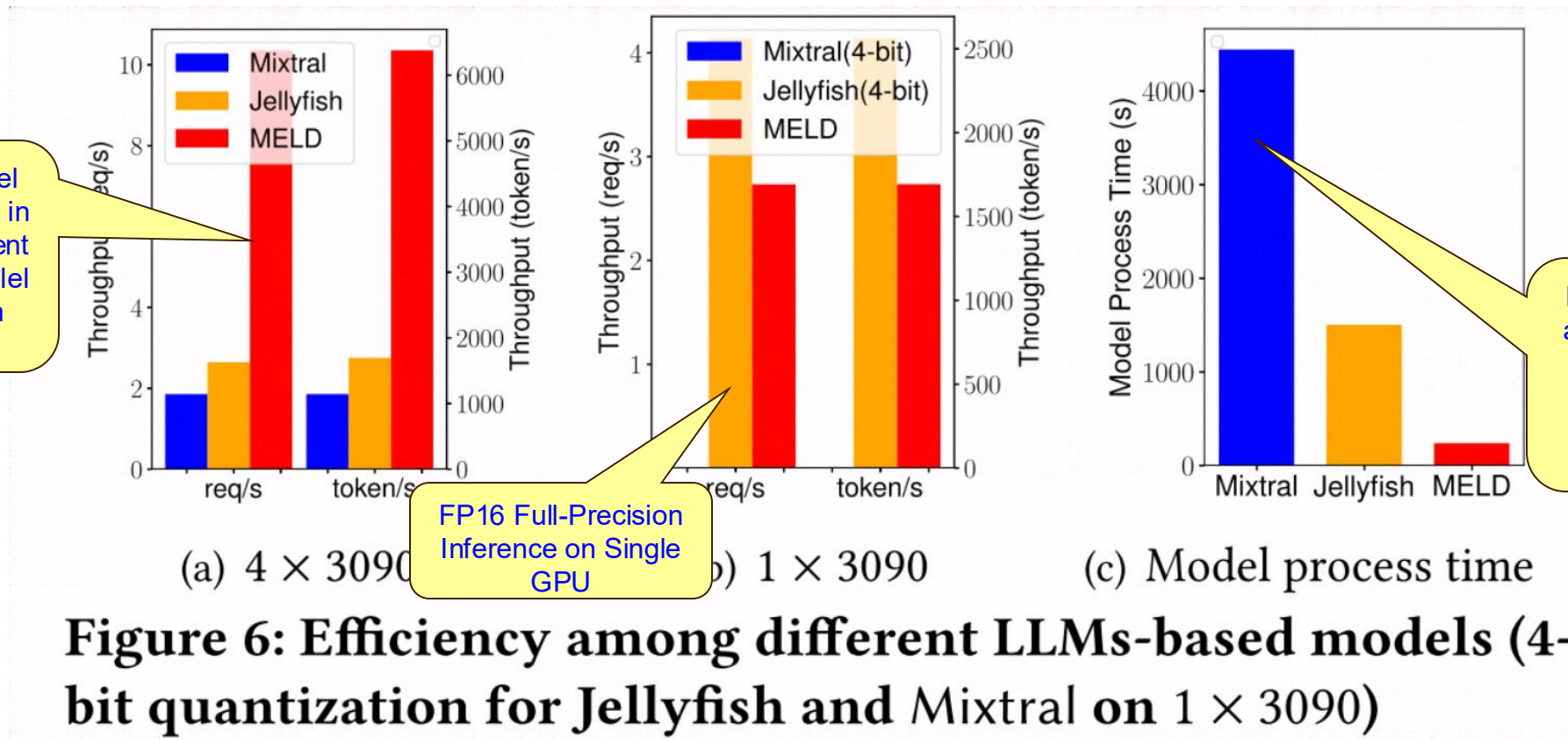
Task	Dataset	MELD Few-shot	Non-LLM Baseline Few-shot	LLM Baseline Few-Shot	Mixtral Few-shot
ED	Hospital	<b>98.51</b>	95.23	89.41	69.14
	Rayyan	<b>90.37</b>	80.21	69.67	31.96
	Beer	99.10	<b>100.00</b>	81.64	70.23
CTA	SemTab19	<b>89.35</b>	69.70	87.77	<b>89.35</b>
	WebTables	<b>96.30</b>	90.93	94.77	80.16
RE	WikiGS-RE	<b>89.30</b>	73.50	60.38	65.88
EL	WikiGS-EL	<b>87.05</b>	60.55	82.20	73.25
SM	CMS	<b>60.27</b>	50.00	59.29	31.01
	Synthea	<b>56.00</b>	38.50	40.00	23.53
DI	Walmart	<b>87.50</b>	65.70	57.69	79.82
	Amazon	<b>75.12</b>	60.35	60.05	62.62
	Restaurant	<b>93.10</b>	37.50	68.97	72.41
AVE	OA-mine	74.62	67.00	65.70	<b>77.36</b>

MoE Framework is suitable for few-shot learning and multi-task learning

Search and retrieval across different domains and task, can alleviate biased distribution and few-shot labelling



# Experiment 2 Inference Efficiency



- Based on vLLM 0.40.0 in January 2024, maybe changed due to MoE kernel optimization.

# Experiment 3 Cross-Domain and Cross-Task

**Table 2: Cross-Dataset(C-D) and Cross-Task(C-T)**

Task	Dataset	MELD C-D	MELD C-T	LLM Baseline C-D	LLM Baseline C-T	Mixtral C-D	Mixtral C-T
EM	Amazon-Google	<b>69.05</b>	67.95	18.58	18.58	43.23	43.23
	Semi-Text-Watch	<b>65.07</b>	51.13	20.52	20.51	37.12	37.12
CTA	SemTab19	<b>76.84</b>	61.21	15.79	7.96	64.83	61.64
	WebTables	86.76	<b>88.95</b>	38.92	14.29	79.72	67.64
DI	Walmart	54.80	54.80	43.26	17.86	<b>79.82</b>	<b>78.85</b>
	Restaurant	<b>75.86</b>	<b>75.86</b>	68.96	6.95	72.43	58.62

MELD have the least performance drop in domain adaptation

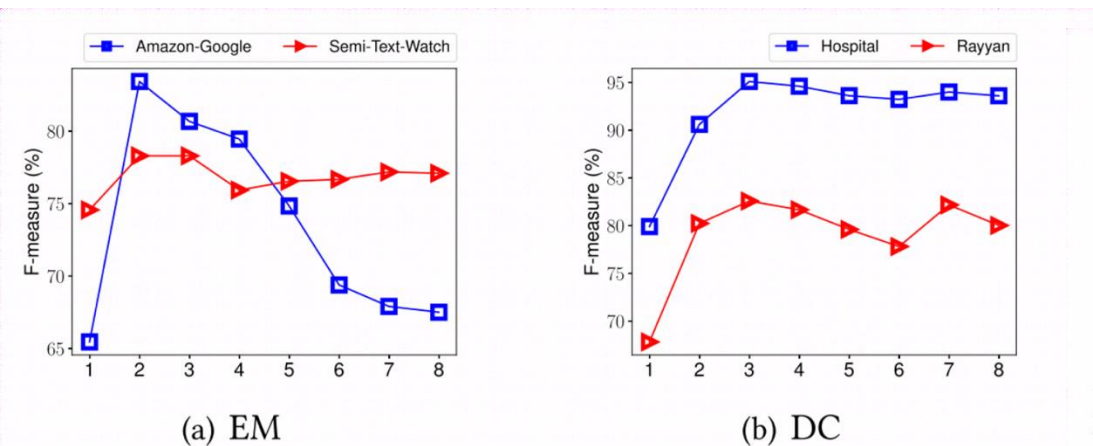
# Experiment 4 Ablation Study

**Table 3: Performance for Ablation Study**

Task	Dataset	MELD w/o MoE	MELD w/o RAG	MELD w/o Meta-Path	MELD with Mixtral
EM	Amazon-Google	76.70	69.21	62.52	77.85
	Walmart-Amazon	87.66	81.44	79.55	91.03
	WDC-All	90.38	83.16	91.73	91.32
	Ant-Buy	87.58	85.75	90.12	85.26
	Semi-Text-Watch	70.78	55.07	39.89	75.42
	Semi-Text-Computer	79.49	42.02	63.74	81.98

- MELD w/o MoE: Delete Router Network, directly apply task-corresponding expert. (Decrease Parameter-Level Diversify )
- MELD w/o RAG: Delete RAG Module, each task is trained by excluding cross-task and cross-dataset samples. (Decrease Distribution-Level Diversify )
- MELD w/o meta-path: Delete Meta-Path based data augmentation (Decrease Information-Level Diversify )
- MELD with Mixtral: Replace expert model with Mixtral, replace Router Network with Mixtral build-in layer.

# Experiment 5 Visualization



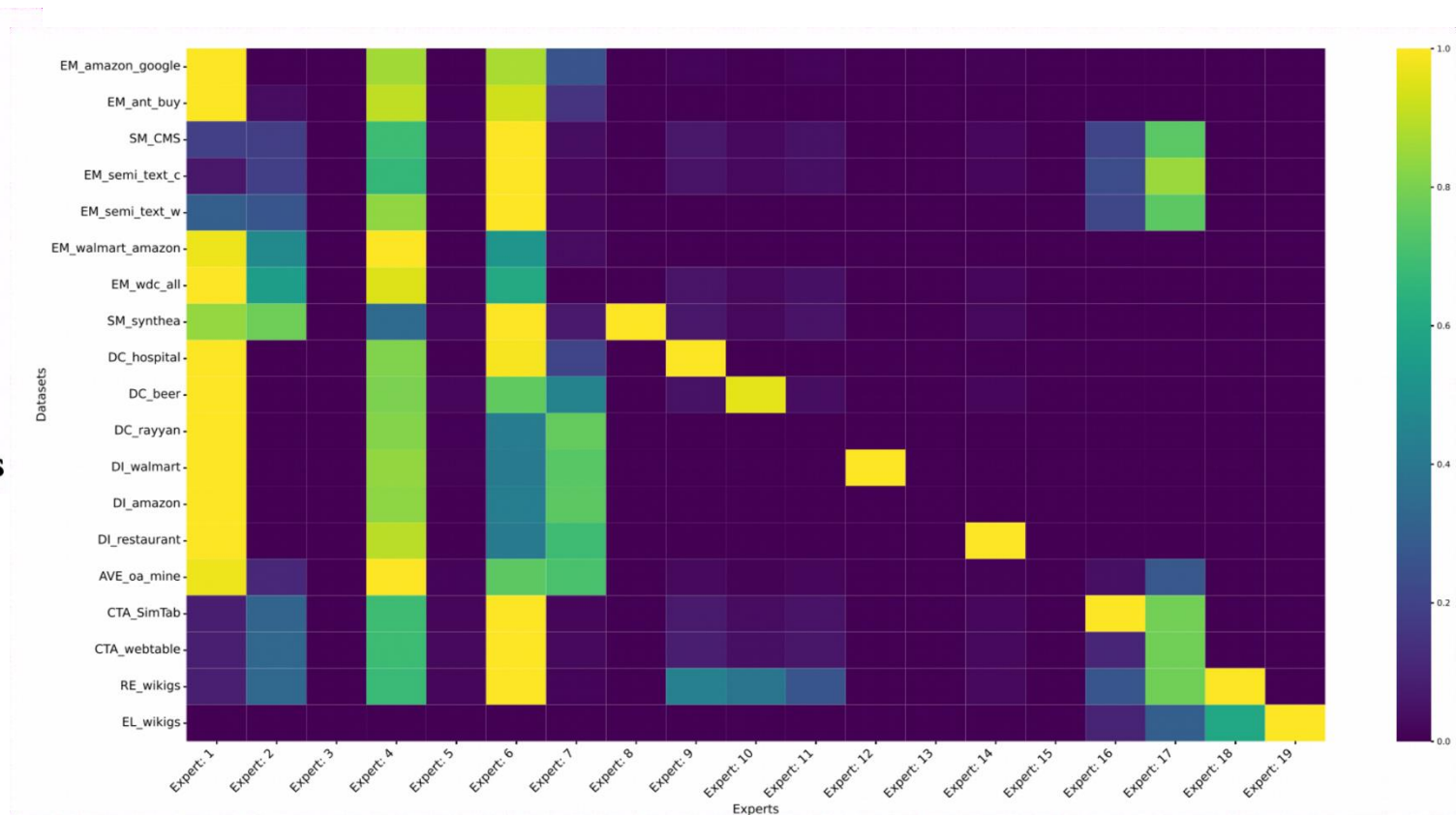
**Figure 7: Performance for different number of experts**

Increasing expert number may lead to more noise in parameter-level

**Table 6: Performance compared with GPT-4**

Task	Dataset	MELD Few-shot	GPT-4	LLM Baseline Few-Shot	Mixtral Few-shot
EM	Amazon-Google	<b>83.41</b>	74.21	65.98	51.28
	Walmart-Amazon	<b>91.42</b>	90.27	42.03	39.78
	Ant-Buy	91.12	<b>92.77</b>	71.40	60.42
SM	CMS	<b>60.27</b>	59.29	59.29	31.01
	Synthea	56.00	<b>66.67</b>	40.00	23.53
DI	Restaurant	93.10	<b>97.75</b>	68.97	72.41
AVE	OA-mine	74.62	<b>80.20</b>	65.70	77.36

Comparison with Online Model

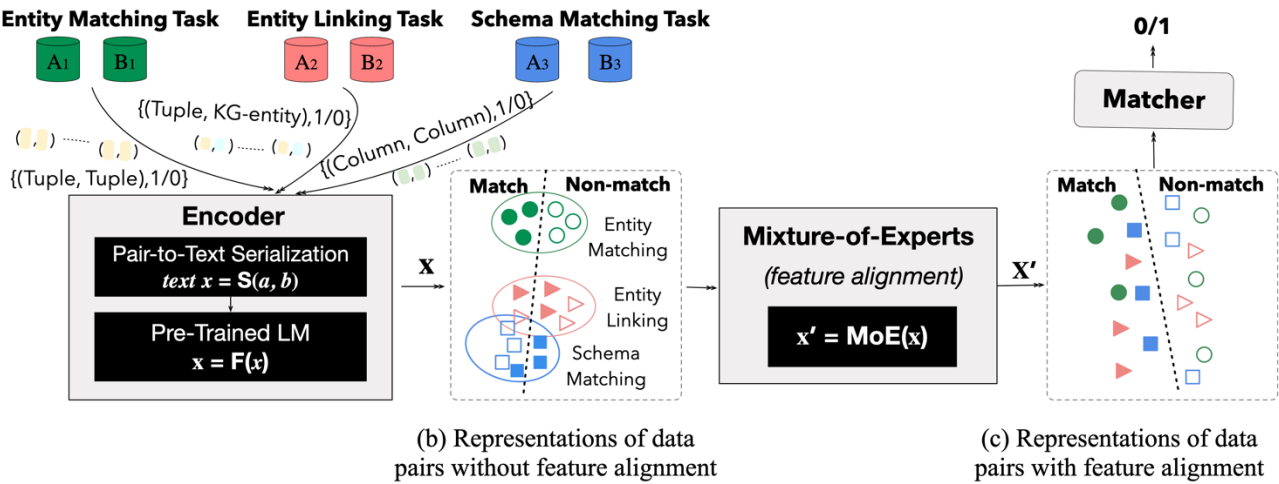


Heatmap for expert assignment weights

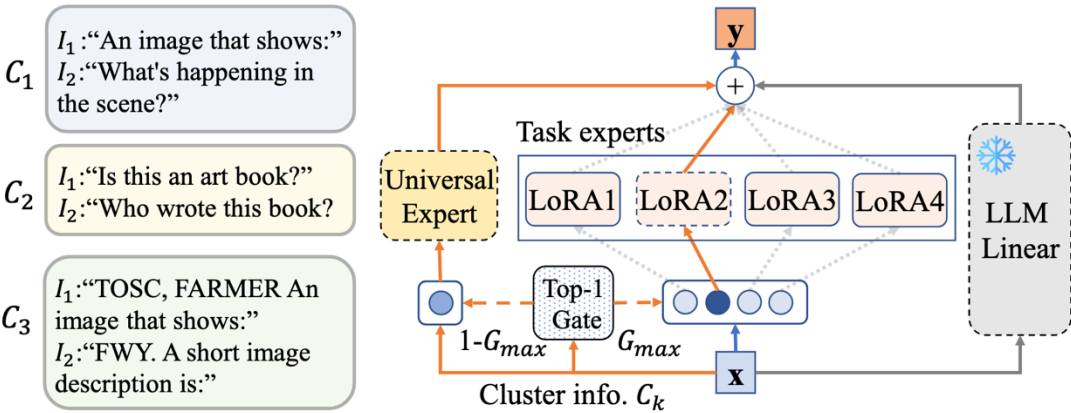
Code/Full Version Paper is available at: <https://github.com/athurlord/MELD>

# Future Work and Discussion

- In which level should we apply expert assignment?
  - Token-Level(Switch Transformer/ Mixtral/ Qwen-MoE)
  - Sentence/Query Level(Unicorn/ MELD)
  - Cluster Level (MoCLE)
  - Task Level



Unicorn<sup>1</sup>



MoCLE<sup>2</sup>

1. Unicorn: A Unified Multi-Tasking Matching Model
2. Mixture of Cluster-conditional LoRA Experts for Vision-language Instruction Tuning