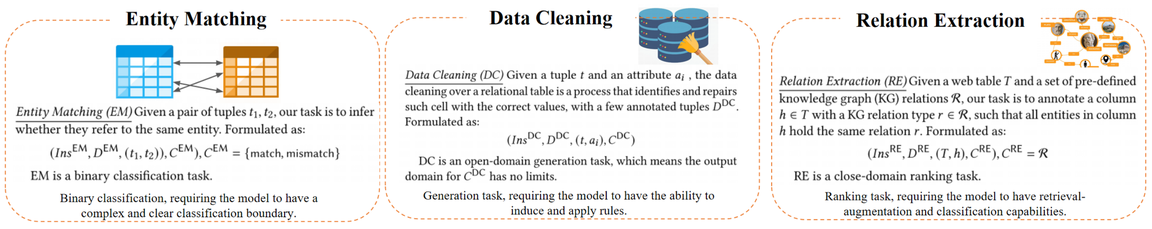
Efficient Mixture of Experts based on Large Language Models for Low-Resource Data Preprocessing

Mengyi Yan[1] Yaoshu Wang[*2] Kehan Pang[1] Min Xie[2] Jianxin Li[*1]

[1]Beihang University [2]Shenzhen Institute of Computing Sciences

A. Introduction

- **Context:** Data preprocessing (DP) is essential for transforming raw, erroneous data into a usable format, serving as the backbone in the data mining process.
- **Challenges**
  - **Manual Design per Task:** Traditional DP methods require extensive manual intervention to craft domain-specific rules or task-specific model design, leading to scalability issues.
  - **Labelling Cost:** These manual and model-based approaches are not only time-consuming but also incur significant labelling costs.
  - **Inflexibility:** Current methods struggle to adapt to the diverse requirements of different data types and downstream tasks in dynamic environments.



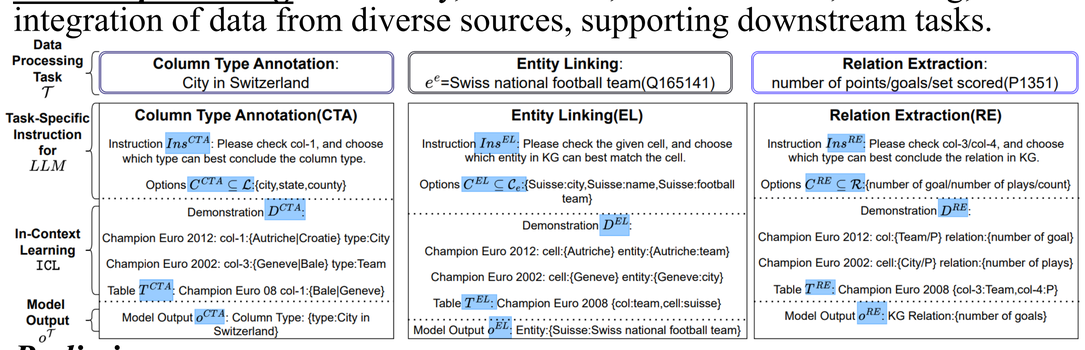
1. Can different preprocessing tasks be unified into a common framework in **generative manner**? (**Multi-Task**)
2. Can **few-shot** labelling data from different tasks mutually boost each others performance?(**Data-Efficient**)
3. Can Sparse-Activated Mixture of Expert models(SMoE) outperform single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

B. Motivation

1. **Increasing Data Complexity**: The variety and complexity of data increase, necessitating more sophisticated preprocessing techniques that can handle such diversity efficiently.
2. **Scalability Issues**: Existing DP methods often do not scale well with increasing data volumes or variety, making them unsuitable for large-scale applications in real-world scenarios.
3. **Utilizing One-for-All LLM for diverse tasks**: Despite the recent progress in machine learning and large language models for data quality, their potential in various DP tasks, especially in low-resource settings, remains largely untapped.
4. **Resource Constraints**: There's a critical need for cost-effective DP solutions that require minimal resources, making advanced data preprocessing accessible to smaller organizations or projects with limited labelling and computational budgets, also regarding privacy concerns.

C. Problem Definition

- ***Data Preprocessing*:** Discovery, extraction, transformation, cleaning, and integration of data from diverse sources, supporting downstream tasks.



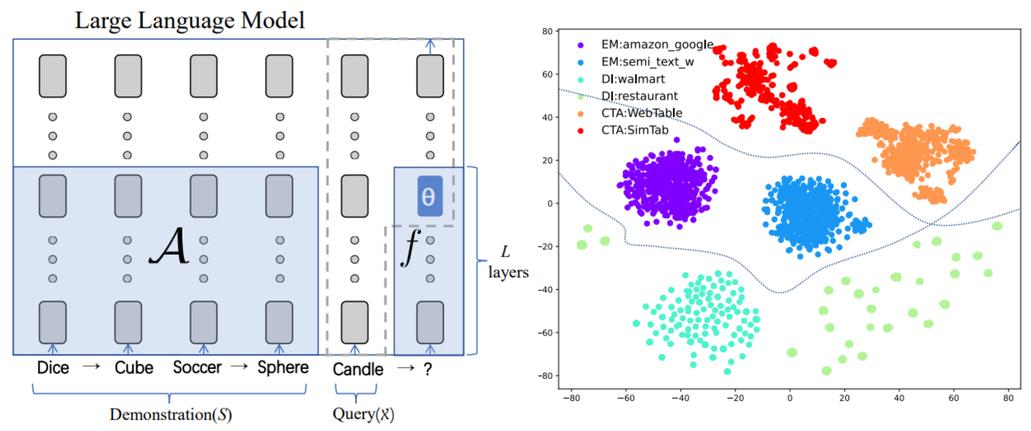
- ***Preliminary***
  - **Mixture of Experts (MoE):** Architecture that dynamically assigns tasks to specialized networks, optimizing processing efficiency and adaptability.
  - **Multi-task Learning(MTL):** Leverages shared information across multiple tasks to improve generalization and performance.
- ***Problem***. The problem studied in this paper is stated as follows.
  - **Input**: A set of tasks $\{\mathcal{T}_1, \cdots, \mathcal{T}_n\}$ with few-shot training data $\mathcal{X}$ in the low-resource DP setting. (w.r.t <=10% labelling data, deployed in consumer-level hardware.)
  - **Output**: An universal LLM-based system under the MoE architecture that is able to answer the (unseen) query of all $\mathcal{T}_i$

D. Contribution

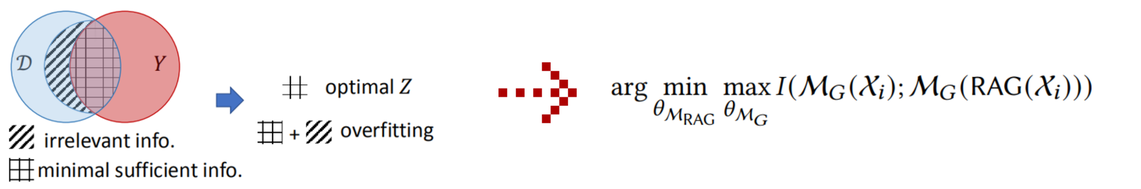
- ✓ **MELD Framework**: Introduction of a Mixture of Experts on Large Language Models for Data Preprocessing (MELD) optimized for low-resource environments.
- ✓ **Data Augmentation and Expert Optimize Techniques**: Development of novel tuning methods for expert, and cross-task data augmentation method, including retrieval-augmented generation and meta-path data augmentation.
- ✓ **Theoretical Advancements**: Proof of concept that MoE in MELD outperforms single expert setups and demonstrates efficient expert allocation.
- ✓ **Resource Optimization**: Demonstrates reduced computational overhead while maintaining high accuracy.
- ✓ **Enhanced Specialization**: Improved task-specific performance through expert specialization, and mixture of different experts.

E. Theoretical Analysis

1. **Task Subspace:** Different Task $\mathcal{T}_i$ can be compressed to low-dimension Task Vector θ_i for LLM.

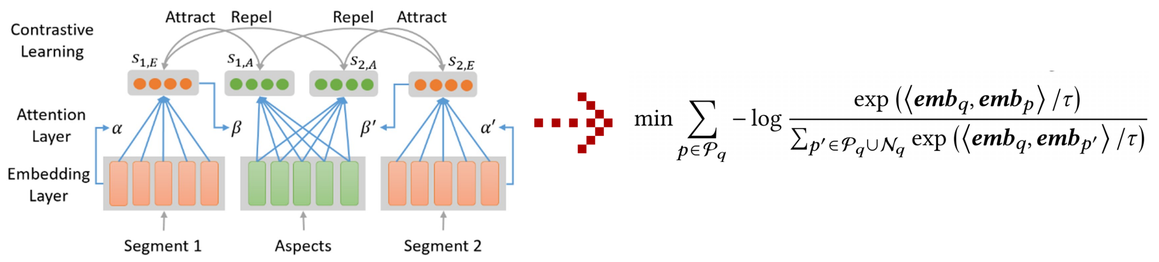


2. **Error Bound:** In same parameter size, dense model falls short in multi-task learning than MoE model in error bound.
3. **Router Network Efficacy:** Demonstrating that the router network effectively directs data to the most suitable experts, optimizing resource allocation.
4. **Information Bottleneck:** Utilization of an information bottleneck in training enhances the model's focus and efficiency by reducing irrelevant data processing.

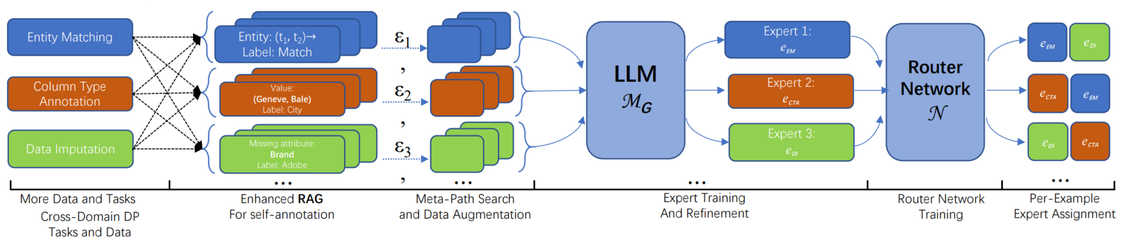


F. MoE architecture based on LLMs

- **Enhanced RAG Component:** Using a fine-tuned Sentence-BERT model with contrastive learning, to support effective cross-task and cross-domain information retrieval for LLM.



- **Meta-Path Search Component:** Processing expert sets and enlarged data to establish a meta-path that augments and optimizes training data for specific tasks in sequence.
- **Expert Refinement:** Applying information bottleneck theory to refine expert accuracy and relevance, enhancing the performance of each expert.
- **Router Network:** Implementing a multi-gate network to select top experts for each query, optimizing resource allocation and improving query response efficiency.



G. Experiment

- **Performance**: Evaluated across 19 datasets and 10 tasks, MELD outperforms state-of-the-art methods in terms of effectiveness and efficiency in low-resource environments.
- **Theoretical Validation**: Empirical evidence supports theoretical claims about MoE superiority and effective data routing via the router network.
- **Resource Utilization**: MELD demonstrates efficient use of computational resources, optimizing both time and hardware constraints.
- **Adaptability**: MELD shows significant improvements in handling domain-specific tasks with limited annotated data.

Task	Dataset	MELD Few-shot	Non-LLM Baseline Few-shot	LLM Baseline Few-shot	Mixtral Few-shot
EM & (BLK)	Amazon-Google	**83.41(74.12)**	61.88(50.47)	65.98(/)	51.28(/)
	Walmart-Amazon	**91.42(78.80)**	79.09(58.21)	42.03(/)	39.78(/)
	WDC-All	**91.97(31.50)**	34.35(1.70)	49.80(/)	48.97(/)
	Ant-Buy	**91.12(86.20)**	84.89(40.66)	71.40(/)	60.42(/)
	Semi-Text-Watch	**78.28(59.23)**	23.60(2.66)	54.27(/)	40.55(/)
	Semi-Text-Computer	**86.46(30.85)**	33.90(8.09)	76.80(/)	73.15(/)
DC	Hospital	**95.01**	67.10	49.30	53.20
	Rayyan	**82.15**	28.50	9.39	6.68
	Beer	**97.30**	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	**98.51**	95.23	89.41	69.14
	Rayyan	**90.37**	80.21	69.67	31.96
	Beer	99.10	**100.00**	81.64	70.23
CTA	SemTab19	**89.35**	69.70	87.77	**89.35**
	WebTables	**96.30**	90.93	94.77	80.16
RE	WikiGS-RE	**89.30**	73.50	60.38	65.88
EL	WikiGS-EL	**87.05**	60.55	82.20	73.25
SM	CMS	**60.27**	50.00	59.29	31.01
	Synthea	**56.00**	38.50	40.00	23.53
DI	Walmart	**87.50**	65.70	57.69	79.82
	Amazon	**75.12**	60.35	60.05	62.62
	Restaurant	**93.10**	37.50	68.97	72.41
AVE	OA-mine	74.62	67.00	65.70	**77.36**

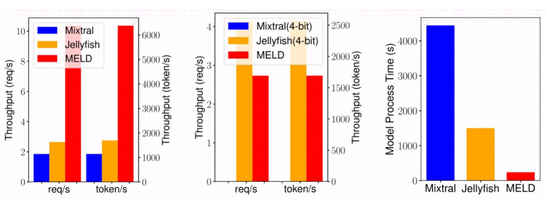
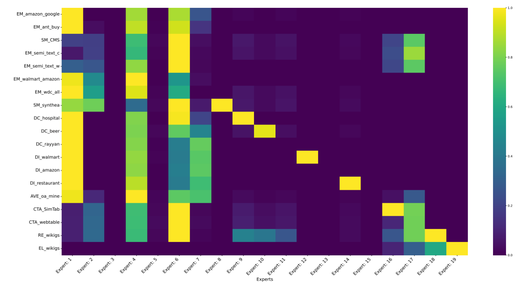


Table 2: Cross-Dataset(C-D) and Cross-Task(C-T)

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	**69.05**	67.95	18.58	18.58	43.23	43.23
	Semi-Text-Watch	**65.07**	51.13	20.52	20.51	37.12	37.12
CTA	SemTab19	**76.84**	61.21	15.79	7.96	64.83	61.64
	WebTables	86.76	**88.95**	38.92	14.29	79.72	67.64
DI	Walmart	54.80	54.80	43.26	17.86	**79.82**	**78.85**
	Restaurant	**75.86**	**75.86**	68.96	6.95	72.43	58.62



Full version, resource and code available at https://github.com/authurlord/MELD