Towards Unified Task Embeddings Across Multiple Models: Bridging the Gap for Prompt-Based Large Language Models and Beyond

Xinyu Wang, Hainiu Xu, Lin Gui, and Yulan He

WARWICK
THE UNIVERSITY OF WARWICK

KING'S College LONDON

The Alan Turing Institute

Scan to check the paper!

Task Embedding

- Task Embedding is a meta-learning tool for capturing the task-specific information of a task.
- Existing task embedding methods rely on fine-tuned, task-specific language models. Such approach is limited to the single-model scenarios, and is not applicable for LLMs.
- In this paper, we introduce a new framework, capable of learning unified task embeddings from diverse models, including language models of different architectures, and LLMs with various prompts, within a single vector space.

Background

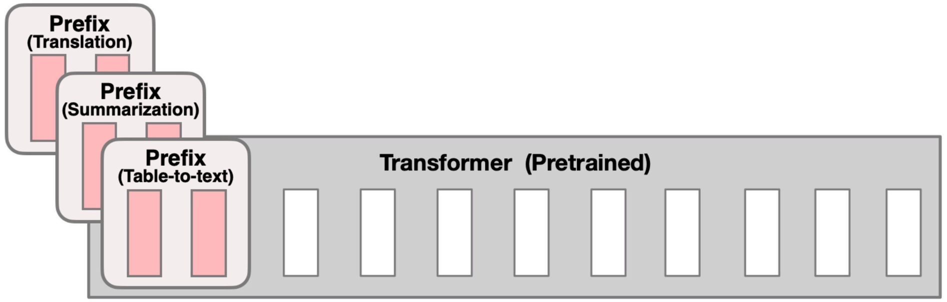
TaskEmb:

- computes the empirical Fisher on a fine-tuned model as task embedding

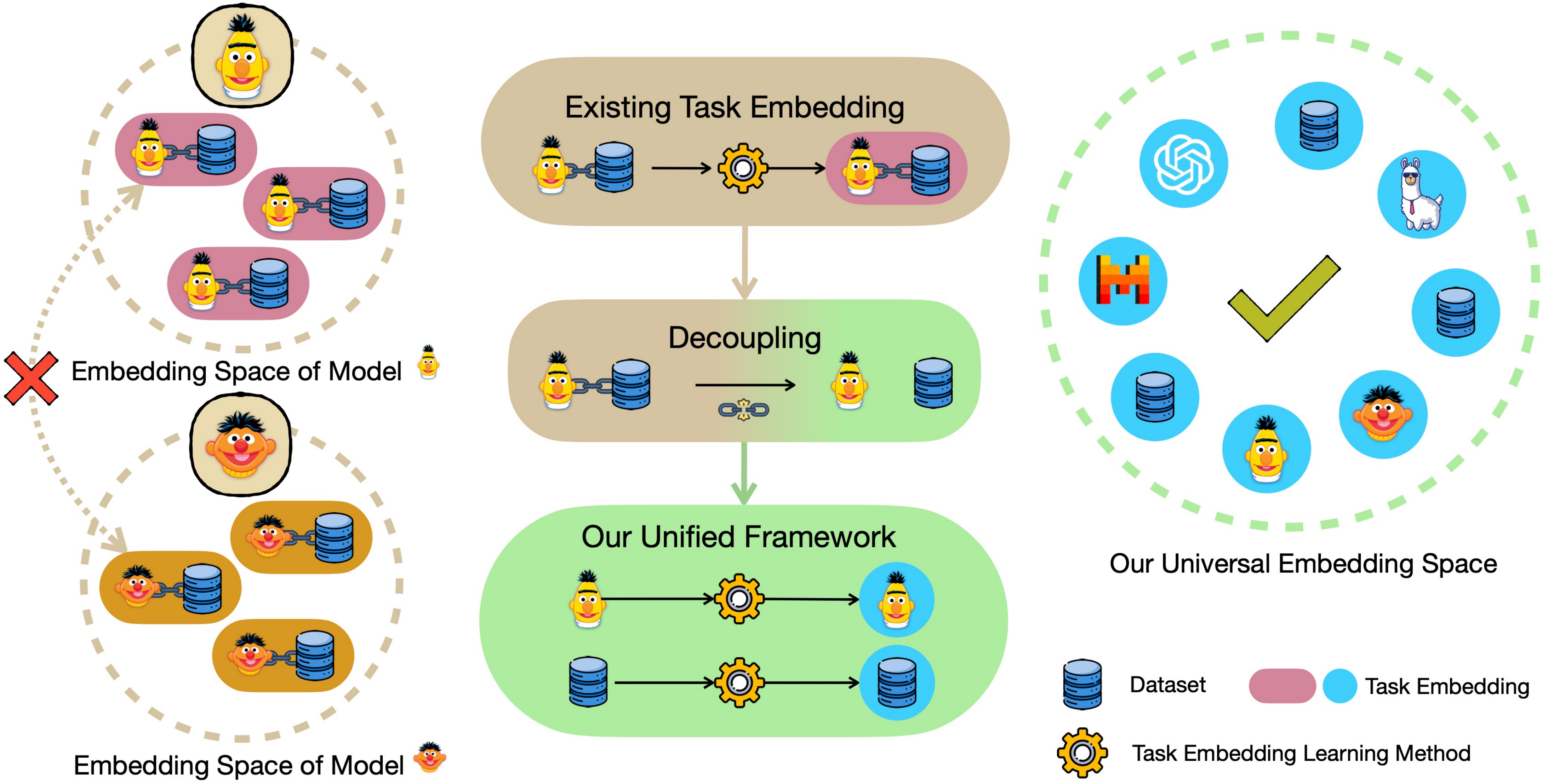
$$F_\theta = \frac{1}{n}\sum_{i=1}^{n}\left[\nabla_\theta \log P_\theta(y_i|x_i)\nabla_\theta \log P_\theta(y_i|x_i)^T\right]$$

TuPaTE:

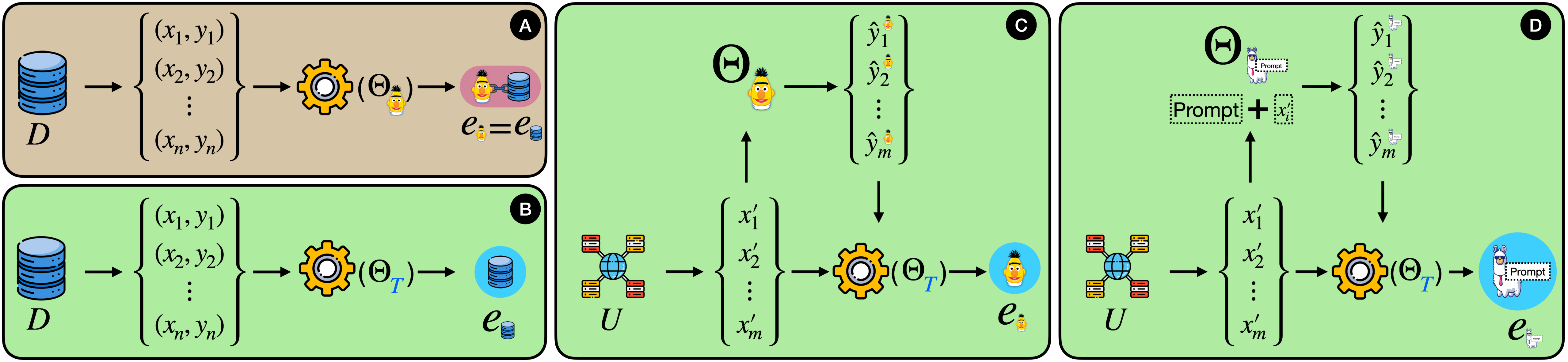
- utilizes Parameter-Efficient FineTuning (PEFT) methods on a language model and extract the tuned parameters as the task embedding.



Our Framework



Our Methods



D: Dataset; U: Unsupervised Dataset;

T: surrogate base model; e: Task Embedding

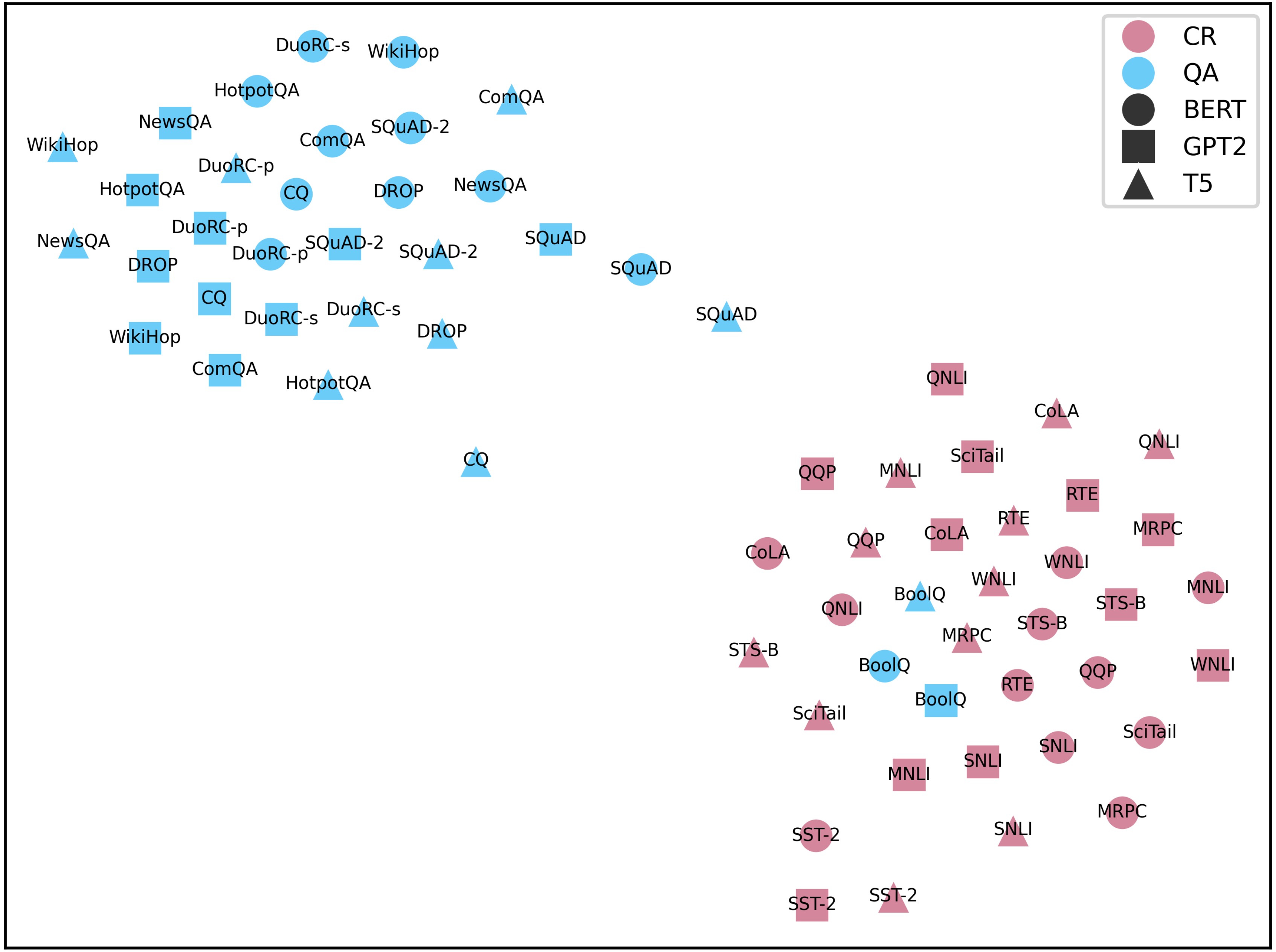
Visualization

Task embeddings from our framework extracted from different language model fine-tuned on different datasets.

CR: Classification or Regression task.

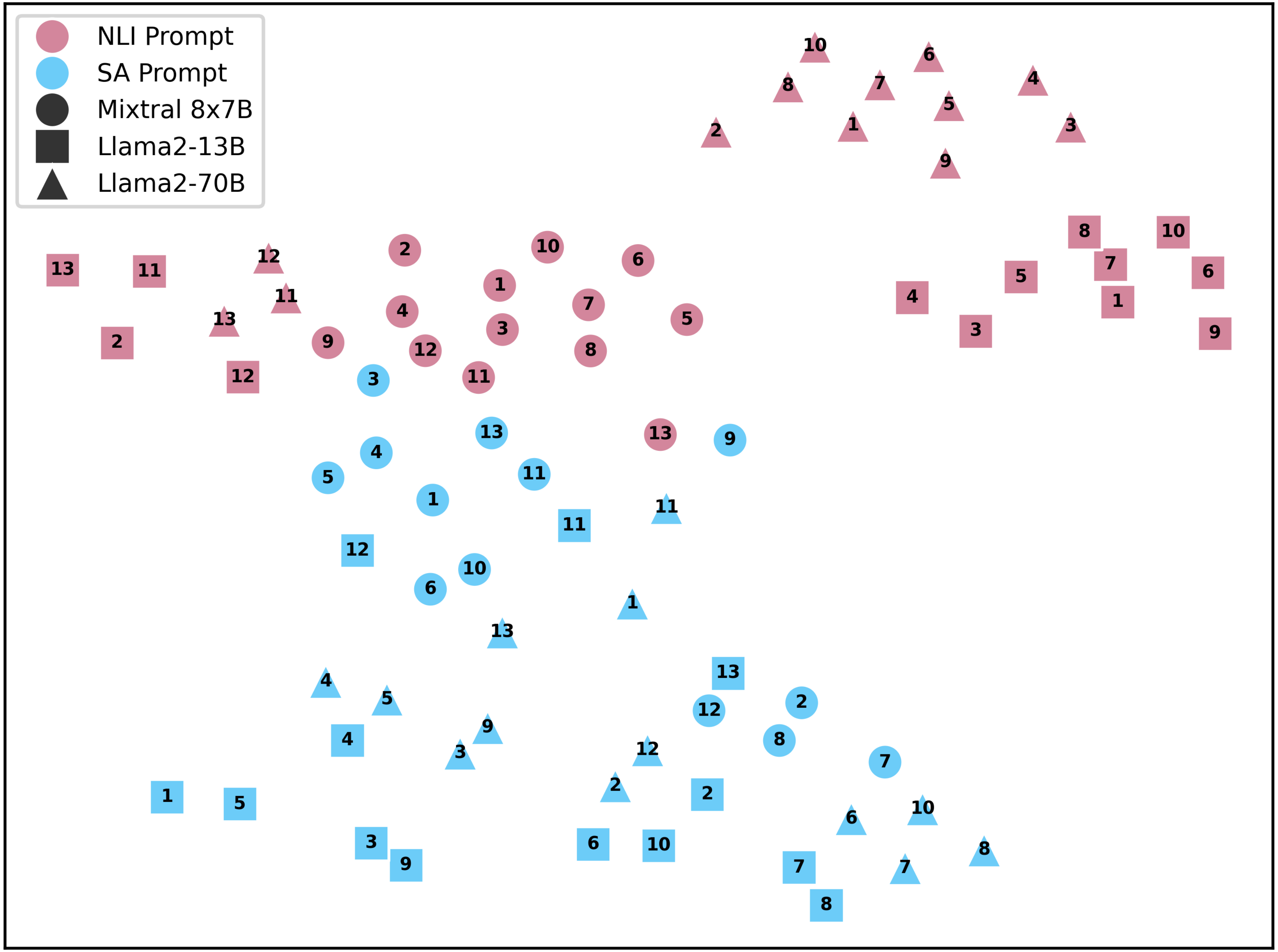
QA: Question Answering task.

(BoolQ is a boolean answer task, which is more similar to CR task.)



Visualization

Task embeddings from our framework extracted from different LLMs guided by different prompts.



NLI: Natural Language Inference.

SA: Sentiment Analysis.

1-10: Vanilla prompts.

11-13: CoT prompts.

(Check paper for detailed prompts)

Experiments

- Transferability experiments: selecting the best source dataset transferred to the target dataset based on the task embedding.
- Our framework retains a performance to be comparable to the existing model-specific methods

Method	CR				QA			
	in-class		*all-class*		*in-class*		*all-class*	
	$\rho\downarrow$	NDCG $\uparrow$	$\rho\downarrow$	NDCG $\uparrow$	$\rho\downarrow$	NDCG $\uparrow$	$\rho\downarrow$	NDCG $\uparrow$
DataSize	3.6	80.4	7.8	75.2	3.2	84.4	11.4	65.8
CurveGrad	5.5	68.6	-	-	8.3	64.8	-	-
TextEmb	5.2	76.4	9.8	74.7	4.1	81.1	5.8	82.0
TaskEmb	2.8	82.3	5.4	78.3	3.2	84.5	5.4	82.8
TuPaTE	**2.5**	83.7	**4.5**	81.0	**3.0**	**85.7**	4.8	83.3
FUTE + TaskEmb	4.4	79.4	7.0	77.9	4.5	83.5	5.3	84.3
FUTE + TuPaTE	3.3	**83.8**	6.2	**82.0**	3.3	85.6	**4.1**	**84.8**

Experiments

- Prompts selection experiments: selecting the best prompts based on the task embedding.
- Our framework also shows comparable performance to other prompts selection methods.

Category	Method	Llama 2 13B Performance	Llama 2 13B Rate	Llama 2 13B NDCG	Llama 2 70B Performance	Llama 2 70B Rate	Llama 2 70B NDCG	Mixtral 8x7B Performance	Mixtral 8x7B Rate	Mixtral 8x7B NDCG
SA	MI	88.0	94.4	59.0	85.3	90.9	72.9	87.2	85.1	65.1
	LocalE	84.3	90.2	47.5	88.3	88.7	57.5	**88.5**	**96.7**	78.6
	GlobalE	89.2	95.7	88.8	91.9	97.9	**82.7**	88.4	96.6	**86.7**
	ZPS-Log	54.3	58.2	38.5	78.0	83.0	54.0	57.0	62.1	33.9
	ZPS-Prob	54.3	58.2	38.5	78.0	83.0	50.8	57.0	62.1	33.9
	ZPS-Vote	54.3	58.2	38.5	78.0	83.0	50.8	57.0	52.1	33.9
	Self-Select	54.3	58.2	42.8	85.3	90.9	69.3	57.0	62.1	38.5
	SPELL	89.2	95.7	**89.6**	79.6	84.6	65.4	57.0	62.1	38.2
	FUTE + TaskEmb	**89.6**	**96.1**	89.4	**93.0**	**99.0**	74.5	86.9	94.9	71.1
	FUTE + TuPaTE	89.2	95.7	55.6	92.4	98.4	67.9	87.9	95.8	52.2
NLI	MI	**46.8**	**90.1**	61.7	48.6	94.8	74.6	37.2	73.6	36.7
	LocalE	37.5	74.4	56.4	43.9	84.6	66.6	39.1	78.6	43.5
	GlobalE	40.4	80.4	65.3	48.2	93.7	76.9	40.2	79.1	44.1
	ZPS-Log	34.8	70.2	48.1	34.9	66.8	49.5	39.0	76.5	41.7
	ZPS-Prob	32.6	65.6	39.7	38.0	73.2	51.2	39.5	78.9	39.3
	ZPS-Vote	32.6	65.6	39.7	33.7	64.3	48.4	39.5	78.9	39.3
	Self-Select	33.9	68.1	39.5	39.7	76.7	53.6	39.1	78.6	43.9
	SPELL	42.4	84.4	**78.1**	48.6	94.8	77.2	39.1	78.6	41.5
	FUTE + TaskEmb	35.8	72.0	47.4	41.1	78.8	60.8	**43.2**	**85.6**	**49.1**
	FUTE + TuPaTE	37.0	75.0	71.8	**50.6**	**98.4**	**81.8**	40.8	81.3	44.4

Thank you!