Prompt Optimization with EASE? Efficient Ordering-aware Automated Selection of Exemplars

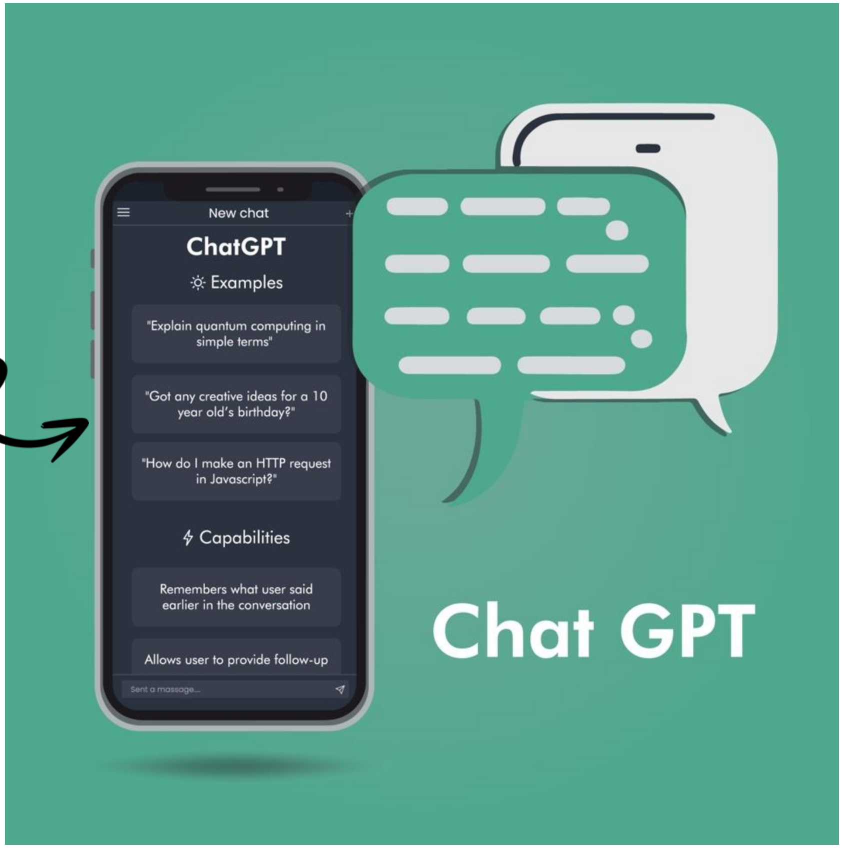
Zhaoxuan Wu[*1,2], Xiaoqiang Lin[*1], Zhongxiang Dai[3], Wenyang Hu[1], Yao Shu[4], See-Kiong Ng[1], Patrick Jaillet[5], Bryan Kian Hsiang Low[1]

Equal Contribution[*]

NUS[1], SMART[2], CUHK(SZ)[3], Guangdong Lab of AI and Digital Economy(SZ)[4], MIT[5]

Large Language Models

Instructions
+ Exemplars



In-Context Learning

The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

The answer is 27.

In-Context Learning

Instruction

Exemplar

Instruction: Answer the question step by step.

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

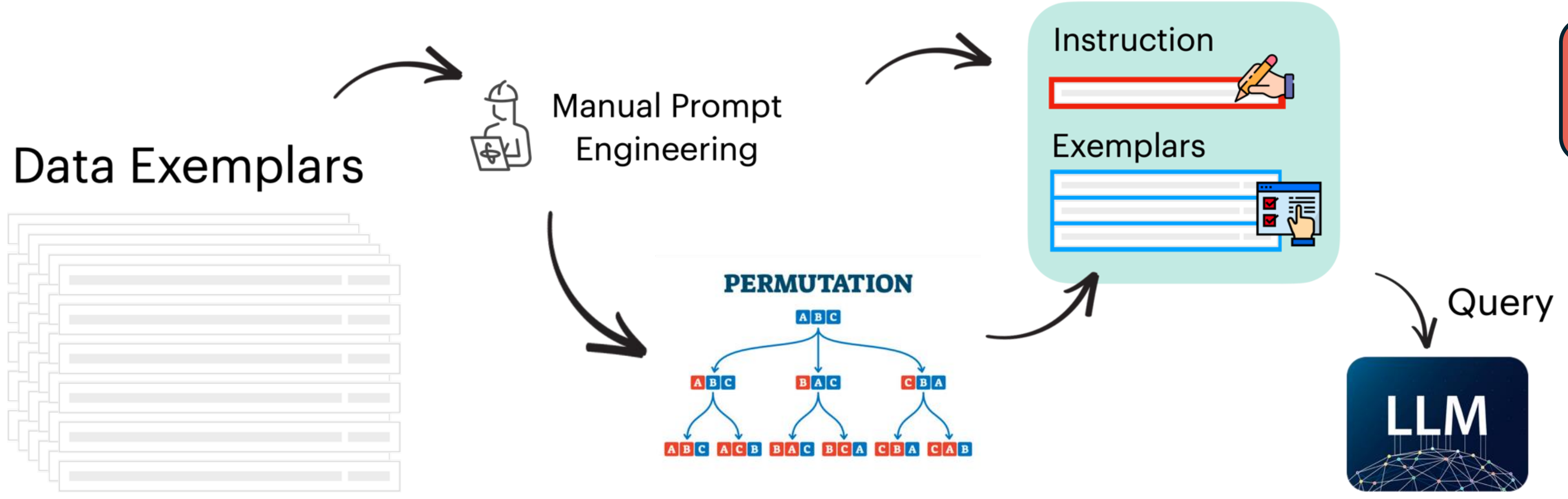
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23– 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9.

Motivation

- Good exemplars and instructions are vital to the performance
- The quality, relevance and even the order of exemplars matters!

How do we design a data selection method for LLM in-context prompting?



Challenge: Best performing LLMs are usually black-box!

Goal: Automate the above!

Formulation

$$\max_{E\in\Omega} F(E) \triangleq \mathbb{E}_{(x,y)\in D_V}[s(f(E,x),y)],$$

$$f([E,x]) = f([\underbrace{e_1,e_2,\ldots,e_k}_{\text{context}},x])$$

where D_V is the validation set, f is a black-box LLM, k exemplars, E is the exemplar sequence and $s(\cdot,\cdot)$ is a score function

How to Optimize?

- We propose to use *neural bandit algorithms*
  - Selects the next input query based on the belief of the objective given all past observations $O_{t-1} := \{(E_i, s_V(E_i))\}_{i=1}^{t-1}$

$$E_t = \arg\max_{E \in \Omega} \mathrm{NeuralUCB}_t(E)$$

a trained neural network

a pretrained embedding model

$$\mathrm{NeuralUCB}_t(E) := \boxed{m(h(E);\theta_t)} + \nu_t \boxed{\sigma_{t-1}(h(E);\theta_t)}$$

exploitation of current score predictions

VS

exploration based on uncertainties of the prediction

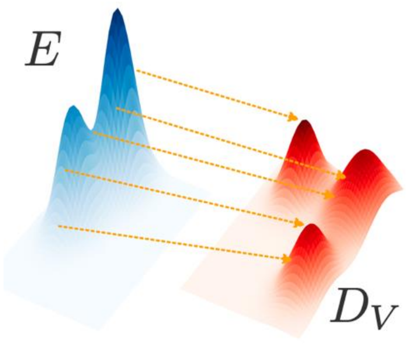
Speeding Up

- Each evaluation of the NeuralUCB acquisition function requires
  1. A forward pass of the embedding model $h(E)$
  2. A forward pass of the NN $m(h(E);\theta_t)$
  3. Computing the uncertainty $\sigma_{t-1}(h(E);\theta_t)$ which involves inverting the NTK matrix, and taking the gradient for the current $h(E)$

Costly!

Speeding Up

- We instead employ a ***filter-then-compute*** strategy
- **Stage 1**: Filter based on the ***inductive bias*** that using exemplars similar to the validation exemplars performs better
  - Optimal Transport distance between $\{e\}_{e \in E}$ and D_V
    - Pre-computations of $h(e)$ is possible
    - Cosine similarity cost function is easy-to-compute $c(h(e), h(e')) = 1 - sim_{cos}(h(e), h(e'))$



- **Stage 2**: Compute NeuralUCB acquisition for the filtered exemplars

Joint Optim. of Exemplars + Instructions

- Naturally extend to

$$E = (p, e_1, e_2, \ldots, e_k)$$

  where instruction $p \in P$

- Intuitively, the instruction p is just "another type of exemplar"

Experiments

	Subset selection methods			Retrieval methods				Heuristics / Uniform		
	DPP	MMD	OT	Cosine	BM25	Active	Inf	Evo	Best-of-N	EASE
antonyms	70.0±0.0	80.0±0.0	81.7±1.7	85.0±0.0	85.0±0.0	80.0±0.0	86.7±1.7	88.3±1.7	**90.0±0.0**	**90.0±0.0**
auto_categorization	3.3±1.7	8.3±1.7	0.0±0.0	25.0±0.0	16.7±1.7	10.0±2.4	21.7±1.7	21.7±1.7	20.0±0.0	**30.0±0.0**
diff	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**
larger_animal	70.0±0.0	91.7±1.7	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**	66.7±1.4	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**
negation	**95.0±0.0**	**95.0±0.0**	**95.0±0.0**	**95.0±0.0**	**95.0±0.0**	**95.0±0.0**	**95.0±0.0**	**95.0±0.0**	**95.0±0.0**	**95.0±0.0**
object_counting	55.0±2.9	56.7±1.7	48.3±1.7	61.7±1.7	66.7±1.7	51.7±1.4	63.3±4.4	70.0±0.0	70.0±0.0	**73.3±1.7**
orthography_starts_with	20.0±2.9	35.0±0.0	61.7±1.7	78.3±1.7	70.0±0.0	43.3±1.4	70.0±2.9	75.0±0.0	78.3±1.7	**80.0±0.0**
rhymes	60.0±0.0	51.7±1.7	0.0±0.0	**100.0±0.0**	80.0±0.0	65.0±8.2	70.0±13.2	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**
second_word_letter	10.0±2.9	30.0±0.0	28.3±1.7	50.0±0.0	50.0±0.0	26.7±8.3	40.0±0.0	46.7±1.7	50.0±0.0	**53.3±1.7**
sentence_similarity	20.0±0.0	21.7±3.3	40.0±2.9	46.7±1.7	53.3±1.7	5.0±4.1	18.3±6.7	45.0±0.0	51.7±1.7	**56.7±1.7**
sentiment	85.0±0.0	90.0±0.0	85.0±0.0	96.7±1.7	**100.0±0.0**	85.0±4.1	91.7±1.7	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**
sum	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**
synonyms	10.0±0.0	25.0±0.0	20.0±0.0	**35.0±0.0**	30.0±0.0	3.3±1.4	26.7±1.7	30.0±0.0	30.0±0.0	30.0±0.0
taxonomy_animal	43.3±4.4	40.0±2.9	46.7±1.7	85.0±2.9	80.0±0.0	45.0±6.2	70.0±5.0	80.0±0.0	80.0±0.0	**88.3±1.7**
translation_en-de	**90.0±0.0**	80.0±0.0	80.0±0.0	**90.0±0.0**	85.0±0.0	56.7±13.0	**90.0±0.0**	**90.0±0.0**	**90.0±0.0**	**90.0±0.0**
translation_en-es	90.0±0.0	**100.0±0.0**	96.7±1.7	**100.0±0.0**	**100.0±0.0**	96.7±1.4	98.3±1.7	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**
translation_en-fr	76.7±1.7	76.7±1.7	81.7±1.7	85.0±0.0	85.0±0.0	81.7±1.4	85.0±0.0	86.7±1.7	85.0±0.0	**88.3±1.7**
word_sorting	26.7±1.7	88.3±1.7	88.3±1.7	90.0±0.0	71.7±1.7	80.0±0.0	88.3±1.7	**93.3±1.7**	91.7±1.7	90.0±0.0
word_unscrambling	68.3±1.7	56.7±1.7	71.7±1.7	75.0±0.0	76.7±1.7	63.3±3.6	66.7±1.7	75.0±0.0	75.0±0.0	**78.3±1.7**
# best-performing tasks	2	2	2	6	4	1	5	9	9	**17**

Experiments

Type	Task	Noise	DPP	MMD	OT	Cosine	BM25	Active	Inf	Evo	Best-of-N	EASE
Rule-based tasks	LR	0%	$31.7_{\pm1.7}$	$38.3_{\pm3.3}$	$50.0_{\pm0.0}$	$71.7_{\pm1.7}$	$70.0_{\pm0.0}$	$36.7_{\pm1.4}$	$56.7_{\pm7.3}$	$61.7_{\pm1.7}$	$66.7_{\pm1.7}$	$\mathbf{80.0}_{\pm2.9}$
		10%	$8.3_{\pm1.7}$	$36.7_{\pm1.7}$	$48.3_{\pm1.7}$	$61.7_{\pm1.7}$	$61.7_{\pm1.7}$	$0.0_{\pm0.0}$	$58.3_{\pm4.4}$	$60.0_{\pm0.0}$	$65.0_{\pm2.9}$	$\mathbf{73.3}_{\pm1.7}$
		30%	$10.0_{\pm0.0}$	$28.3_{\pm1.7}$	$46.7_{\pm1.7}$	$63.3_{\pm1.7}$	$60.0_{\pm0.0}$	$40.0_{\pm2.4}$	$35.0_{\pm2.9}$	$53.3_{\pm1.7}$	$50.0_{\pm0.0}$	$\mathbf{76.7}_{\pm1.7}$
		50%	$0.0_{\pm0.0}$	$38.3_{\pm1.7}$	$45.0_{\pm0.0}$	$65.0_{\pm0.0}$	$53.3_{\pm1.7}$	$0.0_{\pm0.0}$	$53.3_{\pm1.7}$	$46.7_{\pm1.7}$	$45.0_{\pm0.0}$	$\mathbf{78.3}_{\pm4.4}$
		70%	$0.0_{\pm0.0}$	$55.0_{\pm0.0}$	$38.3_{\pm3.3}$	$65.0_{\pm0.0}$	$50.0_{\pm0.0}$	$26.7_{\pm5.4}$	$30.0_{\pm5.8}$	$33.3_{\pm1.7}$	$33.3_{\pm1.7}$	$\mathbf{66.7}_{\pm1.7}$
		90%	$0.0_{\pm0.0}$	$21.7_{\pm1.7}$	$26.7_{\pm1.7}$	$46.7_{\pm1.7}$	$3.3_{\pm1.7}$	$0.0_{\pm0.0}$	$6.7_{\pm3.3}$	$8.3_{\pm1.7}$	$15.0_{\pm0.0}$	$\mathbf{53.3}_{\pm1.7}$
	LP-variant	0%	$48.3_{\pm3.3}$	$40.0_{\pm2.9}$	$41.7_{\pm1.7}$	$65.0_{\pm0.0}$	$58.3_{\pm1.7}$	$30.0_{\pm0.0}$	$61.7_{\pm1.7}$	$75.0_{\pm2.9}$	$71.7_{\pm1.7}$	$\mathbf{75.0}_{\pm0.0}$
		10%	$0.0_{\pm0.0}$	$36.7_{\pm1.7}$	$40.0_{\pm0.0}$	$63.3_{\pm3.3}$	$60.0_{\pm0.0}$	$36.7_{\pm2.7}$	$65.0_{\pm2.9}$	$70.0_{\pm2.9}$	$73.3_{\pm1.7}$	$\mathbf{80.0}_{\pm2.9}$
		30%	$0.0_{\pm0.0}$	$48.3_{\pm3.3}$	$40.0_{\pm2.9}$	$60.0_{\pm0.0}$	$55.0_{\pm0.0}$	$40.0_{\pm7.1}$	$53.3_{\pm6.0}$	$65.0_{\pm2.9}$	$65.0_{\pm0.0}$	$\mathbf{75.0}_{\pm0.0}$
		50%	$0.0_{\pm0.0}$	$65.0_{\pm0.0}$	$35.0_{\pm2.9}$	$63.3_{\pm3.3}$	$60.0_{\pm0.0}$	$38.3_{\pm3.6}$	$48.3_{\pm4.4}$	$61.7_{\pm1.7}$	$65.0_{\pm0.0}$	$\mathbf{78.3}_{\pm1.7}$
		70%	$0.0_{\pm0.0}$	$46.7_{\pm3.3}$	$35.0_{\pm0.0}$	$70.0_{\pm0.0}$	$60.0_{\pm0.0}$	$25.0_{\pm8.2}$	$60.0_{\pm5.0}$	$56.7_{\pm1.7}$	$56.7_{\pm1.7}$	$\mathbf{71.7}_{\pm1.7}$
		90%	$0.0_{\pm0.0}$	$35.0_{\pm2.9}$	$50.0_{\pm0.0}$	$65.0_{\pm2.9}$	$0.0_{\pm0.0}$	$30.0_{\pm12.5}$	$50.0_{\pm2.9}$	$38.3_{\pm1.7}$	$55.0_{\pm2.9}$	$\mathbf{66.7}_{\pm1.7}$
Re-mapped label tasks	AG News Remap	0%	$20.0_{\pm2.9}$	$15.0_{\pm0.0}$	$26.7_{\pm1.7}$	$43.3_{\pm1.7}$	$43.3_{\pm3.3}$	$5.0_{\pm2.4}$	$25.0_{\pm5.0}$	$40.0_{\pm0.0}$	$40.0_{\pm0.0}$	$\mathbf{50.0}_{\pm0.0}$
		10%	$5.0_{\pm0.0}$	$15.0_{\pm0.0}$	$15.0_{\pm0.0}$	$41.7_{\pm1.7}$	$38.3_{\pm1.7}$	$3.3_{\pm1.4}$	$26.7_{\pm3.3}$	$36.7_{\pm1.7}$	$40.0_{\pm0.0}$	$\mathbf{51.7}_{\pm1.7}$
		30%	$10.0_{\pm0.0}$	$5.0_{\pm0.0}$	$5.0_{\pm0.0}$	$40.0_{\pm0.0}$	$36.7_{\pm1.7}$	$1.7_{\pm1.4}$	$10.0_{\pm0.0}$	$40.0_{\pm0.0}$	$43.3_{\pm1.7}$	$\mathbf{55.0}_{\pm0.0}$
		50%	$5.0_{\pm0.0}$	$10.0_{\pm0.0}$	$5.0_{\pm0.0}$	$43.3_{\pm1.7}$	$35.0_{\pm0.0}$	$3.3_{\pm1.4}$	$20.0_{\pm5.0}$	$35.0_{\pm0.0}$	$35.0_{\pm0.0}$	$\mathbf{55.0}_{\pm2.9}$
		70%	$5.0_{\pm0.0}$	$25.0_{\pm0.0}$	$8.3_{\pm1.7}$	$50.0_{\pm0.0}$	$35.0_{\pm0.0}$	$1.7_{\pm1.4}$	$11.7_{\pm6.7}$	$38.3_{\pm1.7}$	$46.7_{\pm1.7}$	$\mathbf{58.3}_{\pm3.3}$
		90%	$5.0_{\pm0.0}$	$18.3_{\pm1.7}$	$5.0_{\pm0.0}$	$40.0_{\pm0.0}$	$10.0_{\pm0.0}$	$15.0_{\pm6.2}$	$35.0_{\pm0.0}$	$35.0_{\pm0.0}$	$41.7_{\pm1.7}$	$\mathbf{53.3}_{\pm1.7}$
	SST5 Reverse	0%	$20.0_{\pm0.0}$	$10.0_{\pm0.0}$	$13.3_{\pm1.7}$	$40.0_{\pm0.0}$	$40.0_{\pm0.0}$	$15.0_{\pm2.4}$	$33.3_{\pm6.7}$	$35.0_{\pm2.9}$	$40.0_{\pm0.0}$	$\mathbf{50.0}_{\pm2.9}$
		10%	$16.7_{\pm1.7}$	$10.0_{\pm0.0}$	$15.0_{\pm0.0}$	$\mathbf{48.3}_{\pm1.7}$	$40.0_{\pm0.0}$	$13.3_{\pm2.7}$	$23.3_{\pm6.7}$	$33.3_{\pm3.3}$	$40.0_{\pm0.0}$	$\mathbf{48.3}_{\pm1.7}$
		30%	$23.3_{\pm1.7}$	$6.7_{\pm1.7}$	$25.0_{\pm2.9}$	$40.0_{\pm0.0}$	$40.0_{\pm0.0}$	$21.7_{\pm3.6}$	$26.7_{\pm1.7}$	$30.0_{\pm0.0}$	$31.7_{\pm1.7}$	$\mathbf{46.7}_{\pm3.3}$
		50%	$21.7_{\pm1.7}$	$15.0_{\pm0.0}$	$15.0_{\pm0.0}$	$43.3_{\pm1.7}$	$33.3_{\pm1.7}$	$21.7_{\pm1.4}$	$23.3_{\pm1.7}$	$28.3_{\pm1.7}$	$30.0_{\pm0.0}$	$\mathbf{46.7}_{\pm3.3}$
		70%	$25.0_{\pm0.0}$	$23.3_{\pm1.7}$	$23.3_{\pm1.7}$	$40.0_{\pm0.0}$	$30.0_{\pm0.0}$	$20.0_{\pm2.4}$	$25.0_{\pm2.9}$	$36.7_{\pm1.7}$	$36.7_{\pm1.7}$	$\mathbf{45.0}_{\pm5.0}$
		90%	$20.0_{\pm0.0}$	$15.0_{\pm2.9}$	$20.0_{\pm0.0}$	$30.0_{\pm0.0}$	$30.0_{\pm0.0}$	$13.3_{\pm2.7}$	$21.7_{\pm1.7}$	$30.0_{\pm0.0}$	$30.0_{\pm0.0}$	$\mathbf{31.7}_{\pm1.7}$

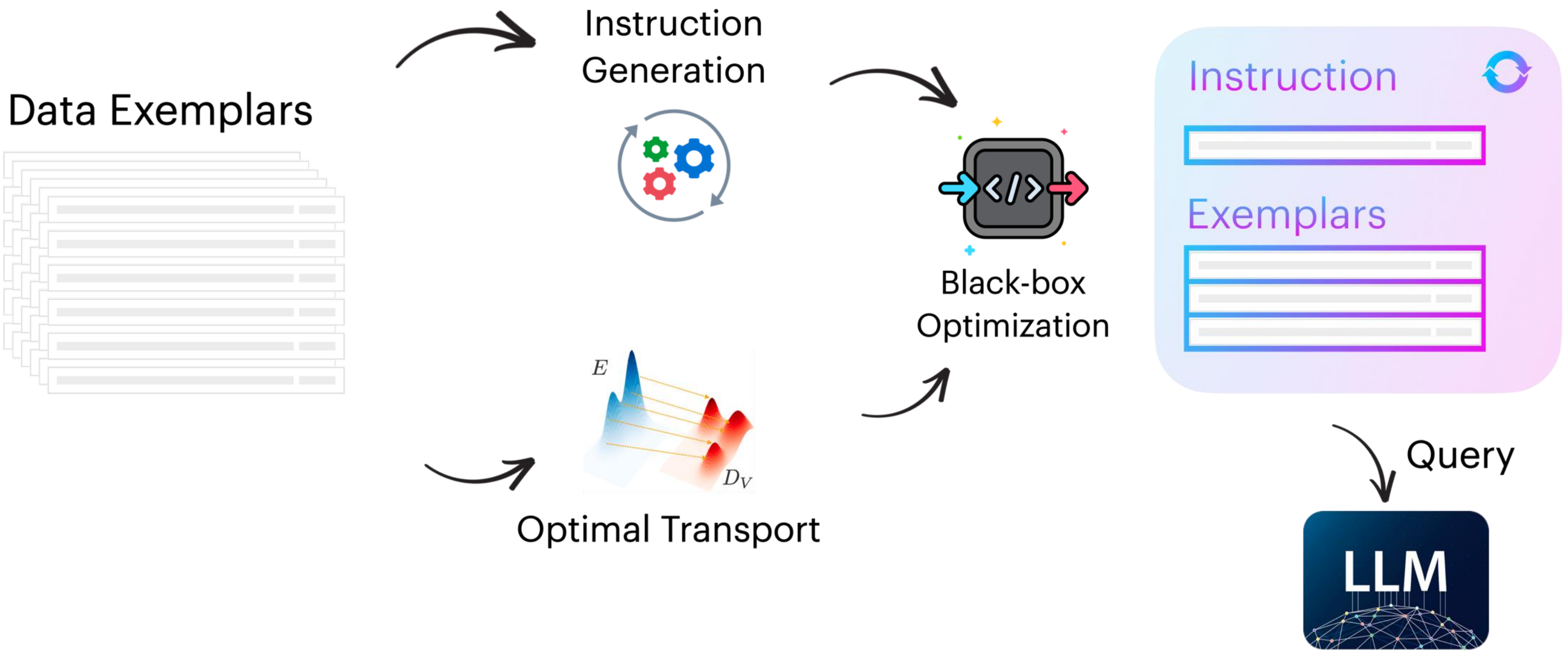
Effective!

Further Improvement with Instructions

	EASE	EASE with instructions	Improve -ment
antonyms	**90.0** $_{\pm 0.0}$	85.0 $_{\pm 0.0}$	-5.0 ↓
auto_categorization	30.0 $_{\pm 0.0}$	**56.7** $_{\pm 1.7}$	26.7 ↑
negation	95.0 $_{\pm 0.0}$	**100.0** $_{\pm 0.0}$	5.0 ↑
object_counting	73.3 $_{\pm 1.7}$	**75.0** $_{\pm 0.0}$	1.7 ↑
orthography_starts_with	80.0 $_{\pm 0.0}$	**81.7** $_{\pm 1.7}$	1.7 ↑
second_word_letter	53.3 $_{\pm 1.7}$	**100.0** $_{\pm 0.0}$	46.7 ↑
sentence_similarity	56.7 $_{\pm 1.7}$	**58.3** $_{\pm 1.7}$	1.7 ↑
synonyms	30.0 $_{\pm 0.0}$	**31.7** $_{\pm 1.7}$	1.7 ↑
taxonomy_animal	88.3 $_{\pm 1.7}$	**100.0** $_{\pm 0.0}$	11.7 ↑
translation_en-de	**90.0** $_{\pm 0.0}$	**90.0** $_{\pm 0.0}$	0.0 ∘
translation_en-fr	**88.3** $_{\pm 1.7}$	85.0 $_{\pm 0.0}$	-3.3 ↓
word_sorting	90.0 $_{\pm 0.0}$	**93.3** $_{\pm 1.7}$	3.3 ↑
word_unscrambling	78.3 $_{\pm 1.7}$	**80.0** $_{\pm 0.0}$	1.7 ↑
LR (10% noise)	**73.3** $_{\pm 1.7}$	45.0 $_{\pm 15.0}$	-28.3 ↓
LP-variant (10% noise)	80.0 $_{\pm 2.9}$	**86.7** $_{\pm 1.7}$	6.7 ↑
AG News Remap (10% noise)	51.7 $_{\pm 1.7}$	**65.0** $_{\pm 0.0}$	13.3 ↑
SST5 Reverse (10% noise)	48.3 $_{\pm 1.7}$	**53.3** $_{\pm 1.7}$	5.0 ↑

Joint optimization further improves performance!

Summary of EASE



EASE Conclusion

- A novel algorithm that selects the optimal ordered set of exemplars for in-context learning of black-box LLMs in an automated fashion
  - Proposed a ***query-efficient*** neural bandit approach
  - Made ***computationally feasible*** through a technique based on optimal transport
  - Extended to a fully automated pipeline that ***jointly optimize*** instructions and exemplars

- **Data selection is also important in the era of LLM!**
- **Highly practical to use data selection for improving downstream usage of black-box LLMs!**