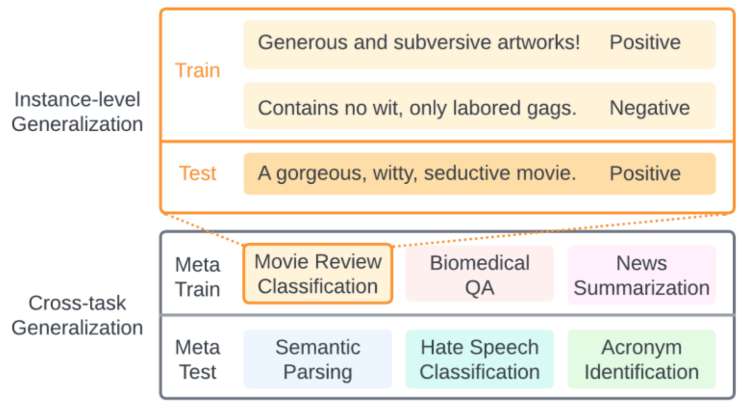
Cross-task Generalization Abilities of Large Language Models

Qinyuan Ye qinyuany@usc.edu @qinyuan_ye
USC University of Southern California · USC Information Sciences Institute · ink

Instance-level vs. Cross-task Generalization

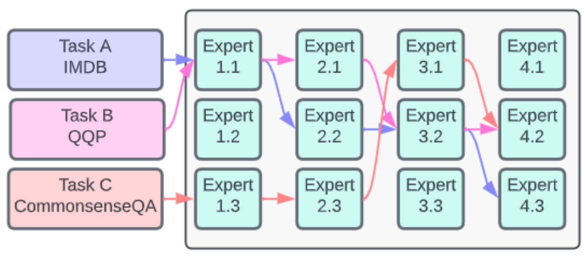


What? **Gain knowledge and experience from seen tasks. Learn more efficiently when encountering new tasks.**

Why? ① It can help **reduce task-specific efforts** when we develop new NLP applications in the future.
② We should **evaluate intelligent systems** not only on their skills, but also **on skill-acquisition efficiency**.

Modeling: Cross-Task MoE for Modularity

Task-level Mixture-of-Experts

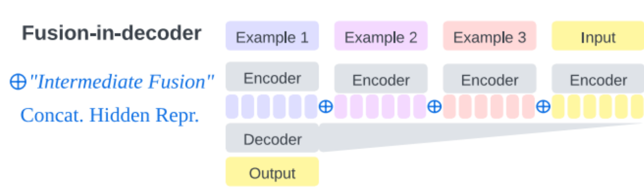


We train **task-level MoE models** to multi-task on NLP tasks.

- Naïve multi-task learning is sub-optimal due to task interference. Task-level MoE addresses this issues and improves generalization to unseen tasks.
- The MoE model partly rediscovers human categorization of NLP tasks (by itself!). Certain experts are strongly associated with *extractive* tasks, some with *classification* tasks, and some with tasks requiring *world knowledge*.

Modeling: FiD-ICL for Inference Efficiency

FiD-ICL



We adapt **fusion-in-decoder models** (Izacard et al., 2020; originally designed for open-domain QA) to perform **in-context learning**.

- **Strong ICL performance on unseen tasks**
  - FiD-ICL outperforms Concat-ICL and Ensemble-ICL.
  - The gap between FiD-ICL and fine-tuning is <3% on P3 meta-test tasks.
- **Faster Inference**
  - FiD-ICL is faster than Concat-ICL and Ensemble-ICL
  - More efficient than fine-tuning when considering optimization costs.

Benchmarking: The CrossFit Challenge

The CrossFit Challenge

Large-scale Pre-training (e.g., BART, T5 models)
+ Upstream Learning on a set of meta-train (seen) tasks
with: Multi-task learning · First-order MAML · Model-agnostic Meta-learning · Reptile
+ Downstream Fine-tuning on meta-test (unseen) tasks

NLP Few-shot Gym

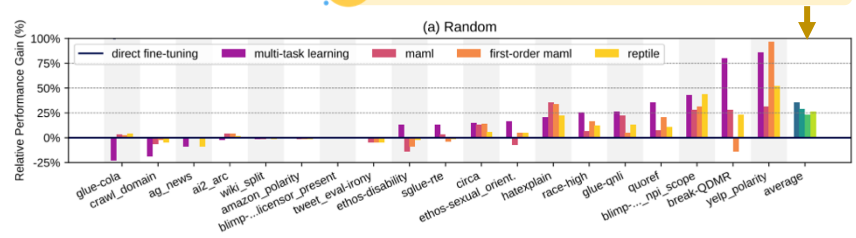
- **160 few-shot NLP tasks**, covering four task categories
  - Classification
  - Question Answering
  - Conditional Generation
  - Others
- Accessed and processed with 🤗 huggingface datasets
- Converted to a **unified text-to-text format**

Evaluation Metric

- We measure the success with **average relative gain (ARG)**: How much does the performance change *with/without* the **upstream learning phase**? (averaged over all test tasks)

Result Snippet

On average, ~**25%** gain on unseen tasks!



Key Takeaways

- Upstream learning on diverse NLP tasks enables cross-task generalization.
- Multi-task learning matches or outperforms more complex meta-learning algorithms.
- Similarity in task format does not fully explain how models learn transferable skills.
- Applying task-specific prompts to *only* meta-test tasks leads to worse performance.
  - *Both* meta-train and meta-test tasks should be formatted with prompts. → Instruction Tuning

Analysis: Predicting LLM Generalization Landscape

Model Family	# param	Task	# shot	Perf.
GPT-3	3B	strategy_qa	0	0.48
BIG-G T=1	8B	elementary_math	3	0.19
PaLM	64B	code_line_desc	2	0.23
GPT-3	6B	elementary_math	1	?

We train regression models to **predict LLM performance on unseen experiment configurations**.

- LLMs' performance follows predictable patterns. Our model achieves an **RMSE<0.05** in a random train-test split.

Ongoing and Future Work

Pushing the limit of in-context learning

Current research efforts mainly focus on ICL *with examples of one single task*. Will LLMs benefit from diverse and heterogeneous contexts?

From data-sufficient learners to self-sufficient learners

So far, we prepare the few-shot examples for the LLMs. Can we enable them to learn in the open-endedness by themselves?