

# From Instance Training to Instruction Learning: Task Adapters Generation from Instructions

NeurIPS 2024

Huanxuan Liao,   Shizhu He,   Yao Xu,   Yuanzhe Zhang,   Yanchao Hao,   
Shengping Liu,  Kang Liu,   Jun Zhao,  

## ◆ Cross-task Generalization


- Develop models capable of effectively transferring knowledge across a diverse range of tasks. This involves training on a set of tasks (**meta-train phase**) and evaluating the model's ability to perform on unseen tasks (**meta-test phase**).

## ◆ Challenges

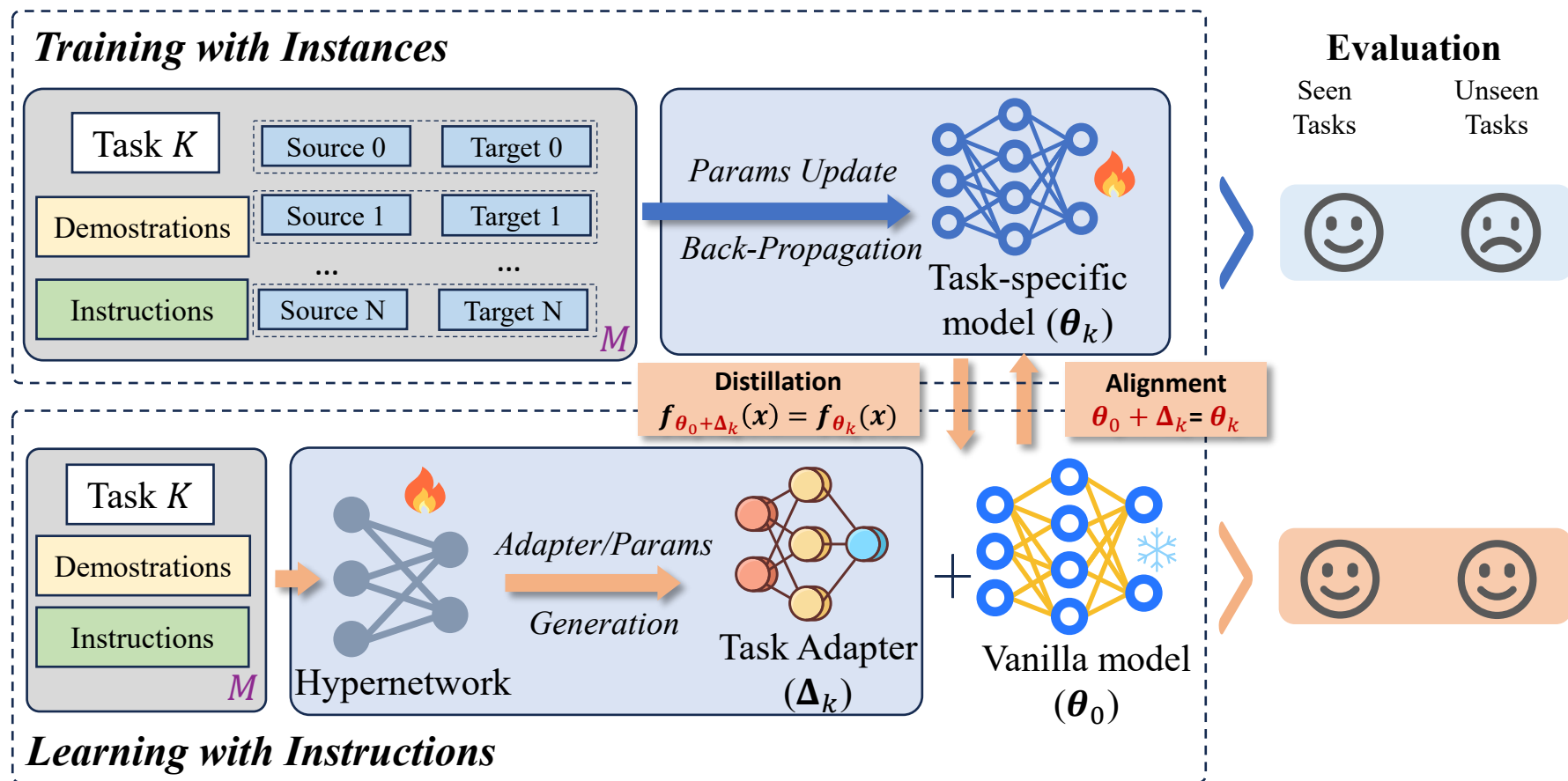
- Traditional instance training methods require **extensive task-specific data** , limiting adaptability in real-world scenarios.
- The need to process instructions repeatedly for the same task (all need to concatenate the task description and demonstrations) leads to **high computational costs**.

**Can we mimic the way humans understand and follow instruction descriptions when learning new skills and tasks to assist in solving new problems?**

# Instruction Learning

 **Instruction Learning:** By learning from instructions, a class of tasks is treated **as a whole**, allowing for a higher-level understanding of instructions and the ability to address problems effectively.

- The hypernetwork automatically transform instructions and demonstrations into **efficient and lightweight task adapters**, seamlessly integrating them into LLMs.
- **Knowledge distillation** reinforces generated adapters' consistency with task-specific models developed through instance training.



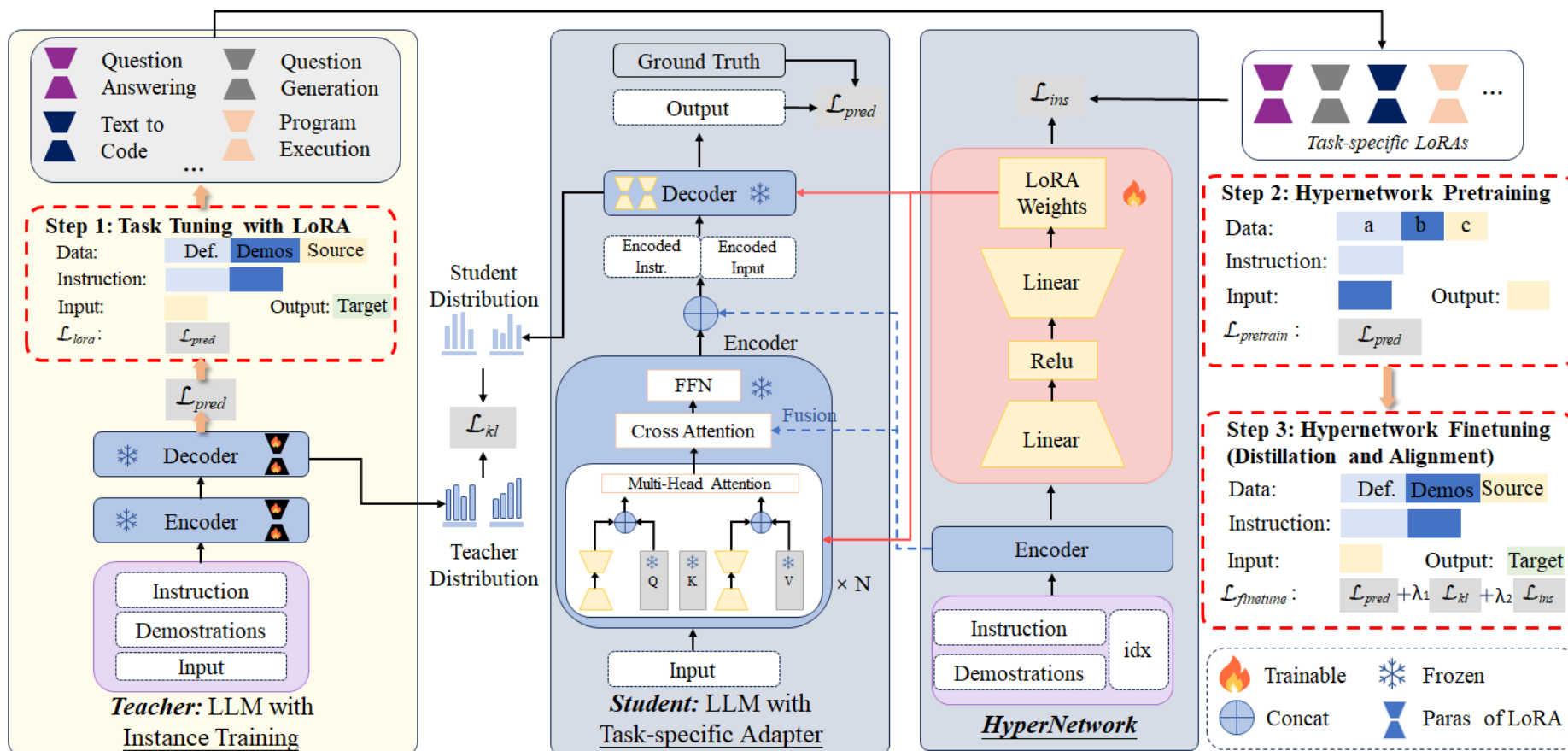
# Comparison of TAGI and Baselines

- **Instruction Fusion:** TAGI incorporates instruction fusion, allowing dynamic interaction between inputs and instructions, enriching the model's understanding and enhancing performance on unseen tasks.
- **Low Inference Cost:** TAGI processes task instructions **only once**, significantly reducing computational overhead during inference, especially beneficial as the number of samples or instruction length increases.
- **Update with Low Parameters:** TAGI requires minimal parameter updates, leveraging a hypernetwork to generate task-specific adapters.
- **PreTraining:** TAGI benefits from pretraining, which enhances the model's ability to comprehend and execute task instructions effectively.
- **Perform Unseen Tasks Well:** TAGI is designed to learn and perform well on unseen tasks.

Method	Meta-Train	Pre-Train	Instr. Concat.	Instr. Fus.	Low Up. Params	Low Infer. Cost	Instr. Learning	Unseen Task
Simple FT	X	X	✓	✓	X	X	X	X
T0 / Tk-Instruct	✓	X	✓	✓	X	X	X	✓✓✓
Hypter	✓	X	X	X	✓	✓	X	✓
HyperTuning	✓	✓	X	X	✓	✓	X	✓
HINT	✓	✓	✓	X	✓	✓	X	✓✓
<b>TAGI (Ours)</b>	✓	✓	✓	✓	✓	✓	✓	✓✓✓

# TAGI Architecture

- ◆ **Hypernetwork Pretraining:** Pretrains the hypernetwork on standard text data to enhance its ability to recognize and respond to instructions.
- ◆ **Hypernetwork Finetuning (Distillation and Alignment):** Finetunes the hypernetwork on meta-training tasks to learn the generation of optimal parameters from task instructions. Utilizes **knowledge distillation** to align the task-specific model (acting as the teacher) with the vanilla LLM combined with the generated task adapters (acting as the student).



Instruction fusion:

$$\mathbf{F}_l = \text{CrossAttentionLayer}_l(\mathbf{S}_l, \mathbf{I}_x)$$

LoRA Generation:

$$\text{LoRA}_l^{\{Q, V\}} = \text{MLP}_l(\mathbf{I}_{x_k}; \text{idx}_l^{\{Q, V\}} | \text{idx}_l^Q = 2l, \text{idx}_l^V = 2l + 1)$$

# Main Results: SNI



RougeL results on **Super-Natural Instructions (SNI)**. The best results are in bold, while the second-best are underlined. The Average Relative FLOPs cost is calculated relative to Tk-Instruct. We use the number of FLOPs required by each model to process one task (containing 100 examples).

Method	Def (Zero-shot)			Def + 2 Pos. (Few-shot)			Avg. Rel.
	Base (250M)	XL (3B)	XXL (11B)	Base (250M)	XL (3B)	XXL (11B)	FLOPs
No FT	8.8	14.3	26.2	9.4	13.6	30.5	×1.0
Tk-Instruct <sup>†</sup>	<b>35.3</b>	<u>48.0</u>	<b>53.6</b>	<u>42.1</u>	54.0	<b>62.0</b>	×1.0
<i># Decoder-only model</i>							
GPT-2 XL (1.5B)*	-	38.2	-	-	45.3	-	× <b>0.33</b>
OPT (13B)*	-	-	44.8	-	-	51.5	×0.36
<i># Hypernetwork-based model</i>							
Hypter*	12.1	16.8	15.5	10.6	14.2	13.4	×0.35
HyperTuning <sup>†</sup>	-	38.9	-	-	48.6	-	× <u>0.34</u>
HINT*	<u>33.3</u>	47.2	51.1	41.8	53.2	56.4	×0.37
<b>TAGI (Ours)</b>	<b>35.3</b>	<b>48.4</b>	<u>52.3</u> ‡	<b>42.5</b>	<b>56.3</b>	<u>58.4</u> ‡	×0.39

# Main Results: P3

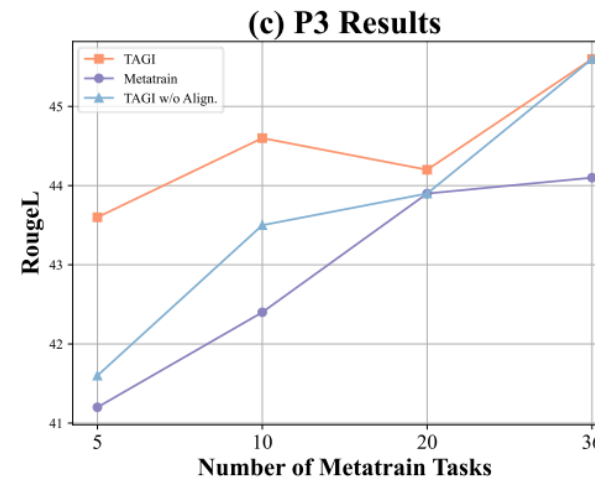
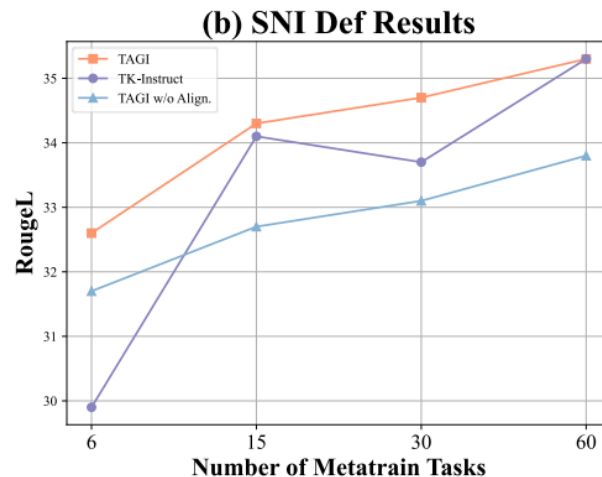
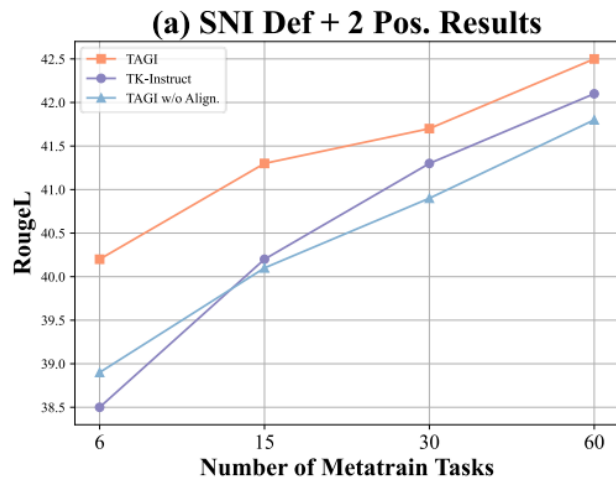


Average accuracy results over T0 evaluation tasks after training on the T0 P3 train set. Our method uses only template inputs without demonstrations yet achieves competitive performance with ICL-based methods using 16 shots, with much-reduced inference overhead. The Average Relative Inference Time is calculated relative to the Metatraining. We use the inference time required by each model to process all 11 test tasks with batch\_size of 1.

Method	T5-LM			T0			Avg. Rel.
	Base (250M)	Large (800M)	XL (3B)	Base (250M)	Large (800M)	XL (3B)	Infer. Time
<i># MTest11 Avg.</i>							
Zero-shot	43.9	41.5	42.6	49.1	52.4	57.6	×1.0
Full FT	44.6	45.5	47.2	<b>51.9</b>	<b>56.6</b>	<b>61.4</b>	×1.0
Metatraining <sup>♡</sup>	44.1	52.4	53.1	50.1	52.4	56.8	×1.0
<i># ICL-based method</i>							
Concat-ICL <sup>α</sup>	44.2	47.6	-	48.6	53.2	-	×4.1
FiD-ICL <sup>α</sup>	<b>47.0</b>	<b>55.2</b>	<b>60.0</b>	<u>51.0</u>	53.4	58.2	×1.9
Ensemble-ICL <sup>α</sup>	44.6	54.5	52.6	49.9	53.7	57.7	×13.2
<i># Hypernetwork-based model</i>							
Hypster*	-	-	-	-	-	56.2	-
HINT*	-	-	-	-	-	60.3	-
<b>TAGI (Ours)</b>	<u>45.6</u>	<u>54.7</u>	<u>58.9</u>	50.8	<u>53.8</u>	58.8	× <b>0.88</b>
<i># HyperT5 Avg. (Without SCloze dataset)</i>							
FiD-ICL <sup>α</sup>	<b>46.9</b>	<b>55.8</b>	60.6	<b>51.7</b>	<u>53.9</u>	58.5	×1.9
HyperTuning <sup>†</sup>	-	54.6	59.6	-	-	-	-
<b>TAGI (Ours)</b>	<u>46.7</u>	<u>56.0</u>	<u>59.8</u>	<b>51.7</b>	<b>54.6</b>	<b>59.2</b>	× <b>0.88</b>

# Ablation Results

The performance of different numbers of meta-training tasks.



**Ablation Study:** Demonstrate that the inclusion of **pretraining**, **instruction fusion**, and **alignment** are crucial for enhancing cross-task generalization, with each component significantly contributing to the model's overall efficacy, leading to a **5%** improvement over baselines in **cross-task** performance.

Method	Def	Def + 2Pos.	P3
TK-Instruct	48.0	54.0	-
TK-Instruct-LoRA	47.5	54.6	-
TK-Instruct-Prefix	42.6	54.2	-
Hypertuning	38.9	48.6	59.6
HINT	47.2	53.2	60.3
TAGI	<b>48.4</b>	<b>56.3</b>	<b>60.6</b>
<b>Ablation Study</b>			
w/o pretraining	47.1	55.6	58.3
w/o Instr. Fus.	35.1	40.6	44.2
w/o $\mathcal{L}_{ce}$	47.6	55.4	59.8
w/o $\mathcal{L}_{kl}$	45.7	53.9	57.3
w/o $\mathcal{L}_{ins}$	47.5	55.2	59.4
w/o Hypernetwork	43.8	50.7	-