Improving In-Context Few-Shot Learning via Self-Supervised Training

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov,
Srini Iyer, Veselin Stoyanov, Zornitsa Kozareva

In-Context Few-Shot Learning

Solve unseen tasks at inference time
while forgoing any weight updates

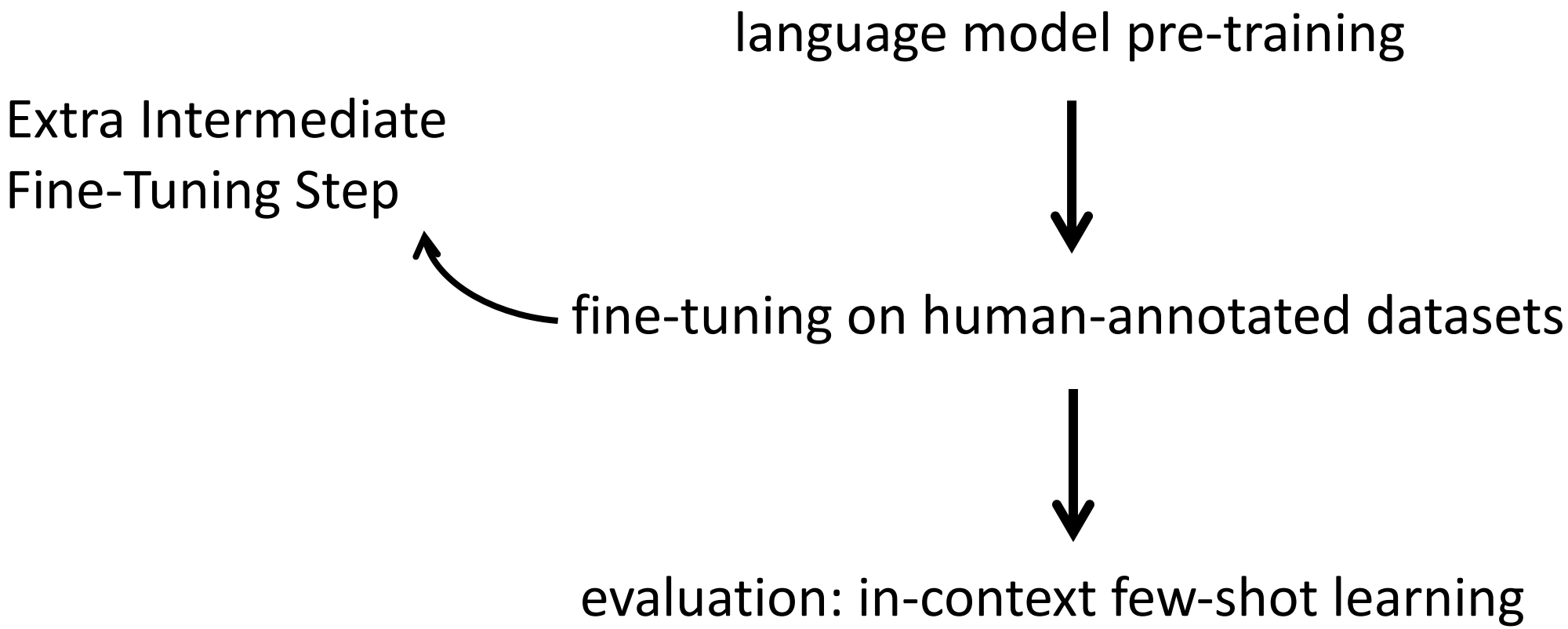
In-Context Few-Shot Learning

Task demonstration

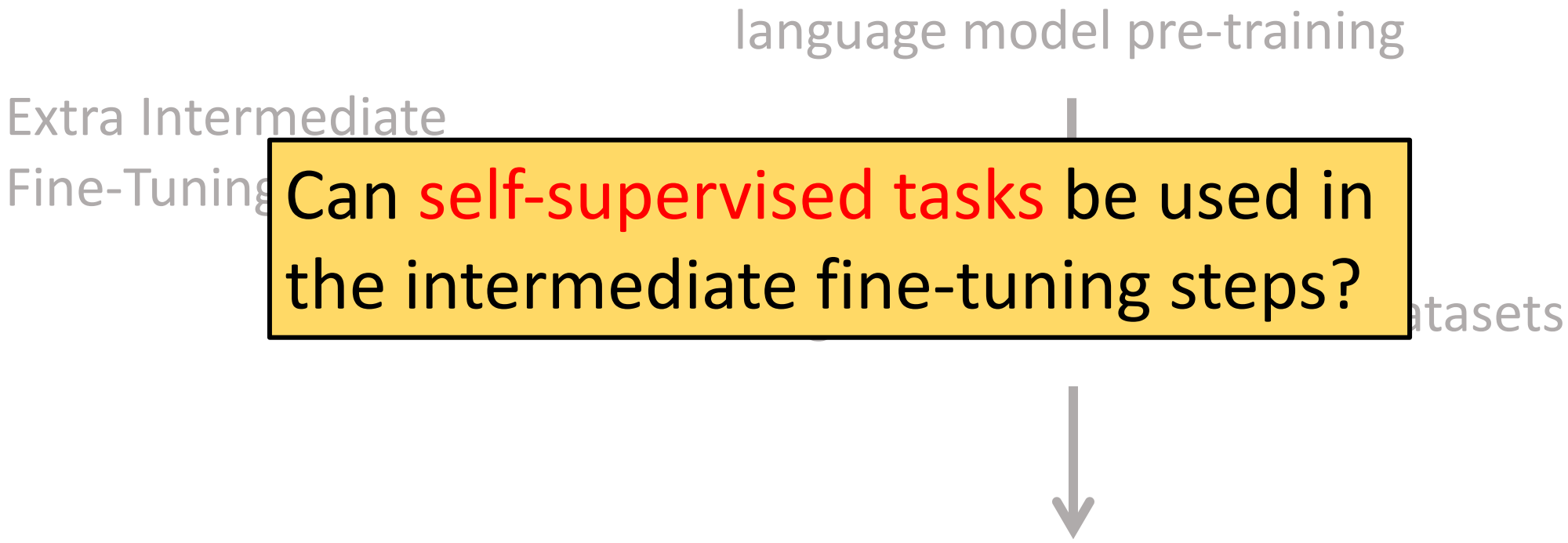
Input: Context word: fit. Question: The trophy doesn't fit into the brown suitcase because __ is too large.
Output: trophy.
… (extra input-output pairs for the same task)
Input: Context word: water. Question: I poured water from the bottle into the cup until the __ was empty.
Output: bottle.

Model prediction

Prior Work


language model pre-training
Extra Intermediate
Fine-Tuning Step
fine-tuning on human-annotated datasets
evaluation: in-context few-shot learning

Prior Work



Can self-supervised tasks be used in the intermediate fine-tuning steps?

Self-Supervised Tasks

- Next Sentence Generation: generating next sentence conditioned on previous sentences
- Masked Word Prediction: generating masked out words
- Last Phrase Prediction: generation last phrases or classifying whether the shown last phrase is the correct one
- Classification: classifying whether the input has the correct properties: e.g., next sentence prediction

Baselines

- ExtraLM: Perform additional LM pre-training on the portion of the original raw text used in our self-supervised training
- CrossTask: Using human-annotated datasets in the intermediate fine-tuning step
- See our paper for more baseline results!

Experimental Results

	BoolQ	MultiRC	COPA	RTE	CB	Avg.
LM	48.6	5.5/53.7	83.4	51.9	53.6	51.8
ExtraLM	49.6	4.9/54.8	82.6	52.9	51.4	51.7
CrossTask	53.4	1.2/57.2	76.2	**54.3**	44.6	49.6
SelfSup	**61.7**	**5.2/62.1**	**84.0**	53.1	**54.3**	**55.6**

SuperGLUE Results

	QG	AG	MM	VF	Avg.
GPT3	43.0	50.0	70.0	32.0	48.8
LM	40.9	32.5	74.0	27.8	43.8
ExtraLM	41.1	32.7	**75.9**	25.2	43.7
CrossTask	38.1	**41.6**	69.2	23.0	42.9
SelfSup	**43.9**	37.5	72.3	**28.6**	**45.5**

Natural-Instructions Results

Experimental Results

	BoolQ	MultiRC	COPA	RTE	CB	Avg.
LM	48.6	5.5/53.7	83.4	51.9	53.6	51.8
ExtraLM	49.6	4.9/54.8	82.6	52.9	51.4	51.7
CrossTask	53.4	1.2/57.2	76.2	**54.3**	44.6	49.6
SelfSup	**61.7**	**5.2/62.1**	**84.0**	53.1	**54.3**	**55.6**

SuperGLUE Results

	QG	AG	MM	VF	Avg.
GPT3	43.0	50.0	70.0	32.0	48.8
LM	40.9	32.5	74.0	27.8	43.8
ExtraLM	41.1	32.7	**75.9**	25.2	43.7
CrossTask	38.1	**41.6**	69.2	23.0	42.9
SelfSup	**43.9**	37.5	72.3	**28.6**	**45.5**

Natural-Instructions Results

Experimental Results

- We also conducted analysis finding that the downstream task performance can be affected by
  - The amount of self-supervised training data
  - The choice of self-supervised tasks
  - The templates we used to format the self-supervised tasks
  - …
- See our paper for more details!

Summary

- Experimentally, we evaluate four self-supervised tasks on two benchmarks.
- We showed that self-supervised tasks can improve model performance on in-context few-shot learning
- Our paper has more detailed experiments and analysis, including experiments characterizing the benefits of self-supervised tasks