Fine-tuning by Self-improvement

Fine-tuning a Source model to exhibit the behavior of a Target model

- requires a number of training examples of the Target task

Obtaining these examples can be difficult

- human labeling: labor intensive
- limited number of naturally-occurring examples

Thus, it is often necessary

- to augment the initial, limited number of examples
- with synthetic examples

One way to generate the synthetic examples

- ask a strong LLM (not necessarily the one for the Source task) to create them for you !

Although these may be less than perfect, they may still help to advance the fine-tuning toward the Target task.

This suggests an iterative process of *Self-Improvement*

- train the Target model $\backslash\mathrm{model}$ in stages
  - creating a sequence of fine-tuned Target models $\backslash\mathrm{model}_{(0)}, \backslash\mathrm{model}_{(1)}, \ldots$
  - of increasing power
- base case
  - fine-tune initial Target $\backslash\mathrm{model}_{(0)}$
  - using a mixture of strong (human-generated) and weak (LLM generated) fine-tuning examples of the Target task
  - resulting in weak Target model $\backslash\mathrm{model}_{(1)}$
- inductive case
  - create improved Target $\backslash\mathrm{model}_{(+1)}$
  - by fine-tuning $\backslash\mathrm{model}_{\backslash\mathrm{tp}}$
    - with the strong examples we already have
    - augmented with examples created as outputs of Target model $\backslash\mathrm{model}_{\backslash\mathrm{tp}}$

To make this more powerful

- filter the outputs of model $\backslash\mathrm{model}_{\backslash\mathrm{tp}}$
- using some metric of quality

so that the augmented examples used to obtain fine-tuned model $\backslash\mathrm{model}_{(+1)}$ are higher quality.

Self Improvement of an LLM

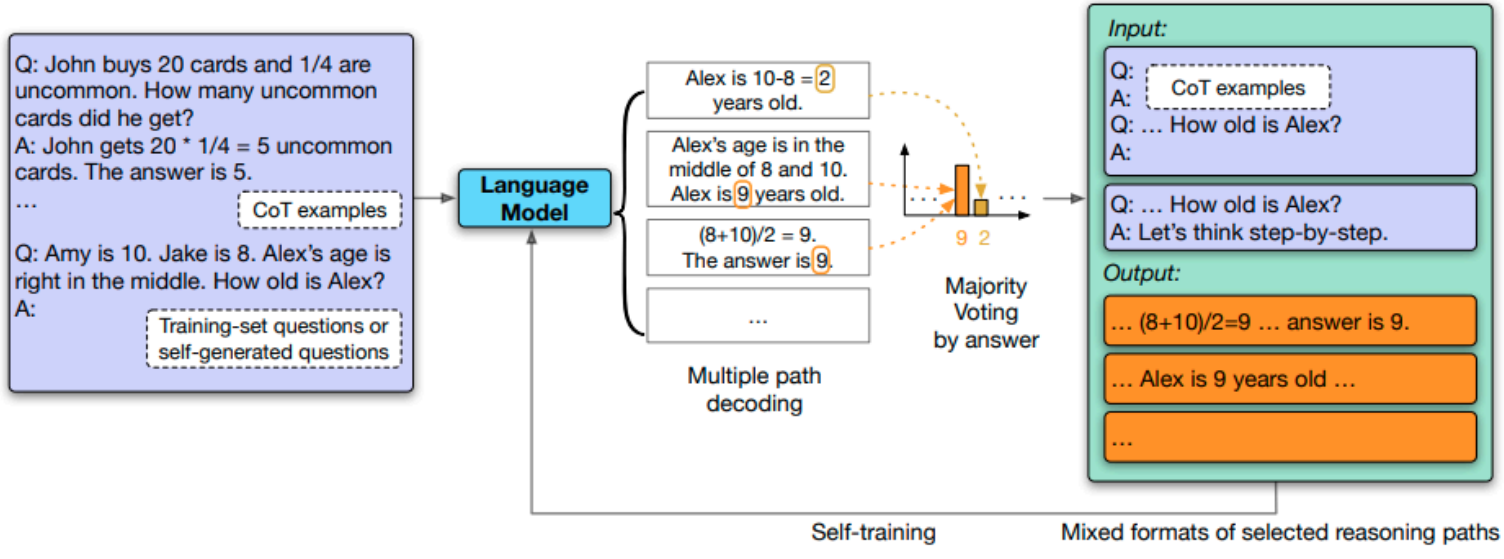

Figure 1: Overview of our method. With Chain-of-Thought (CoT) examples as demonstration (Wei et al., 2022b), the language model generates multiple CoT reasoning paths and answers (temperature $T > 0$) for each question. The most consistent answer is selected by majority voting (Wang et al., 2022b). The "high-confidence" CoT reasoning paths that lead to the majority answer are augmented by mixed formats as the final training samples to be fed back to the model for fine-tuning.

Attribution: https://arxiv.org/pdf/2210.11610.pdf#page=2

In [2]:
```
print("Done")
```

```
Done
```