ALERT: Adapting language models to reasoning tasks

Ping Yu ♠, Tianlu Wang ♠, Olga Golovneva ♠, Badr AlKhamissi△, Siddharth Verma△
Zhijing Jin△❤, Gargi Ghosh ♠, Mona Diab ♠, Asli Celikyilmaz ♠
♠ Meta AI △ Work done at Meta AI ❤ Max Planck Institute & ETH

A **benchmark** and suit of **analyses** for evaluating reasoning skills of language models. Our benchmark provides a test bed to assess any language model on fine-grained reasoning skills, which spans over 20 datasets and covers 10 different reasoning skills.

We investigate the ***role of finetuning***, discovered **3 findings**:

- Our experiments indicate that there is no strong correlation between high vocabulary overlap (between finetuning and evaluation datasets) and performance gain on reasoning evaluation datasets. **This means that LLMs are not simply memorizing the training data during the finetuning.**
- Finetuning helps improve **reasoning capabilities of LLMs** (e.g. analogical and abductive) but not all of them (e.g. commonsense reasoning);
- Finetuning can cause **overfitting towards data format**, which makes it harder for LLMs to generalize to other prompt templates, while CoT-finetuning helps to mitigate this issue as it incorporates a variety of explanations.

Reasoning benchmark

Reasoning Skills	Datasets
Logical	bigbench repeat copy logic, mmmlu answer generation
Causal	plausable result generation, anli r2 entailment, anli r3 entailment, cb entailment
Commonsense	piqa answer generation, commongen sentence generation, sciq answer generation, openbookqa question answering
Entailment	nli r2 entailment, anli r3 entailment, cb entailment, lue entailment classification
Mathematics	semeval closed vocabulary math, semeval geometric math, mmmlu formal logic
Abductive	tellmewhy
Spatial	babi t1 single supporting fact, piqa answer generation, toqa find location easy clean
Analogical	commongen sentence generation, bard analogical reasoning causation
Argument	argument stance classification, argument consequence classification
Deductive	rocstories correct answer generation

Table 1: ALERT benmark consists of 20 datasets covering 10 different reasoning skills.

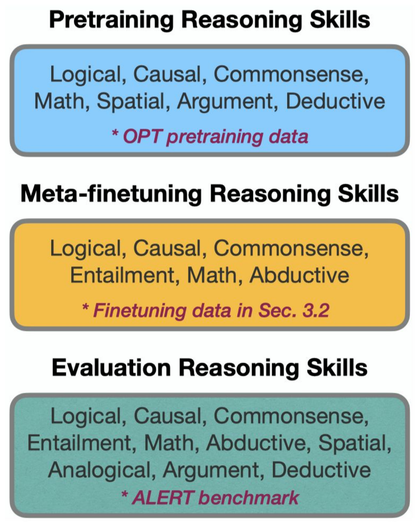

Figure 4: Reasoning skills learned during pretraining and meta-finetuning stages, as well as tested through ALERT .

Experiments

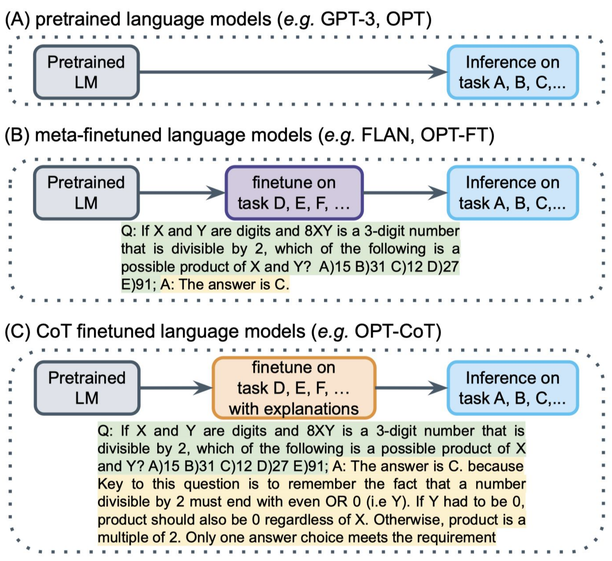

Figure 1: Three types of models are compared.

- 3 types of models: (A) directly apply pretrained LLMs on reasoning tasks; (B) finetune LLMs on a set of tasks; (C) finetune LLMs on tasks with explanations (CoT-finetuning). Finetuning data contains source and target parts, and the language modeling loss only applied to the target part.
- 5 different templates during finetuning and inference.
- 2 scales: 1.3B and 13B

Does finetuning help?

In Figure 2, Rationale-based finetuning (OPT-CoT) improves the performance of both 1.3B and 13B models. However, finetuning (OPT-FT) sometimes yields worse results than the vanilla pre-trained model.

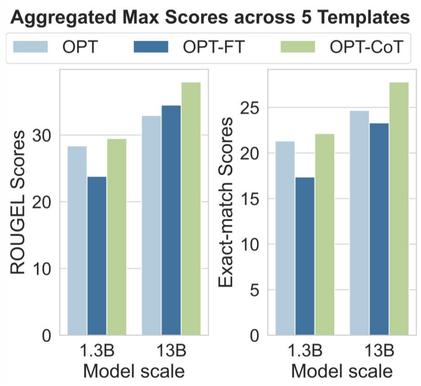

Figure 2: Three types of model performance on ALERT benchmark.

What does LLMs learn during finetuning?

Data Memorization

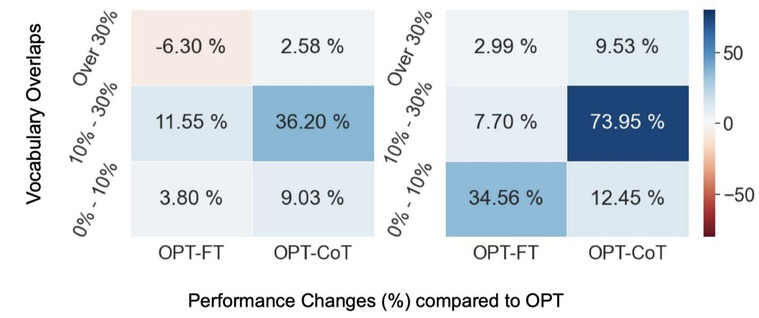

Figure 3: Correlation between vocabulary overlap and performance improvement using 13B parameter models. The left chart shows ROUGE-L while the right shows relaxedmatch score.

Does the performance improvement due to the increased amount of data seen during the finetuning stage?

We measure the dissimilarity between the finetuning data and evaluation data. If higher similarity leads to better performance, it may indicate that the improvements of finetuned LLMs are due to seeing more similar data during the finetuning stage. We present both ROUGEL score (left) and relaxed-match score (right) in Figure 3.

Result: No strong correlation between the vocabulary overlap between fineuning and evaluation datasets. Overall, for these challenging tasks, seeing similar data during finetuning stage does not guarantee performance improvement.

Reasoning Skill Transfer

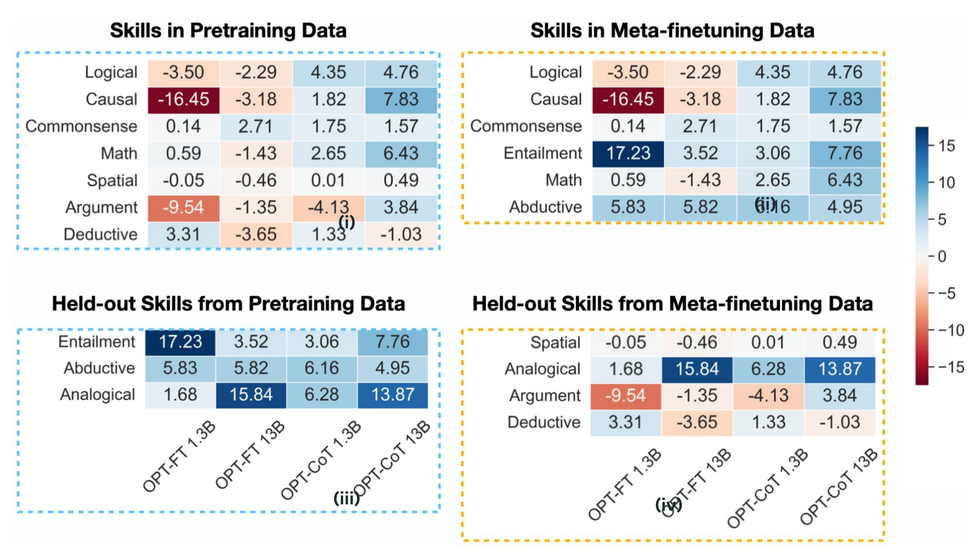

Figure 5:The ROUGE-L scores illustrating the difference between OPT-FT and OPT, as well as OPT-CoT and OPT models within each reasoning skill. Left: skills split by pretraining data; Right: skills split by meta-finetuning data.

- **Figure 5 (ii)**: All four of the LLMs demonstrate enhanced reasoning capabilities on textual entailment, abductive reasoning, and analogical reasoning tasks.
- Skills such as commonsense reasoning or spatial reasoning can be gained during the pretraining stage, while the benefits of further finetuning are not as pronounced.
- **Figure 5 (iii):** The reasoning skills gained during the meta-finetuning stage may not necessarily transfer to the improvement of the same skills on the evaluation datasets.

Data Format Memorization

- When measuring exact-match score in Figure 2, OPT-FT is worse than OPT;
- When measuring relaxed-match score in Figure 6, OPT-FT outperforms OPT;

If we decouple performance from format adherence by using soft-match score, OPT-FT performs better than OPT. Finetuning is helpful but it can make the output more noisy. This explains the reason for the performance drop when exact-match is used as the metric.

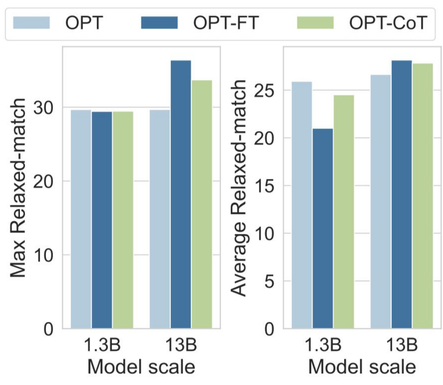

Figure 6: Comparing pretraining and finetuning models with **relaxed match score**. Left: aggregated best (max) performance across 5 Templates; Right: aggregated average performance across 5 Templates.

Take a photo to learn more: