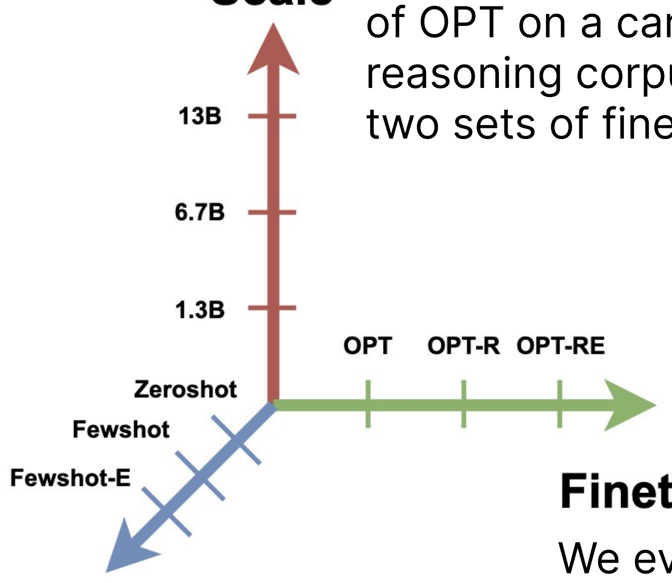
OPT-R: Exploring the Role of Explanations in Finetuning and Prompting for Reasoning Skills of Large Language Models

Badr AlKhamissi[1,2] Siddharth Verma[2,4] Ping Yu[2] Zhijing Jin[2,3] Asli Celikyilmaz[2] Mona Diab[2]

1. EPFL 2. Meta AI 3. ETH Zurich, Max Plank Institute 4. Square

Introduction

We investigate the reasoning capabilities of LLMs, focusing on the OPT models. We ablate across the following dimensions.



Scale

We finetune three different sizes of OPT on a carefully curated reasoning corpus, resulting in two sets of fine-tuned models.

Finetuning

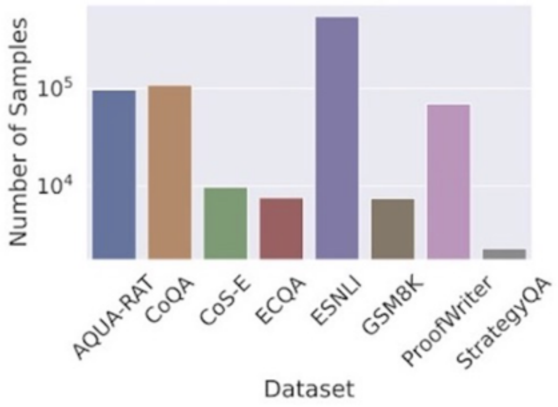
We evaluate all models on out-of-domain reasoning skills zero-shot, few-shot and few-shot with explanations.

Prompting

OPT-R, finetuned without explanations, and OPT-RE, finetuned with explanations.

This results in a comprehensive grid of 27 configurations and 6,156 test evaluations.

Finetuning Corpus



{Task Definition}

Provide your answer followed by a brief reasoning.

{In-Context Examples}

Input: {input}

Options: {options}

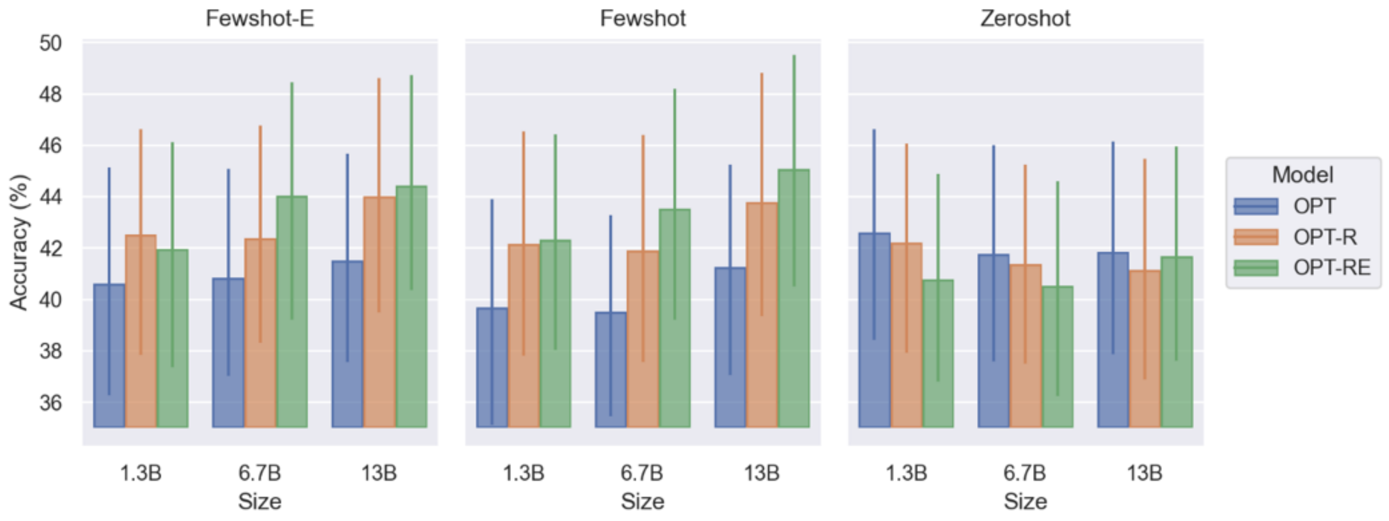
Output: The answer is {answer} because {explanation}

The finetuning corpus is comprised of 8 reasoning datasets with explanations that are either free-form or step-by-step. We format all datasets using the above prompt and generate two training corpuses, one with explanations for OPT-RE and one without explanations for OPT-R.

Evaluation

We evaluate OPT-R on 26 reasoning skills, sampled from a held-out set of 57 `Super-NaturalInstructions` tasks. Performance is measured by calculating the most likely output from a set of candidates (i.e. rank classification).

Explanations had a small effect on performance as we incorporate it during finetuning, prompting and as model size increases.



Skill Analysis

Classification accuracy achieved by different models as a function of the reasoning skill and few-shot prompting method employed. The cells are shaded to according to score, with the highest bolded. Results are grouped by which model performs best.

	OPT		OPT-*R*		OPT-*RE*	
Skill	**Fewshot**	**Fewshot-E**	**Fewshot**	**Fewshot-E**	**Fewshot**	**Fewshot-E**
Numerical	39.9	49.7	65.1	**65.3**	64.7	64.8
Analogical	51.9	46.2	**63.3**	62.5	60.7	60.9
Objects	53.5	55.1	61.4	**63.8**	60.0	59.7
Social Interactions	33.6	34.7	**43.8**	42.3	40.2	40.0
Textual Entailment	43.3	42.0	47.1	47.3	**51.9**	51.2
Grammatical	54.4	55.1	61.2	60.0	62.0	**63.1**
Multihop	36.6	31.7	38.9	**39.9**	39.5	37.0
Symbols	44.2	47.2	51.7	51.8	51.9	**52.4**
Spatial	44.1	47.1	49.8	**51.8**	49.6	49.2
Social Situations	46.3	46.6	53.2	53.2	51.9	52.3
Counting	19.6	20.0	13.5	12.7	29.8	**32.9**
Physical	35.8	40.6	36.9	38.8	48.1	**50.0**
Logical	31.7	33.4	33.7	34.1	36.9	**38.4**
Temporal	**50.7**	49.7	43.4	46.5	48.5	38.5
Argument	55.8	**60.1**	46.3	45.9	48.6	48.8
TE - Deductive	33.7	**38.3**	27.9	30.1	29.0	29.9
Relational	47.4	**51.1**	47.6	47.9	44.8	44.6
Commonsense	**35.0**	31.8	29.8	29.5	28.5	29.2
TE - Analogical	16.3	18.7	18.6	**20.7**	18.7	18.1
Abductive	33.9	36.1	**36.9**	34.4	34.2	35.3
Ethics	26.8	25.8	26.5	25.9	26.2	**27.6**
Deductive	39.4	40.4	39.4	40.4	40.0	**41.1**
Causal	50.2	**50.6**	49.1	48.9	50.1	50.5
Scientific	23.4	23.3	24.3	24.5	**25.0**	24.5
Numerical Commonsense	**59.5**	59.2	59.0	59.0	59.2	59.4
Strings	60.7	60.7	61.1	**61.2**	60.7	60.7

Further Analysis

Model	**Std(|F-FE|)**	**Avg(F)**	**Avg(FE)**
OPT	2.31	40.68	41.82
OPT-*R*	0.84	43.44	43.68
OPT-*RE*	0.78	44.49	44.86

Explanations during prompting does not significantly impact finetuned models on reasoning datasets but makes a difference for vanilla OPT.

Skill	**OPT**	**OPT-*R***	**OPT-*RE***
Numerical	44.8	**65.2***	64.7*
Analogical	49.0	**62.9***	60.8*
Counting	19.8	13.1	**31.3***
Physical	38.2	37.8	**49.1***
Entailment	42.6	47.2	**51.6***
Social Int	34.1	**43.0***	40.1
Objects	54.3	**62.6***	59.9*

Reasoning skills where either OPT-RE or OPT-R are **significantly better than the vanilla OPT**. Explanations help Counting, Physical Reasoning and Entailment.

Skill	**OPT**	**OPT-*R***	**OPT-*RE***
Argument	**57.9**	46.1^{-}	48.7^{-}
TE - Deductive	**36.0**	29.0^{-}	29.4^{-}
Commonsense	**33.4**	29.7	28.8^{-}

Reasoning skill where **OPT performs significantly better** than either OPT-R or OPT-RE

^ Signficance is measured by Welch's t-test ($p < 0.05$) denoted by the * symbol.