

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
{pingyu,aslic}@meta.com

Abstract

Recent advancements in large language models have enabled them to perform well on complex tasks that require step-by-step reasoning with few-shot learning. However, it is unclear whether these models are applying reasoning skills they have learned during pre-training, or if they are simply memorizing their training corpus at finer granularity and have learned to better understand their context. To address this question, we introduce ALERT, a benchmark and suite of analyses for evaluating reasoning skills of language models. ALERT enables comparing pre-trained and finetuned models on complex tasks that require reasoning skills to solve them. 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. To prove the efficacy of ALERT we investigate *the role of finetuning*. Our extensive empirical analysis shows that language models acquire reasoning skills such as textual entailment, abductive reasoning, and analogical reasoning during the finetuning stage compared to pretraining stage. Another finding is when language models are finetuned they tend to overfit to the prompt template, which hurts the robustness of models resulting in generalization problems.

1 Introduction

Large language models (LLMs) (e.g., GPT-3 (Brown et al., 2020a), PALM (Chowdhery et al., 2022), OPT (Zhang et al., 2022)) have shown increasing in-context learning capabilities with scaling up the model and data sizes. Despite this progress, even the largest of these models still struggle with tasks such as commonsense reasoning (West et al., 2022), and math word problems (Hendrycks et al., 2021b) which require arithmetic reasoning or symbolic manipulation (Rytting and Wingate, 2021). Table 1 presents some examples that require certain reasoning skills. Even

The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

The answer is 29 apples .

Select the best translation into predicate logic. David teaches Chris. (c: Chris; d: David; Txy: x teaches y) (A) Tdc; (B) Tcd; (C) Tcc; (D) dTc. The answer is (B) Tcd .

Isabella entered the hall. Olivia entered the hall. The apple is in the blue_treasure_chest. Olivia exited the hall. Isabella moved the apple to the green_basket. Question: Where does Isabella think that Olivia searches for the apple? The answer is Isabella thinks that Olivia searches for the apple in the green_basket .

Table 1: Examples from tasks that require reasoning skills and generated outputs from GPT-3 series *text-davinci-003* engine. The failed outputs are highlighted in red . Predictions by ChatGPT are shown in Table 9 in Appendix.

the powerful LLMs (such as *text-davinci-003*¹ and ChatGPT²) fail to make correct predictions.

To improve large LLMs’ performance on tasks that require multiple steps of reasoning, recent work used different prompting methods which included a rationale with the final answer in the form of: scratchpad for arithmetic and logical reasoning (Nye et al., 2021), chain-of-thought (CoT) (Wei et al., 2022) for practically any tasks, or adding *let’s think step-by-step* (Kojima et al., 2022) to prompt models to generate explanations. Other works such as Chung et al. (2022) integrated step-by-step explanations into the finetuning stage (CoT-finetuning). While these techniques may improve the accuracy and interpretability, it is not well understood which reasoning skills they rely on or to what degree they require higher-order reasoning. It is also uncertain how frequently the stated reasoning steps actually contribute to the final task predictions. For instance, to correctly answer the questions in Table 1 a combination of logical, commonsense, math and spatial reasoning skills are required.

In this work, to gain a deeper understanding of LLMs reasoning abilities in in-context learning

¹<https://beta.openai.com/docs/models/gpt-3>.

²<https://chat.openai.com/chat>.

settings, we introduce ALERT, a new pipeline to benchmark different LLMs on various reasoning skills and provide analysis to assess reasoning abilities. Unlike existing commonly used benchmarks (e.g., Mishra et al. (2022); Wang et al. (2022c); Srivastava et al. (2022)), ALERT can evaluate LLMs’ fine-grained reasoning skills. It spans over 20 datasets and covers 10 different reasoning skills including logical, causal, commonsense, abductive, spatial, analogical, argument and deductive reasoning as well as textual entailment, and mathematics (see Figure 6). ALERT enables easy benchmarking of any LM (e.g., pre-trained, finetuned, CoT-finetuned) on a rich set of new inference methods including zero-shot, few-shot and CoT.

Using ALERT, we further investigate whether finetuning can improve LLMs’ performance on downstream reasoning tasks. Specifically, we are interested in diagnosing what actually improved when we observe a performance increase on reasoning tasks. Is it because models have seen similar data in the finetuning stage? Or is it because models have seen prompts in a specific template and memorize the template during finetuning such as definitions provided in the NIV2 benchmark (Wang et al., 2022c)? Or does the LLM actually acquired the required reasoning skill? We investigate these three possibilities.

To study the above questions, we compare three different model types (as shown in Figure 2): a pre-trained model and two types of finetuned models. Specifically:

- **OPT** (Zhang et al., 2022): A baseline LLM a pre-trained model with no finetuning (figure (A) in Figure 2);
- **OPT-FT**: Meta-finetuned OPT on reference answers *without* explanations, illustrated in (figure (B) in Figure 2);
- **OPT-CoT**: Meta-finetuned OPT on data with rationales (explanations) (Chung et al., 2022; AlKhamissi et al., 2023) (figure (C) in Figure 2).

Using these three types of models, we investigate *the role of finetuning* on three dimensions:

(1) Data memorization: We investigate whether the performance improvements obtained after finetuning can be attributed to using similar or sometimes the exact same data as in the evaluation datasets. To this end, we use vocabulary overlap to

Reasoning Skills	Datasets
Logical	bigbench repeat copy logic, mmmlu answer generation
Causal	plausible result generation, anli r2 entailment, anli r3 entailment, cb entailment
Commonsense	piqa answer generation, commongen sentence generation, sciqa 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 2: ALERT benchmark consists of 20 datasets covering 10 different reasoning skills. The full list of the reasoning skills and datasets is in Table 4 in Appendix A.1.

measure the extent to which the evaluation data is different from the finetuning data, i.e. We investigate whether the improvement is more significant when evaluation data and finetuning data are more similar.

(2) Reasoning skills transfer: We investigate if certain reasoning skills can be more successfully permeated in LLMs than other reasoning skills. To verify this, we carefully divide the evaluation datasets into groups which require different reasoning skills. We compile held-out datasets as shown in Figure 6 which require skills held-out from any of the training datasets. This way, we expect to see larger improvements on in-domain skills compared to held-out skills if reasoning skills can be transferred during finetuning stages.

(3) Prompt template memorization: Our third hypothesis is that LLMs can overfit to data format used in the finetuning datasets such as training data format used in Figure 2. In other words, the consistency in data format helps LLMs better understand the instruction which then yields better performance after finetuning. To test this, we evaluate finetuned LLMs on datasets with 5 different prompt templates.

Summary of findings: (i) Different from Gururangan et al. (2020), 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

Definition: In this task, we ask you to write an implausible answer to a question that involves event duration, based on a given sentence. Here, event duration is defined as the understanding of how long events typically last. For example, “brushing teeth”, usually takes a few minutes. Even though there exist multiple wrong answers, we only need a single wrong answer.

Example 1-

input: Sentence: Jack played basketball after school, after which he was very tired.

Question: How long did Jack play basketball?

output: 22 hours.

explanation: Typically we play basketball for a couple of hours. So any answer beyond that range is unlikely.

Figure 1: An example from NIV2 (Wang et al., 2022c) that requires a deep understanding of the long task instruction and can be very challenging even for humans.

simply memorizing the training data during the finetuning stage; (ii) Finetuning helps improve certain reasoning capabilities of LLMs (e.g. analogical and abductive) but not all of them (e.g. commonsense reasoning); (iii) 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.

Though many of the aspects that we study have been discussed in prior analyses of LLMs (Chung et al., 2022; Wei et al., 2021a, 2022; Kojima et al., 2022; Cobbe et al., 2021; Sanh et al., 2021), prior work has not evaluated LLMs on different reasoning skills and how these skills can be improved. Overall, by evaluating reasoning skills with ALERT, we gain new insights on how models have or have not succeeded in generalizing beyond their *training* experience.

To summarize our contributions, this paper presents a meticulously designed benchmark for assessing reasoning abilities. Furthermore, a thorough investigation of *the role of finetuning* in the context of reasoning abilities, data memorization, and data format is conducted.

2 Motivation and Our Benchmark

Motivation. The analyses in ALERT are inspired by a scientific question: To what extent do LLMs learn generalizable reasoning abilities? This question motivates our focus on measuring LLMs’ performance on tasks that require contextual understanding and perform multi-step operations, which are crucial to perform well on downstream tasks.

Datasets Construction. To construct the datasets of ALERT, we select datasets from NIV2 benchmark

(Wang et al., 2022c) and perform the following operations:

(1) Omit extremely hard tasks. We design ALERT so that it can be used to benchmark a variety of LLMs, from pre-trained, finetuned to instruction-tuned models. To select such tasks, we apply several heuristics: firstly, we manually omit tasks that heavily rely on instructions. Some tasks are hard to solve when only in-context examples (demonstrations) are provided (e.g., the example in Figure 1). Secondly, we selected only those tasks that achieved a reasonable level of performance (empirically use ROUGE-L > 5.0) when evaluated with a pre-trained model (we use the OPT-13B model). Thirdly, we omit tasks on which humans fail to get decent performance given the ground truth labels from NIV2. For example, *task963_librispeech_asr_next_word_prediction* (Weir et al., 2020) provides a prompt “Joey’s favourite food is ___”, with the ground truth answer “sandwiches”. Without any context or background information, the answer can be any food thus it is extremely hard for humans to accurately predict “sandwiches”.

(2) Remove tasks with long input context. The input sentence length of some tasks can be very long, and currently most LLMs are not designed for solving long text problems. We omit tasks with demonstration length longer than 2048 tokens.

(3) Fix ground truth labels. For each reasoning task, NIV2 provides the reasoning skills required to solve the task, e.g. *task102_commongen_data_to_text* requires relational, analogical and commonsense reasoning. However, we found that some tasks have been labeled with incorrect reasoning skills. For example, *task393_plausible_result_generation* provides a sentence and asks LLMs to complete the sentence. The labels given by NIV2 are causal reasoning and textual entailment, but in fact this task can hardly examine an entailment skill. Accordingly, we manually fix reasoning skill labels. In addition, we only keep the predominant skill. For example, many tasks need more or less commonsense knowledge, therefore we select the related tasks that only heavily rely on commonsense knowledge to assess commonsense reasoning.

Benchmark. After the above steps, we select tasks that represent a variety of reasoning skills and construct ALERT reasoning benchmark, where Table 2 shows details about our benchmark.

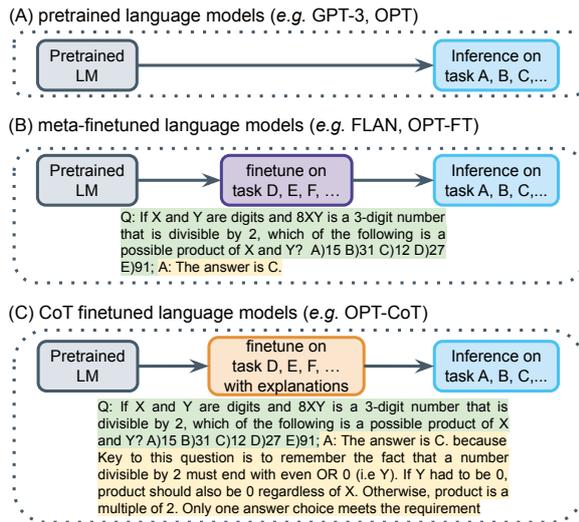


Figure 2: We compare three types of models: (A) directly apply pretrained LLMs on reasoning tasks; (B) finetune LLMs on a set of tasks; (C) finetune LLMs on tasks with explanations (CoT-finetuning). Finetuning data contains `source` and `target` parts, and the language modeling loss only applied to the `target` part.

3 Experiment Setup

3.1 Models

To perform a controlled comparison across training and prompting methods, we focus on three different models: pre-trained, meta-finetuned, and rationale-based meta-finetuned (CoT-finetuned) models. For pre-trained models, we use OPT (Zhang et al., 2022), a suite of decoder-only pre-trained transformers which are reported to yield comparable performance to GPT-3 (Brown et al., 2020b). We benchmark with OPT models of two scales: 1.3B and 13B. For finetuned models (OPT-FT), we finetune OPT models on datasets without explanations. For CoT-finetuned models (OPT-CoT), we finetune OPT models on data with rationales (explanations).

We train all models in Pytorch (Paszke et al., 2017) using OPT-IML (Iyer et al., 2022) codebase³. We initialize model hyper-parameters for each model scale following OPT (Zhang et al., 2022). We pack our training examples into sequences of length 2048, left-truncating examples that overflow. We use AdamW (Loshchilov and Hutter, 2017) with 32-bit state with $(\beta_1, \beta_2) = (0.9, 0.95)$, linearly warming up the learning rate for 6% steps to the maximum, followed by linearly decaying it to 0. For all 1.3B models, we use batch size of 128, and for 13B models, we use batch size of 256.

³<https://github.com/facebookresearch/metaseq/tree/main/projects/OPT-IML>

3.2 Finetuning Data

Our finetuning corpus is comprised of 10 datasets: ProofWriter (Tafjord et al., 2020), StrategyQA (Geva et al., 2021), ECQA (Aggarwal et al., 2021), CoQA (Reddy et al., 2019), GSM8K (Cobbe et al., 2021), AQUA-RAT (Ling et al., 2017), ESNLI (Camburu et al., 2018), MATH (Hendrycks et al., 2021c), CoS-E (Rajani et al., 2019), WinoWhy (Zhang et al., 2020). These 10 finetuning datasets collectively contain 6 different reasoning skills: logical reasoning, causal reasoning, commonsense reasoning, textual entailment, mathematics, abductive reasoning. In addition, these 10 datasets all come with instructions, demonstration examples and explanations. This enables fair comparison of OPT-FT and OPT-CoT models. More details about finetuning corpus can be found in Table 5 in Section A.2. More details about development data selection can be found in the Appendix. A.3.

3.3 Evaluation

Templates Following (Wei et al., 2021b), to control for the effect of variable prompt templates, we adopt different templates (T) during inference stage in our experiments:

- T1:** instruction + demonstration examples with explanations + "let's think step by step";
- T2:** instruction + "Please give a short explanation after the answer" + demonstration examples with explanations + "let's think step by step"
- T3:** instruction + "Please give a short explanation after the answer" + demonstration examples with explanations
- T4:** "Please give a short explanation after the answer" + demonstration examples with explanations + "Let's think step by step"
- T5:** instructions + demonstrations

For each dataset, we report the average and max score among these five templates. The final aggregated results (including aggregated average score and aggregated max score) are reported by further averaging across all datasets. Unless specified otherwise, the default score refers to the aggregated max score among five templates.

Evaluation metrics. Since our benchmark contains both classification and generation tasks, we cannot use classification accuracy to evaluate all the tasks. Following FLAN (Wei et al., 2021b), we append classification choices at the end of prompts and ask models to generate answers. Thus, classification tasks can be treated as a special case of

generation tasks. Accordingly, we use ROUGE-L (Lin, 2004) to measure the performance of both classification and generation tasks and report the aggregated score. Similar to Chung et al. (2022), we also use *exact-match* score which is more suitable for tasks with short answers. Additionally, we compute *relaxed-match* score which is a relaxed version of exact-match. Specifically, we normalize ground truth answers and predictions to have all text in lower case and remove punctuation and extra white spaces.

4 Analysis

4.1 Does finetuning help?

Figure 3 demonstrates the performance averaged across all evaluation tasks in our benchmark. Rationale-based finetuning (OPT-CoT) has been shown to improve the performance of the 1.3B model by 3.89% in terms of the aggregated max ROUGE-L score and 3.83% in terms of the aggregated max exact-match score. As for 13B model, OPT-CoT gains the improvement by 15.22% in regard of aggregated max ROUGE-L score, 12.64% in regard of aggregated max exact-match score. However, finetuning (OPT-FT) sometimes yields worse results than the vanilla pre-trained model.

4.2 What does LLMs learn during finetuning?

We find that CoT-finetuning improves performance on reasoning tasks in general. However, what exactly does the LLMs learn during the finetuning stage is still under explored. Thus, we study the role of finetuning from three perspectives: data memorization, reasoning skill transfer, and prompt template memorization.

4.2.1 Data Memorization

Gururangan et al. (2020) finds that the performance gain is larger when the finetuning dataset is more dissimilar to the pre-training dataset. However, their conclusion is made by a single-task finetuning. They evaluate their model on the same dataset that was used for finetuning. A more thorough evaluation dictates that finetuned models (Wei et al., 2021b; Chung et al., 2022) be evaluated on held-out datasets. As such, in Figure 2 in blocks (B) and (C) we show two potential ways of finetuning and inference as illustrated here in our paper.

To confirm that the improvement in finetuning performance is due to the increased amount of data seen during the finetuning stage, we measure the

dissimilarity between the training data used in finetuning and evaluation, respectively. 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. Following (Gururangan et al., 2020), we use unigram vocabulary overlap to measure the data similarity. More specifically, we divide our tasks into three categories: The first category has 10 datasets which consists of up to 10% overlap between the finetuning data and evaluation data. The second category comprises 3 datasets with an overlap between 10% and 30%. The third category has 7 datasets with an overlap over 30%. Details can be found in Table 7 in appendix A.5.

We measure the performance improvements of OPT-FT and OPT-CoT compared against the pre-trained OPT model. We present both ROUGE-L score (top) and relaxed-match score (down) in Figure 5. The results indicate that there is no strong correlation between the vocabulary overlap between finetuning and evaluation datasets and the performance of the model (neither a higher nor a lower vocabulary overlap always translate to a performance improvement). OPT-CoT achieves the best ROUGE-L and relaxed-match scores both in settings when there is a medium (10%-30%) level of vocabulary overlap. We don't observe a consistent pattern on OPT-FT models either. Overall, for these challenging tasks, seeing similar data during finetuning stage does not guarantee performance improvement.

4.2.2 Reasoning Skill Transfer

Table 6 illustrates the reasoning skills present in each stage. 7 skills can be learned from pretraining data. Appendix A.4 shows more details about pretraining data. 6 skills can be learned from finetuning data (Table 5). Using ALERT we measure a total of 10 reasoning skills in model evaluation.

The average ROUGE-L scores are calculated for each reasoning skill on 6 models (1.3B OPT, 1.3B OPT-FT, 1.3B OPT-CoT, 13B OPT, 13B OPT-FT, 13B OPT-CoT). Figure 7 shows the difference between OPT-FT and OPT, and the difference between OPT-CoT and OPT models' performance. For example, OPT-FT 1.3B model yields on average 3.5 less ROUGE-L points than OPT 1.3B model on the tasks of logical reasoning.

Figure 7 contains 4 sub-figures, showing reasoning skills transfer results: (i) The upper left sub-figure shows 7 skills that are acquired during the

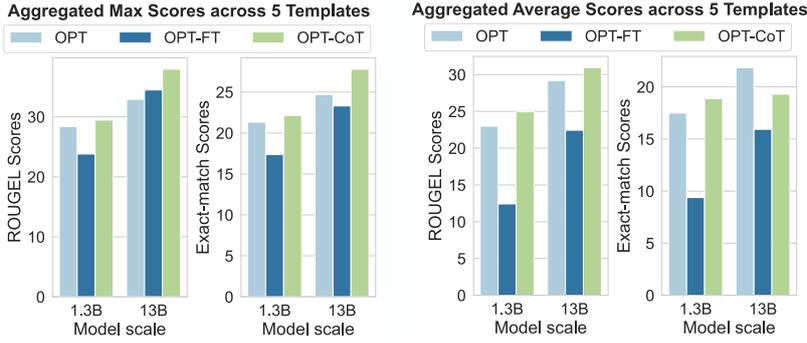


Figure 3: Performance of pre-trained LM (OPT), finetuned LM (OPT-FT) and CoT-finetuned LM (OPT-CoT) on ALERT reasoning benchmark. Left charts show aggregated **max** scores while right are **average** scores across 5 templates. Scores are averaged across 20 tasks.

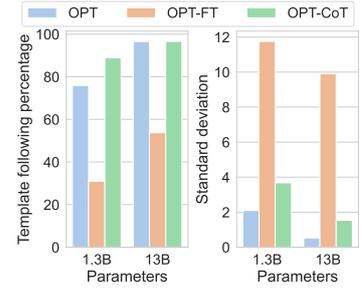


Figure 4: Analyzing the robustness of models in following the templates. **Left:** template following percentage by each model; **Right:** standard deviation of template following percentage.

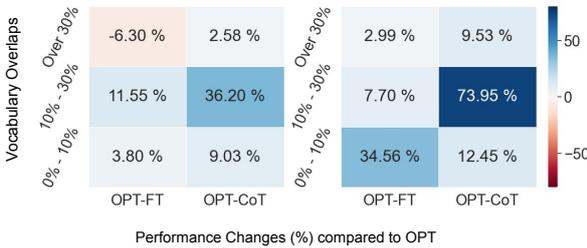


Figure 5: Correlation between **vocabulary overlap** and **performance improvement** using 13B parameter models. The **top** chart shows ROUGE-L while the **bottom** shows relaxed-match score.

pretraining stage (OPT pretraining data), and how much improvement can be obtained through meta-finetuning (OPT-FT and OPT-CoT); (ii) The bottom left sub-figure illustrates that these 3 skills are harder to acquire during the pre-training stage, and the amount of improvement that can be obtained through meta-finetuning; (iii) The upper right sub-figure illustrates that such 7 skills are acquired during the meta-finetuning stage through finetuning datasets (Table 5). Do these skills show improvement measured by evaluation benchmark? (iv) The bottom right sub-figure studies the reasoning skills that were not learned in the finetuning stage, can these skills be improved through meta-finetuning? We study the answers to these questions below.

From figure (ii) We observe that all four of the LLMs demonstrate enhanced reasoning capabilities on textual entailment, abductive reasoning, and analogical reasoning tasks. These abilities are not readily acquired during the pretraining stage, as the pretraining data consists only of plain text. On the other hand, 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. Additionally,

Gururangan et al. (2020) concluded that the more dissimilar the domain between pretraining and finetuning are, the higher the potential for finetuning to yield gains. We see the same trend but the domain in Gururangan et al. (2020) is defined by the vocabulary overlaps, while we define the domains by reasoning skills. From figure (iii) we can see that the reasoning skills gained during the meta-finetuning stage may not necessarily transfer to the improvement of the same skills on the evaluation datasets.

We also observe that finetuning with OPT-CoT enables the model to acquire a wider range of reasoning skills, resulting in stronger performance on logical and causal reasoning tasks, in addition to skills that consistently improve across all finetuned models.

4.2.3 Data Format Memorization

We investigate whether finetuning can simply memorize the template representation of the training data, and the effect of data format on the robustness of the models.

Evaluation with relaxed-match score. We compare two metrics: exact-match and relaxed-match. From Figure 3, we observe that OPT-FT is worse than OPT when exact-match is used as the metric. However, when relaxed-match is used, OPT-FT outperforms OPT as shown in Figure 8. Relaxed-match score ignores punctuation, articles and extra whitespace. This suggests that if we decouple performance from format adherence, OPT-FT performs better than OPT. In other words, 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.

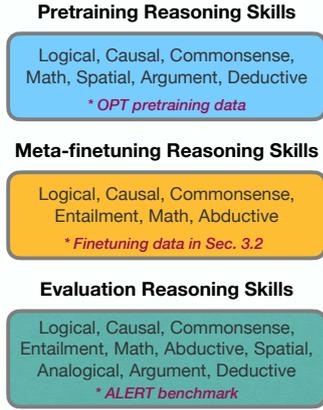


Figure 6: Reasoning skills learned during pretraining and meta-finetuning stages, as well as tested through ALERT .

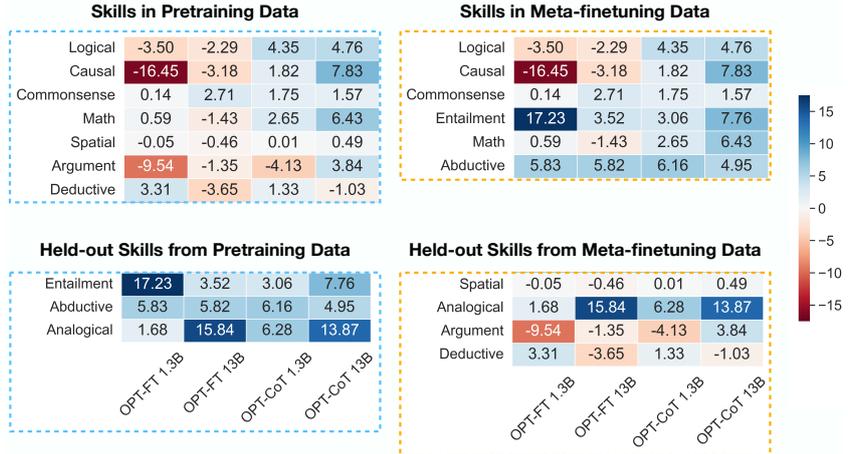


Figure 7: The ROUGE-L scores illustrating the difference between OPT-FT and OPT, as well as OPT-CoT and OPT models within each reasoning skill. Left: skills split by pretraining data; Right: skills split by meta-finetuning data.

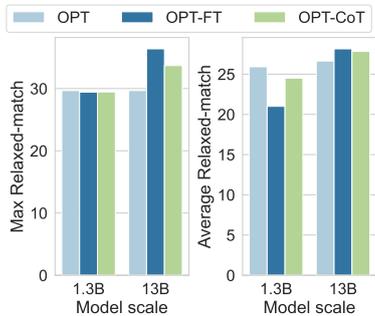


Figure 8: Comparing pretraining and finetuning models with relaxed match score. **Left:** aggregated best (max) performance across 5 Templates; **Right:** aggregated average performance across 5 Templates.

Template following percentage. We check whether the model can follow the template of the demonstrations. For example, if a demonstration uses "the answer is xxx because yyy", then we check what percentage of instances can follow the exact same template as the demonstration. Figure 4 (left) shows the average template following percentage for each model. Both OPT and OPT-CoT consistently show that they can follow demonstrations' even though OPT is not pre-trained on rationales. Compared to 1.3B models, larger models demonstrate a greater overall ability to follow the template of the demonstrations. Compared to OPT and OPT-CoT, OPT-FT lacks the ability to follow diverse templates. This is because the OPT-FT training process does not contain any rationale data. Finetuning causes the model to become more biased towards a particular template representation, while its ability to adapt to other templates becomes impaired. It is worth noting that despite being trained on rationales, the OPT-CoT model performs well when evaluated using non-CoT templates.

Robustness To assess the robustness of each model to various templates, we compute the standard deviation of ROUGE-L scores for each model across five different templates. As we can see from Figure 4 (right), OPT is robust to different templates, while OPT-FT has difficulties adapting to changing templates. In general, finetuning (both OPT-FT and OPT-CoT) adversely affects the robustness of the model and makes the model biased towards a specific data format, however, OPT-CoT is better than general finetuning (OPT-FT).

Reasoning chain quality. Following (Golovneva et al., 2022) we evaluate reasoning abilities of the models using ROSCOE scoring suite (Table 3). Looking at each score in detail (Appendix C), we found that overall across templates OPT-FT models produce shorter, less informative chains, while OPT baseline models produce long chains with high amount of self-repetitions. 13B OPT-CoT chains showed best quality despite some self-consistency and grammar issues. When comparing prompt templates, models prompted with Template 5 produce short chains, often without reasoning at all, even if they were fine-tuned on reasoning chains (OPT-CoT), suggesting overfitting to the prompt template.

In summary, models learn the data format representation and templates during finetuning stage. However, finetuned models contain bias towards the data formats and template it has seen, which potentially reduces the robustness of the model to more generalized settings. When comparing robustness, OPT-CoT is better than OPT-FT, but it is still not as robust as the pre-trained model.

Metrics	1.3B			13B		
	OPT	OPT-FT	OPT-CoT	OPT	OPT-FT	OPT-CoT
ROSCOE-SA	0.936	0.921	<u>0.938</u>	0.936	0.923	0.940
ROSCOE-SS	<u>0.925</u>	0.923	0.920	0.926	0.916	<u>0.925</u>
ROSCOE-LI	0.848	0.953	0.875	0.863	<u>0.944</u>	0.890
ROSCOE-LS	<u>0.725</u>	0.744	0.666	0.688	0.705	0.640

Table 3: Summary of the ROSCOE evaluation results averaged across templates. Each metric is bounded within $[0, 1]$, where 1 indicates the perfect score and 0 corresponds to failure. In each row, values corresponding to the best-performing model are **bolded**, second best are underscored.

5 Related Work

LLMs that Reason. To improve LLMs’ reasoning abilities, Kojima et al. (2022) shows that LLMs can be decent zero-shot reasoners by simply appending “Let’s think step by step” to the prompt. Wei et al. (2022) adds a series of intermediate reasoning steps to improve LLMs’ reasoning abilities. Wang et al. (2022a) further proposes to expand prompts to include rationales in each few-shot example. Fu et al. (2022) discovers that prompting with higher reasoning complexity achieves substantial gains on math word tasks. To tackle problems harder than demonstration examples, Zhou et al. (2022) first reduces a complex problem into a list of subproblems and solve subproblems sequentially. Another line of research is to improve the naive decoding strategy, Wang et al. (2022b) introduces a self-consistency strategy which selects the most consistent answer among a set of reasoning paths.

Existing Reasoning Benchmarks. Many benchmarks are used for evaluating language models’ performance, such as BIG-Bench (Srivastava et al., 2022), Natural Instruction V2 (NIV2) (Wang et al., 2022c), MMLU (Hendrycks et al., 2020). Although they contain some reasoning tasks, none of them are specifically designed to test models’ reasoning skills. For example, NIV2 contains 172 datasets and a total of 1554 tasks, including some reasoning tasks. It has several issues which make it inappropriate to be directly used as a reasoning benchmark: (1) it is designed for instruction-tuned models and some tasks might be unsuitable for evaluating pretrained models or non-instruction finetuned models, as shown in Figure 1; (2) reasoning skills have been divided into 27 categories while some of them have large overlaps, e.g. numerical reasoning, quantitative reasoning, reasoning on numbers; (3) some reasoning labels are wrongly labeled, e.g. *task393_plausible_result_generation* gives textual entailment label but this task can hardly examine the entailment skill.

The Curriculum benchmark (Chen and Gao, 2022) is designed for probing LLMs’ reasoning abilities and covers 8 different reasoning skills. However, this work only focuses on classification tasks and it converts all examples into the Natural Language Inference (NLI) format to fit into a unified framework. We argue that the forced conversion of all datasets into the NLI format does not align with human natural conversational style. We observed that even davinci-003 fails at some simple tasks due to their forced conversion, e.g. examples in Table 1. More discussion and results are shown in the Appendix B.

Finetuning LLMs. LLMs meta-finetuned on a range of NLP tasks have shown improved performance on held-out downstream tasks such as FLAN (Wei et al., 2021b), T0 (Sanh et al., 2021), Tk-Instruct (Wang et al., 2022c) and Instruct-GPT (Ouyang et al., 2022). Following this approach, we finetune OPT models and name this type of models as OPT-FT ((B) in Figure 2). Chung et al. (2022) further adds chain-of-thought data at finetuning stage and shows significant improvements. We also study this type of models and name them as OPT-CoT ((C) in Figure 2). However, from previous research it still remains unclear whether the improvement comes from simply adding more training data or finetuning on rationales actually helps. We conduct rigorous evaluations to address this question.

6 Conclusion

We introduce ALERT, a carefully curated benchmark for evaluating reasoning abilities of LLMs. It comprises over 20 datasets and covers 10 different reasoning skills. Using this benchmark, we further investigate the impact of finetuning on these complex tasks. Our experiments reveal that LLMs do not simply memorize training data, but are capable of learning various reasoning skills, such as textual entailment, abductive reasoning and analogical reasoning. While we found that finetuning generally leads to improved performance, we also discovered some negative effects. LLMs tend to memorize the data template representation and templates seen during finetuning, thus reducing the robustness of the model to generalized settings. CoT-finetuning (OPT-CoT) can alleviate this issue to some extent, but it is still less robust compared to the vanilla pre-trained model.

Limitations

ALERT aims to encompass a wide range of reasoning skills, but some reasoning skills are missing, specifically in regards to symbolic reasoning (last letter concatenation task and coin flip (Wei et al., 2022)) and compositionality reasoning (SCAN (Lake and Baroni, 2018), COGS (Kim and Linzen, 2020) and CFQ (Keysers et al., 2019)). These reasoning skills should be included in future work.

In terms of computing power, we have experimented with models that were accessible to us. We acknowledge that there are larger models that we were not able to train due to the limitations of our computational budget.

During our analysis, we discovered that some datasets contain noise, where even human experts are unable to provide accurate answers for certain instances. While it is important to address this issue, it is a time-consuming process to carefully review and clean each instance in the dataset. We plan to address this in future work.

Ethics Statement

Large language models (LLMs), due to potential bias in the training data, can be prone to generate toxic and unwanted content (Weidinger et al., 2021). However, in this paper, we are focused on reasoning tasks where the model is prompted to explain its decisions, because of which our model falls under contained generation. By providing clear prompts and constraints, we believe that this might help guide the model’s output towards specific, desired outcomes and reduce the likelihood of generating unwanted or harmful content, as opposed to open ended text generation tasks.

References

- Shourya Aggarwal, Divyanshu Mandowara, Vishwa-jeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. [Explanations for CommonsenseQA: New Dataset and Models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.
- Badr AlKhamissi, Siddharth Verma, Ping Yu, Zhijing Jin, Asli Celikyilmaz, and Mona Diab. 2023. [Opt-r: Exploring the role of explanations in finetuning and prompting for reasoning skills of large language models](#). [arXiv preprint arXiv:2305.12001](#).
- Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, et al. 2021. [Efficient large scale language modeling with mixtures of experts](#). [arXiv preprint arXiv:2112.10684](#).
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The pushshift reddit dataset](#). In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. [Piqa: Reasoning about physical commonsense in natural language](#). In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. [Language models are few-shot learners](#). *Advances in neural information processing systems*, 33:1877–1901.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. [Language models are few-shot learners](#). *Advances in neural information processing systems*, 33:1877–1901.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. [e-snli: Natural language inference with natural language explanations](#). *Advances in Neural Information Processing Systems*, 31.
- Zeming Chen and Qiyue Gao. 2022. [Curriculum: A broad-coverage benchmark for linguistic phenomena in natural language understanding](#). [arXiv preprint arXiv:2204.06283](#).
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. [Palm: Scaling language modeling with pathways](#). [arXiv preprint arXiv:2204.02311](#).
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. [Scaling instruction-finetuned language models](#). [arXiv preprint arXiv:2210.11416](#).
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). [arXiv preprint arXiv:2110.14168](#).

- Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. [arXiv preprint arXiv:2210.00720](#).
- Nancy Fulda, Nathan Tibbetts, Zachary Brown, and David Wingate. 2017. Harvesting common-sense navigational knowledge for robotics from uncurated text corpora. In [Conference on Robot Learning](#), pages 525–534. PMLR.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. [arXiv preprint arXiv:2101.00027](#).
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. [Transactions of the Association for Computational Linguistics](#), 9:346–361.
- Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. [Roscoe: A suite of metrics for scoring step-by-step reasoning](#).
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: adapt language models to domains and tasks. [arXiv preprint arXiv:2004.10964](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. [arXiv preprint arXiv:2009.03300](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. [Proceedings of the International Conference on Learning Representations \(ICLR\)](#).
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. [NeurIPS](#).
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the math dataset. [arXiv preprint arXiv:2103.03874](#).
- Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, Gabriel Stanovsky, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. 2019. Semeval-2019 task 10: math question answering. In [Proceedings of the 13th International Workshop on Semantic Evaluation](#), pages 893–899.
- Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-impl: Scaling language model instruction meta learning through the lens of generalization. [arXiv preprint arXiv:2212.12017](#).
- Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. [arXiv preprint arXiv:1912.09713](#).
- Najoung Kim and Tal Linzen. 2020. Cogs: A compositional generalization challenge based on semantic interpretation. [arXiv preprint arXiv:2010.05465](#).
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. [arXiv preprint arXiv:1412.6980](#).
- Jonathan Kobbe, Ioana Hulpuş, and Heiner Stuckenschmidt. 2020. Unsupervised stance detection for arguments from consequences. In [Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing \(EMNLP\)](#), pages 50–60.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. [arXiv preprint arXiv:2205.11916](#).
- Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In [International conference on machine learning](#), pages 2873–2882. PMLR.
- Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. [TellMeWhy: A dataset for answering why-questions in narratives](#). In [Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021](#), pages 596–610, Online. Association for Computational Linguistics.
- Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. [CommonGen: A constrained text generation challenge for generative common-sense reasoning](#). In [Findings of the Association for Computational Linguistics: EMNLP 2020](#), pages 1823–1840, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In [Text summarization branches out](#), pages 74–81.
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. [Program induction by rationale generation: Learning to solve and explain algebraic word problems](#). In [Proceedings of the 55th Annual Meeting of the Association for Computational](#)

- Linguistics (Volume 1: Long Papers), pages 158–167, Vancouver, Canada. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. [arXiv preprint arXiv:1907.11692](#).
- Ilya Loshchilov and Frank Hutter. 2017. [Fixing weight decay regularization in adam](#). [CoRR](#), abs/1711.05101.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. [arXiv preprint arXiv:1809.02789](#).
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In [ACL](#).
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In [Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies](#), pages 839–849.
- Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L Griffiths. 2018. Evaluating theory of mind in question answering. [arXiv preprint arXiv:1808.09352](#).
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. [arXiv preprint arXiv:2112.00114](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. [arXiv preprint arXiv:2203.02155](#).
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. [Automatic differentiation in pytorch](#). In [NIPS 2017 Workshop on Autodiff](#).
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. [arXiv preprint arXiv:1906.02361](#).
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [CoQA: A conversational question answering challenge](#). [Transactions of the Association for Computational Linguistics](#), 7:249–266.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. [arXiv preprint arXiv:2004.13637](#).
- Christopher Rytting and David Wingate. 2021. Leveraging the inductive bias of large language models for abstract textual reasoning. [Advances in Neural Information Processing Systems](#), 34:17111–17122.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. [arXiv preprint arXiv:2110.08207](#).
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. [arXiv preprint arXiv:1909.08053](#).
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. [arXiv preprint arXiv:2206.04615](#).
- Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. [DREAM: A challenge dataset and models for dialogue-based reading comprehension](#). [Transactions of the Association for Computational Linguistics](#).
- Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2020. Proofwriter: Generating implications, proofs, and abductive statements over natural language. [arXiv preprint arXiv:2012.13048](#).
- Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. [arXiv preprint arXiv:1806.02847](#).
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. [arXiv preprint arXiv:1905.00537](#).
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In [Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP](#), pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022a. Rationale-augmented ensembles in language models. [arXiv preprint arXiv:2207.00747](#).

- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. [arXiv preprint arXiv:2203.11171](#).
- Yizhong Wang, Swaroop Mishra, Pegah Alipoor-molabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022c. Super-naturalinstructions: generalization via declarative instructions on 1600+ tasks. In [EMNLP](#).
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021a. Finetuned language models are zero-shot learners. [arXiv preprint arXiv:2109.01652](#).
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021b. Finetuned language models are zero-shot learners. [arXiv preprint arXiv:2109.01652](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. [arXiv preprint arXiv:2201.11903](#).
- Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. [arXiv preprint arXiv:2112.04359](#).
- Nathaniel Weir, João Sedoc, and Benjamin Van Durme. 2020. Cod3s: Diverse generation with discrete semantic signatures. [arXiv preprint arXiv:2010.02882](#).
- Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. [arXiv preprint arXiv:1707.06209](#).
- Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. [Symbolic knowledge distillation: from general language models to commonsense models](#). In [Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies](#), pages 4602–4625, Seattle, United States. Association for Computational Linguistics.
- Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. [arXiv preprint arXiv:1502.05698](#).
- Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. [Anlizing the adversarial natural language inference dataset](#). In [Proceedings of the 5th Annual Meeting of the Society for Computation in Linguistics](#), pages 23–54. Association for Computational Linguistics.
- Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020. [WinoWhy: A deep diagnosis of essential commonsense knowledge for answering Winograd schema challenge](#). In [Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics](#), pages 5736–5745, Online. Association for Computational Linguistics.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. [arXiv preprint arXiv:2205.01068](#).
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. [arXiv preprint arXiv:2205.10625](#).
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In [Proceedings of the IEEE international conference on computer vision](#), pages 19–27.

A More Details about Data Usage

A.1 Reasoning Benchmark

Table 4 shows detailed reasoning benchmark.

A.2 Training Corpus (cont. from §3.2)

We used 10 datasets for finetuning, which contain 6 different reasoning skills.

A.3 Development Data Details

Our finetuning models are tuned on pretrained LLMs on the finetuning corpus with the goal of improving the performance of unseen tasks. For example, blocks (B) and (C) in Figure 2 are showing models that are finetuned on tasks B,C,D and the goal is to achieve good results on task A.

Checkpoint selection can determine the final performance of the LLMs to a very large extent. There are several ways to select checkpoints: (i) select checkpoint of the last iteration; (ii) select checkpoint based on perplexity or loss from validation datasets of finetuning corpus (validation datasets of task B, C, D); (iii) select checkpoint based on perplexity or loss from validation datasets of evaluation corpus (validation datasets of task A);

In order to achieve a better performance on evaluation corpus, a common approach is to use methods like (iii) to select a checkpoint. However, we would like to prevent LLMs overfitting to the distribution of our final evaluation corpus. We initially used the method (ii) but found that it didn't work well. However, this resulted in a distribution mismatch issue. We speculate this to the fact that some tasks in our finetuning corpus do not have a validation set. We thus select 3 tasks from NIV2 benchmark and compile a development set that does not have any overlaps with our finetuning data or evaluation data. There are 3 datasets used as our development set for checkpoint selection: task 247 dream answer generation (Sun et al., 2019), task 118 semeval and task 10 open vocabulary mathematical answer generation (Hopkins et al., 2019) and anli r1 entailment (Williams et al., 2022)

A.4 Pretraining Data Analysis

The pre-training corpus of OPT model (Zhang et al., 2022) contains a concatenation of datasets used in RoBERTa (Liu et al., 2019), the Pile (Gao et al., 2020), and PushShift.io Reddit (Baumgartner et al., 2020; Roller et al., 2020).

RoBERTa Three datasets in RoBERTa (Liu et al., 2019) are used as pretraining corpus: BookCorpus (Zhu et al., 2015), Stories (Trinh and Le, 2018), and CCNews (Liu et al., 2019). Deductive reasoning skill and spatial reasoning skill can be learned from stories dataset. Logical reasoning skill can be learned from these three datasets.

Pile A subset of the Pile (Gao et al., 2020) are used as pre-training corpus, including CommonCrawl, DM Mathematics, Project Gutenberg, HackerNews, OpenSubtitles, OpenWebText2, USPTO, and Wikipedia. Mathematics reasoning skill can be learned from DM Mathematics dataset. Causal Reasoning can be learned widely from OpenWebText2. Commonsense reasoning skill can be learned from Wikipedia.

PushShift.io Reddit The longest chain of comments in each thread are extracted from PushShift.io Reddit (Baumgartner et al., 2020). Argument reasoning skill can be learned from this dataset.

A.5 Vocabulary Overlaps (Cont. from § 4.2.1)

We measure unigram vocabulary overlaps between our finetuning corpus and the evaluation corpus (reasoning benchmark).

B Curriculum Benchmark Results (Cont. from §5)

We randomly selected one dataset from each reasoning skill and reported the results of GPT-3 (Brown et al., 2020b) (text-davinci engine). Since all of the data has been converted to NLI format, we measure classification accuracy of GPT-3 model. From Table 8, we can see that even GPT-3 achieves a pretty random results on these datasets. Through our analysis, we found that it is not because those tasks are too difficult for GPT-3, it is because curriculum benchmark forcing all the data to be NLI format, resulting in unnatural data expression, which made GPT-3 fail on it. We conclude that the curriculum benchmark may be suitable for classification finetuned models, but it is not suitable for language models for in-context learning.

C Evaluating reasoning chains (Cont. from §5)

Following (Golovneva et al., 2022) we evaluate reasoning abilities of the models using ROSCOE scoring suite (Table 10). Chains are evaluated

Reasoning Skills	Task ID	Datasets
Logical Reasoning	62	bigbench repeat copy logic (Srivastava et al., 2022)
	697	mmmlu answer generation formal logic (Hendrycks et al., 2021a)
Causal Reasoning	393	plausible result generation (Weir et al., 2020)
	1386	anli r2 entailment (Williams et al., 2022)
	1387	anli r3 entailment (Williams et al., 2022)
	1388	cb entailment (Wang et al., 2019)
Commonsense Reasoning	80	piqa answer generation (Bisk et al., 2020)
	102	commongen sentence generation (Lin et al., 2020)
	591	sciq answer generation (Welbl et al., 2017)
	1286	openbookqa question answering (Mihaylov et al., 2018)
Textual Entailment	1386	anli r2 entailment (Williams et al., 2022)
	1387	anli r3 entailment (Williams et al., 2022)
	1388	cb entailment (Wang et al., 2019)
	1344	glue entailment classification (Wang et al., 2018)
Mathematics	104	semeval closed vocabulary math answer generation (Hopkins et al., 2019)
	119	semeval geometric math answer generation (Hopkins et al., 2019)
	697	mmmlu answer generation formal logic (Hendrycks et al., 2021a)
Abductive Reasoning	332	tellmewhy answer generation (Lal et al., 2021)
Spatial Reasoning	83	babi t1 single supporting fact answer generation (Weston et al., 2015)
	80	piqa answer generation (Bisk et al., 2020)
	151	tomqa find location easy clean (Nematzadeh et al., 2018)
Analogical Reasoning	102	commongen sentence generation (Lin et al., 2020)
	1152	bard analogical reasoning causation (Fulda et al., 2017)
Argument Reasoning	513	argument stance classification (Kobbe et al., 2020)
	514	argument consequence classification (Kobbe et al., 2020)
Deductive Reasoning	216	rocstories correct answer generation (Mostafazadeh et al., 2016)

Table 4: Details about ALERT benchmark.

Datasets	Train Size	Val Size	Test Size	Reasoning Skills
ProofWriter	69,810	10,190	20,030	Logical Reasoning, Causal Reasoning
StrategyQA	2,290	-	490	Commonsense Reasoning
ECQA	7,598	1,090	2,194	Commonsense Reasoning
CoQA	10,8647	7,983	-	Textual Entailment
GSM8K	7,473	-	1,319	Mathematics
AQUA-RAT	97,467	254	254	Mathematics
ESNLI	549,367	9,842	9,824	Commonsense Reasoning, Logical Reasoning, Textual Entailment
MATH	7,500	-	5,000	Mathematics
CoS-E	9,741	1,221	-	Commonsense Reasoning
WinoWhy	273	-	-	Abductive Reasoning, Commonsense Reasoning

Table 5: Training corpus for meta-finetuning OPT-FT and OPT-CoT. (Cont. from § 3.2)

Task ID	Datasets	Reasoning Skills
247	dream answer generation (Sun et al., 2019)	Logical Reasoning Commonsense Reasoning
118	semeval open vocabulary mathematical answer generation (Hopkins et al., 2019)	Commonsense Reasoning Mathematics
1385	anli r1 entailment (Williams et al., 2022)	Textual Entailment Commonsense Reasoning Causal Reasoning

Table 6: Dev set for checkpoint selection

Category	Datasets	Vocabulary Overlaps
0% to 10%	bigbench repeat copy logic (Srivastava et al., 2022)	1.59%
	babi t1 single supporting fact answer generation (Weston et al., 2015)	0.38%
	semeval closed vocabulary math answer generation (Hopkins et al., 2019)	7.90%
	semeval geometric math answer generation (Hopkins et al., 2019)	5.84%
	tomqa find location easy clean (Nematzadeh et al., 2018)	0.94%
	plausible result generation (Weir et al., 2020)	3.72%
	argument stance classification (Kobbe et al., 2020)	6.04%
	argument consequence classification (Kobbe et al., 2020)	6.11%
	mmmlu answer generation formal logic (Hendrycks et al., 2021a)	5.35%
bard analogical reasoning causation (Fulda et al., 2017)	0.45%	
10% to 30%	commongen sentence generation (Lin et al., 2020)	29.31%
	tellmewhy answer generation (Lal et al., 2021)	28.05%
	cb entailment (Wang et al., 2019)	20.97%
over 30%	piqa answer generation (Bisk et al., 2020)	42.51%
	rocstories correct answer generation (Mostafazadeh et al., 2016)	57.45%
	sciq answer generation (Welbl et al., 2017)	32.54%
	openbookqa question answering (Mihaylov et al., 2018)	48.2%
	glue entailment classification (Wang et al., 2018)	55.19%
	anli r2 entailment (Williams et al., 2022)	43.37%
	anli r3 entailment (Williams et al., 2022)	53.13%

Table 7: Vocabulary overlap. Dissimilarity has been measured between training data (in Table 4) and evaluation data (in Table 5).

Datasets	Random score	GPT-3 Davinci score
Boolean	34.50%	31.80%
Physical	49.08%	50.00%
Entailment Tree	50.88%	54.41%
Event Semantic	54.08%	55.89%
Spatial	49.80%	49.30%

Table 8: Classification accuracy by GPT-3 davinci-engine on the Curriculum benchmark.

using *facebook/roscoe-512-roberta-base* sentence embedding model. Evaluation results are detailed in Table 10. We found that the chain quality varies between models, in particular some reasoning aspects correlate with chain length as seen in Table 11. Similar to (Chung et al., 2022), we noticed that non-finetuned models (i.e. OPT-1.3B and OPT-13B) tend to produce long chains of reasoning, often repeating themselves, which significantly affects the quality of the chains and final scores (Figure 9).

Below we explore the differences between models’ outputs under four perspectives: semantic alignment, semantic similarity, logical inference and language coherence.

C.1 Semantic Alignment

Despite the fact that model 13B OPT-CoT on average outperforms other models in almost all semantic alignment scores (*Faithfulness-Step*, *Faithfulness-Token*, and *Info-Step*, see Table 10),

The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
The answer is 29 apples. The cafeteria started with 23 apples and bought 6 more, for a total of $23 + 6 = \langle 23+6=29 \rangle$ 29 apples.
Select the best translation into predicate logic. David teaches Chris. (c: Chris; d: David; Tx: x teaches y)
(A)Tdc;(B)Tcd;(C)Tcc;(D)dTc.
The answer is (B) Tcd. In predicate logic, the predicate (in this case, "T") is written before the subject and object of the sentence.
Since David (d) is the teacher and Chris (c) is the student, the correct translation is Tcd, which reads as "David teaches Chris."
Isabella entered the hall. Olivia entered the hall. The apple is in the blue_treasure_chest. Olivia exited the hall. Isabella moved the apple to the green_basket. Question: Where does Isabella think that Olivia searches for the apple? The answer is It is impossible for me to answer this question as I have no information about the thoughts or actions of Isabella or Olivia.
Additionally, I do not have access to browse any information beyond my training data, which ends in 2021.

Table 9: Examples from tasks that require reasoning skills and generated outputs from ChatGPT. The failed outputs are highlighted in red.

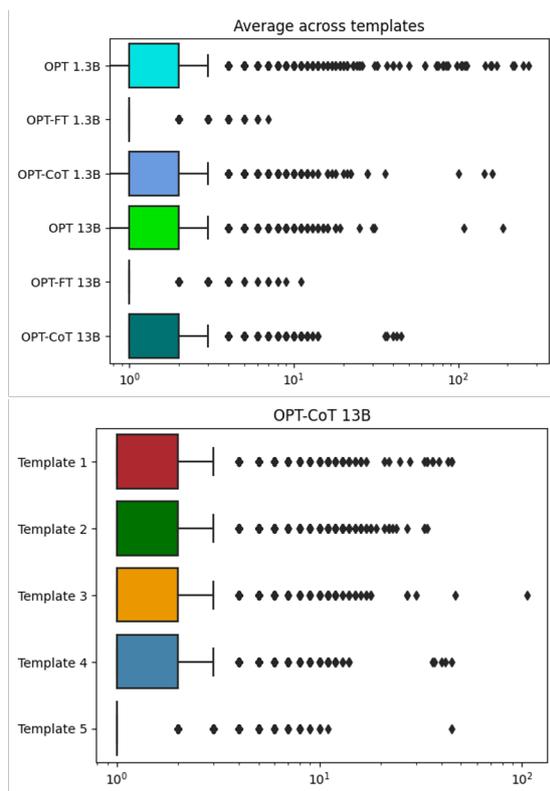


Figure 9: Distribution of the steps' number across all tasks and templates varying between models (top) and between templates for OPT-CoT 13B model.

there is no common pattern across tasks (Fig 10). The performance change between finetuned models and corresponding pretrained versions are significant⁴ on half of the tasks (11 tasks out of 20 for *Faithfulness*-* scores, and 9 out of 20 for *InfoStep*).

Repetition-Token score variations exhibit different behavior. Half of the tasks have higher number of repetitions between reasoning steps for pretrained models, with OPT-FT models generally out-

⁴Significance is determined using T-test comparison, where *p*-value is below 0.05.

performing others (all performance improvements are significant). Generations produced by these models tend to be shorter in terms of the number of steps (Figure 9), so they contain less repetitions, but also less semantic overlap with the context, thus in general having lower faithfulness and informativeness. Some examples reflecting this behavior are provided in Table 12.

Scores are mostly aligned across Templates (Figure 11), except Template 5, that stands out in having less aligned scores with respect to the context, but also more self-consistent across the task. This is the only template that did not have any explanation in its prompt. Manual review showed that despite CoT-finetuning, OPT-COT models tend to produce 1-step answer-only generations (see example in the Table 12, and Figure 9 for chains' length distribution), thus overfitting to the template rather than learning from finetuning.

In summary, ROSCOE-SA is able to identify aligned information, but it does not guarantee high-quality output. It will favor model with short explanations and high semantic overlap with the reference. We found that often OPT-FT-1.3B simply repeats one sentence from the input, instead of producing reasoning, and thus will get highest ROSCOE-SA scores on these chains, while other models that produce some sort of reasoning will be punished.

C.2 Semantic Similarity

Semantic similarity scores support previous conclusions: models, finetuned on final answers (OPT-FT) exhibit lower similarity with respect to the baseline and CoT-finetuned models, while having less repetitions (Figure 12). Again, we attribute that to the fact that these models produce short chains that lack detailed reasoning steps.

	OPT 1.3B	OPT-FT 1.3B	OPT-CoT 1.3B	OPT 13B	OPT-FT 13B	OPT-CoT 13B
ROSCOE-SA						
Faithfulness-Step	<u>0.863</u>	0.841	0.862	<u>0.863</u>	0.858	0.870
Faithfulness-Token	0.936	0.921	<u>0.938</u>	0.936	0.923	0.940
Info-Step	0.857	0.829	0.854	<u>0.858</u>	0.846	0.861
Repetition-Token	0.618	0.920	0.683	0.582	<u>0.857</u>	0.701
ROSCOE-SS						
Info-Chain	<u>0.925</u>	0.909	0.920	0.926	0.916	<u>0.925</u>
Repetition-Step	0.627	0.923	0.692	0.591	<u>0.859</u>	0.708
ROSCOE-LI						
Source Consistency	0.550	<u>0.604</u>	0.573	0.584	0.617	0.598
Self-Consistency	0.848	0.953	0.875	0.863	<u>0.944</u>	0.890
ROSCOE-LS						
Perplexity-Step	0.016	0.006	<u>0.015</u>	0.010	0.006	0.009
Perplexity-Chain	0.022	0.006	<u>0.020</u>	0.016	0.006	0.013
Grammar	<u>0.725</u>	0.744	0.666	0.688	0.705	0.640

Table 10: ROSCOE evaluation results averaged across templates. Each metric is bounded within $[0, 1]$, where 1 indicates the perfect score and 0 corresponds to failure. Values corresponding to the best performing model are **bolded**, second best are underscored.

	Kendall's τ score	Kendall's τ p-value
Faithfulness-Step	-0.101	0.000
Faithfulness-Token	0.039	0.000
Info-Step	0.054	0.000
Repetition-Token	-0.869	0.000
Info-Chain	0.009	0.000
Repetition-Step	-0.867	0.000
Source Consistency	-0.119	0.000
Self-Consistency	-0.553	0.000
Perplexity-Step	0.000	0.960
Perplexity-Chain	0.369	0.000
Grammar	0.013	0.000

Table 11: Kendall correlation between evaluation perspective and number of steps in chain across all generated reasoning chains. Strong correlations ($|\tau| > 0.4$) are **bolded**.

C.3 Logical Inference

In general, finetuned models are more self- and source-consistent than respective baselines (Figure 13, significantly outperforming nonfinetuned models on 14 out of 20 tasks. We further looked into the task 083, which is a task to find a right answer given s given single supporting fact, potentially amongst a set of other irrelevant facts. Manual review showed that although in this task finetuned models tend to produce answers that are more consistent, they often fail to select the fact that is relevant to the question asked (see "Spatial Reasoning" example in Table 12.

C.4 Language Coherence

Despite the variations in the values, *Perplexity*-* score changes between models are mostly insignificant (15 out of 20 tasks, see Figure 14). Manual review showed that all models produce mostly

grammatically correct content.

D Licenses

D.1 Data in ALERT

- task62: Apache 2.0
- task697: MIT
- task393: MIT
- task1386: CC BY-NC 4.0
- task1387: CC BY-NC 4.0
- task1388: CC BY-SA 3.0
- task080: AFL 3.0
- task102: MIT
- task591: CC BY-NC-3.0

- task1286: Apache 2.0
- task1344: CC BY 4.0
- task104: Please refer to: <https://github.com/allenai/semEval-2019-task-10#terms-and-conditions>
- task119: Please refer to: <https://github.com/allenai/semEval-2019-task-10#terms-and-conditions>
- task332: Please refer to: <https://github.com/StonyBrookNLP/tellmewhy>
- task083: CC BY 3.0
- task151: Please refer to: <https://github.com/kayburns/tom-qa-dataset>
- task1152: Apache 2.0
- task513: Please refer to: <https://github.com/dwslab/StArCon>
- task514: Please refer to: <https://github.com/dwslab/StArCon>
- task216: Please refer to: <https://www.microsoft.com/en-us/research/publication/a-corpus-and-cloze-evaluation-for-deeper-understanding-of-commonsense-stories/>
- ECQA: Literature and Wikipedia passages are shared under CC BY-SA 4.0 license. Middle/High school exam passages are collected from RACE which comes with its own license.
- GSM8K: MIT. Downloaded from https://raw.githubusercontent.com/openai/grade-school-math/master/grade_school_math/data/train.jsonl.
- AQUA-RAT: Apache License, Version 2.0. Downloaded from: <https://raw.githubusercontent.com/deepmind/AQUA/master/train.json>
- ESNLI: please refer to <https://github.com/OanaMariaCamburu/e-SNLI/commit/bab0fa0212be9e5c6737da70c639a596f882e931>. Downloaded from: https://raw.githubusercontent.com/OanaMariaCamburu/e-SNLI/master/dataset/esnli_train_1.csv
- MATH: MIT. Downloaded from: <https://people.eecs.berkeley.edu/~hendrycks/MATH.tar>
- CoS-E: BSD-3-Clause license. Downloaded from: https://raw.githubusercontent.com/salesforce/cos-e/master/data/v1.11/cose_train_v1.11_processed.jsonl
- WinoWhy: MIT. Downloaded from: <https://raw.githubusercontent.com/HKUST-KnowComp/WinoWhy/master/winowhy.json>

D.2 Data in Dev set

- task247: Dream dataset is intended for non-commercial research purpose only. <https://github.com/nlpdata/dream>.
- task118: Please refer to: <https://github.com/allenai/semEval-2019-task-10#terms-and-conditions>
- task 1385: CC BY-NC 4.0

D.3 Data in Training set

- ProofWriter: CC BY. Downloaded from <http://s3.amazonaws.com/proofwriter/proofwriter-dataset-V2020.12.3.zip>
- StrategyQA: MIT. Downloaded from https://storage.googleapis.com/ai2i/strategyqa/data/strategyqa_dataset.zip.

E More Details about Model Training

We finetune our 1.3B models on 32 V100s with batch size 8 on each GPU with totally 38 hours and 21 minutes. We finetune our 13B models on 128 V100s with batch size 4 on each GPU with totally 13 hours and 26 minutes.

Following OPT-IML (Iyer et al., 2022), we use Fully Sharded Data Parallel (Artetxe et al., 2021) and the Megatron-LM Tensor Parallelism (Shoeybi et al., 2019). We inherit most model hyper-parameters for each model scale following OPT-IML. We pack our training examples into sequences of length 2048, left-truncating examples that overflow. We use Adam (Kingma and Ba, 2014) with 32-bit state with $(\beta_1, \beta_2) = (0.9, 0.95)$, linearly warming up the learning rate for 60 steps to the maximum, followed by linearly decaying it to 0.

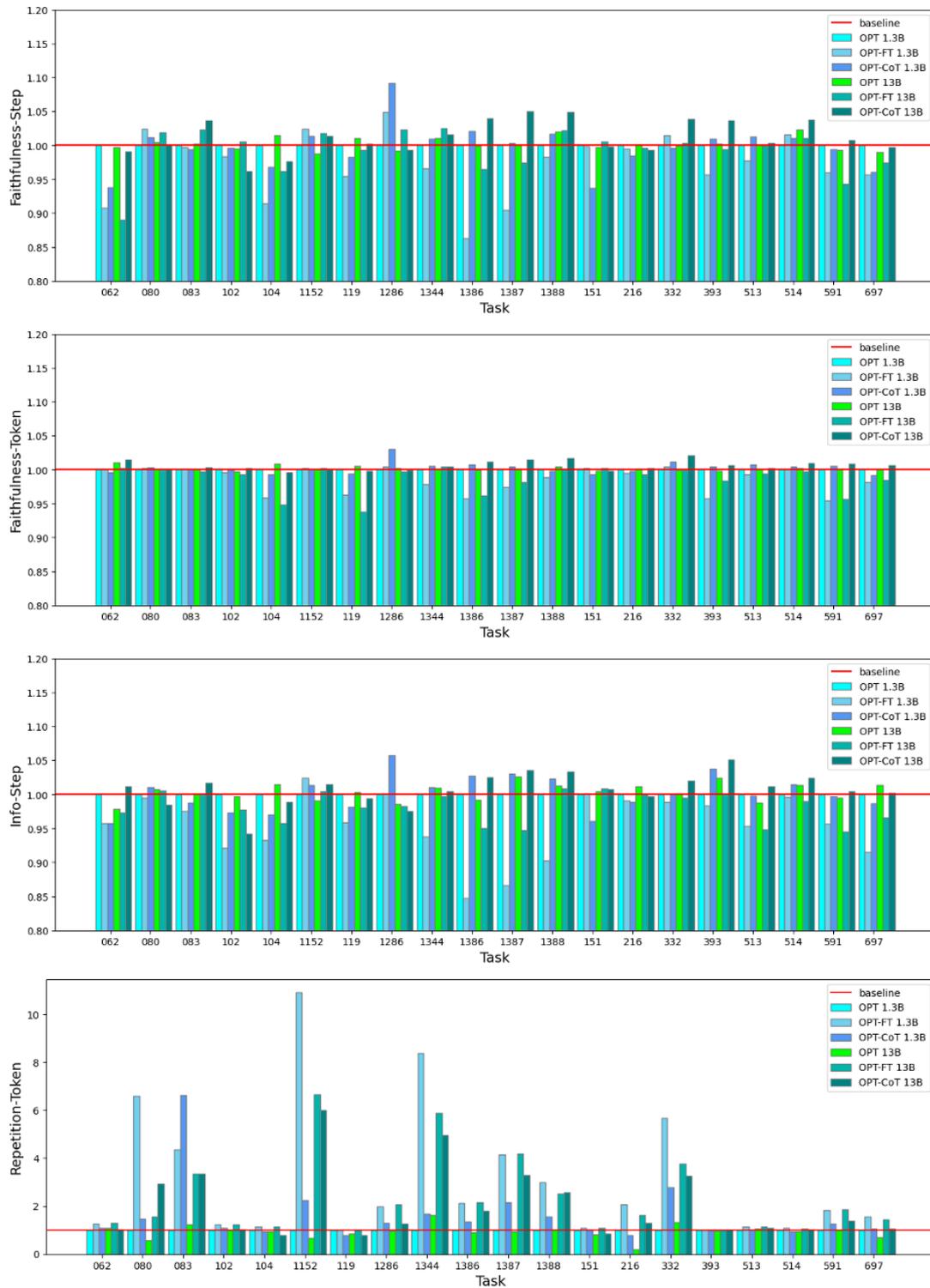


Figure 10: Normalized ROSCOE-SA scores per task, averaged across templates. Scores are normalised by their mean value across OPT 1.3B model’s generations.

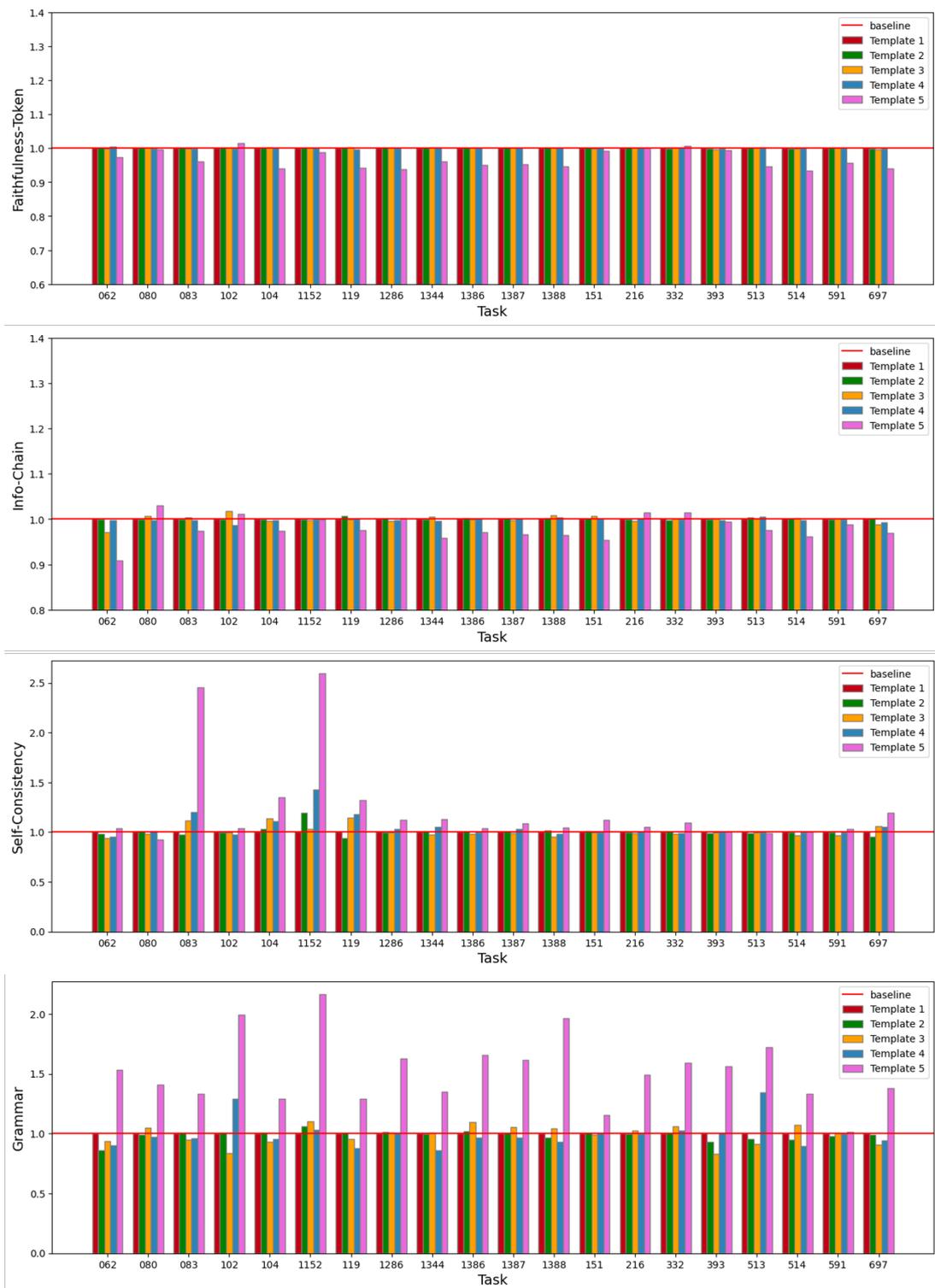


Figure 11: Selected scores per task for OPT-CoT 13B model. Scores are normalised by their mean value across Template 1 generations.

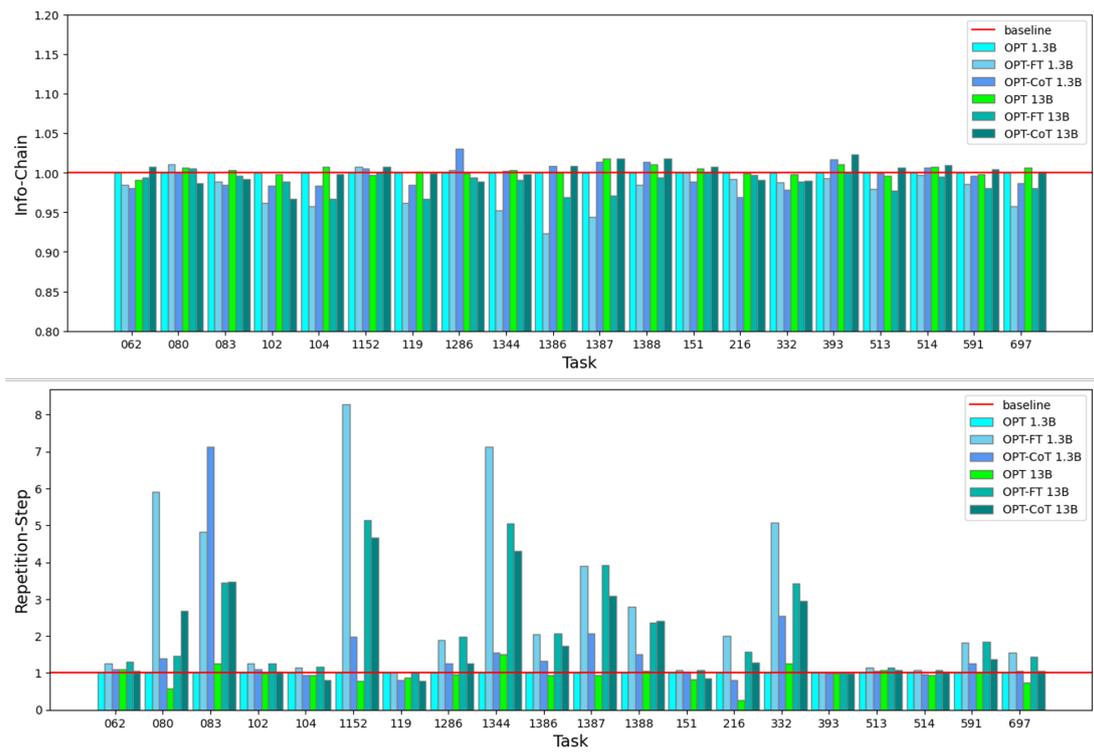


Figure 12: Normalized ROSCOE-SS scores per task, averaged across templates. Scores are normalised by their mean value across OPT 1.3B model's generations.

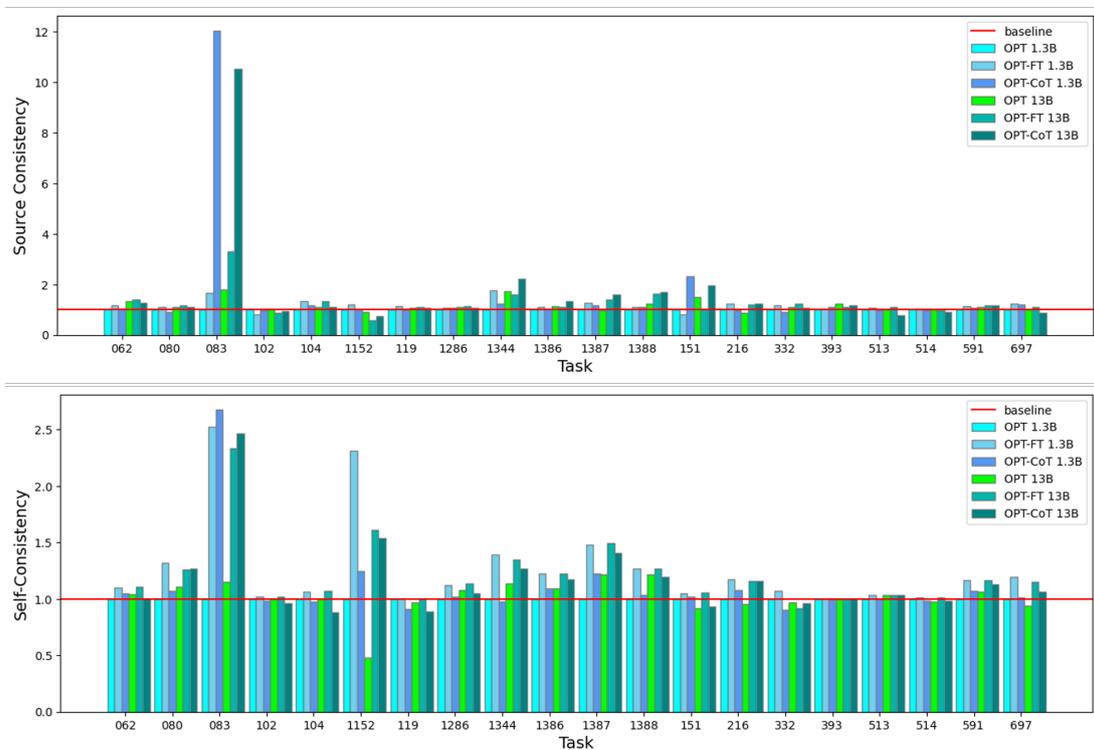


Figure 13: Normalized ROSCOE-LI scores per task, averaged across templates. Scores are normalised by their mean value across OPT 1.3B model's generations.

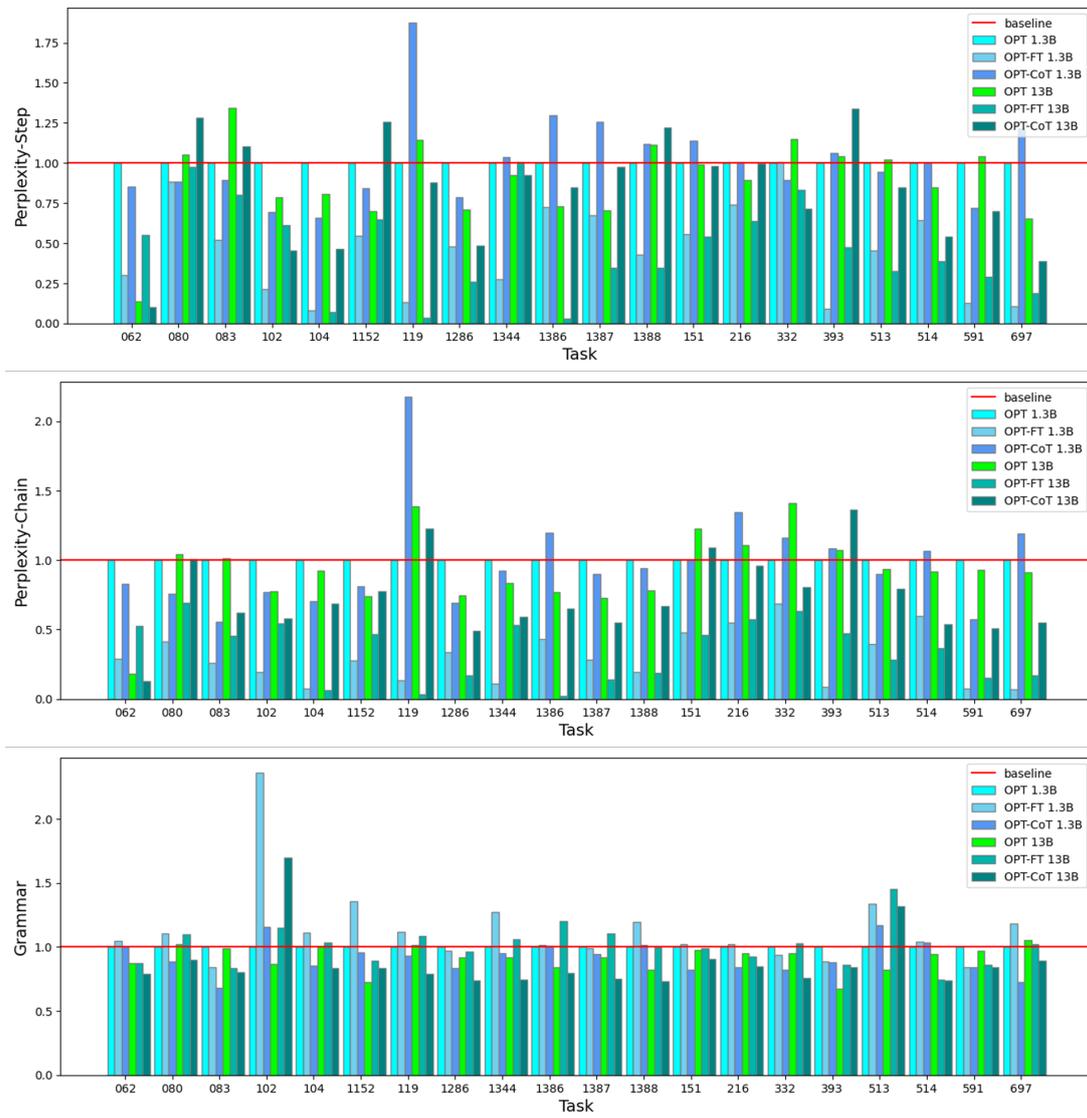


Figure 14: Normalized ROSCOE-LC scores per task, averaged across templates. Scores are normalised by their mean value across OPT 1.3B model’s generations.

REASONING SKILL: Logistic Reasoning, Mathematics

PROMPT: Please give a short explanation after the answer. Input: Identify the conclusion of the following argument. It is hard not to verify in our peers the same weakened intelligence due to emotions that we observe in our everyday patients. The arrogance of our consciousness, which in general, belongs to the strongest defense mechanisms, blocks the unconscious complexes. Because of this, it is difficult to convince people of the unconscious, and in turn to teach them what their conscious knowledge contradicts. (Sigmund Freud, The Origin and Development of Psychoanalysis) (A)It is hard not to verify in our peers the same weakened intelligence due to emotions that we observe in our everyday patients. (B)The arrogance of our consciousness, which in general, belongs to the strongest defense mechanisms, blocks the unconscious complexes. (C)Because of this, it is difficult to convince people of the unconscious, and in turn to teach them what their conscious knowledge contradicts. (D)It is difficult to convince people of the unconscious, and in turn to teach them what their conscious knowledge contradicts. Let's think step by step. Output: The answer is D because Because of the reasons specified in the argument, it is difficult to convince people of the unconscious, and in turn to teach them what their conscious knowledge contradicts. Input: Select the best translation into predicate logic: No people drive on Mars. (Dx: x drives on Mars) (A) $\sim Pd$ (B) $(\forall x)(Px \vee \sim Dx)$ (C) $(\forall x)(Px \supset \sim Dx)$ (D) $\sim Dp$ Let's think step by step. Output: The answer is C because Answer C means for everything, if that thing is a person, it doesn't drive on Mars. So that's the correct translation. Input: Select the best translation into predicate logic. David teaches Chris. (c: Chris; d: David; Txy: x teaches y) (A)Tdc (B)Tcd (C)Tcc (D)dTc Let's think step by step. Output: The answer is A because David teaches Chris, so in the relation Txy x should be 'd' for davis and y should be 'c' for Chris. Input: Which of the given formulas of PL is the best symbolization of the following sentence? Either England's importing beef is not a necessary condition for France's subsidizing agriculture, or China promotes human rights if South Africa supplies diamonds. (A) $(\sim F \supset E) \vee (C \equiv \sim S)$ (B) $(\sim E \supset F) \vee (C \supset S)$ (C) $(\sim F \supset E) \vee (C \equiv S)$ (D) $\sim(F \supset E) \vee (S \supset C)$ Let's think step by step. Output: The answer is D because When not defined in the question, entities, relationships or statements are shown with their first letter. In this example 'France subsidizing agriculture' is shown with F for simplicity. The correct answer is D, because it uses the 'or' operator between the two 'if' clauses. Input: Select the best translation into predicate logic. Kevin is introduced to José by Wilma. (j: José; k: Kevin; w: Wilma; Ixyz: x introduces y to z) (A)Iwjk (B)Ijkw (C)Ikjw (D)Iwkj Let's think step by step. Output:

REFERENCE: D

OPT 1.3B: The answer is **A** because all statements need to be presented with an argument.

OPT-FT 1.3B: The answer is **B**

OPT-CoT 1.3B: The answer is **A** because Using the 'or' operator between the statements, we have : Kevin is introduced to José, because of its 'e' factor in the equation. Wilma is introduced to Kevin because of its 'j' factor in the equation. José is introduced to Wilma because of its 'y' factor in the equation. In both cases, Wilma accepts 'x' and 'y'.

OPT 13B: The answer is **A** because introducing a person to another person is the action of introducing. It is possible that Kevin has been introduced to Wilma but not Jose.

OPT-FT 13B: The answer is **A** because when w is introduced, it says 'Wilma introduced Kevin to José'

OPT-CoT 13B: The answer is **D** because the correct term is 'introduces', the relation is 'j: kevin and w: wilma to k: x', it is an 'and' relation. Using the operator 'and' allows us to choose the greatest of the 5 possibilities.

REASONING SKILL: Commonsense Reasoning

PROMPT: Please give a short explanation after the answer. Input: Deserts are in extremely dry environments, so liquids will naturally be (A) rainbow (B) plentiful (C) meager (D) talented Let's think step by step. Output: The answer is C because This is a good example and the output correctly answers the question. Input: Dairy has things in it that would make humans sick, so in order for dairy to be edible, this process has to take place (A) mixing (B) boiling (C) freezing (D) pasteurized Let's think step by step. Output: The answer is D because in order for dairy to be edible, this process has to take place pasteurized. So, the correct answer is "D". Input: Digestion is when stomach acid breaks down (A) a pizza (B) a house (C) a rock (D) a car Let's think step by step. Output: The answer is A because Digestion is when stomach acid breaks down a food and only option A is a food. So, this example is correct. Input: With a ruler I can measure (A) feelings (B) love (C) air (D) a phone Let's think step by step. Output:

REFERENCE: D

OPT 1.3B: **A is a feeling, and all feelings come from feelings (A)**. Therefore, feelings are products of emotions (A). Therefore, emotions are created by emotions (A). Therefore, feelings and emotions are directly related (A). Therefore, emotions are created by emotions (A)... [repeats 249 times]

OPT-FT 1.3B: **A** because feelings and love are measurements and air is a phone

OPT-CoT 1.3B: The answer is D because With a ruler I can measure an object or range of objects and options are all to do with measuring. So, this example is correct.

OPT 13B: The answer is **B** because With a ruler I can measure a length and only options B and D are in front of me. So, the answer is B.

OPT-FT 13B: The answer is D because a phone is an object

OPT-CoT 13B: The answer is **A** because Answer: Measure feelings With a ruler I can measure.

REASONING SKILL: Casual Reasoning, Textual entailment

PROMPT: Please give a short explanation after the answer. Input: Premise: If you can dream it, you can achieve it — unless you're a goose trying to play a very human game of rugby. In the video above, one bold bird took a chance when it ran onto a rugby field mid-play. Things got dicey when it got into a tussle with another player, but it shook it off and kept right on running. After the play ended, the players escorted the feisty goose off the pitch. It was a risky move, but the crowd chanting its name was well worth it. Video credit: Storyful / Nathan Swarbrick @swazza1990 <sep> Hypothesis: The crowd believed they knew the name of the goose running on the field. Let's think step by step. Output: The answer is Entailment because This is a good example, the hypothesis sentence entails the given premise. Because the crowd was chanting its name, the crowd must have believed they knew the goose's name. Input: Premise: Notley hopeful attacks on women politicians will fade in coming year After a politically nasty year that saw women politicians face a barrage of attacks on social media and angry demonstrators shouting to have them put behind bars, Premier Rachel Notley is optimistic there will be a return to civil discourse. In a year-end interview with CBC News, Notley reflected on 2016, a year that saw hateful social-media attacks against MLA Sandra Jansen, and chants of "Lock her up" directed at the premier at an anti-carbon tax rally in Edmonton. <sep> Hypothesis: Notley reflected on the chants of "lock her up" from previous to the year 2016. Let's think step by step. Output: The answer is Contradiction because The Premise states Notley reflected on the chants from 2016, not before 2016. In this example hypothesis sentence contradicts the given premise. Input: Premise: Mazda Sales (Thailand), a unit of Japanese carmaker Mazda Motor Corp, remains bullish about its sales prospects this year despite falling figures in the first six months, the Bangkok Post reported, citing its President Hidesuke Takesue. (bit.ly/1gr7HMd) — NOTE: Reuters has not verified this story and does not vouch for its accuracy. (Bangkok Newsroom; Editing by Prateek Chatterjee) <sep> Hypothesis: Mazda Sales will be fired next week Let's think step by step. Output: The answer is Neutral because In this example the hypothesis can't be proven one way or the other based on the context. It's set in the future and no one knows the future. So, hypothesis sentence is neutral with respect to a given premise. Input: Premise: How to clear nasal congestion quickly
Gently blow your nose. Perhaps the easiest and quickest way to reduce nasal congestion is to simply blow your nose into a soft tissue. Unfortunately, it's not always possible to completely relieve congestion by merely blowing, but it's always a good starting point. <sep> Hypothesis: strike a pose to fix your nose Let's think step by step. Output:

REFERENCE: Contradiction

OPT 1.3B: The answer is **Modified Contradiction** because In this example the hypothesis sentence suggests the method is "Do not hit your head." In fact, that's the phrase a person who is rolling around on the ground, or who has had their head hit a hard surface while trying to get a good sleeper on the sheets should try. Why is this? A head-on collision will ... [*continues to hallucinate for 23 steps*]

OPT-FT 1.3B: Answer: **Blow your nose into a soft tissue.**

OPT-CoT 1.3B: The answer is Contradiction because The hypothesis states that Nastai will solve his and nasal congestion only by blowing into his nose, but that's not necessarily true, because he could have blown into his nose and caused irritation as well.

OPT 13B: The answer is **Yes** because The answer is Yes. All the above information about clearing the nose is proven by the given premise, but if you still don't know the answer, then we suggest you go here and click on "How to clear your nose naturally" page.

OPT-FT 13B: The answer: contradiction

OPT-CoT 13B: The answer is contradiction because One cannot blow their nose if they have struck a pose.

REASONING SKILL: Analogical Reasoning

PROMPT: Please give a short explanation after the answer. Input: throw : fly. aspire : ? Let's think step by step. Output: The answer is attain because Fly is a causation of throw. Hence, the inferred analogy being causation, attain is the causation of aspire. Input: listen : hear. drop : ? Let's think step by step. Output: The answer is fall because Hear is a causation of listen. Hence, the inferred analogy being causation, the causation of drop is fall. Input: hit : shatter. nourish : ? Let's think step by step. Output: The answer is grow because Shatter is a causation of hit. Hence, the inferred analogy being causation, the causation of nourish is grow. Input: throw : fly. ignite : ? Let's think step by step. Output:

REFERENCE: burn

OPT 1.3B: The answer is **move** because fly is a causation of ignite. Hence, the inferred analogy being causation, the causation of move is fly.

OPT-FT 1.3B: The answer is **throw**

OPT-CoT 1.3B: The answer is **ignite** because Fire is a causation of throw. Hence, the inferred analogy being causation, the causation of ignite is fly.

OPT 13B: The answer is **evaporate** because Fly is a causation of throw. Hence, the inferred analogy being causation, the causation of evaporate is ignite.

OPT-FT 13B: The answer is **fly** because Ignite is a causation of throw

OPT-CoT 13B: The answer is **grow** because Throw is a causation of grow.

REASONING SKILL: Commonsense reasoning, Spatial Reasoning

PROMPT: Please give a short explanation after the answer. Input: How do you remove keys from a computer keyboard for cleaning? Let's think step by step. Output: The answer is Place the tip of a knife under the corner of a key and gently twist to pop it off. because The process of removing keys from a keyboard for cleaning is descriptive and logical. Input: How can you dry wet replacement computer keyboard keys? Let's think step by step. Output: The answer is by arranging paper towels on the counter and evenly spacing out each key to dry, face side up, on top of them. because The steps involved in the process of drying wet replacement keyboard keys are logical. Input: To coat small plastic objects like replacement keyboard keys with soap Let's think step by step. Output: The answer is place them, water and liquid dishsoap in a sealed container. Shake vigorously until the container is completely full of foam. because The process of coating small plastic objects like replacement keys is descriptive and logical. Input: How do you cut gnocchi without a knife? Let's think step by step. Output: The answer is Press the edge of a spoon down into the dough to cut it. because The answer is quite detailed about the method to cut gnocchi when knife is not available. Input: How do you taste something? Let's think step by step. Output: The answer is Place it in your mouth to taste. because To taste something one has to keep it in contact with the tongue. Hence it is logical to put it in the mouth to taste. Input: How do you wear a bracelet? Let's think step by step. Output: The answer is Put it on around your wrist. because Bracelet is worn around wrists. Hence it is correct answer. Input: To find an easy way to complete levels on a video game, Let's think step by step. Output:

REFERENCE: look up cheat codes to make the rules of the game different or null.

OPT 1.3B: The answer is **Enter key without pressing**. because How to find the perfect way to complete levels on a video game is logical and descriptive.

OPT-FT 1.3B: **It tells you the easiest way to get all the coins and medals and the fastest way to get through the level.**

OPT-CoT 1.3B: The answer is **play the game**. because If you play a game to finish the levels then you can call it a game.

OPT 13B: The answer is **Complete the level as fast as possible to get a high score**, by jumping through the stage as quickly as possible. because The answer is explicit and very clear that how a user can complete levels on a video game with just a hint.

OPT-FT 13B: The answer is **try harder or use a guide**, because The process of finding an easy way to complete levels on a video game is descriptive and logical.

OPT-CoT 13B: The answer is **YouTube to find out how to get past a level**. because YouTube is used to find out ways to complete a game easily.

REASONING SKILL: Spatial Reasoning

PROMPT: Instructions: You will be given a passage consisting of set of facts and a question as input. The task is to answer a question of form 'Where is <person_name>?' using one of the given facts to determine the latest location of the person. Answer should be a word/phrase describing the location from the supporting fact. Avoid answers that are incomplete or incorrect. Positive Examples: Input: Passage: Mary moved to the bathroom. John went to the hallway. Question: Where is Mary? Output: bathroom Input: Passage: Mary moved to the bathroom. John went to the hallway. Daniel went back to the hallway. Sandra moved to the garden. Question: Where is Daniel? Output: hallway Input: Passage: Mary moved to the bathroom. John went to the hallway. Daniel went back to the hallway. Sandra moved to the garden. John moved to the office. Sandra journeyed to the bathroom. Question: Where is Daniel? Output: hallway Input: Passage: John moved to the hallway. Mary travelled to the hallway. Sandra went to the bedroom. Sandra went back to the hallway. Question: Where is Mary? Output:

REFERENCE: hallway

OPT 1.3B: **bathroom**

OPT-FT 1.3B: hallway

OPT-CoT 1.3B: hallway

OPT 13B: **bathroom**

OPT-FT 13B: **bathroom**

OPT-CoT 13B: hallway

Table 12: Examples from tasks that require different reasoning skills and generated outputs. The failed outputs are highlighted in red. Outputs that are different from the reference, but can be valid because the question is ambiguous are highlighted in blue. All except the one are examples generated using prompt Template 4, last example used Template 5.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Section 7
- A2. Did you discuss any potential risks of your work?
We did not see any potential risks in our paper
- A3. Do the abstract and introduction summarize the paper’s main claims?
Section 1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Our data and pretrained model may contain scientific artifacts. Section 2 (data part). Section 3 (model part)

- B1. Did you cite the creators of artifacts you used?
Section 2 and Section 3
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section D in appendix
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2 and Section D in appendix
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No, we use public datasets.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No, we use public datasets.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix

C Did you run computational experiments?

Section 3

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Section 3

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Section 4

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Section 3 and 4

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.