Summary of LIMA– Less Is More for Alignment and Finetuned Language Models Are Zero-Shot Learners

member_1(netid_1),member_2(netid_2), member_3(netid_3)

Problem and Motivation

Although large-scale pre-trained models (e.g., GPT-3) have demonstrated impressive language abilities through few-shot learning, their zero-shot performance often falls short, particularly on tasks that differ significantly from pre-training data, including reading comprehension, question answering, and language inference. To improve the zero-shot performance, Wei et al. proposed *instruction tuning*, a method that finetunes the model on a large-scale dataset designed to express natural language processing tasks via natural language instructions. By applying instruction tuning to the pre-trained model with a new dataset – Finetuned Language Net (FLAN), the zero-shot model can outperform few-shot GPT-3 significantly on various tasks. Their findings suggest that instruction tuning is a promising technique for aligning zero-shot models to respond in ways preferred by humans.

Nevertheless, existing alignment methods, including instruction tuning and reinforcement learning from human feedback, require large-scale computing and specialized data. To address this challenge, Zhou et al. hypothesized that alignment can be achieved via finetuning on a small set of carefully curated training examples. To test this hypothesis, they curated 1,000 examples that mimic real user prompts paired with high-quality responses. After fine-tuning the LLaMa model on this curated dataset, they demonstrated that finetuning on small-scale high-quality data is sufficient to align the model's output with human preference, thereby significantly reducing the cost of post-training fine-tuning.

In the rest of this summary, we review previous research, detail the methods used, discuss limitations, and explore future directions for this topic.

Related Works

[The related work before these two papers]

Multitask-prompted training enables zero-shot task generalization [1]. This paper fine-tuned T5 in an instruction-tuning setup, demonstrating improved zero-shot learning in a model with 11B parameters. To explore whether zero-shot generalization can be directly induced through explicit multitask learning, the paper develops a system to map natural language tasks into human-readable prompts. A large set of supervised datasets is converted, each with multiple prompts in varied wording, enabling the model to be benchmarked on entirely unseen tasks specified through natural language.

Reinforcement Learning from Human Feedback (RLHF). Large language models can generate toxic or useless outputs. To address it, previous work proposes to finetune the model by reinforcement learning from human feedback (RLHF) [2, 3]. They train a reward model (RM) on a human-labeled dataset and use it as a reward function to fine-tune large

language models. These methods help large language models generate more human-preferred outputs and reduce toxic text. However, they require a large number of training examples, which demands significant human labeling effort.

[The related work after these two papers]

Scaling Instruction-Finetuned Language Models [4]. Finetuning language models on datasets framed as instructions has been shown to enhance performance and generalization to unseen tasks. This paper investigates instruction finetuning with a focus on three key aspects: (1) increasing the number of tasks, (2) scaling model size, and (3) finetuning on chain-of-thought data. The results show that incorporating these elements significantly boosts performance across various model architectures (PaLM, T5, U-PaLM), prompt setups (zero-shot, few-shot, CoT), and evaluation benchmarks (MMLU, BBH, TyDiQA, MGSM, open-ended generation, RealToxicityPrompts).

Solution Overview

Finetuned Language Models Are Zero-Shot Learners.

To enhance the zero-shot performance of large language models, this paper investigates a straightforward approach—instruction tuning, which means finetuning language models on a collection of datasets described via instructions, as shown at the top of Figure 1. This method builds on the idea that NLP tasks can be framed as natural language instructions, such as “Is this movie review’s sentiment positive or negative?”.

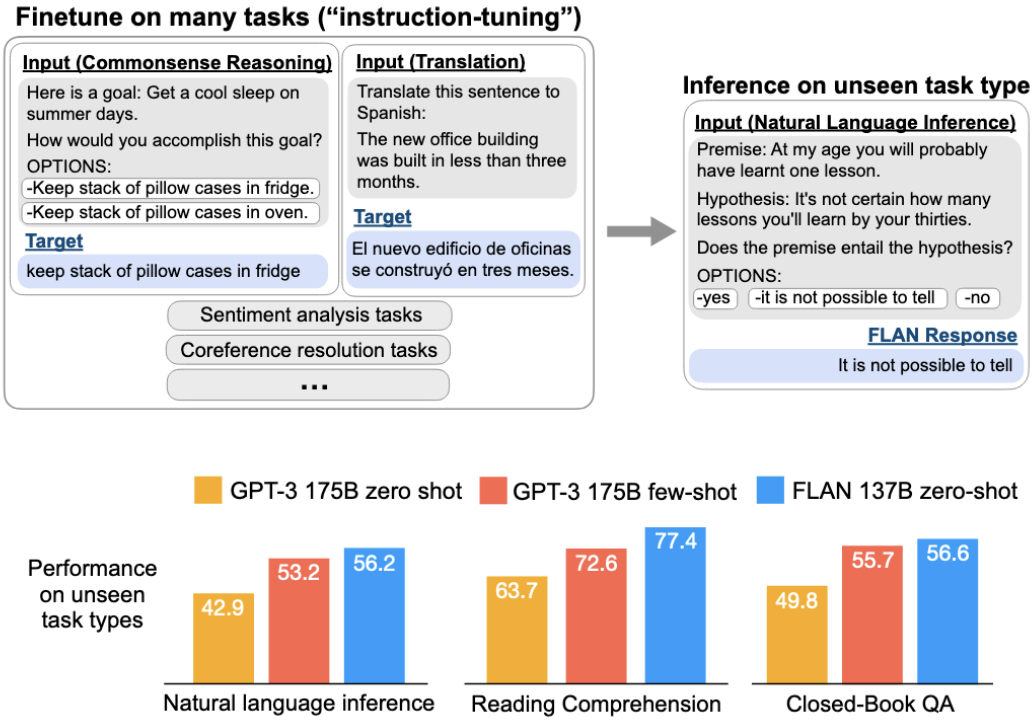

Figure 1: Top: overview of instruction tuning and FLAN. Bottom: performance of zero-shot FLAN, compared with zero-shot and few-shot GPT-3, on three unseen task types.

Instruction tuning is a straightforward technique that, as shown in Figure 2, merges the strengths of both the pretrain-finetune approach and the prompting paradigm. It achieves this by using supervised finetuning to enhance the language model's ability to respond more effectively during inference-time text interactions.

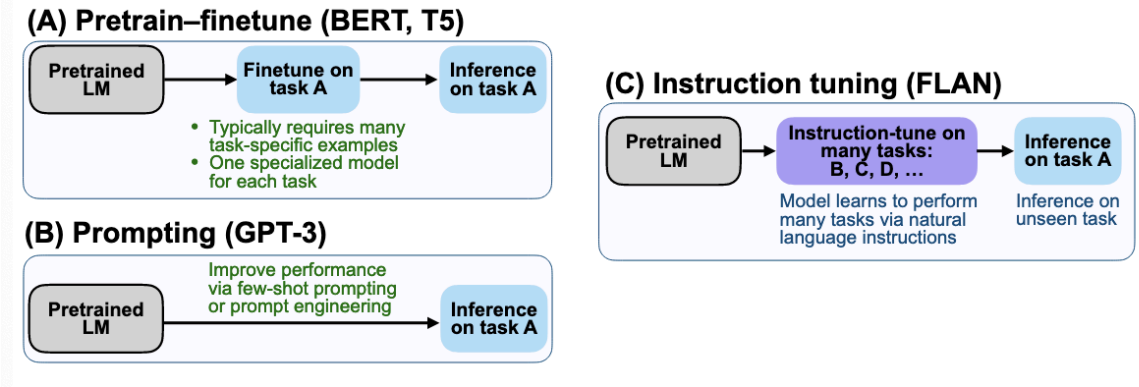

Figure 2. Comparing instruction tuning with pretrain–finetune and prompting.

This paper addresses the resource-intensive process of creating an instruction-tuning dataset from scratch by transforming existing datasets into an instructional format. It aggregates 62 publicly available text datasets from Tensorflow Datasets, covering both language understanding and generation tasks, into a unified mixture. The datasets are organized into twelve task clusters, each representing a specific task type. For each dataset, the paper manually designs ten unique templates, using natural language instructions to describe the task. To enhance diversity, up to three templates per dataset "turn the task around," such as by generating a movie review for sentiment classification. The pretrained language model is then instruction-tuned on this mixture, with examples formatted according to randomly selected templates. Figures 3 illustrate the multiple instruction templates describing a natural language inference task.

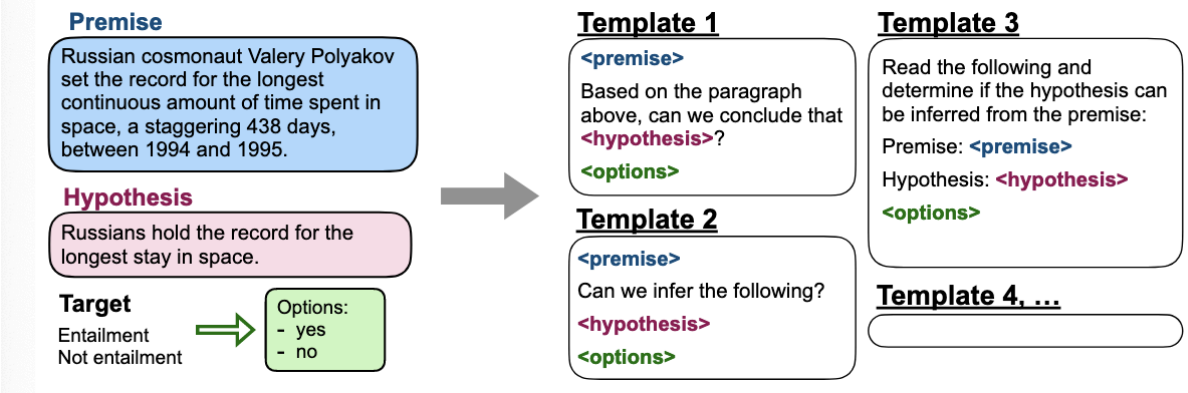

Figure 3: Multiple instruction templates describing a natural language inference task.

Results. As shown in the bottom of Figure 1, the performance of zero-shot FLAN, performs better than zero-shot and few-shot GPT-3, on three unseen task types where instruction tuning improved performance.

LIMA– Less Is More for Alignment.

The paper first proposes a hypothesis: the power of a model is gained almost entirely during pretraining, and the alignment is a simple process that helps the model learn the style of interacting with users.

To verify this hypothesis, they created a dataset of 1000 prompts and responses, in which the inputs are diverse, but the outputs have similar styles, including 750 samples from three community Q&A websites and 250 manually authored samples. They finetune LLaMa 65B on this 1,000-example dataset and compare the performance with several baseline models which are finetuned with a large-scale training set. Although it has fewer training examples, LIMA can outperform some models, which verifies the author's hypothesis. From the ablation study, they find only scaling up the quantity is not enough, prompt diversity is also important.

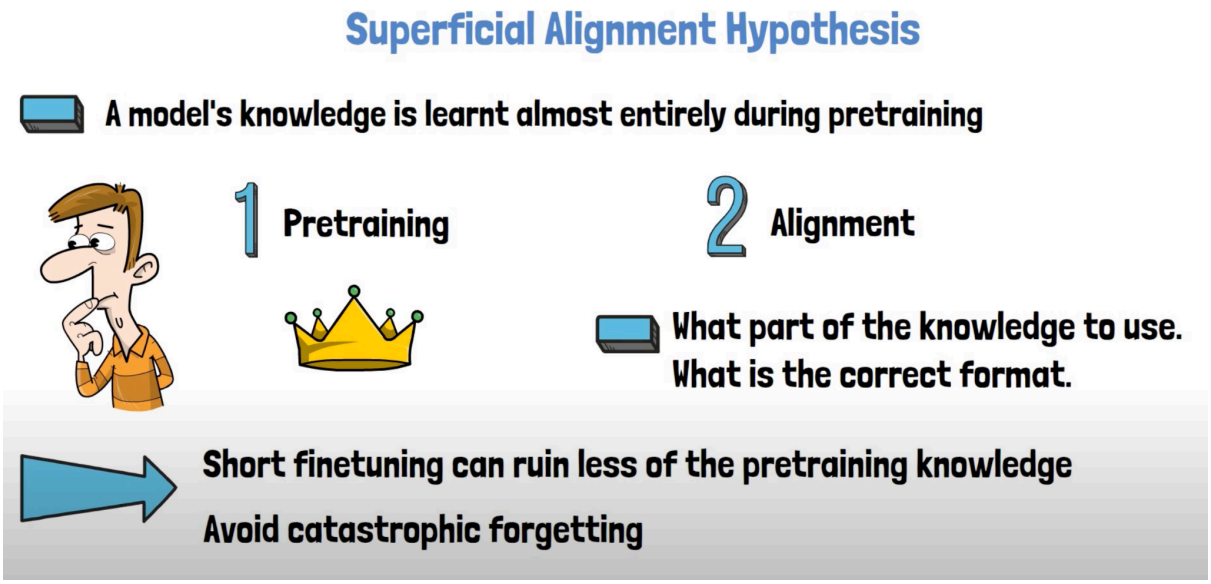

Figure 4: [Superficial Alignment Hypothesis](#).

Limitations

Instruction tuning

1. It is challenging and expensive to create a large-scale fine-tuning dataset (e.g., with more than 60 instruction tasks) with high quality.
2. There is a degree of subjectivity in assigning tasks to clusters for instruction-tuning.
3. It only explores the use of relatively short instructions typically a single sentence.
4. It may influence the performance due to OOD issue, resulting in forgeting knowledge learned during pre-training that is unrelated to the fine-tuned tasks.

LIMA

1. It is difficult to obtain a response with some personalized style if LLM is fine-tuned in a fixed response style.
2. Curating high-quality datasets may require a large amount of human labor.
3. It is unclear how the small-scale finetuning method scales to tasks with significantly different contexts.

Future Research Directions

1. We can develop measurements to quantify the effect of the diversity and quantity of the training (finetuning) dataset on the final model performance.
2. We can automatically synthesize high-quality finetuning datasets for the personalized use of LLMs.
3. We can combine finetuning methods and few-shot methods to further improve the performance of LLMs on challenging tasks, such as questions with complex reasoning or questions related to timely information.

Summary of Class Discussion

We first list five questions that are most interesting and widely discussed in the class.

Q1: **[Method]** *Based on LIMA, both the diversity of topics and the consistency of response style are important to the finetuning performance. Which one is more important?*
A1: The LIMA paper does not answer this question in detail. To address it in the future, we need more ablation studies to compare between (1) inconsistent response style with high topic diversity and (2) consistent response style with low topic diversity.

Q2: **[Evaluation]** *Given two different generations of the LLM, how can we evaluate which one is better?*
A2: Both papers do not provide details beyond just asking “Which one is better?” However, evaluating generation models is inherently challenging. Using a different LLM to judge can be an approach that is easy to implement and efficient to execute. However, studies show that the order of the answer given to the judger may affect the results. In this case, we need to make multiple runs with different orders and take an average as the result.

Q3: **[Limitation]** *Given that the instruction tuning will adjust all the parameters in the original pre-trained language model, how can the tuning process affect the model’s knowledge of tasks not presented in the finetuning dataset? Specifically, will the tuning process drop knowledge of any tasks, which can be a disadvantage when the task's labeled data is available?*
A3: Finetuning can cause the out-of-distribution issue. This can occur especially when, after fine-tuning, the mode loses its generalization capabilities or forgets knowledge learned during pre-training that is unrelated to the fine-tuned tasks.

Q4: **[Evaluation]** *Is it better to have diverse and inconsistent finetuning data as in the instruction tuning or data with consistent format as in LIMA?*
A4: Diverse templates are better in terms of improving zero-shot prompting ability. However, for alignment with small datasets, it is better to use a consistent format.

Q5: **[Future work]** *How can we generalize the instruction tuning method to stable diffusion or vision-language models? How to construct the image and text pairs?*
A5: We can generalize finetuning to models with other modalities, including vision models. However, there are two challenges as we discussed in Q1 and Q2. First, how can we evaluate the quality of image generation? Is using another vision-language model as a judger a good solution? Second, how can we ensure the diversity of finetuning data is high?

Does K-means with pre-trained image encoding work well? Those questions can be interesting future work directions.
To summarize, the class discussion started from the concepts and methods proposed in the paper, including finetuning procedures and the diversity of finetuning datasets, and extended to many interesting future work, such as the guideline for choosing finetuning procedures and finetuning non-text models. Those inspiring discussions demonstrate that many aspects of LLM finetuning are unexplored, resulting in many opportunities in future research.

Appendix: Additional Class Questions

Q1: **[Method]** *Does LIMA train the entire model or the Lora version?*
A1: The entire model.

Q2: **[Future work]** *Are instruction tuning and few-shot learning complementary to each other? Can we combine them to make the model perform even better?*
A2: Yes

Q3: **[Method]** *What are the benefits of packing to combine multiple training examples in instruction tuning?*
A3: It is not computationally efficient to always pad inputs to the maximum length. In contrast, packing inputs is more efficient. However, this raises the challenge of efficient heterogeneous computing with GPUs. For example, how to reduce energy bubbles when we pack inputs during training.

Q4: **[Evaluation]** *What is the difference between LIMA and instruction-turning in designing templates?*
A4: In instruction tuning, datasets are constructed manually by authors. In LIMA, datasets are constructed based on the community Q&A.

Reference

1. Sanh, Victor, et al. "Multitask prompted training enables zero-shot task generalization." *International Conference on Learning Representations*. 2022.

2. Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.

3. Bai, Yuntao, et al. "Training a helpful and harmless assistant with reinforcement learning from human feedback." *arXiv preprint arXiv:2204.05862* (2022).

4. Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." Journal of Machine Learning Research 25.70 (2024): 1-53.