AgentInstruct: Agent Instructs Large Language Models to be General Zero-Shot Reasoners

Nicholas Crispino, Kyle Montgomery, Fankun Zeng, Dawn Song, Chenguang Wang

https://arxiv.org/abs/2310.03710

Code

Introduction

Zero-Shot

Passage: Alan Rickman & Emma Thompson give good performances with southern/New Orleans accents in this detective flick. It's worth seeing for their scenes- and Rickman's scene with Hal Holbrook.

These three actors mannage to entertain us no matter what the movie, it seems. The plot for the movie shows potential, but one gets the impression in watching the film that it was not pulled off as well as it could have been.

The fact that it is cluttered by a rather uninteresting subplot and mostly uninteresting kidnappers really muddles things.

The movie is worth a view- if for nothing more than entertaining performances by Rickman, Thompson, and Holbrook.

Sentiment: Positive ✗

Zero-Shot CoT

Passage: Alan Rickman & Emma Thompson...
Let's think step by step.

...What is the author's conclusion?

The author concludes that the movie is worth a view, mainly for the entertaining performances of the three main actors, despite some flaws in the plot.

Therefore, the correct label among {'Negative', 'Positive'} (just the label) to the original task is below.

Sentiment: Positive ✗

Zero-Shot AgentInstruct

Task Information
Name: *IMDB*; Labels: *Positive, Negative*; Input Examples: *"Passage: ..."*

Agent

Agent Instructions

Instructions:
1. Understand the Dataset: The IMDB Movie Reviews dataset contains 50,000 reviews... These reviews are labeled as either positive or negative...
2. Analyze the Passage: Pay attention to the language, the tone of the review, and any explicit or implicit expressions of sentiment towards the movie...

Chain of Thought (CoT) Reasoning

Passage: Alan Rickman & Emma Thompson...
Follow the instructions...

The passage provides a mixed assessment of the movie, praises the performances of Alan Rickman, Emma Thompson, and Hal Holbrook but also expresses disappointment... the reviewer's use of language and tone suggests that they are leaning towards a negative sentiment...

Additionally, the reviewer mentions that the movie is "worth a view" only for the performances of the three actors, implying that the movie itself is not particularly noteworthy.

Therefore, the correct label among {'Negative', 'Positive'} (just the label) to the original task is below.

Sentiment: Negative ✓

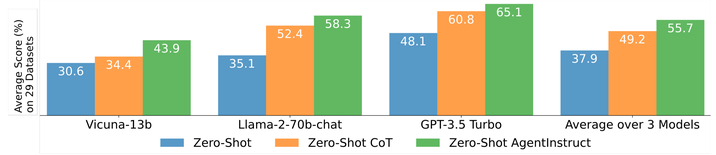

Figure 1: Summary of our approach and results. Top: Our zero-shot AgentInstruct generalizes the zero-shot reasoning abilities of large language models to a wide set of language understanding tasks including generation, classification, and reasoning. Our agent produces task-specific instructions to instruct the reasoning process of large language models to solve a task. Both the agent instructions and task-specific reasoning process are highlighted. Bottom: Performance of zero-shot AgentInstruct compared with standard zero-shot and zero-shot chain of thought (CoT). Zero-shot AgentInstruct improves the performance of three large language models substantially on the 29 datasets we evaluate.

Approach

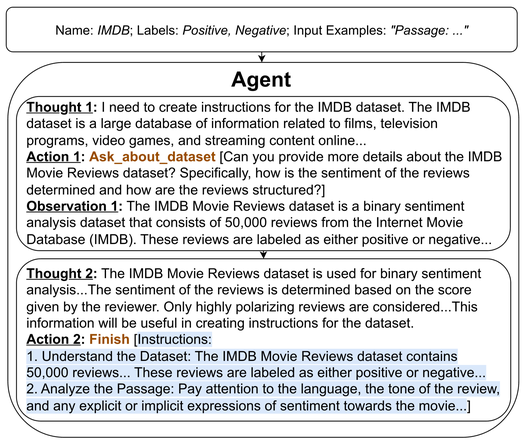

Figure 2: An example of our agent producing instructions for a classification dataset IMDB. The task-specific instructions are highlighted.

There are two main parts of zero-shot AgentInstruct:

1. Agent Instructions: We build an agent to automatically synthesize high-quality instructions using a powerful language model with access to a wide range of existing task knowledge on the web (Figure 2).
2. Chain of Thought Reasoning: LLMs follow our task-specific instructions to decompose the task into a chain of more specific intermediate steps to solve the task.

Experiment

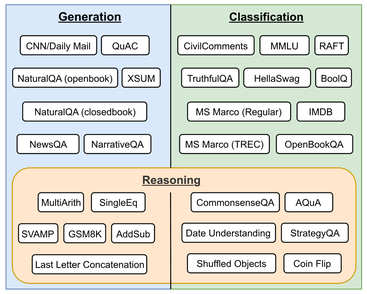

Figure 3: Datasets for generation (blue), classification (green), and reasoning (orange). Reasoning contains generation and classification tasks.

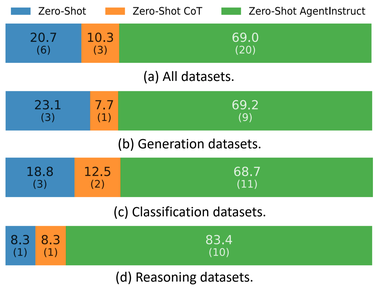

Figure 4: Winning rate (%) between zero-shot, zero-shot CoT, and zero-shot AgentInstruct based on the average results over three models.

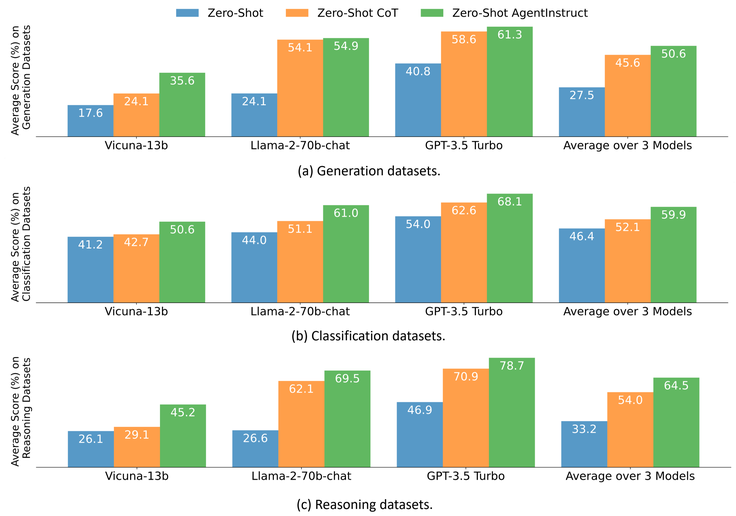

Figure 5: Results on Vicuna-13b, Llama-2-70b-chat, and GPT-3.5 Turbo across tasks. Top: generation. Middle: classification. Bottom: reasoning.

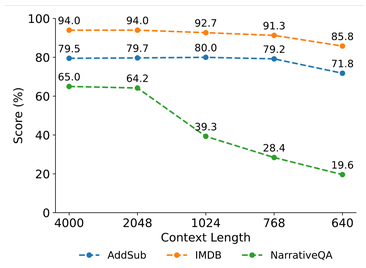

Figure 6: Truncating context lengths on Llama-2-70b-chat with zero-shot AgentInstruct on AddSub, IMDB, and NarrativeQA.

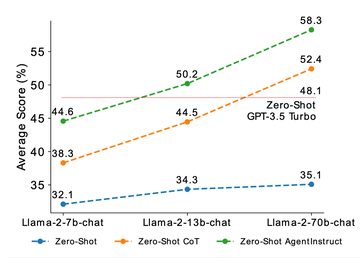

Figure 7: Model scaling results of zero-shot, zero-shot CoT, and zero-shot AgentInstruct with Llama-2-chat on all datasets.

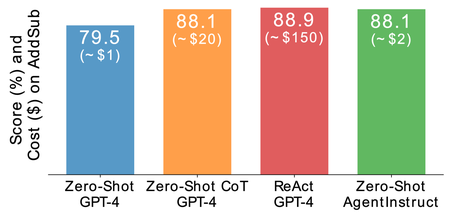

Figure 8: Comparison on GPT-4 using zero-shot, zero-shot CoT, ReAct, and zero-shot AgentInstruct on AddSub.

	AddSub	IMDB	NarrativeQA
Zero-Shot AgentInstruct	**79.5**	**94.0**	**65.0**
w/o Agent Instructions	73.2	89.0	62.3
w/o Input Examples	72.4	88.0	60.1
w/o Labels	74.9	93.8	63.9
w/o GPT-4	75.2	92.6	63.5

Table 1: Ablation over different facets of zero-shot AgentInstruct with Llama-2-70b-chat.

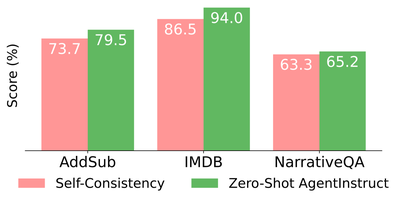

Figure 10: Comparison between zero-shot AgentInstruct and zero-shot self-consistency on Llama-2-70b-chat.

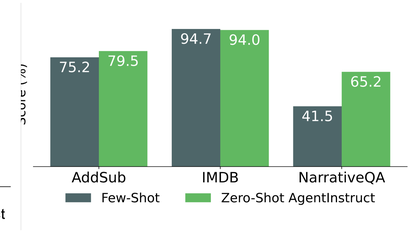

Figure 9: Comparison between zero-shot AgentInstruct and few-shot on Llama-2-70b-chat on AddSub, IMDB, and NarrativeQA.

Passage: This is a good movie. Terrance Stamp is great, the music is sweet, Carol White is very believeable. The only thing that marred this was the shakey acting of Carol's first husband, but if you can get past that, you're OK. Donovan provides some of the most languid, mellow, bittersweet lyrics from the 60s.

CoT Reasoning: The reviewer mentions that the movie is "good" and that it has a "sweet" sound track. They also mention that the acting by Carol White is "believeable". However, they also mention that the acting by Carol's first husband is "shakey."

Answer: Positive ✓

Figure 11: Case study example for Llama-2-70b-chat with zero-shot AgentInstruct on IMDB. Here, the answer is correct and the task-specific reasoning is helpful for finding the answer (highlighted).

Conclusion

Our work proposes a new way of improving the zero-shot reasoning abilities of large language models on general language understanding tasks:

- Our agent automatically generates task-specific instructions for a wide set of tasks, guiding LLMs to reason better across these tasks.
- Our method is zero-shot so no input-output examples are required to solve the task.
- Our approach leads to substantial improvements across various NLP tasks spanning generation, classification, and reasoning.
- Our method wins on 20 of the 29 datasets used for evaluation.

We believe zero-shot AgentInstruct's style of human-understandable reasoning, along with its utilization of an autonomous agent, can replace more traditional styles of zero or few-shot prompting as models become equipped with stronger reasoning capabilities.