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

There are two main parts of zero-shot large language models substantially on the 29 datasets we evaluate.

1. **AgentInstruct**

   - **Introduction**
     - **Approach**
       - **Agent**
         - **Thought 1:** I need to create instructions for the IMDB dataset. The IMDB dataset is a large database of information related to films, television programs, video games, and streaming content online...
         - **Action 1:** Ask about dataset [Can you provide more details about the IMDB Movie Reviews dataset? Specifically, how is the sentiment of the reviews determined and how are the reviews structured?]
         - **Observation 1:** The IMDB Movie Reviews dataset is a binary sentiment analysis dataset that consists of 50,000 reviews from the Internet Movie Database (IMDB). These reviews are rated as either positive or negative.

   - **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:** Winning rate (%) between zero-shot, zero-shot CoT, and zero-shot AgentInstruct on AddSub.

     - **Figure 6:** Results on Vicuna-1.3B, Llama-2-70b-chat, and GPT-3.5 Turbo across tasks. Top: generation. Middle: classification. Bottom: reasoning.

     - **Figure 7:** Model scaling results of zero-shot, zero-shot CoT, and zero-shot AgentInstruct on Llama-2-70b-chat.

   - **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.

   - **Figure 8:** Comparison between zero-shot AgentInstruct and few-shot on Llama-2-70b-chat on AddSub, IMDB, and NarrativeQA.

   - **Table 1:** Ablation over different facets of zero-shot AgentInstruct.

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

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

   - **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).

   - **Table 2:** The IMDB Movie Reviews dataset is used for binary sentiment analysis. The sentiment of the reviews is determined based on the score given by the reviewers. Only highly polarizing reviews are considered. This information will be useful in creating instructions for the dataset.

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

   - **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.

   - **Table 3:** Ablation over different facets of zero-shot AgentInstruct.