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

Nicholas Crispino, Kyle Montgomery, Fankun Zeng, Dawn Song, Chenguang Wang
Large language models exhibit advanced performance on language understanding tasks.
Zero-shot reasoning methodologies have proven useful for specific task domains.
Language Models as Agents

Agents utilize planning to accomplish more complex tasks
Our agent produces task-specific instructions to instruct the reasoning process of LLMs.
Zero-Shot AgentInstruct: Agent Instructions

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

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 labeled as either positive or negative...

Thought 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 reviewer. Only highly polarizing reviews are considered...This information will be useful in creating instructions for the dataset.

Action 2: Finish [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...]

An example of our agent creating task-specific instructions for the IMDB dataset
We evaluated our method on 29 datasets spanning generation, classification, and reasoning.
Results

Zero-shot AgentInstruct shows a significant improvement over all 29 datasets on average.
Results

Zero-shot AgentInstruct shows an improvement on each category of tasks, especially reasoning.
### Results

Zero-shot AgentInstruct achieves SoTA results on 20 of the 29 datasets.

#### (a) Overall

<table>
<thead>
<tr>
<th></th>
<th>Zero-Shot</th>
<th>Zero-Shot CoT</th>
<th>Zero-Shot AgentInstruct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>20.7 (6)</td>
<td>10.3 (3)</td>
<td>69.0 (20)</td>
</tr>
</tbody>
</table>

#### (b) Generation Datasets

<table>
<thead>
<tr>
<th></th>
<th>Zero-Shot</th>
<th>Zero-Shot CoT</th>
<th>Zero-Shot AgentInstruct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>23.1 (3)</td>
<td>7.7 (1)</td>
<td>69.2 (9)</td>
</tr>
</tbody>
</table>

#### (c) Classification Datasets

<table>
<thead>
<tr>
<th></th>
<th>Zero-Shot</th>
<th>Zero-Shot CoT</th>
<th>Zero-Shot AgentInstruct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>18.8 (3)</td>
<td>12.5 (2)</td>
<td>68.7 (11)</td>
</tr>
</tbody>
</table>

#### (d) Reasoning Datasets

<table>
<thead>
<tr>
<th></th>
<th>Zero-Shot</th>
<th>Zero-Shot CoT</th>
<th>Zero-Shot AgentInstruct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>8.3 (1)</td>
<td>8.3 (1)</td>
<td>83.4 (10)</td>
</tr>
</tbody>
</table>
### Ablation

All components of zero-shot AgentInstruct are important

<table>
<thead>
<tr>
<th></th>
<th>AddSub</th>
<th>IMDB</th>
<th>NarrativeQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot AgentInstruct</td>
<td>79.5</td>
<td>94.0</td>
<td>65.0</td>
</tr>
<tr>
<td>w/o Agent Instructions</td>
<td>73.2</td>
<td>89.0</td>
<td>62.3</td>
</tr>
<tr>
<td>w/o Input Examples</td>
<td>72.4</td>
<td>88.0</td>
<td>60.1</td>
</tr>
<tr>
<td>w/o Labels</td>
<td>74.9</td>
<td>93.8</td>
<td>63.9</td>
</tr>
<tr>
<td>w/o GPT-4</td>
<td>75.2</td>
<td>92.6</td>
<td>63.5</td>
</tr>
</tbody>
</table>
Comparison on GPT-4

Zero-shot AgentInstruct is a cost-effective alternative to using agents directly.
Zero-shot AgentInstruct is sensitive to the context window
Scaling

Llama-2-70b-chat with Zero-shot AgentInstruct outperforms zero-shot ChatGPT by 10.2%
Comparison Methods: Few-Shot

Zero-shot AgentInstruct performs near the level of few-shot prompting.
Comparison Methods: Self-Consistency

Zero-shot AgentInstruct exceeds the performance of self-consistency
Conclusion

• Zero-Shot AgentInstruct: Combine an autonomous agent generating instructions with CoT reasoning

• AgentInstruct outperforms zero-shot and zero-shot CoT on generation, classification, and reasoning tasks

• State-of-the-art on 20 of 29 datasets

• Code: https://github.com/wang-research-lab/agentinstruct