

Peer Review

Review of: "ScribeAgent: Towards Specialized Web Agents Using Production-Scale Workflow Data"

Ross Gore¹

1. Old Dominion University, United States

This paper introduces ScribeAgent, a specialized web agent developed by fine-tuning open-source language models on a large dataset of real-world web interactions. The authors present specialized fine-tuning in developing capable web agents, demonstrating improved performance compared to alternatives on various benchmarks, including Mind2Web and WebArena.

The strengths of this work lie in its comprehensive approach to developing a web agent that outperforms existing methods using proprietary models. Fine-tuning open-source models using LoRA is a pragmatic choice that balances performance gains with computational efficiency. The extensive ablation studies provide valuable insights into the impact of various design choices on the agent's performance.

However, the paper has several opportunities for improvement.

The use of a proprietary dataset for training raises significant reproducibility concerns. While the authors release code and a version trained on open data, the full results cannot be independently verified. This lack of transparency hinders scientific progress and collaboration. Even if the authors can create an argument to not release the dataset, there needs to be more information provided about the data collection process, potential demographic biases in the dataset, or how representative it is of real-world web tasks. Addressing these issues would improve the paper.

The evaluation methodology can also be improved. The comparison to baseline LLMs like GPT-4 is not fair, as these models were not specifically fine-tuned for web tasks. A more apt comparison would be to other specialized web agents.

A notable omission in the paper is the lack of statistical significance testing for the reported performance improvements. As it stands, it is unclear if the extent to which ScribeAgent outperforms alternatives is due to statistical noise or an actual signal of improved capabilities. Including statistical significance testing for the reported performance improvements would provide stronger evidence for the claimed improvements and enhance the scientific rigor of the evaluation. Applying the Wilcoxon signed-rank test could be used to compare the performance of ScribeAgent against the alternatives.

The paper could benefit from a more thorough discussion of potential limitations and ethical considerations. The agent's performance on completely new websites or web structures not represented in the training data is not extensively explored. The ability to handle dynamic content and single-page applications, which are common in modern web development, is not addressed. Moreover, the potential security risks and privacy implications of deploying autonomous web agents at scale are not sufficiently discussed.

Declarations

Potential competing interests: No potential competing interests to declare.