Dear Reviewers,

In this submission of paper, we have made the following changes:

- Extended the paper from a short paper to a long paper
- Extended the data set sizes
- Performed another round of data quality check to address ambiguity in the first submission (Feb 2025 ACL ARR) and rerun the experiments with VLM and Agents on this new set of data.
- Added a whole new experiment with GUI Agents
- Added an analysis of tools being used
- Added an qualitative analysis.

Thanks,
On behalf of all authors