AgentMerge: Enhancing Generalization in Fine-Tuned LLM Agents

Megh Thakkar[1,2], Léo Boisvert[1,2,3], Thibault Le Sellier de Chezelles[1,3], Alexandre Piché[1], Maxime Gasse[1,2,3], Alexandre Lacoste[1], Massimo Caccia[1]

[1] servicenow [2] Mila [3] Polytechnique Montréal Université d'ingénierie

Introduction

- Behavior cloning, where models learn from expert-generated data to replicate decision-making processes, has shown potential in enhancing accuracy
- However, fine-tuning still faces significant challenges, such as catastrophic forgetting and degradation of reasoning abilities learned during pretraining
- **AgentMerge** enhances generalization in fine-tuned LLM agents by merging agentic fine-tuning with instruction-tuning

Contributions

- Empirical evidence demonstrating the effectiveness of model merging to alleviate issues like catastrophic forgetting in LLM fine-tuning
- Insights into the disconnect between expert trajectory prediction and downstream task success, highlighting the need for more robust fine-tuning
- An open-source 140M token dataset of successful expert traces and a complete fine-tuning pipeline for further research and experimentation

Fine-tuning Pipeline and WorkArena

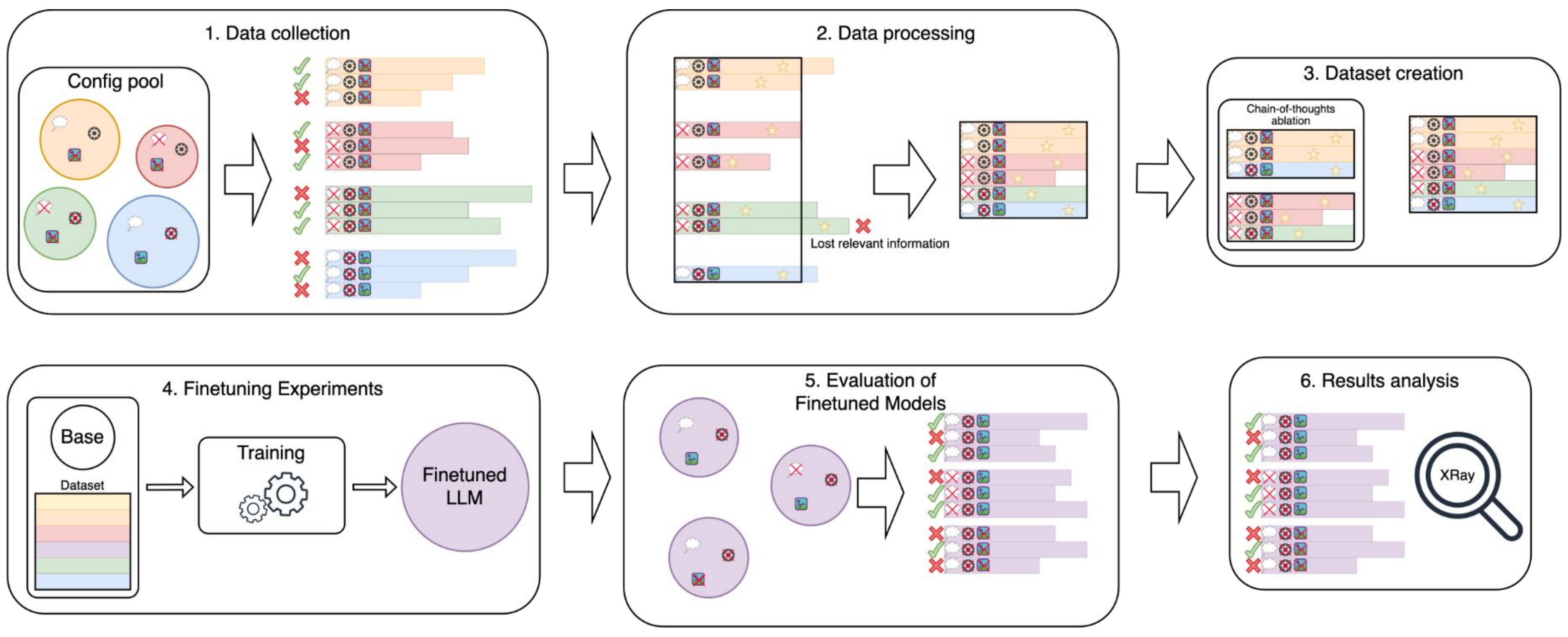
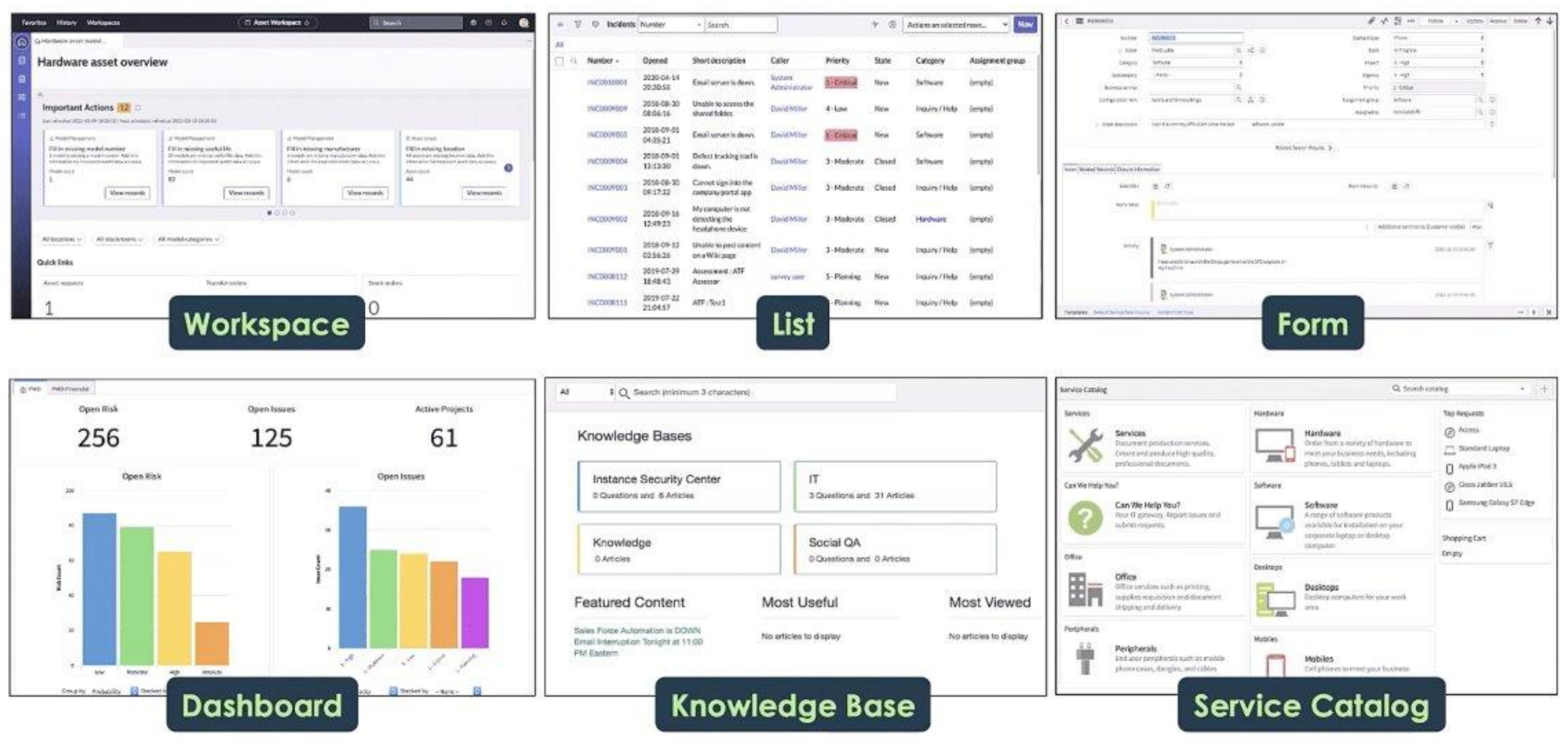


Figure 2. Our generic pipeline: 1) Trajectories are generated using different configurations (Chain-of-thoughts: 💬, use error logs: ⚙, use screenshot: 🖼) and different LLMs (highlighted by different colors). 2) Only the successful trajectories are kept. As each prompt is truncated to fit in our trained model's window, some key information (☆) might get lost in the process. Those samples are discarded. 3) The pipeline now has a pool of data, which can be used to build training sets with different properties. Here, we build an ablation dataset that separates data with and without chain-of-thoughts, and a dataset that merges both. 4) After selecting a dataset, we train our model starting from a base model to make a stronger finetuned LLM. 5) The latter is used along with different agent configurations to assess the finetuning quality. 6) Finally, we can leverage AgentLab's tools to manually analyze the traces produced by the model.



WorkArena is a collection of tasks which measure the ability of web agents to interact with basic UI components in the **ServiceNow** platform

Results with standard fine-tuning vs using AgentMerge

While standard fine-tuning seems to suffer from catastrophic forgetting while fine-tuning, **AgentMerge** shows some more resilience and overall leads to better performance on the downstream task

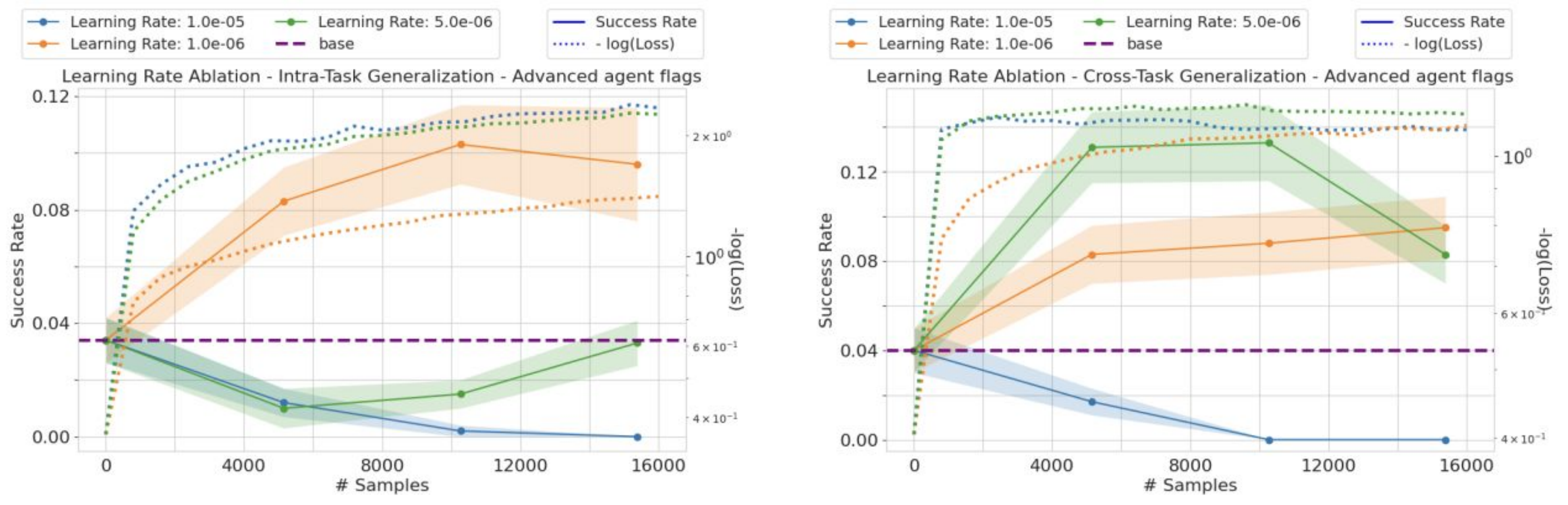


Figure 2: Success rate (left y-axis) and modified likelihood of expert trajectories (right y-axis) in the inter-task (left) and cross-task (right) generalization setup, throughout the fine-tuning phase. Interestingly, the model's improved ability to predict expert trajectories does not directly translate to better performance on downstream tasks.

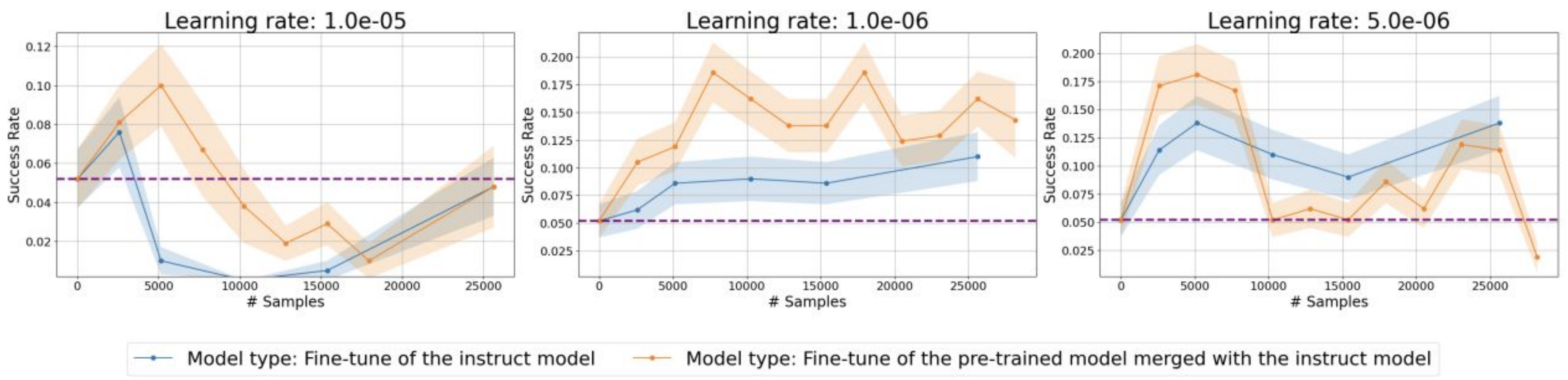


Figure 3: Success rate on WorkArena in the cross-task generalization setup throughout the fine-tuning phase for agentic fine-tuning of the instruction-tuned model vs AgentMerge. Merging consistently provides the fastest learning and achieves the highest peak performance. However, it eventually experiences some forgetting, likely due to the growing divergence between the agentic fine-tuning and instruction-tuning, which becomes increasingly difficult to merge effectively.

- **AgentMerge** enhances generalization in fine-tuned LLM agents by merging agentic fine-tuning with instruction-tuning.
- It interpolates agentic vectors from expert trajectories with instruction-tuned models, reducing catastrophic forgetting and improving task performance.
- Experiments demonstrate superior results over standard fine-tuning and model ensembling.

LR	Method	Test reward$_{\pm \text{std err}}$
5.0e-6	Instruct-fine-tuned	0.136 ± 0.03
	Model soup	0.156± 0.04
	AgentMerge	**0.178**± 0.03
1.0e-6	Instruct-fine-tuned	0.123± 0.03
	Model soup	0.156± 0.04
	AgentMerge	**0.171**± 0.03
1.0e-5	Instruct-fine-tuned	0.078± 0.02
	AgentMerge	**0.103**± 0.02

Table 1: Comparing fine-tuning the instruction-tuned model and the merged model with a model soups of the fine-tuned versions of the instruction-tuned models.

Future Work

- More sophisticated merging anchored for agentic fine-tuning
- Elaborate fine-tuning strategies involving counterfactual generation and preference optimization in sync with model merging
- Increase scale of training data with more in-the-wild traces

Cool Resources!

WorkArena

AgentLab