Trusted AI Symposium

Integrating Safety Testing into GenAI Development: Lessons from Amazon Nova and Red Hat – Chatterbox Collaboration

Amazon Nova

Christophe Dupuy[1], Stuart Battersby[2], Rahul Gupta[1], Danny Coleman[2]

[1]Amazon Nova Responsible AI
[2] Red Hat – Chatterbox Labs

1. Introduction

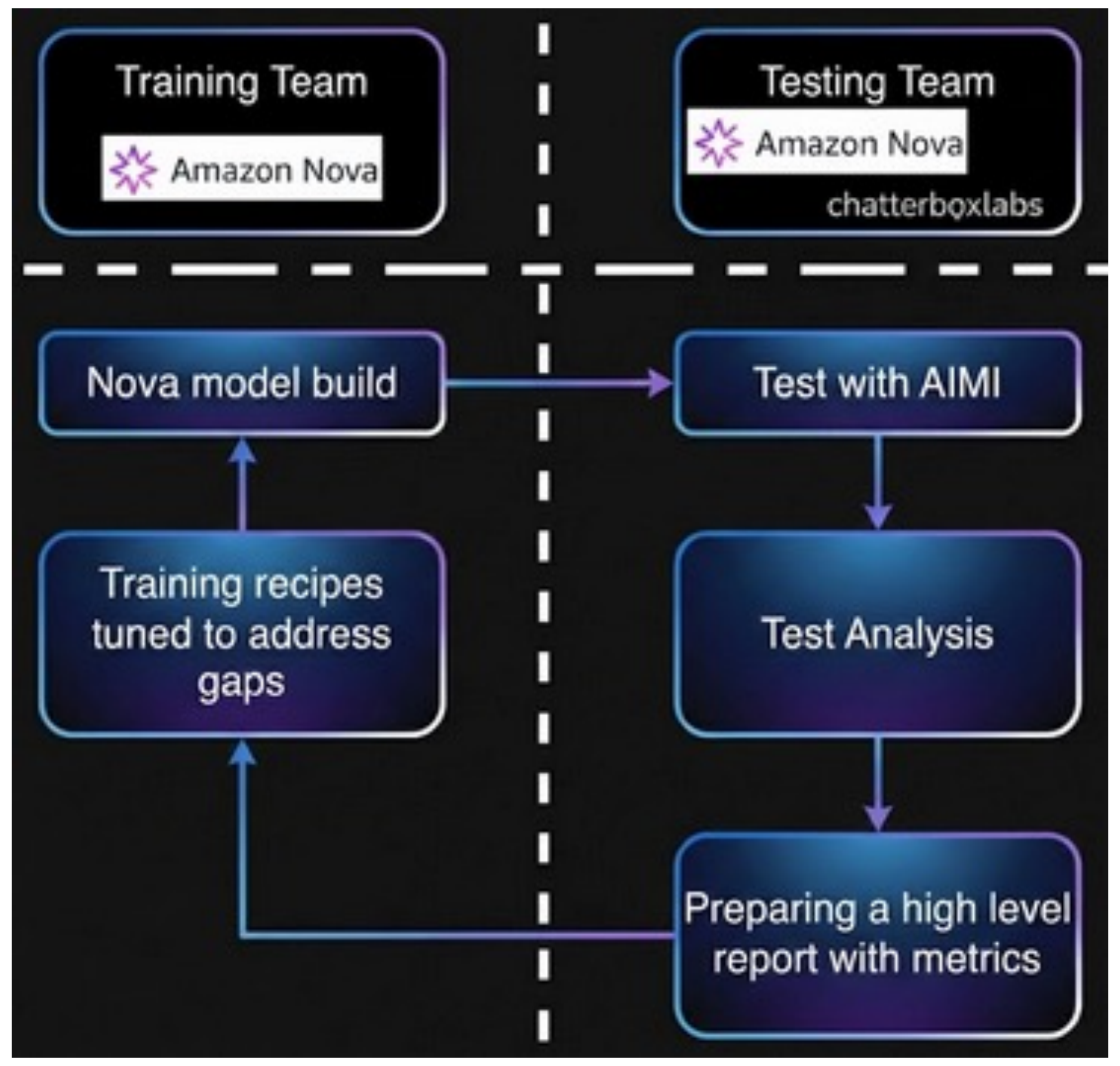
A typical collaboration model between testers and model builders may entail evaluations on final or near-final model checkpoints. In this poster, we present learnings from a collaboration between Amazon Nova Responsible AI and Chatterbox Labs to enhance the safety and security of Nova 2 models. We tested an approach where we frequently ran responsible AI evaluations during the model build process including early model checkpoints.

- Amazon Nova models: The collaboration extended to Amazon's Nova 2 family, which includes four specialized models: Lite, Pro, Sonic, and Omni. These models are designed to balance speed, cost, and intelligence across reasoning, multi-modal processing, and real-time conversational AI.

- Chatterbox AI Model Insights (AIMI) platform Labs provides a range of products to assess vulnerabilities in machine learning models. Their AIMI platform is specifically designed for automated security and safety evaluation of GenAI models.

2. Operational Setup

The figure below illustrates our overall operational setup.



- **Training/test team setup**. As presented above, we separated testing from training. The testing team consisted of members from the Amazon Nova team and Chatterbox team. For each model version, high level findings (such as overall performance of the model, description of jailbreak techniques) was made available to a separate training team. Throughout the process, we ensured that the test prompts were not made available to the Amazon teams (both the Nova training team and Nova members in the testing team).
- **Setup for speed**. We worked with Chatterbox where the AIMI software was deployed for testing within Amazon network. This allowed testing for several checkpoints quickly after internal availability.
- **Testing cadence**. The testing team ran evaluations with each version of the model. An evaluation sync between the Nova and Chatterbox members happened bi-weekly to assess the performance of the model.

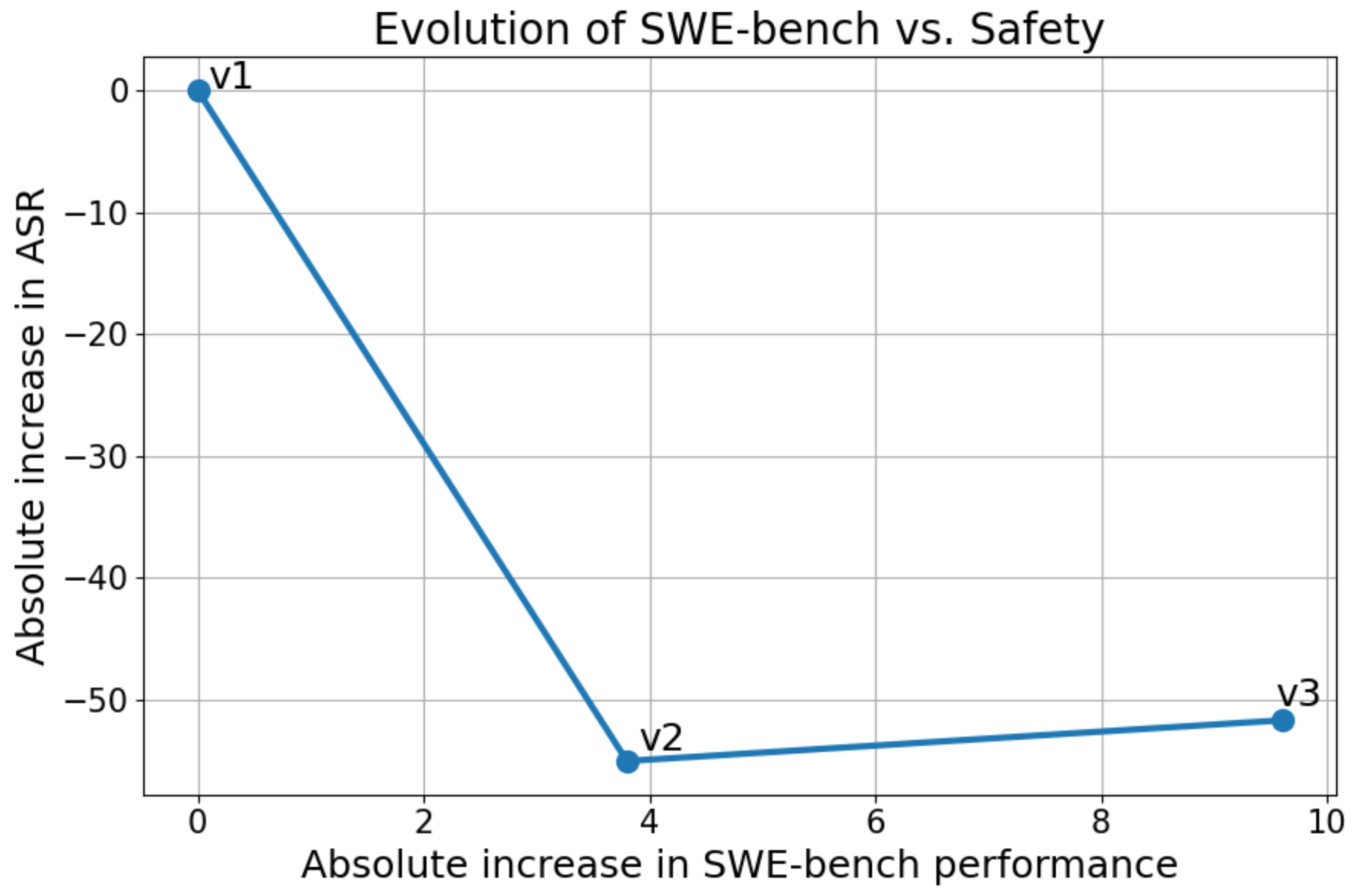
3. Learnings

Given that we ran the Chatterbox tests on multiple versions of the model, we observed certain dynamics on model performance. We summarize them below.

3.1 Performance for safety and other verticals may evolve in opposite directions

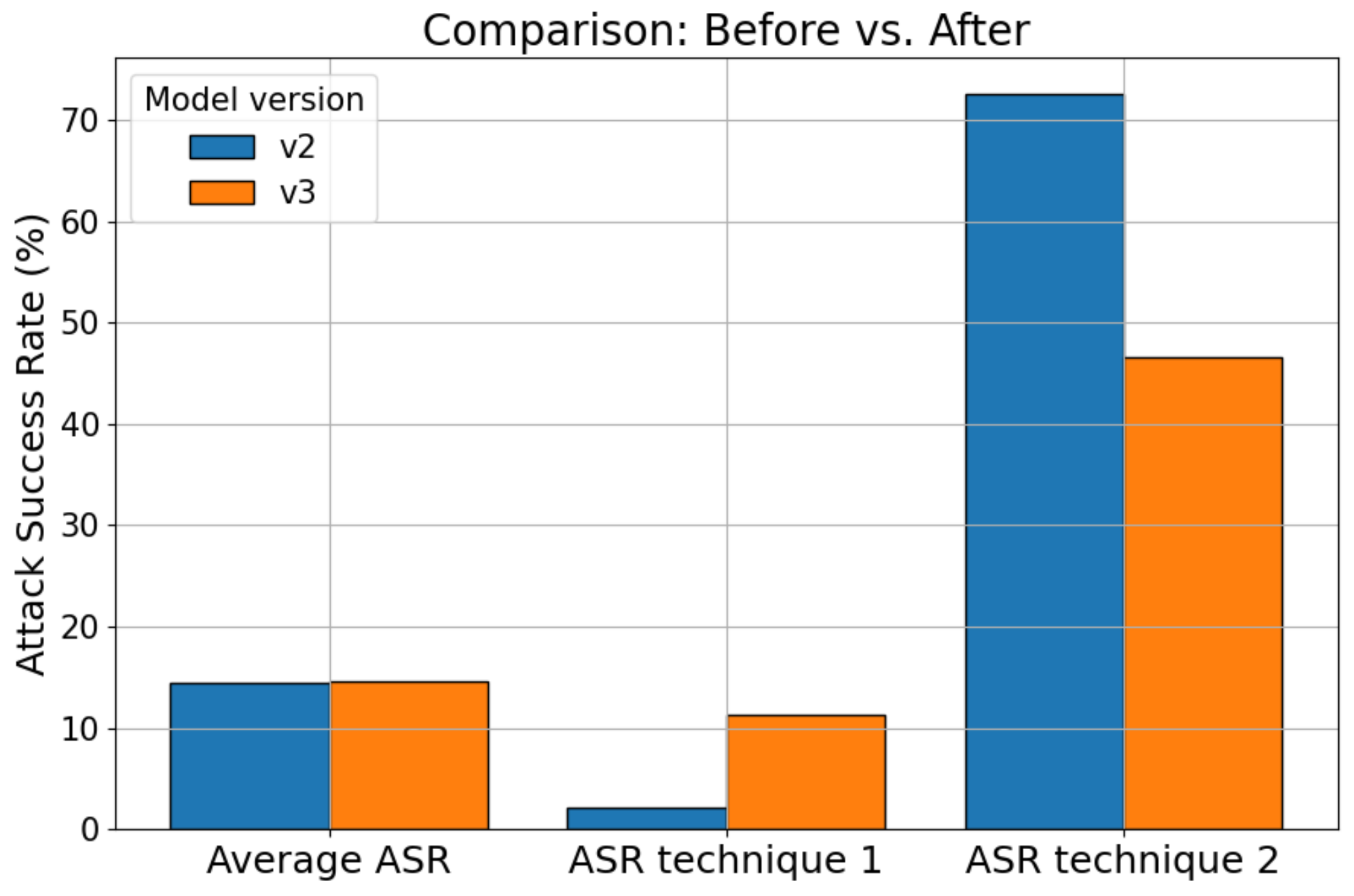
- Improving capability of a model in a certain domain often involves emphasis on curating training data or conducting special operations during RL (e.g. building specialized reward models and/or conducting RL runs for the specific domain).
- However, we observed that efforts to improve safety performance demonstrated a different dynamics. E.g., continuous addition of more data did not necessarily lead to a linear increase in defense.
- This was due to reasons such as adding calibrated boundary data to avoid over-deflection or balancing refusing jailbreaks with following instructions.

In the figure below, we present the relative performance on SWE-Bench as model training progressed against Attack Success Rate (ASR) on a specific Chatterbox evaluation. We observe that while we achieve better performance than earlier checkpoint, an intermediate checkpoint (v2) had slightly lower ASR.



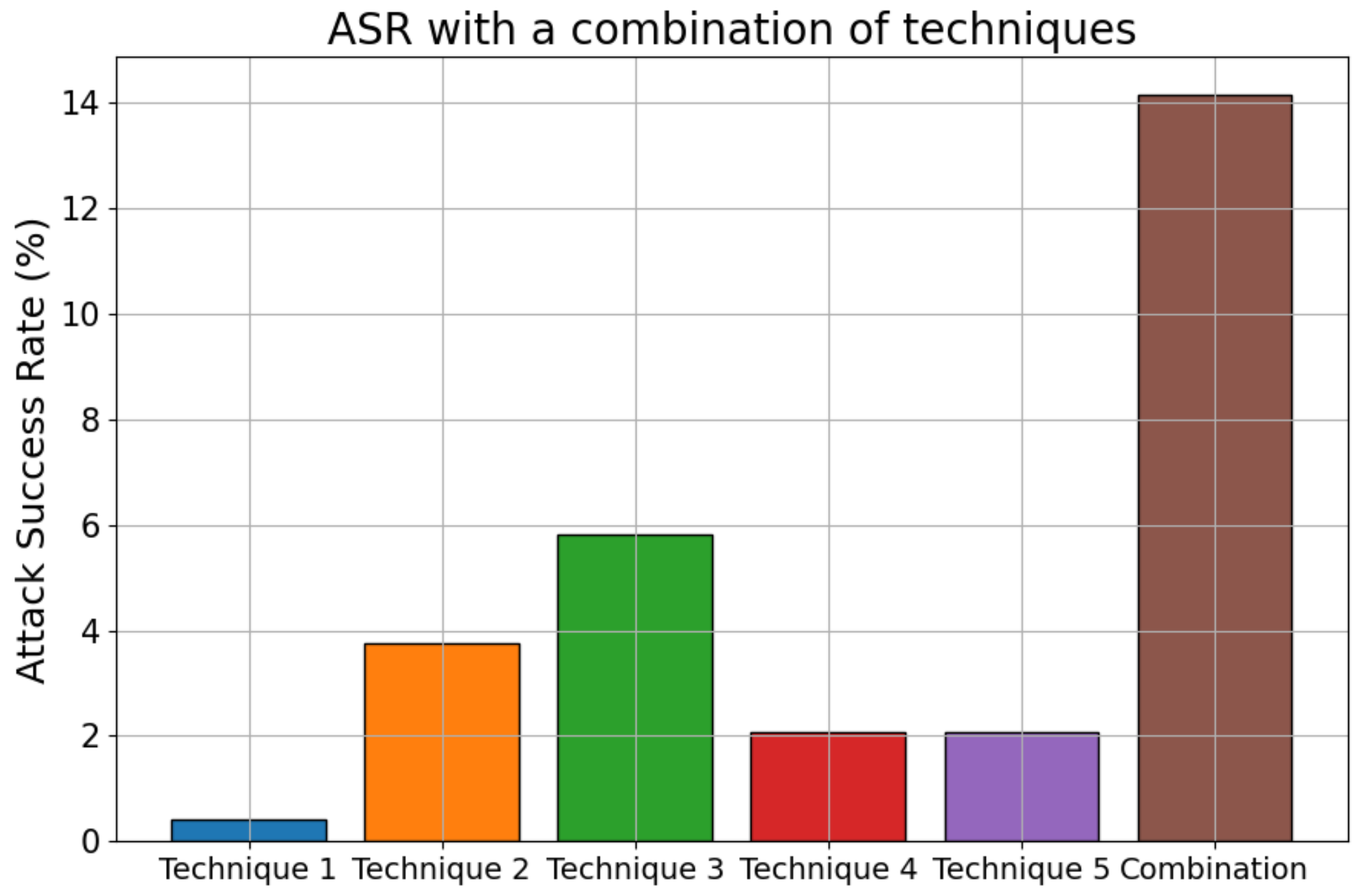
3.2 Average safety robustness does not imply across-the-board enhancement

Building on the trends highlighted earlier, our models demonstrated significant progress across diverse attack techniques compared to previous versions. While these advancements are encouraging, certain techniques presented opportunities for further refinement, underscoring the value of focused evaluation and tailored mitigation strategies in the final phases of model development. This targeted approach ensures comprehensive robustness and maximizes overall performance potential. As illustrated in the figure below, our two latest model versions present similar average ASR values but different ASR profiles across attack techniques.



3.3 Progressive attack escalation and attack combination helped stress test Nova's defenses

- AIMI implemented progressive escalation, wherein we can combine different attack methodologies together.
- This helped discover prompt combinations that can lead to higher ASR and combinatorically expand the number of attacks against our models. The figure below demonstrates performance of one of our checkpoints on a subset of individual techniques and then stacked together.
- Being able to evaluate across combination of techniques allowed testing of Nova models at scale and improving Nova's robustness against complex attacks.



3.4 Evaluations may need to evolve as the model capability/outputs evolve

- As model capabilities evolved, we observed that the judge that assessed model capabilities also needed updates.
- On certain category of model responses, we observed violation judge needed updates.
- We updated the judge based on model capabilities to maintain a higher than 95% accuracy on evaluations.

4. Conclusion

Safety robustness through internal software integration
We experimented with a paradigm where Chatterbox was involved as the model evolved and we informed training while isolating the test samples from training. We demonstrated that through a tighter software integration and collaboration setup, models can be evaluated for safety more frequently.

Safety improvements through anti-contamination framework
The separation between training and testing helped us achieve improvements in general strength of the Nova models. On separate internal test sets, we observed that the model performance improved. While not all of it can be attributed to actions taken due to observations on AIMI alone, general strategy of performance evaluation on a held-out set helped improve our model's performance.

*Corresponding author: dupuychr@amazon.com