Testing GenAI models alongside model builds: A case study of collaboration between Amazon Nova and Chatterbox

Christophe Dupuy ✳ **Stuart Battersby*** **Rahul Gupta** ✳ **Danny Coleman***

✳ **Amazon Nova Responsible AI** * **Chatterbox Labs**

Abstract

As the capabilities of Generative AI (GenAI) models advance, teams must assess security & safety risks before deployment to prevent potential harm. Teams developing GenAI models face the dual challenge of improving both performance and security & safety through specialized testing protocols. The rapid pace of AI development means teams often lack proper security & safety evaluation tools or prioritize performance metrics over security & safety concerns. This report presents findings from an ongoing collaboration between the Amazon Nova Responsible AI team and Chatterbox Labs security & safety specialists, demonstrating how this partnership has automated and enhanced Nova models' safety protocols. We discuss our operating model and present learning from a collaboration model that placed evaluation using Chatterbox's testing software closer in loop during the model build.

1 Introduction: Nova models and Chatterbox

In this work, we present the results of a collaboration between Chatterbox Labs and Amazon Nova Responsible AI team and how this collaboration impacted the development pipeline for Nova 2 models. Earlier, an assessment of Nova models via chatterbox demonstrated the strength of Nova models and in order to further guardrail our models, we collaborated with Chatterbox during the training of Nova 2 models. Amazon's Nova 2 family comprises four specialized models—Lite, Pro, Sonic, and Omni – designed to balance speed, cost, and intelligence across reasoning, multi-modal processing, and real-time conversational AI1. These models support up to 1 million token contexts, process text, images, video, and speech inputs, and include features like extended thinking controls and built-in tools for advanced enterprise workloads such as complex reasoning, speech-to-speech conversations, and unified multi-modal generation.

Chatterbox Labs offer a wide range of products to assess vulnerabilities of machine learning models. For our collaboration arond Nova, *AIMI for GenAI*[1] is a platform to conduct automated security & safety evaluation specifically on GenAI models. While sharing a model under development to an external vendor is tedious or even unfeasible – for instance, due to limited capacity or restricted access (or both) – AIMI software deploys directly onto the customer's infrastructure. The local deployment helps assessing early version(s) of GenAI models, leaving more time and opportunities to mitigate the identified risks through subsequent model builds.

This initial static set of adversarial prompts allows us to compare the security & safety performance of different target models.

[1] `https://chatterbox.co/aimi-for-gen-ai/`

2 Experimental setup

We instituted an evaluation loop with following salient steps: 1/ Post every major checkpoint development, we ran evaluations with the AIMI software, 2/ A testing team evaluated the strengths/weaknesses of the model in collaboration with the Chatterbox team. 3/ The evaluation results were presented to the technical leads on training end with high level findings. The training team took steps to address issues identified in testing based on high level instructions.

3 Learnings

Based on the collaboration between the two teams, we

T1. On safety, a monotonic improvement in model performance is not guaranteed

Historically we have seen improvements in performance of models as their capabilities scale. With larger models and better data, we see improvements in capabilities and corresponding benchmarks. However, for Responsible AI, we observe that performance has to be calibrated during training to ensure refusals on harmful queries without leading to over-refusals.

T2. Strength in one aspect of model usage may not translate to strength in other aspect

Is there a datapoint where a model is strong in one technique, but weaker on another.

T3. Chatterbox's Effectiveness of Progressive Attack Escalation was an ultimate stress The AIMI software implements a technique where a prompt is progressively modified till either the model is jailbroken to provide a harmful response or the software exhausts the jailbreak mutations. This provided a comprehensive way to test the model and guaranteed that an improvement in model performance translated to actual real world model robustness.

T4. Evaluations may need to evolve as the model capablity/outputs evolve Finally, during the course of our experimentation, we realized that the

Next, we describe our operational learnings.

O1. Insulating testing analysis from training is a practice to build more resilience Since, the objective of this collaboration was to improve our models during the course of the training, we had to ensure sufficient insulation to prevent over-fitting to the testing. We ensured this by not making the test prompts available to Amazon team. This enabled high level pass on of model sterngths/weakness. This enabled generalized strengths in the models.

O2. Scalable Evaluation Infrastructure is critical to integrated testing and training In order to

O3. Joint analysis syncs provides detailed insights Finally, the collaboration involved periodic syncs between scientists from the two teams. This enabled extraction of key insights which improved quality of both testing and model training.

4 Conclusion

The collaboration between Chatterbox Labs and Amazon Nova Responsible AI team continues to be crucial for secure & safe GenAI development for the following reasons: (i) The partnership combines diverse perspectives and expertise from both organizations, strengthening the overall development approach. (ii) Independent assessment from Chatterbox Labs' AIMI software provides unbiased evaluation and reduces the risk of overlooking potential issues. (iii) The automated and customized data integration enables direct flow into training workstreams, allowing early risk identification and mitigation.

As Chatterbox Labs continues to enhance the AIMI platform with new features, Amazon Nova will leverage these improvements to advance AI security & safety standards, notably around multi-turn conversations and agentic frameworks.