servicenow

ServiceNow Now Assist Guardian Model Card

VirtueGuard-Text-Lite

GPT-oss-120b

Model Name

VirtueGuard-Text-Lite

Model Provider

Virtue AI

Version

VirtueGuard-Text-Lite

License

Please refer to your agreement with ServiceNow for all license information.

Usage in Our System

VirtueGuard-Text-Lite is used in our system to power content moderation (prompt injection & offensiveness) in Now Assist Guardian along with other Gen AI models.

Key Capabilities

VirtueGuard-Text-Lite is a specialized transformer-based language model designed specifically for text guardrail. The model implements an advanced attention mechanism with both global and local (sliding window) attention patterns, optimized for detecting and analyzing potentially harmful or inappropriate content across various contexts.

VirtueGuard's architecture is specifically optimized for text moderation with emphasis on:

- Comprehensive Coverage: Hybrid attention mechanism captures both explicit harmful content and subtle - contextual violations
- Robustness: Multiple normalization layers ensure stable performance even with adversarial or edge-case inputs
- Efficiency: Balanced architecture enables real-time moderation at scale without compromising accuracy
- Adaptability: Modular design allows fine-tuning for specific platforms, communities, or content policies

Model Name
GPT-oss-120b

Model Provider
Microsoft Azure

Version
GPT-oss-120b

Source
OpenAI GPT-oss-120b Documentation

License
Please refer to your agreement with ServiceNow for all license information.

Usage in Our System
We are using gpt-oss-120b in conjunction with Virtue AI's VirtueGuard-Text-Lite for powering up content moderation (offensiveness & prompt injection identification) for Now Assist Guardian.

Key Capabilities
gpt-oss-120b is an open-weight reasoning models available under the Apache 2.0 license and gpt-oss usage policy.

Developed with feedback from the open-source community, these text-only models are compatible with Open AI's Responses API and are designed to be used within agentic workflows with strong instruction following, tool use like web search and Python code execution, and reasoning capabilities—including the ability to adjust the reasoning effort for tasks that don't require complex reasoning.

The models are customizable, provide full chain-of-thought (CoT), and support Structured Outputs.