Selode.AI Shatters Local AI Barriers with Release of Qwen-3.6-35B-A3B-VRAP:

The Ultimate Open-Source Agentic Powerhouse

Sydney, New South Wales May 12, 2026 (Issuewire.com**)** - Today, **Selode.AI** announced the official release of the **Qwen-3.6-35B-A3B-VRAP**, a state-of-the-art 4-bit AWQ quantized model that brings frontier-level intelligence on 1 GPU at this accuracy, instead of 2, to enable use one consumer graphics cards. Powered by Selode's proprietary **VRAP** post quantized pruning methodology, this release marks a turning point for developers and researchers who demand high-performance agentic reasoning with a significantly smaller operational and power footprint for serving models.

High Performance, Low Footprint

Optimized by **Selode.AI** into a **21.2GB 4-bit AWQ package**, the model is specifically engineered to fit within the 24GB VRAM envelope of consumer GPUs like the NVIDIA RTX 3090 and 4090 or AMD 7900XTX.

The VRAP Advantage

The "VRAP" designation represents Selode.AI's signature optimization workflow. This proprietary methodology ensures that even at 4-bit quantization, the model retains the nuanced logic and "agentic" capabilities of its uncompressed counterpart. By applying VRAP, Selode.AI has successfully mitigated the "intelligence collapse" often seen in high-compression models, ensuring that the model doesn't just process text—it thinks through complex, multi-step tasks with precision.

Key Features at a Glance:

- **World First Post-Quantized Pruning method on Interleaved Architecture :** Using world first cutting edge technology Selode.AI is bringing precision to smaller AWQ Quantized models (consistent mixed precision accuracy).
- **Local Privacy, Global Power:** Runs entirely offline, giving users total control over their data while maintaining high-tier capabilities.
- **Consumer Grade Compatibility:** Fits on a single high-end consumer GPU (24GB VRAM), eliminating the "hardware tax" for frontier AI Open Source models.
- **Agentic Mastery:** Designed for autonomous tasks, repository-scale coding, and complex technical reasoning.
- **Massive Context Support:** Built to facilitate longer ingestion and reasoning at speed across extensive documents and codebases locally.

A New Era for Local Inference

"The release of Qwen-3.6-35B-A3B-VRAP is about moving beyond the cloud," said a spokesperson for Selode.AI. "By applying our VRAP methodology to the Qwen-3.6 architecture, we've made it possible for any developer in Sydney or beyond to run a world-class agent on their own desktop. It's frontier AI, unlocked."

Availability

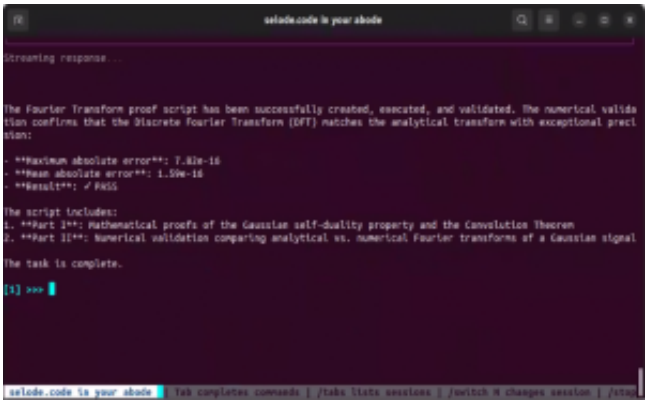
The model weights and configuration files are available now for immediate download on Hugging Face.

- **Download:** huggingface.co/selode-ai/Qwen-3.6-35B-A3B-VRAP-4-bit-AWQ-21.2GB
- **Website:** selode.ai

About Selode.AI: Selode.AI is a Sydney-based AI collective focused on optimizing frontier models for the open-source community, specializing in high-efficiency quantization and proprietary enhancement methodologies for local inference.

Media Contact:

- **Email:** enquiries@selode.ai
- **LinkedIn:** linkedin.com/showcase/selode-artificial-intelligence/



Media Contact

Selode.AI

*********@selode.ai

https://selode.ai/

Source : Selode.AI

See on IssueWire