



Running LLMs Locally: From Terabytes to Tokens

Dr Dean Brand (Stellenbosch University)

Friday, 28 August 2026 | 12h00-12h30 SAST

Attend in the NITheCS Seminar Room, Stellenbosch University, or online

ABSTRACT

Most large language models are accessed through hosted services such as ChatGPT and Claude, but capable open-weight models can instead run on hardware you control. This micro-school will explain what “open-weight” means, why it does not always mean “open-source”, and how downloadable models differ from subscription and API services. We will see why the largest checkpoints are terabyte-scale and how quantisation makes local inference more practical, with trade-offs in quality and hardware use. Through a live demonstration, I will show how to choose a model that fits your hardware, run a quantised build with Ollama, and connect it to a coding assistant such as OpenCode or Claude Code. We will close by comparing local and hosted models in terms of privacy, cost, control and capability. By the end, attendees will know what their hardware can realistically run and be ready to try a local LLM in their own work.

BIOGRAPHY

Dean is a postdoctoral researcher at Stellenbosch University (SU) in the Quantum Research Group under the supervision of Prof Francesco Petruccione. His research is on merging quantum and neuromorphic computing principles and paradigms for novel algorithms and architectures. His research has historically evolved through astrophysics, theoretical physics, and open quantum systems.

Dean received his Bachelor of Science and Honours degrees from the University of the Witwatersrand, his Master's from the University of KwaZulu-Natal, and his Doctorate from SU.



REGISTER: <https://bit.ly/4gvH88v>

