Luca Lowndes

Founding engineer. Large-scale inference researcher. Builds things that ship.
Luca.Lowndes.net

Email : Luca@Lowndes.net
Mobile : +61-456-264-606
LinkedIn · GitHub

Summary

Founding and lead engineer at **EvryStay**, where I built the product as the only engineer, helped take it through a **A$4.5M raise at a A$17M post-money valuation**, and now lead a team of five. First author and speaker at **ICPP 2026** on serving DeepSeek-R1 (671B) with SGLang, work that grew out of placing **3rd in the APAC HPC-AI Competition**. Credited SGLang contributor. Youngest ever **UniHack** winner. Deferred my double degree at Monash to do this properly.

Research & Speaking

- **Paper, first author**: **Luca Lowndes**, Nathan Culshaw. *Less is More: Optimising SGLang Distributed DeepSeek-R1 Inference on a Two-Node H200 Cluster.* Accepted at Benchmarking in the Data Center (BID '26), a workshop of the 55th International Conference on Parallel Processing (ICPP 2026). To appear in the ACM Digital Library. Shows that keeping collectives on NVLink and replicating within a node beats sharding across InfiniBand: 17,417 tok/s, 81% over the 16-GPU baseline.
- **Talk**: Presenting *Less is More* at BID '26, ICPP 2026, Singapore, September 2026, in a session alongside speakers from NVIDIA and the HPC-AI Advisory Council.
- **Panel**: Panellist on *Accelerating the Future: Overcoming Bottlenecks in GPU-Based AI and HPC*, ICPP 2026, with Jeff Adie (NVIDIA), Pengzhi Zhu (Director, HPC-AI Lab, HPC-AI Advisory Council) and Sandeep (IIT Palakkad). Moderated by Dr Samar Aseeri and Dr Jens Domke.

Experience

- **EvryStay** — Mornington, VIC
  Founding and Lead Engineer — *Dec. 2025 – Present*
  - **Built the company's product from zero**: Joined as the first and only engineer. Designed and built Eve, an AI concierge that handles guest conversations end to end for short-term rental hosts, property managers and hotels, along with the property management system integrations and AI service architecture underneath it.
  - **Fundraise**: The platform I built was the product investors backed: EvryStay raised **A$4.5M at a A$17M post-money valuation**, opening and closing the round inside a week.
  - **Grew the team**: Hired and now lead five full-time engineers in person. I own the architecture, the AI service boundary, code review and the final technical call, while still shipping daily: **300+ merged PRs** in my first seven months.
- **Firmus Technologies & HPC-AI Advisory Council** — Melbourne, VIC
  AI Engineer, SGLang Optimisation Partnership — *Nov. 2025 – Present*
  - **Upstream contributions**: Contributed optimisations to SGLang, the inference framework behind hundreds of thousands of GPUs worldwide, and was credited in the v0.5.5 release notes.
  - **Best practice documentation**: Wrote deployment guidance for running SGLang on H200 clusters as part of a joint effort between Firmus, the HPC-AI Advisory Council, Monash eResearch and Monash DeepNeuron.
- **Monash DeepNeuron & Monash eResearch** — Melbourne, VIC
  Lead AI Researcher, APAC HPC-AI Competition Team — *Mar. 2025 – Nov. 2025*
  - **The problem**: Maximise inference throughput for DeepSeek-R1 (671B parameters, 37B active) across two 8×H200 nodes, using infrastructure from NSCC Singapore, NCI Australia and Firmus.
  - **What I found**: Fewer GPUs arranged deliberately beat more GPUs arranged by default: a tuned single node (PP2 · TP4 · DP4 attention) hit **17,417 tok/s** at 2.05× the per-GPU efficiency of 16 GPUs. The boring stuff mattered most: CUDA version alignment alone was worth +35%, and 40+ hand-tuned NCCL configs never beat the defaults.
  - **Outcome**: **3rd of 49 university teams** in the Asia-Pacific region, presented on stage at SupercomputingAsia 2026 in Osaka. The work became the Firmus partnership above and the ICPP paper.

- **Monash DeepNeuron** Melbourne, VIC
  AI & HPC Technical Team, Outreach *Feb. 2025 – Present*
  - **Neural Cellular Automata**: Implemented NCA methods from recent papers in PyTorch to model self-organising systems such as embryonic gastrulation, plus on-device and cost-function optimisations using WebGL.
  - **Outreach**: Ran hands-on workshops in Victorian high schools on how deep learning models work and how to use generative AI well.
- **Freelance AI Engineer** Hybrid
  Scag Australia, International Mowers *Jan. 2025 – Nov. 2025*
  - **Customer-facing AI agents**: Built the Scag Mechanic and IM Advisor assistants for model comparison, troubleshooting and service recommendations. Structured proprietary product data, shipped embeddable chat widgets for web and mobile, and put brand persona and legal guardrails around both.

Awards

- **APAC HPC-AI Competition**, 3rd of 49 university teams, presented at SupercomputingAsia, Osaka *2025 – 2026*
- **UniHack 2025**, 1st of 850+ participants, youngest winner on record *2025*
- **Engineering Dean's Award for Academic Excellence**, Monash, top of the FIT1045 cohort *2025*
- **Dux of Mathematical Methods**, Elwood College *2024*

Education

- **Monash University** Melbourne, VIC
  BEng (Hons) Mechatronics / BCS Algorithms & Software. HD WAM. Deferred 2026 for EvryStay. *2025 – Present*
- **Academy of Interactive Entertainment** Melbourne, VIC
  Certificate III in Information Technology, completed alongside VCE *2023 – 2024*

Selected Projects

- **Dummy (EvryStay)**: Built so the non-technical side of the company can see their ideas before engineers spend time on them. A product person or customer-facing teammate describes a feature in a Slack thread and Dummy builds a working version of it: real branches across our three repos, a full-stack preview with its own seeded data at a permanent URL, desktop and mobile QA screenshots, and draft PRs. They play with it, react in the thread, and it iterates. By the time an engineer picks it up, the feedback is about something concrete rather than a description in a ticket, which has made the cycle from idea to shipped feature much shorter. An independent review agent checks the code, sensitive changes need an allowlisted human to approve in Slack, and nothing it does can merge, deploy, or touch production. Built on the open source QM harness with local whisper.cpp voice transcription, running four workers on one Mac.
- **Swarm Inference Lab**: Contributor to a self-configuring cluster runtime for heterogeneous LLM inference over a LAN: immutable content-addressed model snapshots, artifact leases, and an Apple Silicon multi-worker stage ring.
- **APAC-AI-HPC-2025**: The full open experiment record behind the ICPP paper: launch scripts, Slurm logs and workbooks. Every number is reproducible.
- **PiVoice**: Local speech-to-speech voice interface for terminal coding agents, running entirely on Apple Silicon.

Skills

- **Inference & HPC**: SGLang, CUDA, NCCL, tensor/pipeline/data parallelism, Slurm, Singularity, NVIDIA H100/H200 **ML**: PyTorch, TensorFlow **Product**: TypeScript, Next.js, Python, Postgres, LLM agents and tool use **Other**: C, C++, Git, CI/CD