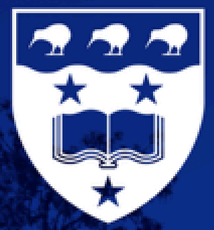


Waipapa Taumata Rau
University of Auckland

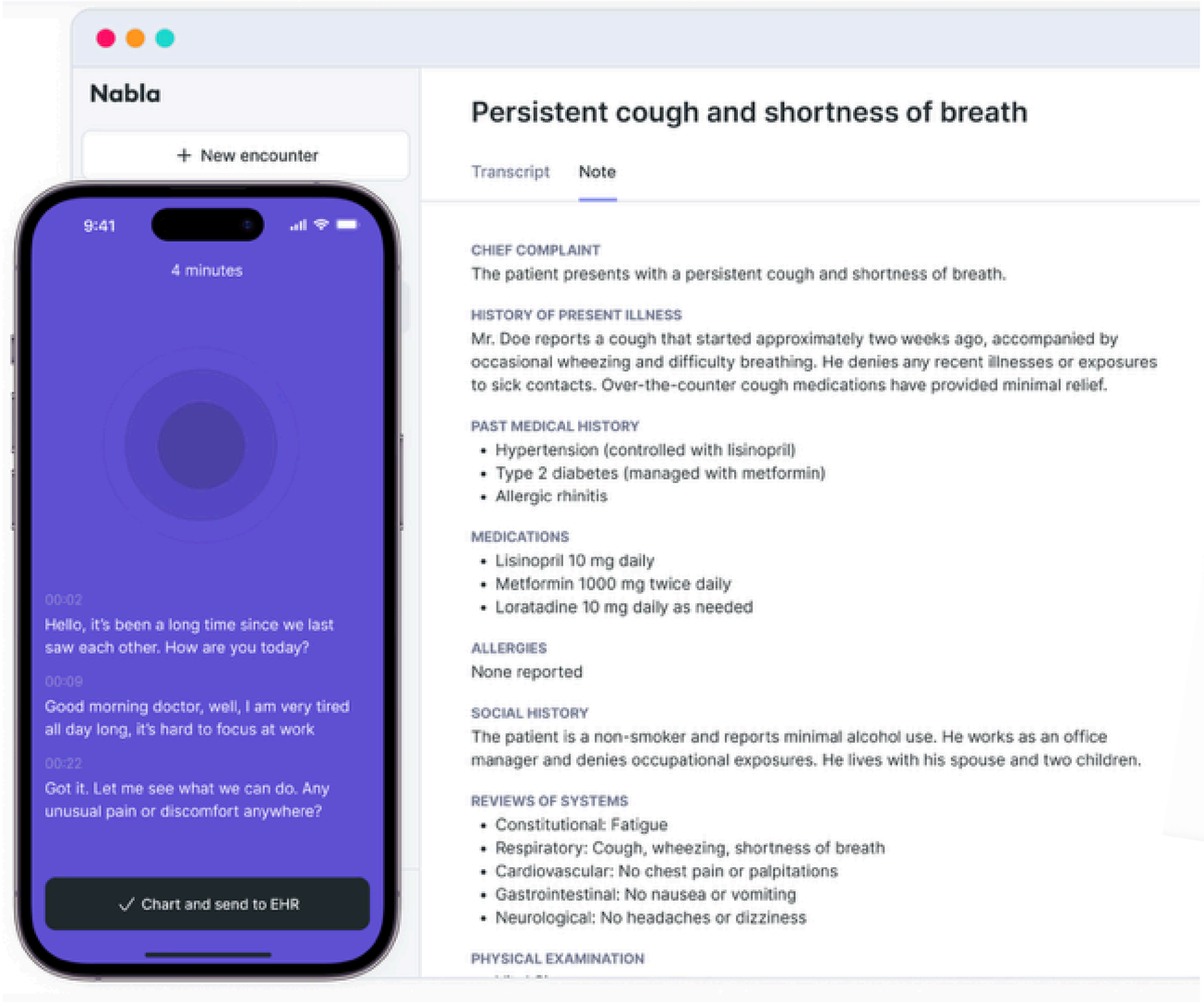
Open Source, On-Device AI for Improving Healthcare Delivery

Chris Paton BMBS BMedSci MBA DPhil FBCS

19 March 2026

Liggins Institute

What is an AI Scribe?



Accurate Health AI is expensive

GPT-5

Plus

$20 USD / month

More access to advanced intelligence

Your current plan

- Access to GPT-5 with advanced reasoning
- Expanded messaging and uploads

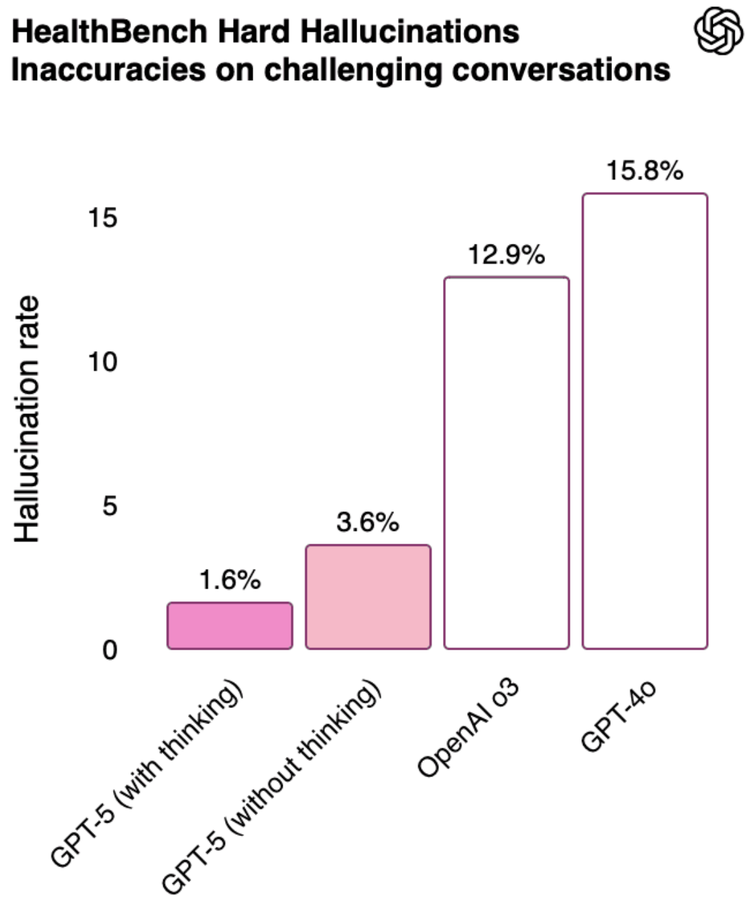
Pro

$200 USD / month

Full access to the best of ChatGPT

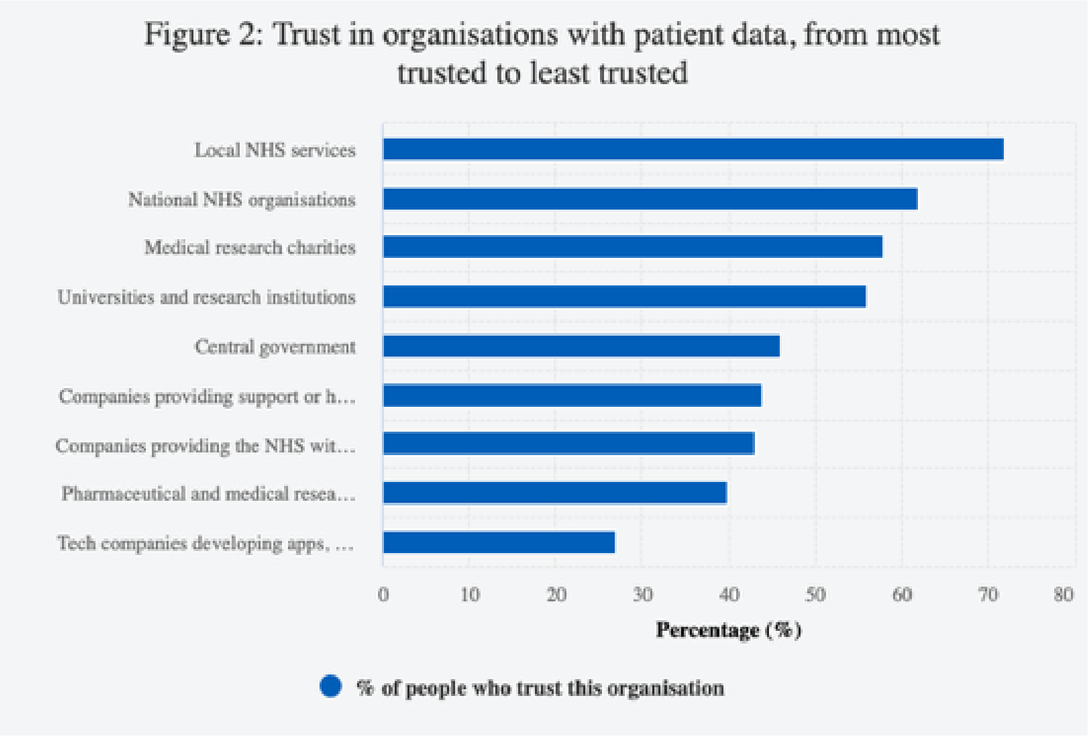
Get Pro

- Access to GPT-5 with pro reasoning
- Unlimited messages and uploads

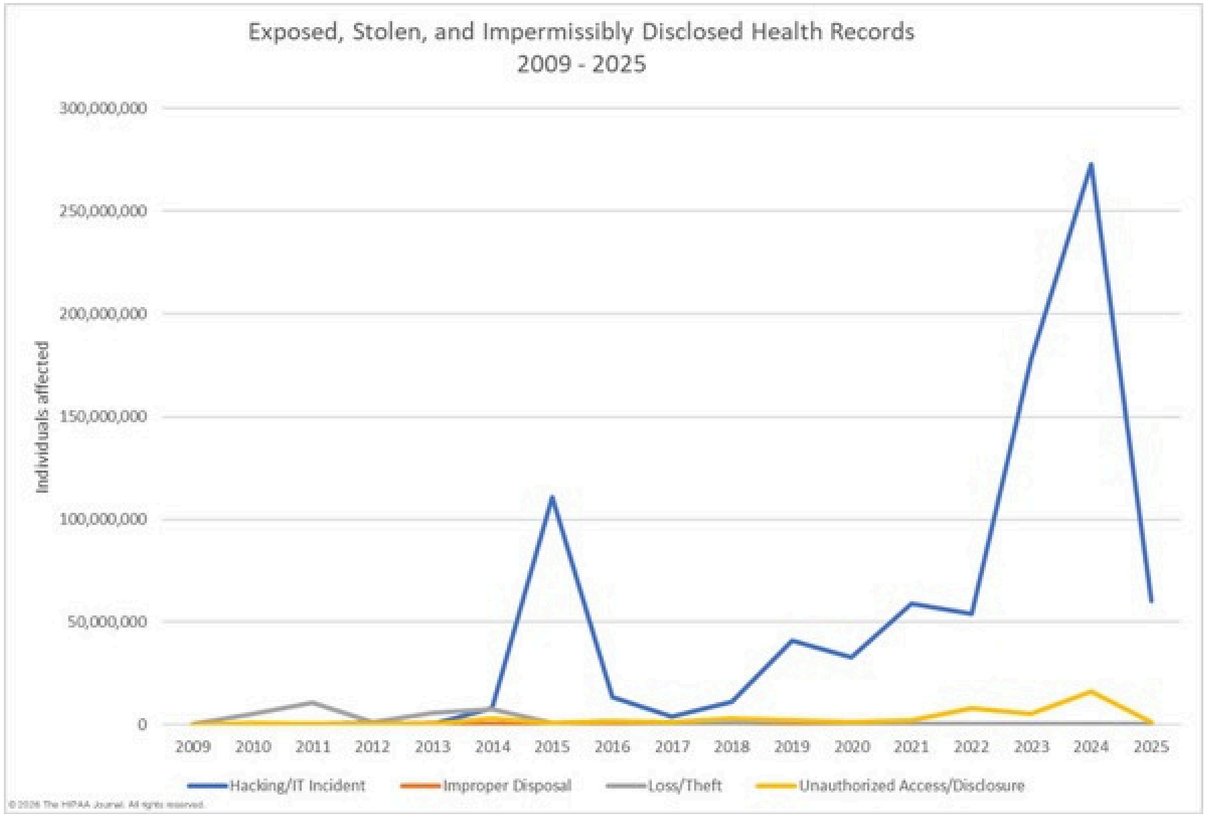


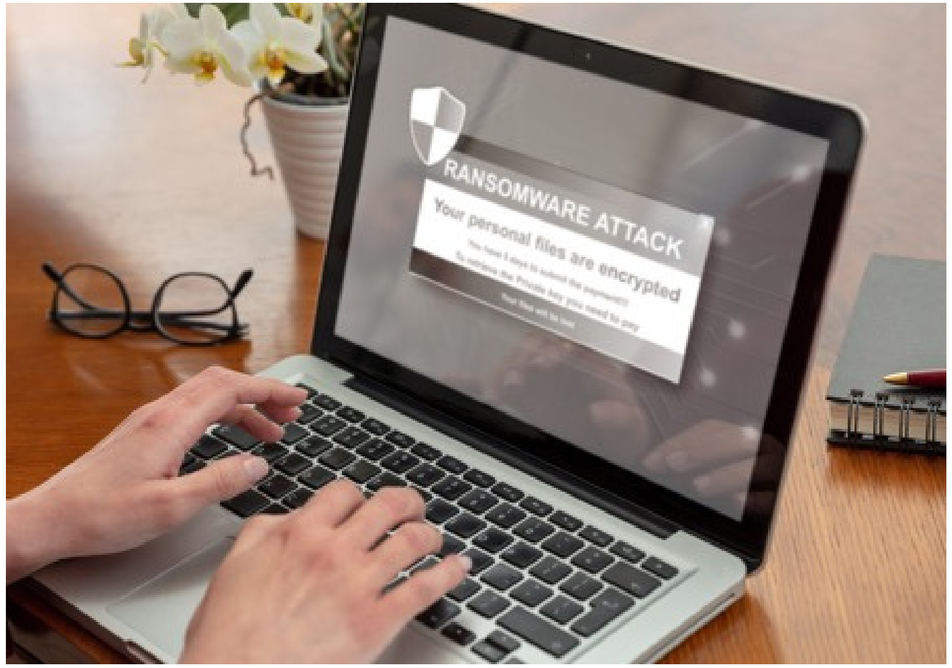
Patients are worried about their data

- Only 27% patients surveyed by the NHS said they trust AI companies with their data
- Because LLMs create the summary by processing transcripts, they are like a medical device processing data (e.g. ECG or ultrasound machine)
- **Regulators (FDA, MRHA, etc.) need to rapidly reassure clinicians and patients and certify high-quality AI scribes**



Cyber-attacks are increasing





MEDCHAT: Offline, On-Device AI for Health

Sera Cawanibuka
16 yrs · Female · Lautoka

Date of Birth
23 Feb 2010

Island
Viti Levu

Phone
+679 9165469

Patient ID
a8a01312-4eec-46d2-b…

Consult | Visit | Vitals

Overview | Vitals | Visits | Meds | Conditio

LATEST VITALS

BP Sys	BP Dia	HR
123 mmHg	84 mmHg	89 bpm
Temp	SpO2	RR
38.4 °C	88 %	21 /min

2:21

MEDCHAT
Clinical Decision Support System

Sign In

Username
Enter your username

Password
Enter your password

Sign In

Contact your system administrator if you need access.

localhost

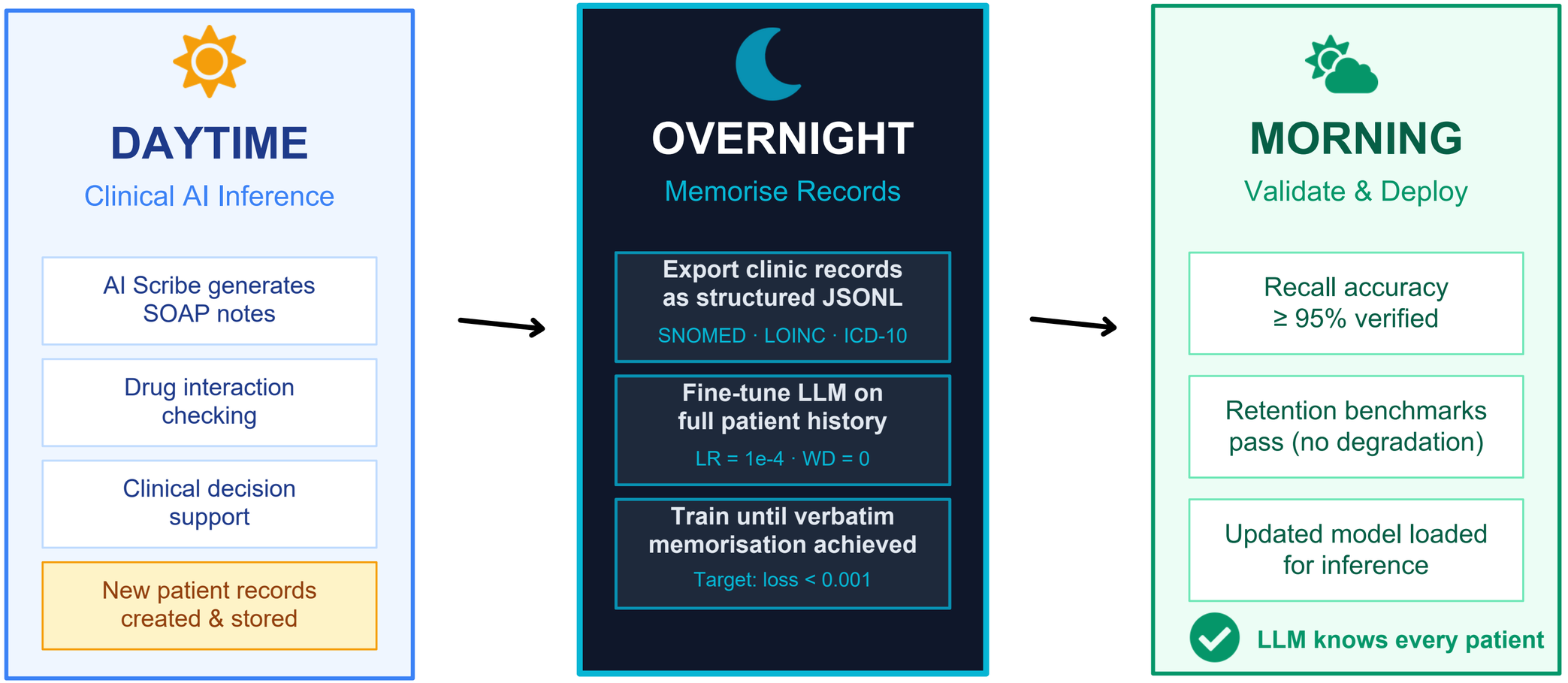
Clinical Scribe

Patient
Search patient by name...

Note Template
General Consultation

Ready to Record
Tap the microphone to start recording the consultation

On-Device AI-based Memorisation



System Architecture

- **Backend:** FastAPI (Python 3.12) + SQLite WAL mode
- **Frontend:** SvelteKit PWA, mobile-first, works offline
- **LLM:** Qwen3.5-35B via vLLM (15.8 tok/s)
- **Speech-to-text:** Whisper large-v3-turbo (3.2x real-time)
- **Infrastructure:** Docker Compose, Caddy reverse proxy, self-signed TLS

SVELTE

FastAPI

Qwen3.5

docker®

Key Principals

- All patient data stays On-Device
- No internet needed
- Access over LAN WiFi connection
- GB10 device lightweight and portable
- Desktop device could be powered by batteries/solar
- Clinical data trains LLM for inference
- Access via mobile phone, tablet or laptop

NVIDIA

Thanks

- Email me: chris.paton@auckland.ac.nz
- Linkedin: https://www.linkedin.com/in/drchrispaton