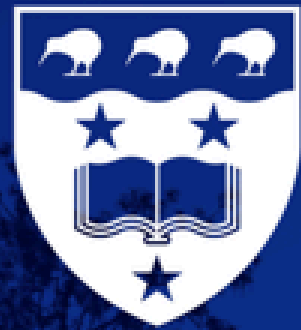


This is not an ADB material. The views expressed in this document are the views of the author/s and/or their organizations and do not necessarily reflect the views or policies of the Asian Development Bank, or its Board of Governors, or the governments they represent. ADB does not guarantee the accuracy and/or completeness of the material's contents, and accepts no responsibility for any direct or indirect consequence of their use or reliance, whether wholly or partially. Please feel free to contact the authors directly should you have queries.



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?

The screenshot displays the Nabla AI scribe interface. On the left, a smartphone shows a voice recording interface with a 4-minute timer and a 'Chart and send to EHR' button. The main desktop view shows a patient encounter titled 'Persistent cough and shortness of breath' with a 'Note' tab selected. The note contains the following sections:

- CHIEF COMPLAINT**  
The patient presents with a persistent cough and shortness of breath.
- HISTORY OF PRESENT ILLNESS**  
Mr. Doe reports a cough that started approximately two weeks ago, accompanied by occasional wheezing and difficulty breathing. He denies any recent illnesses or exposures to sick contacts. Over-the-counter cough medications have provided minimal relief.
- PAST MEDICAL HISTORY**
  - Hypertension (controlled with lisinopril)
  - Type 2 diabetes (managed with metformin)
  - Allergic rhinitis
- MEDICATIONS**
  - Lisinopril 10 mg daily
  - Metformin 1000 mg twice daily
  - Loratadine 10 mg daily as needed
- ALLERGIES**  
None reported
- SOCIAL HISTORY**  
The patient is a non-smoker and reports minimal alcohol use. He works as an office manager and denies occupational exposures. He lives with his spouse and two children.
- REVIEWS OF SYSTEMS**
  - Constitutional: Fatigue
  - Respiratory: Cough, wheezing, shortness of breath
  - Cardiovascular: No chest pain or palpitations
  - Gastrointestinal: No nausea or vomiting
  - Neurological: No headaches or dizziness
- PHYSICAL EXAMINATION**

The image shows a snippet of a Fierce Healthcare article. The headline reads: "Medical AI scribe startup Nabla rolling out tool to Kaiser Permanente docs in Northern California". The article is by Anastassia Gladkovskaya, dated Oct 5, 2023 12:00pm. The article is categorized under "voice-enabled technology" and "generative AI". The background of the snippet shows a doctor in a white coat typing on a laptop.

# Accurate Health AI is expensive



### Plus

**\$20** USD / month

More access to advanced intelligence

Your current plan

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

### Pro

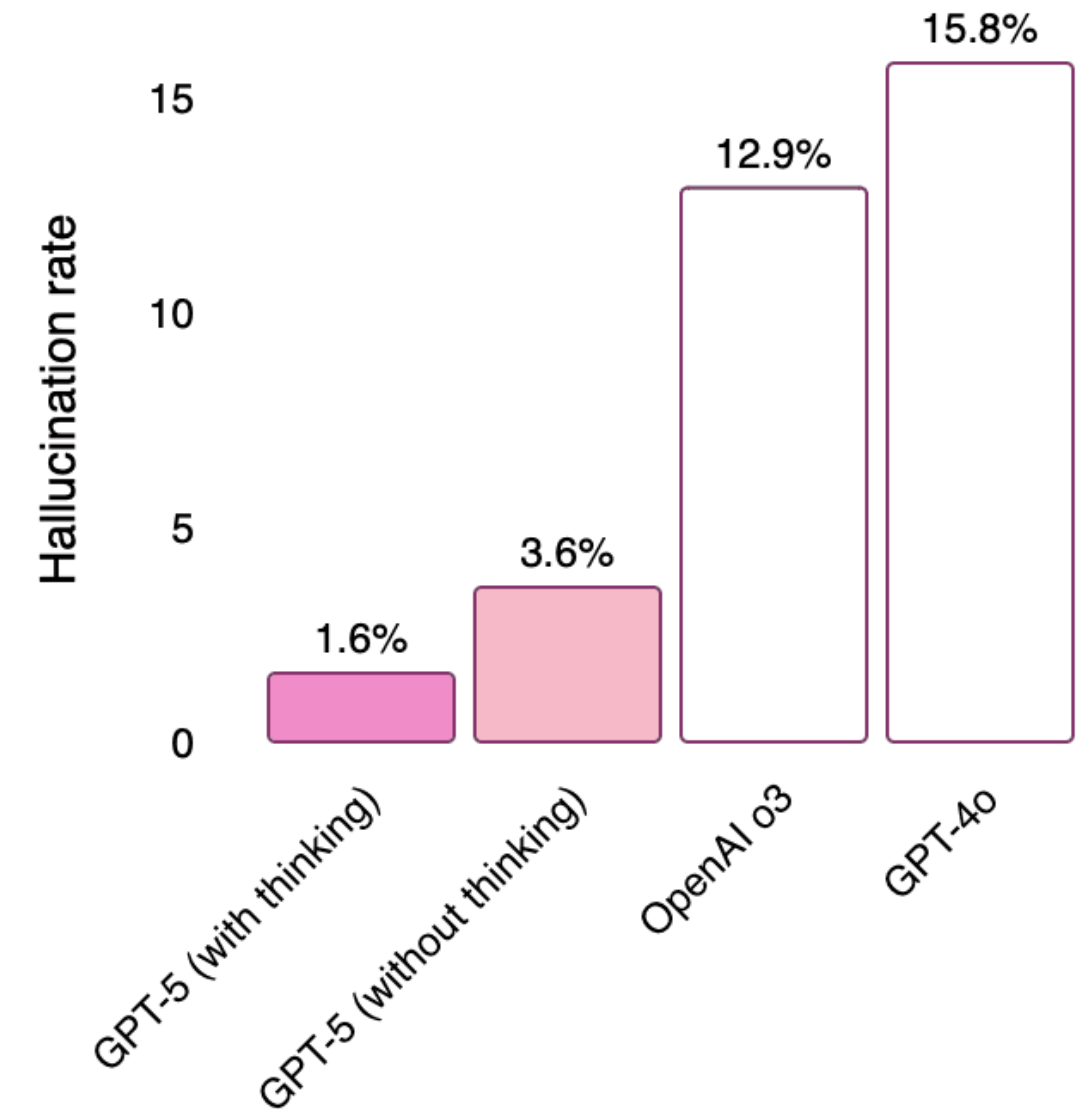
**\$200** USD / month

Full access to the best of ChatGPT

Get Pro

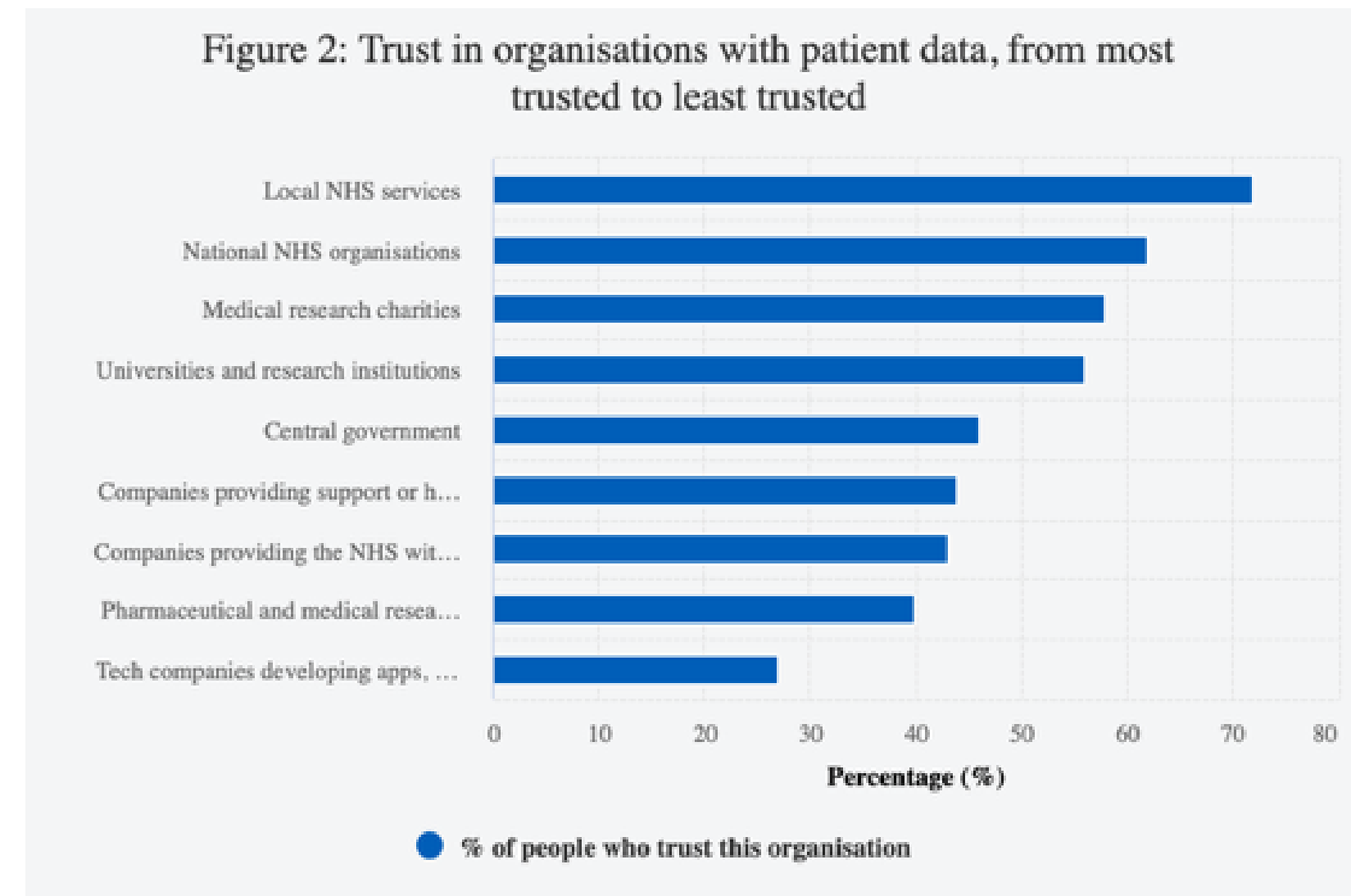
- ✦ Access to GPT-5 with pro reasoning
- 🔄 Unlimited messages and uploads

HealthBench Hard Hallucinations  
Inaccuracies on challenging conversations

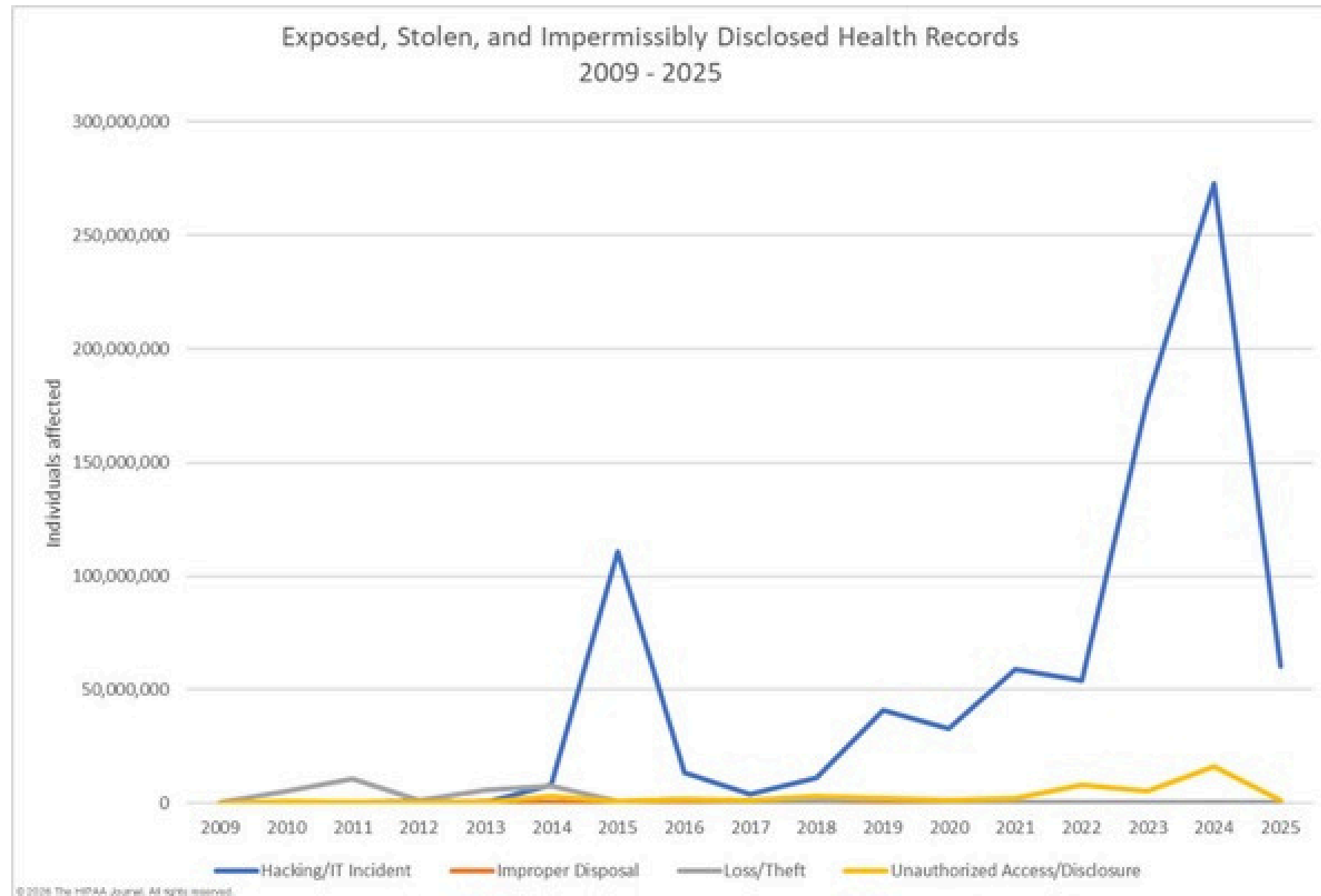


# Patients are worried about their data

- 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 Condition

**LATEST VITALS**

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

**SUMMARY**

2:21

**MEDCHAT**  
Clinical Decision Support System

**Sign In**

Username

Password

**Sign In**

Contact your system administrator if you need access.

localhost


**Clinical Scribe**

Patient

Note Template  
General Consultation

**Ready to Record**  
Tap the microphone to start recording the consultation

# On-Device AI-based Memorisation




**DAYTIME**  
Clinical AI Inference

- AI Scribe generates SOAP notes
- Drug interaction checking
- Clinical decision support
- New patient records created & stored



**OVERNIGHT**  
Memorise Records

- Export clinic records as structured JSONL  
SNOMED · LOINC · ICD-10
- Fine-tune LLM on full patient history  
LR = 1e-4 · WD = 0
- Train until verbatim memorisation achieved  
Target: loss < 0.001



**MORNING**  
Validate & Deploy

- Recall accuracy ≥ 95% verified
- Retention benchmarks pass (no degradation)
- Updated model loaded for inference
- LLM knows every patient

# 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



# 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](mailto:chris.paton@auckland.ac.nz)
- LinkedIn: <https://www.linkedin.com/in/drchrisspaton>