I wanted to drop an update (< 4-minute read time) for everyone who's been following us or working with us on the Tacit Intelligence journey to date.

For those who need a refresher, we're building specialised language models for health, insurance & the public sector, leading to our release of a self-serve platform for SLMs in November.

I've split this update out into:

- **Customer Examples**
  > Examples of the type of POCs we're doing
- **Distribution**
  > How we're getting customers
  > What's making them engage with us
- **Sales**
  > Time to POC from first contact
  > Common objections
- **Technology, Adoption, Research & Governance**
  > AI maturity of orgs we're working with
  > New ways of using LLMs
  > Platform build

Customer Examples

We're selling small language models that are trained on the human judgement bottlenecks:

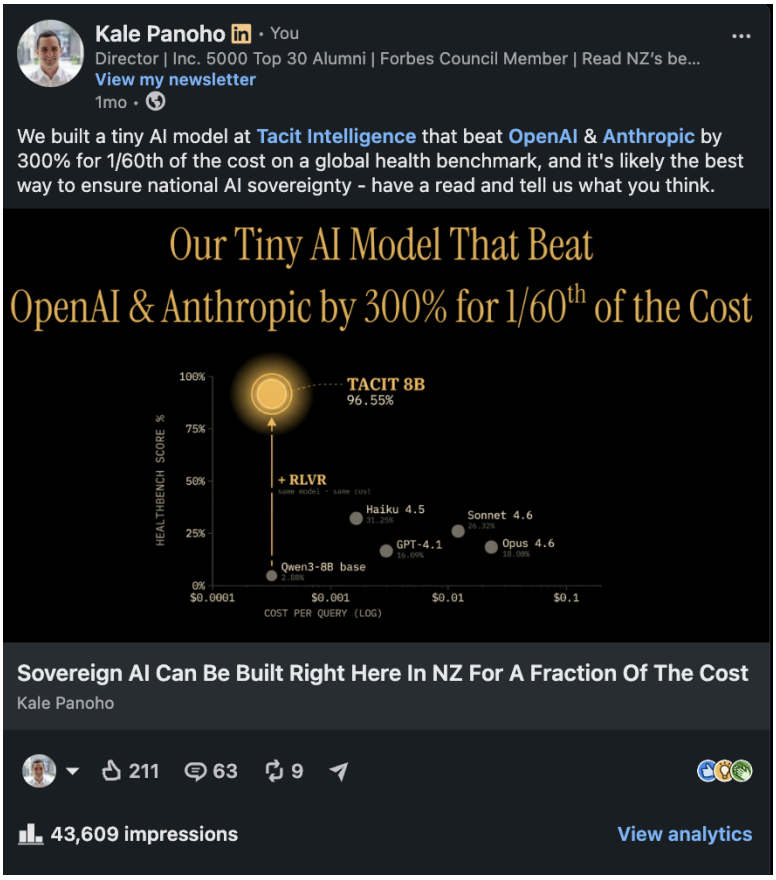
- Clinicians who don't have the bandwidth to review all clinical surgeries for approvals of insurance payments
- Capturing the judgement of the top 1% of any organisation's sellers and using that to train new staff and deploy on first calls with clients

Or training SLM's on high-volume tasks (10,000+ queries a day) that require some level of human judgement:

- An insurance customer asked how they could monitor their team's advice in a regulated industry by reviewing all of the transcripts and judging how far the advice given was from the law

Distribution

Our go-to-market strategy has been content via LinkedIn. Here's one of our top-performing posts:



Impressions: 43,609
In-network impressions: 12%
Out-of-network impressions: 88%
Members reached: 26,165
Article views: 1,224
Profile viewers from post: 166
Followers gained from post: 68
Total social engagements: 356
Reactions: 211
Comments: 63
Reposts: 9
Saves: 58
Sends on LinkedIn: 15
Link engagements: 4

We run this from one of the founders' profiles, then target our prospective buyers (CTOs, CIOs, CAIOs & board members) using paid targeting via LinkedIn ads ($212 worth of spend).

This article generated 38 inbound DM's & 4 website enquiries. The key theme for engagement is that people want to host open-weight models without leaking IP back to the labs. The uptick sprang from Fable 5 being cut off to users by the US government (example from a customer conversation below).

Wed, Jul 15, 1:12 PM

That works well Kale -

Might be good to meet somewhere where we can throw up slides and visuals to facilitate the conversation.

My goal - understand (or begin to) the complexities and opportunities provided by a native AI solution compared to the risks, costs and viability of a commercial provider - currently we use Claude but their CEO may become dazzled by their own brilliance at any stage much like every other tech bro seems to. Also, we are targeting UK, European, Middle Eastern and Asian customers and there is no appetite for further involvement with anything to do with the USA.

We need to build our own dev team capabilities but that does not mean we don't have space for a specialist contractor organisation working with us on the topic above.

Other things that are starting the sales conversation:

- Cost of tokens is getting too high even with using old versions of the frontier models

- Small language models (SLM) allow for constant monitoring with applied intelligence, e.g. an insurance company that needs to monitor 10,000+ conversations per day for regulated financial advice can now apply an SLM with targeted intelligence that acts as their auditors to these conversations for cents on the dollar, enabling much richer data

- PII (Personally Identifiable Information) & IP-specific workflows that can't be sent offshore due to legal regulations or fear of having their products stolen

- Capturing the IP of high-value individuals who are retiring or leaving the workforce

- RAM costs are skyrocketing, so people are looking for ways to make their hardware work harder by using lower-cost AI

- Local government (NZ & Aussie) varies greatly, with some having dedicated AI engineers and others having data teams dabbling in this space. The core difference from the private sector is that

Sales

We've spoken with 58 prospects from startups like Heidi Health to some of the world's biggest companies, Dai-ichi. We've got multiple POCS and RFPs underway. Here's the highlight reel from the sales pipe:

- Average org size we're working with is 10,222, but this is heavily weighted by three companies with 60,000+ staff - removing these brings the average size of the org down 2223
- Average time from first call to live POC is 3.8 months
- POC budgets range from $10,000 to $250,000
- Three government organisations we're working with (4000+) staff have AI panels that review POC's before being approved
- PII + PHI data is the most difficult to work with; we've physically had to turn up at an organisation to collect a USB stick

The sticking points for these POCs have been:

- We can't allocate the people time for you to work with them to train the model. Every org's superstars are in ultra-high demand
- PII approvals: every time PII is run through an outside vendor, there are NDA's and then vetting of the work we're doing that is external to the buyer who said yes

- The buying process in a number of these orgs can go through as many as 5 levels (Board / C-suite to frontline) of hierarchy

Technology, Adoption, Research & Governance

Org maturity varies wildly among the 58 prospects we've spoken with:

- 3% have CAIOs (chief AI officers)
- 55% have CTO's
- 42% have a combination of a COO/CIO/Head of data leading their AI projects

In terms of deployments:

- 100% are using some form of paid AI tool
- 55% are using Copilot as their primary tool for the majority of their staff
- 60% have some form of post-training workflows happening, mostly RAG and SFT

In our workflows with customers, we've been using a decomposition approach: taking a prompt and transforming it into atomic facts that flow through a decision tree. We're seeing significant jumps in model accuracy using this approach:

Approach + Model	Verdict accuracy	Wrong-block rate	Missed-block rate
Decomposed - Tacit 0.8B	98.6%	0.0%	1.9%
Decomposed - Sonnet 4.5	93.8%	13.9%	3.1%
Monolithic - Sonnet 4.5	70.2%	55.2%	19.5%

Full breakdown of this approach here.

We're building a prototype of our self-serve platform now and will have something live in November. We've got training runs down to 4 hours and continuing to reduce these each week.

If you're interested in yarning with us about our next round, have models that need training, want to dive deeper with our technical team or just want to yarn, you can find a time here.