Privacy on purpose in AI governance

Balancing innovation and protection

Dr Marcin Betkier
marcin@betkier.com
Privacy Week 2025
15 May 2025

betkier.com

AGENDA

- Where we are with the AI implementation in New Zealand
- Why we are looking for implementing AI
- What practical options we have to deploy Generative AI systems
- Integration of LLM into AI systems
- Privacy concerns
- How to purposefully tackle privacy concerns
- Conclusions

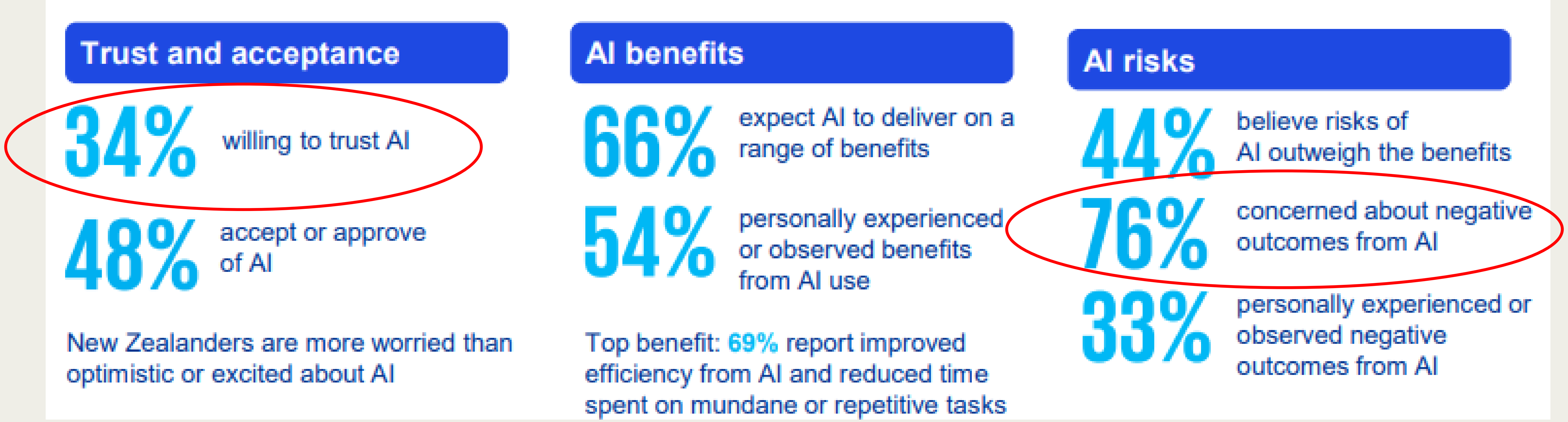
WHAT AI - SCOPE

The presentation:

- Takes the perspective of an organisation that thinks about deploying AI
- Focuses mainly on:
  - Generative AI and Large Language Models
  - the safe deployment of the existing models - adapting them to our needs
  - privacy risks and how to purposefully tackle them
- Excludes Agentic AI (autonomous systems built on top of the existing models)

“An AI system is a machine-based system that, for explicit or implicit objectives, **infers**, from the input it receives, **how to generate outputs** such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.”
(the OECD definition)

WHERE WE ARE WITH AI IMPLEMENTATION



Trust and acceptance

34% willing to trust AI

48% accept or approve of AI

New Zealanders are more worried than optimistic or excited about AI

AI benefits

66% expect AI to deliver on a range of benefits

54% personally experienced or observed benefits from AI use

Top benefit: 69% report improved efficiency from AI and reduced time spent on mundane or repetitive tasks

AI risks

44% believe risks of AI outweigh the benefits

76% concerned about negative outcomes from AI

33% personally experienced or observed negative outcomes from AI

@KPMG 2025 *Trust, attitudes and use of artificial intelligence: a global study*

- We tend to be less trusting and more concerned about AI than much of the world
- Our regular use of AI, knowledge levels, and training are among the lowest globally
- Introducing AI is a change management challenge - we need to bring people along

betkier.com

WHY AI?

Promises of AI:

- Industry-specific - e.g., coding, drug discovery, fraud detection, contract review
- Efficiency and productivity across the business
  - automation of simple, mundane tasks allowing to focus on more important things, augmenting human capacities – e.g., data collection and analysis of unstructured data.
- Improving business processes (automation, simplification, better scalability)
- Improving customer experience (e.g., personalisation, chatbots)
- New business models (?)

Starting point:

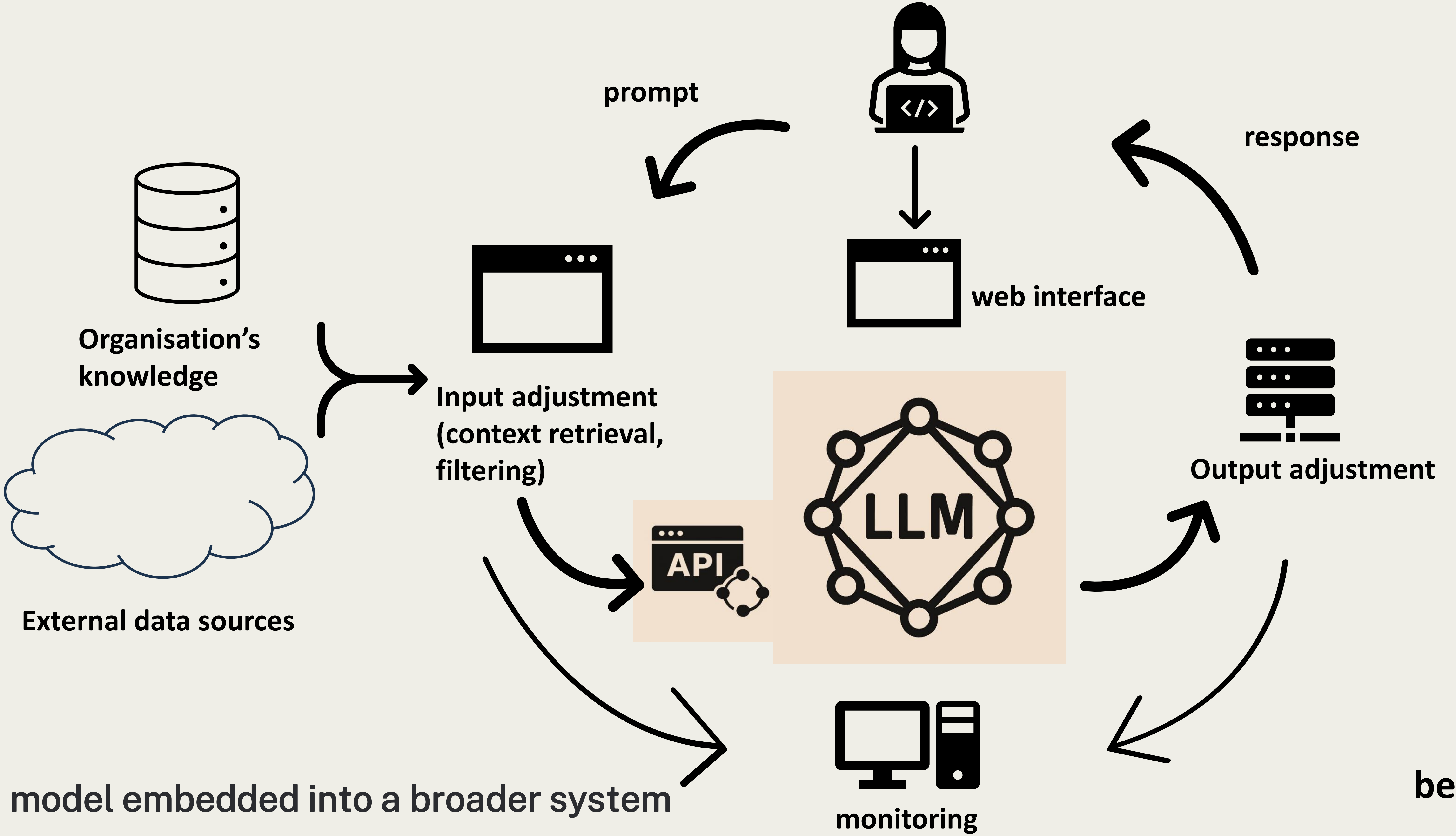
- What is the business goal?
- What are we trying to improve?

THE ORGANISATION’S DILLEMMA

1. **Forbid using Generative AI**
   - Open web AI services are available to anyone
   - Shadow AI – survey shows that 50-80% of users are using their own AI at work
2. **Partially allow using Generative AI** for some goals (like proofreading, ideation), but forbid using it for others
3. **Enable internal LLM tools -** owning and controlling the risks

It is important for organisations so they can monitor and control the use of AI

GENERATIVE AI SYSTEM



LLM model embedded into a broader system

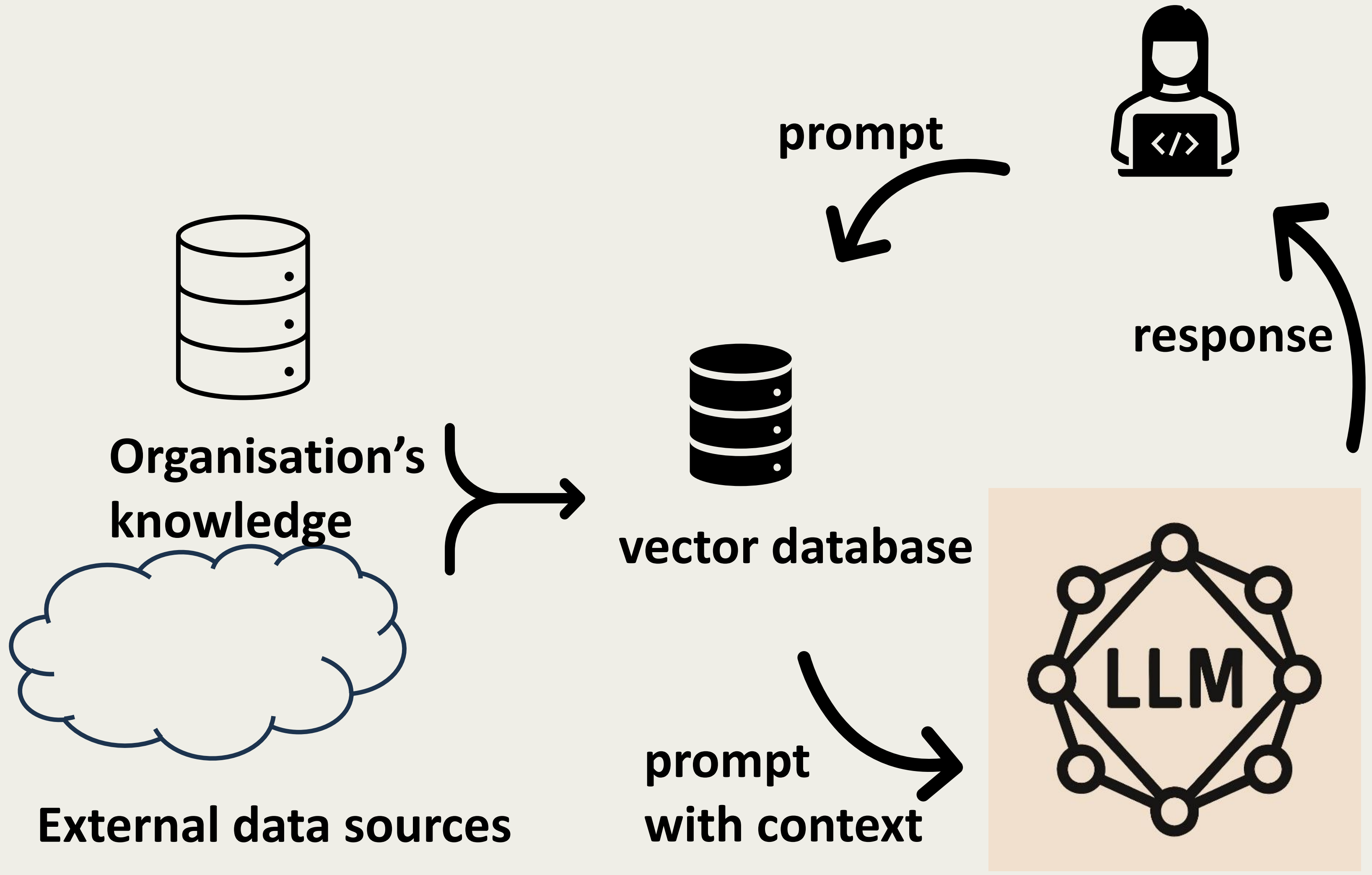
betkier.com

PRACTICAL MODES OF AI USE

1. **Software-as-a-Service** - LLM deployed by the provider
   - Web interface, AI embedded in an application
   - Examples: ChatGPT & OpenAI API, Anthropic Claude & Claude API Access, MS Copilot, Google Gemini
   - Option of API access
2. **Model-as-a-Service / AI-as-a-Service** pretrained model hosted in the cloud
   - Examples - MS Azure AI Service, Google Cloud Vertex AI, Amazon Q / Bedrock / SageMaker
   - Cloud-based “model hubs”: Hugging Face / Groq / OpenRouter / Together.ai
   - Includes open-source models like LLaMA or DeepSeek
   - Option of fine tuning a dedicated model
3. **Own deployment**
   - Modification of an open-source model
   - Own development

LLM INTEGRATION - GROUNDING

- Feeding additional information into the LLM's context window.
- Improves relevance, accuracy, and reliability.
- Techniques:
  - Manual inclusion
  - Retrieval-Augmented Generation (RAG): automatically fetching relevant information from external sources to include in the prompt.
- Either an ad-hoc semantic search or by using specialised databases to store ‘embeddings’ (Vector Databases)



Retrieval-Augmented Generation

TRADE-OFFS

- Each of those options has a distinct data flow and a level of control.
- **SaaS** – low level of control

  (only through the use the provided tools - training data, “memory”; no control over logging)
- **Developing the model “surroundings”** gives more control and potentially better risk management
- **Building own model -** gives maximum control but also introduces unique challenges
  - “Owning” all data problems
  - Requires expertise, decisions about training data, full AI governance process, etc.

WHAT ARE THE TYPICAL PRIVACY CONCERNS

1. **Training data** (for training, testing, validating)
   - May convey biases, gaps in data may decrease performance, etc.
   - How it was collected, were personal information there?
   - If so, was it cleansed/anonymised?
   - If you are not building the model – check the model score cards / ask the providers
2. **The model input**
   - May contain personal information
   - May use additional “private” data from the organisation (e.g., through grounding – RAG)
   - The “surroundings” (e.g., vector database) may also store personal information
3. **The model output**
   - The quality of inferences, also whether it contains PI / re-identification
   - User interactions – how they use the models
   - What the users do with the output data
4. **Monitoring data** - model logs, queries, analytics, etc.

HOW TO TACKLE PRIVACY CONCERNS

1. **User guidance and AI literacy**
   - Absolutely critical, especially if we have external models
   - Transparency to model users and data subjects
2. **Limiting the amount of sensitive data put into the model**
   - Less problematic if we have control of the model and surroundings
   - Even having the model isolated from the third parties (e.g., on our infrastructure) does not protect us from problems of lack or lax data governance
3. **Filtering out the outputs**
   - Filters that 'catch' sensitive data
   - Also tackling broader risks – dangerous uses of the model
   - Some model providers do that already – you can use/customize those mechanisms
4. **Controlling the maintenance / monitoring data**
   - Data minimisation (audit trail only)
   - Secure storage

HOW TO TACKLE PRIVACY CONCERNS – EXAMPLES

Risks	Mitigations	Comments
Disclosure to the LLM provider	User guidance, AI literacy trainings, warnings, limiting the amount of sensitive data, software data filtering, DLP,	Provider can log the data for quality, monitoring, but also training the model, or as a 'memory' feature
Unsecure hosting	Contractual performance clauses & warranties, physical/logical isolation of data (e.g., separate 'tenancy'), encryption, risk management program implemented by the provider, control over the copy of the model	For hosted models much depends on model providers and adhesive contracts
The model output discloses PI to the user	User guidance, AI literacy trainings, filtering the output, monitoring, human review,	Might happen if PI used in training (fine-tuning), also with RAG
Insecure RAG/ anonymisation mechanisms	Access control, logging/monitoring, contractual performance clauses and warranties (for third party systems), filtering input/output, testing	Insecure logging or caching, third-party exposure (if external services used)

CONCLUSIONS

1. The options for deploying AI are available for almost all organisations
2. Privacy risks need to be **purposefully** managed (design of the system, controls)
3. Most important limitations:
   - Lack of data governance foundations
     - Internal data needs to be available/accessible, recognisable (metadata!), have proper quality, properly secured by the access control mechanisms
   - Lack of AI Literacy
     - Helping employees to adapt is critically important.
     - They need to know how to use AI and know how to oversee the AI.
     - Only 36% New Zealanders believe they have skills (KPMG)
     - 12% of the respondents in public sector received training (PSA survey)
4. Introducing AI is a change management process. We need to bring people with us. 😊

Thank you!

marcin@betkier.com

+64 27 320 5772

Dr Marcin Betkier
marcin@betkier.com

15 May 2025