GRIFFITH UNIVERSITY

# Failing Forward: Experiments with AI Tools to Build a Policy Finder

Wei Shen, eResearch Services, Griffith University

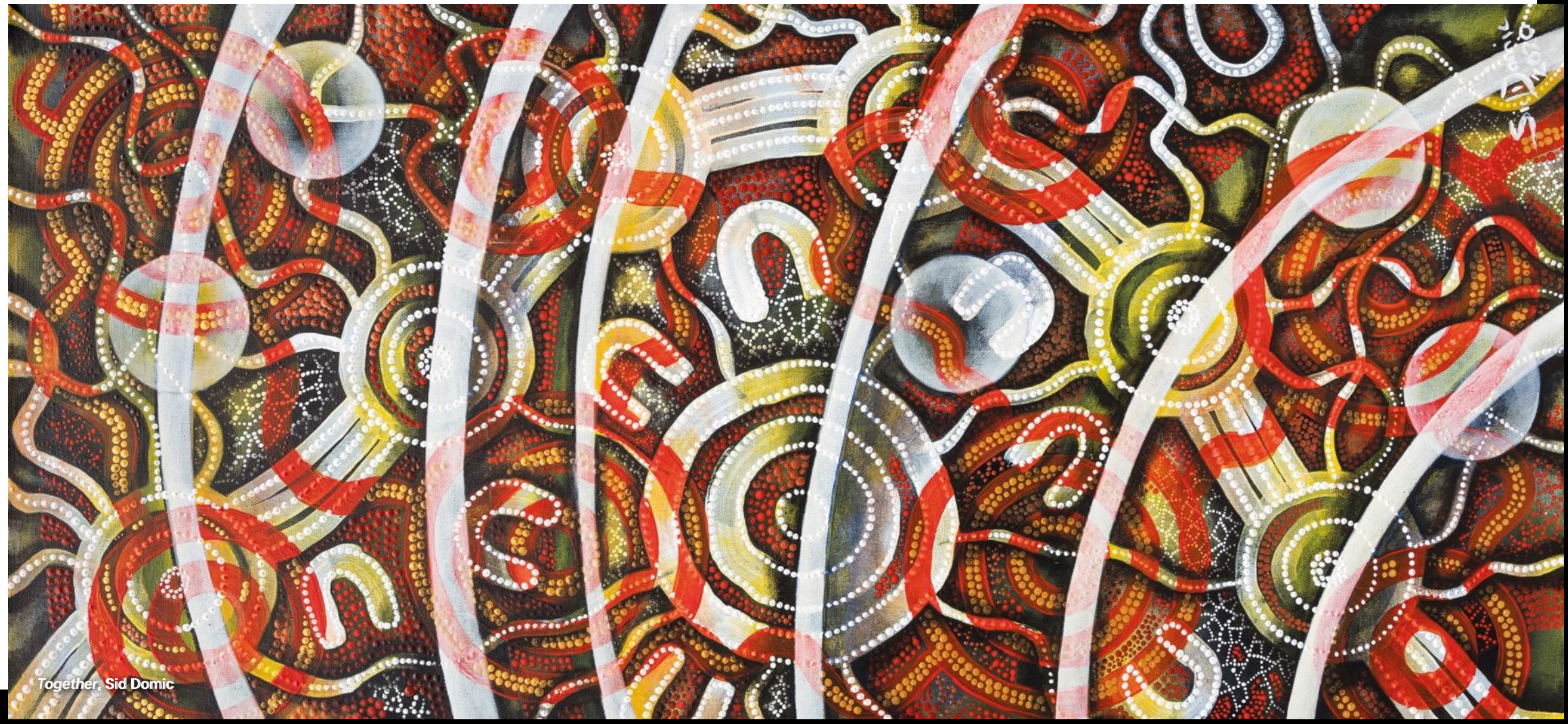
30 October 2025

Queensland Australia

GRIFFITH UNIVERSITY

# ACKNOWLEDGEMENT OF COUNTRY

Griffith University acknowledges the people who are the Traditional Custodians of the land. We pay respect to the Elders, past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples.



*Together*, Sid Domic

# It All Began with Excitement

- **The Challenge**

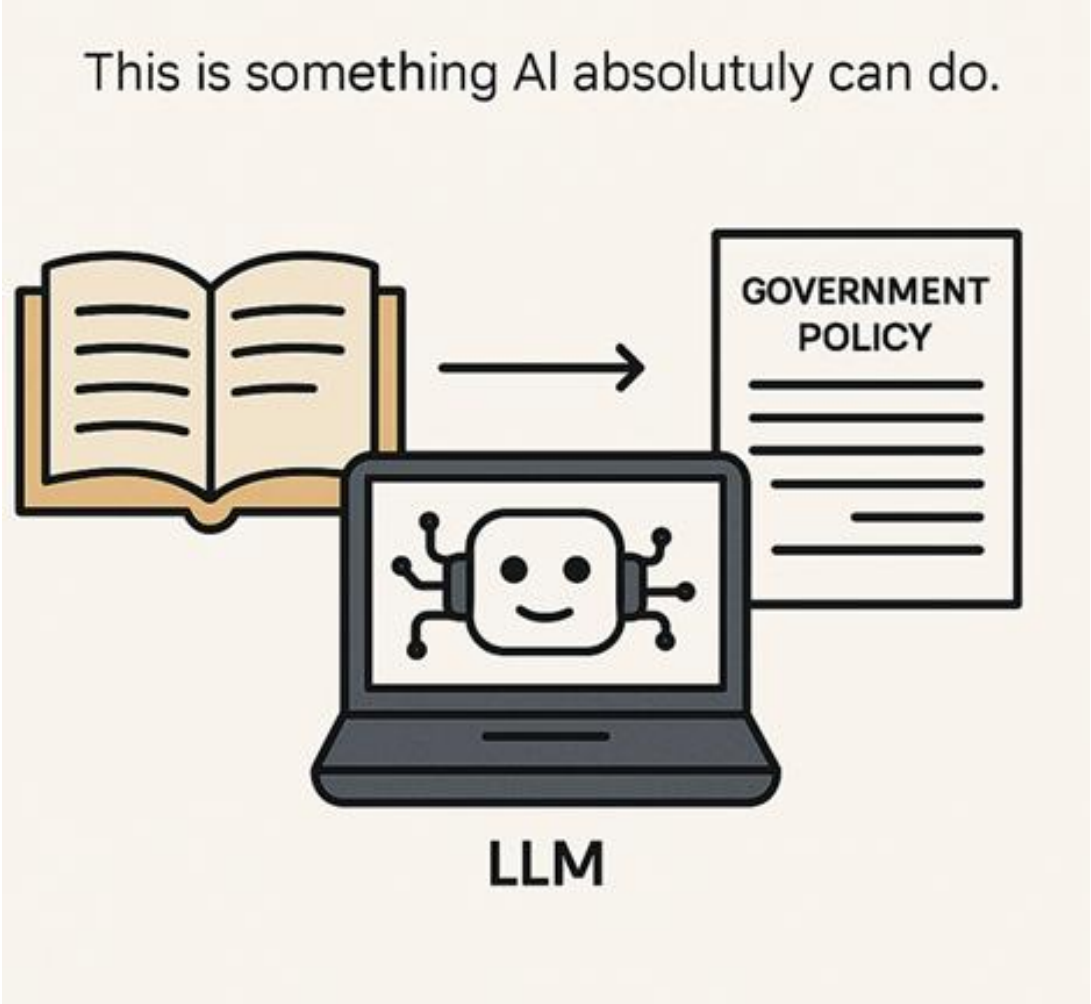
Researchers spend significant time locating and cross-checking relevant government policies.
The process is tedious, fragmented, and easy to miss important details, leading to incomplete or delayed insights.

- **The Opportunity**

An AI assistant that can find, understand, and connect all related policies – helping researchers make faster, more complete, and confident decisions.

- **Bottom Line**

Humans must always review, interpret, and make the final judgment.



GRIFFITH UNIVERSITY

# First Wave of Experiments: Generative Chat Tools

## Tools Tested (in 2024):

Microsoft Copilot · ChatGPT · Gemini ...

## Method:

Uploaded policy documents and asked open-ended questions, such as:

> "I'm a researcher at Griffith University proposing a collaboration with a researcher in United States on Koala. Which policies in the uploaded document apply to this collaboration? Please link to the relevant source sections." **not the best prompt**

## Observation:

- The responses looked promising at first glance — clear language, confident tone.
- But when checked carefully, important policies were missing or inconsistently referenced.
- The model often summarized well, yet failed to provide a complete or verifiable answer.

GRIFFITH UNIVERSITY

# 2025 – The Year of the AI Agent

## Tools Tested:

Microsoft Copilot Studio · Agent in SharePoint

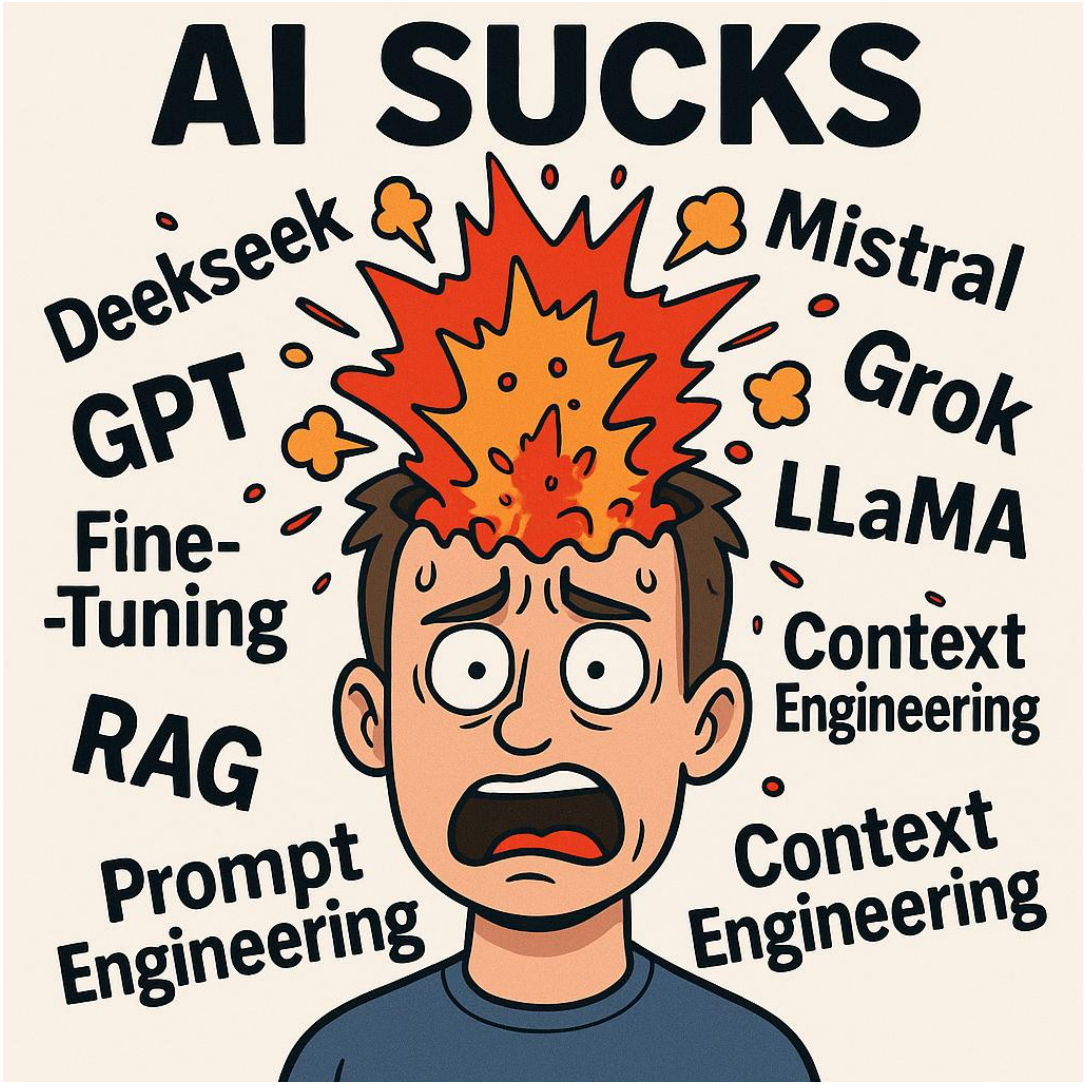
## Method:

Uploaded policy documents to a SharePoint site. Configured the agent's description and instructions to provide the best possible foundation for retrieval and reasoning. Built the agent to retrieve knowledge from these documents.

## Observation:

- Retrieval improved to some degree, but completeness and contextual understanding still lagged.
- The agent framework was powerful and well-integrated, yet the LLM's document comprehension remained the key bottleneck.

CRICOS: 00233E | TEQSA: PRV12076


AI SUCKS
Deekseek
GPT
Fine-
-Tuning
RAG
Prompt
Engineering
Mistral
Grok
LLaMA
Context
Engineering
Context
Engineering

GRIFFITH UNIVERSITY

# Step Back to Go Forward

## Tools Tested :

Microsoft Copilot · ChatGPT · Agent in SharePoint

## Method:

- One key document to focus on, Instead of working with nearly 20 policy documents.
- Asked a simpler, measurable question:

> "Search the provided document for every occurrence of the phrase 'national security', ignoring case and including plural forms ('securities'). Give the total number of occurrences found."

## Observation:

- The answers were inconsistent across runs — counts varied depending on phrasing.
- Showed that LLMs read semantically, not textually — good for summaries, unreliable for literal (keyword) search.

CRICOS: 00233E | TEQSA: PRV12076

GRIFFITH UNIVERSITY

# Exploring Retrieval-Augmented Generation (RAG)

## Tools Tested:

RAGFlow and some other RAG Chatbot stacks

## Method:

- Built RAG prototypes and experimented with lightweight local LLMs & different RAG methods.
- Gained more control over parameters — chunk size, overlap, Top K, Top P, Top N, temperature, context window, embedding model, and reranker model.
- Observed how these tuning choices affected retrieval quality and response completeness.

## Observation:

- I need a better computer (or cloud resource)
- RAG provided greater transparency and control, but it’s not magic.
- It often returned relevant fragments, yet still missed pieces of the full policy narrative.


?
?
RAG
LLM
WHY?
?
?
CHUNKING
PROMPTS
?
?

GRIFFITH UNIVERSITY

# It's Not the Model, It's the Data

## Realisation:

Besides the models, prompting techniques and RAG parameters, it was the quality, structure, and format of the policy documents going into it.

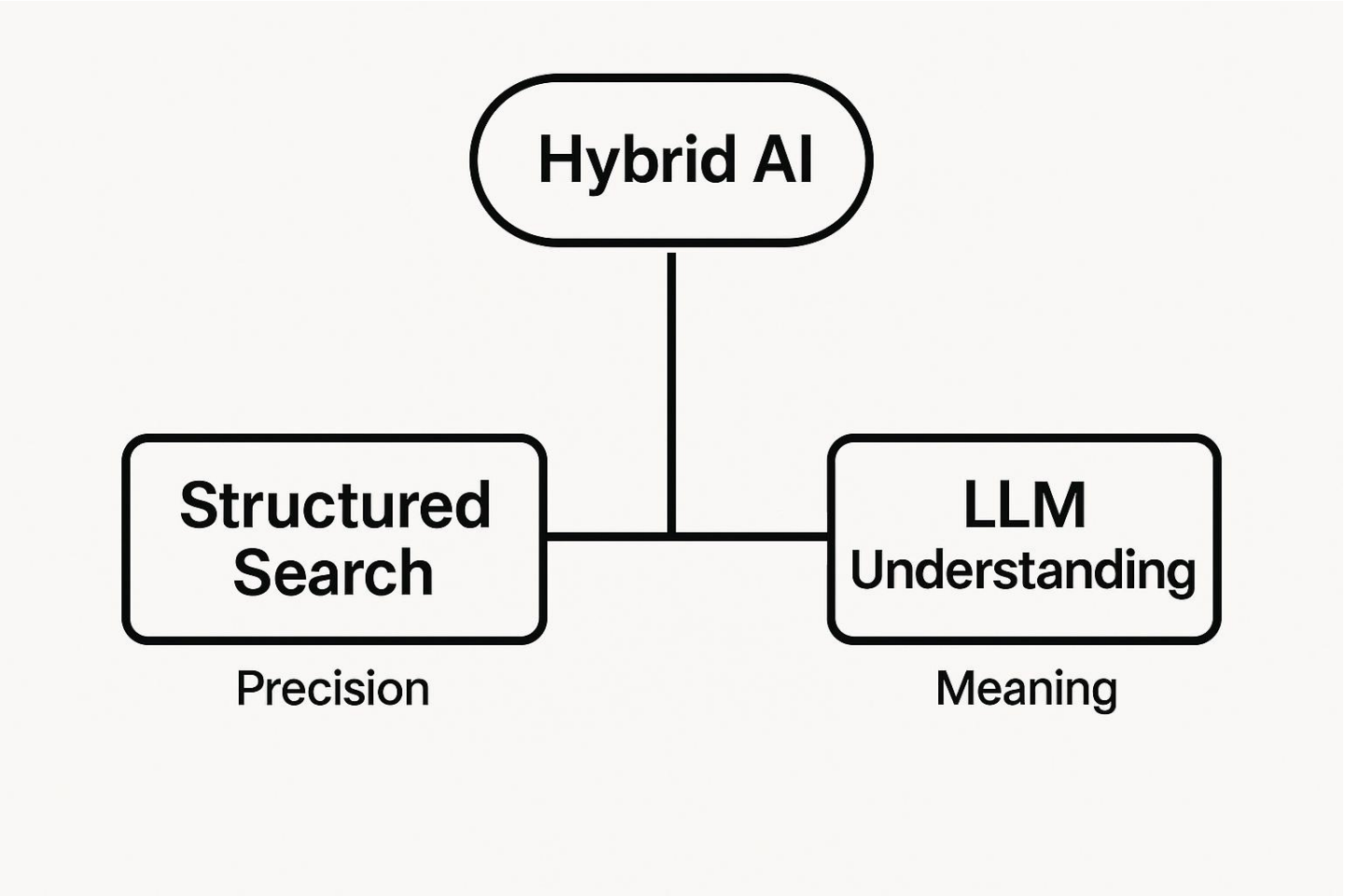
## Observation:

- Many policy documents are long, complex, and inconsistently structured.
- Large context windows were filled with irrelevant text.
- Chunking often broke logical sections — losing the meaning of headings or subclauses.
- OCR and formatting issues from PDFs caused hidden noise and missing text.
- Important context (like document scope, department, or dates) wasn't captured as metadata.
- The answers were inconsistent across runs — counts varied depending on phrasing.
- LLMs are not built for keyword search — they interpret meaning rather than match text literally.

CRICOS: 00233E | TEQSA: PRV12076

GRIFFITH UNIVERSITY

# One Path Forward:



GRIFFITH UNIVERSITY

# THANK YOU

Wei Shen, eResearch Services, Griffith University