Coroner Project

Automated Data Extraction from Coroner's Reports

Group 18

Adam Holt | Divya Mulackal Seetharama | Rahul Sharma | Tahsin Rahman | Aasish Shrestha | Zhenye Zhou

The Problem

The challenge

- A **problem** of data transformation & information extraction
- **Challenge**:

  Given a set of unstructured poor quality documents. Formulate a system which can extract or generate specific and factual information from the documents.

- Documents:

  Formal documents prepared by a coroner following an investigation into an unexpected, unnatural, or unexplained death.

Coroners Act, 1996
[Section 26(1)]

Western *Australia*

RECORD OF INVESTIGATION INTO DEATH

Ref No: 15/15

I, Rosalinda Clorinda Vincenza Fogliani, State Coroner, having investigated the deaths of –

Ruby Natasha NICHOLLS-DIVER *with an Inquest held at the Coroner's Court in Perth on 28 April 2015 – 15 May 2015 find that the identity of the deceased person was* ***Ruby Natasha NICHOLLS-DIVER*** *and that death occurred between 1 and 2 March 2011 at Geo Lithgow Reserve, Bicton, as a result of ligature compression of the neck (hanging); and*

Carly Jean ELLIOTT *with an inquest held at the Coroner's Court in Perth on 28 April 2015 – 15 May 2015 find that the identity of the deceased person was* ***Carly Jean ELLIOTT*** *and that death occurred between 30 and 31 March 2011 at 6 Davies Street, Beaconsfield, as a result of ligature compression of the neck (hanging); and*

Michael Ronald THOMAS *with an inquest held at the Coroner's Court in Perth on 28 April 2015 – 15 May 2015 find that the identity of the deceased person was* ***Michael Ronald THOMAS*** *and that death occurred between 3 June 2011 and 2 September 2011 at Banksia Eucalypt Woodland Reserve Gibbs Road, Aubin Grove as a result of unknown causes; and*

Anthony Ian EDWARDS *with an inquest held at the Coroner's Court in Perth on 28 April 2015 – 15 May 2015 find that the identity of the deceased person was* ***Anthony Ian EDWARDS*** *and that death occurred on 20 March 2012 at 23 Adelaide Street, Fremantle, as a result of multiple injuries ; and*

1

Source: Provided by Dr. Matt Albrecht

The Challenge

Significance & Impact

- Our client:

  *Wanted a method to efficiently get answers to questions about the content of the reports. Eg. "**What was the probable cause of death?**"*

- Our proposal: **R**etrieval **A**ugmented **G**eneration

- Significance:

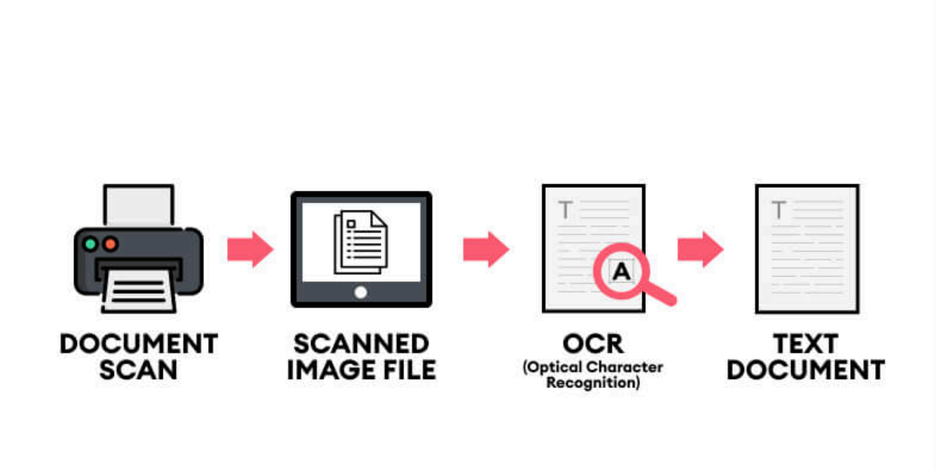
  - Interact with and understand coroners reports
  - Lead to evidence base policy recommendations to prevent the deaths under investigation from occurring in the first place

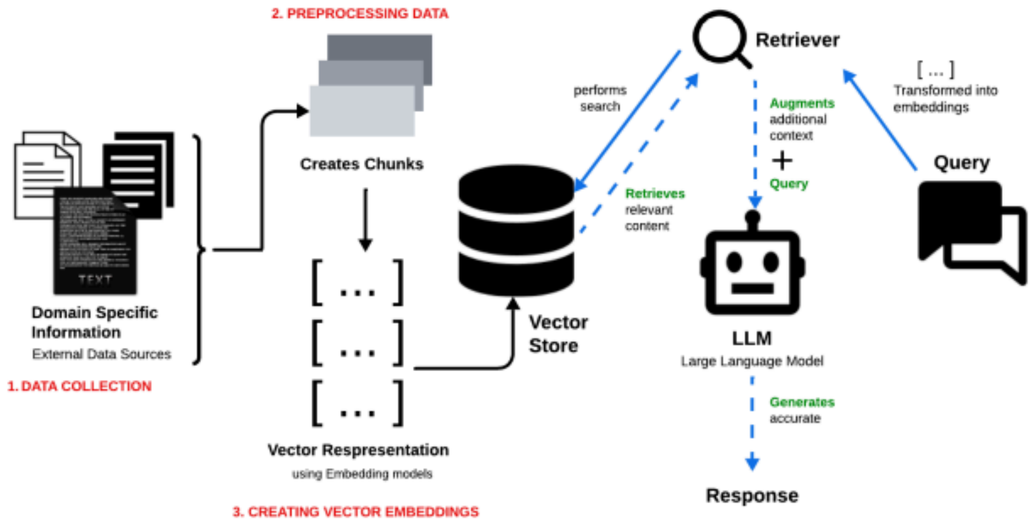
Source: World Health Organisation

Why Data Science?

What *are* the data?

- Data challenge → Data Science + Engineering
- Data:
  - ***Poor quality*** scanned copies of original reports
- Data Science & System Design Engineering:
  - Optical Character Recognition
  - Vector Database
  - Retrieval Augmented Generation
  - Pre-trained LLMs
  - Strictly local execution & hosting





Source: Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report

Surmounting the Challenge

A plan to tackle the problem(s)

- Team proposed to divide up the required components of the system by functionality:
  - **Data Preprocessing** (OCR; serialising; chunking etc)
  - **Vector Database, RAG & LLMs** (locally hosted, local execution)
  - **Interfaces** (textual - terminal; graphical - web app; programmatic - Python script)
- Delineate the project into three separate phases:
  - **Phase 1**: System Design & Implementation (*coding*)
  - **Phase 2**: System Evaluation (*analysis*)
  - **Phase 3**: Reporting & Documentation (*presentation*)

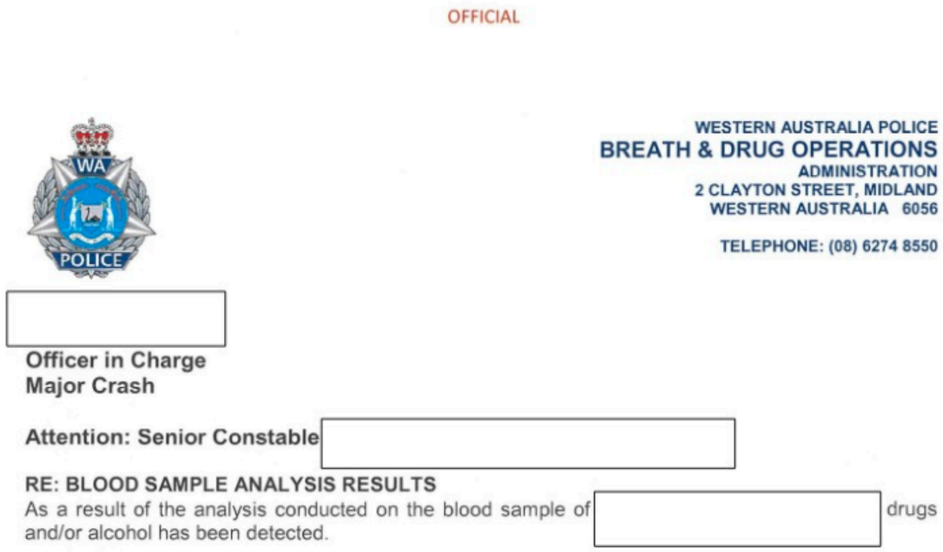
⚠ DANGER

- Phase 1
- System stability
- Dependencies
- Local execution req's

Generalisability

Which parts of the solution are project specific?

- Design of our RAG system is very much project specific:
  - Design features revolve around the nature of the data
  - Strict need for security due to the sensitivity of the data
  - Data is of a particular variety - scanned PDF documents of coroners reports



OFFICIAL

WESTERN AUSTRALIA POLICE
BREATH & DRUG OPERATIONS
ADMINISTRATION
2 CLAYTON STREET, MIDLAND
WESTERN AUSTRALIA 6056

TELEPHONE: (08) 6274 8550

Officer in Charge
Major Crash

Attention: Senior Constable

RE: BLOOD SAMPLE ANALYSIS RESULTS
As a result of the analysis conducted on the blood sample of drugs and/or alcohol has been detected.

- On the other hand:

Provided that the intention of the user is in alignment with those aspects of the project, then in that case, the system is quite generalisable. Our project has the capacity to be used with any documents similar to ours and where security and sensitivity are important considerations.

Example: Police Blood Sample Analysis

The Outcome

What aspects are the team most proud?

- **The system actually works!** ✅

  The outcome of our project is a functional RAG system which is able to generate accurate answers to the questions posed by the user.

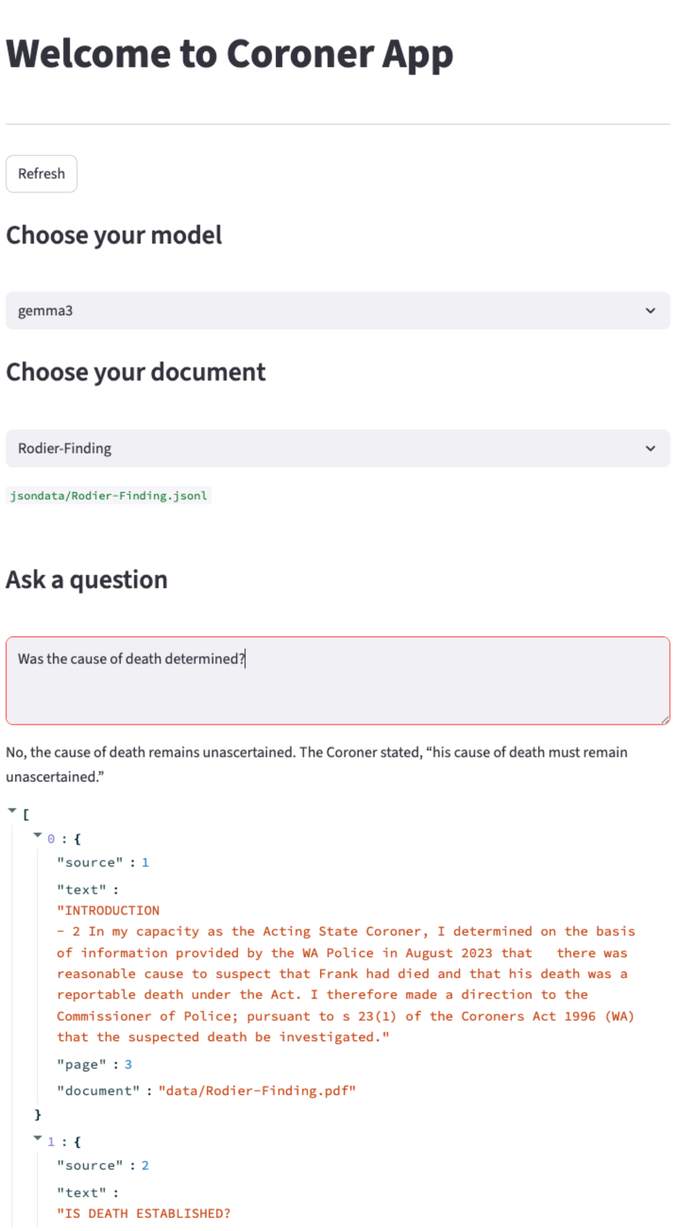
  Importantly, the user is able to check the veracity of the generated answer because included in the output are the sources, i.e., the contextual information that the system used to arrive at the answer.

  We are most proud of the fact that our final product has met our clients expectations and appears to be very much in line with what was envisioned as a good outcome.

Demo

Local web app interface

- In case it doesn't work...
  here is a preprepared example!



Coroner Project

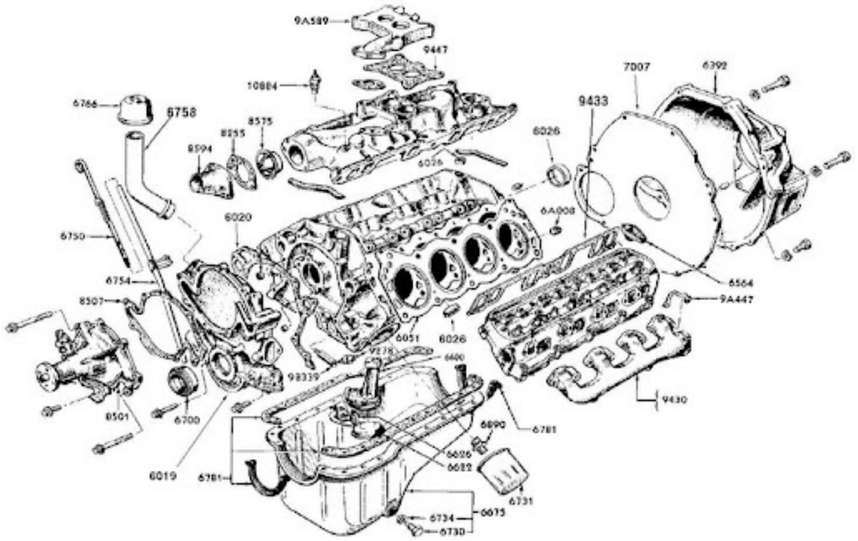
Group 18 Individual Presentation

Adam Holt

What is the Technical Challenge?

System Design & Implementation

- Project was decomposed into phases and Phase 1 was the technically most challenging
- ***Remind me, What was Phase 1?***
  - System Design & Implementation
- Responsibility for Phase 1



Components of a system: 289 Ford V8

*My primary responsibility was to design this system and write the code which would turn the design into a usable application. So, what are the **components** of the system?*

System Design & Implementation

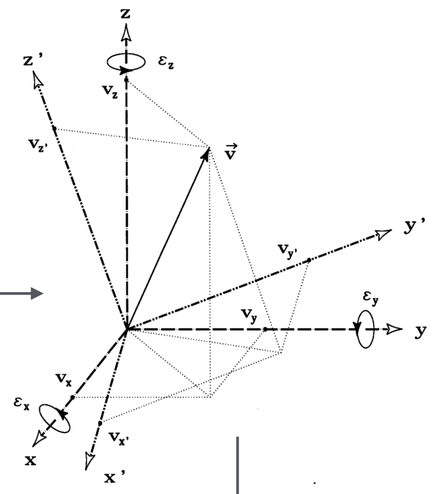
Components

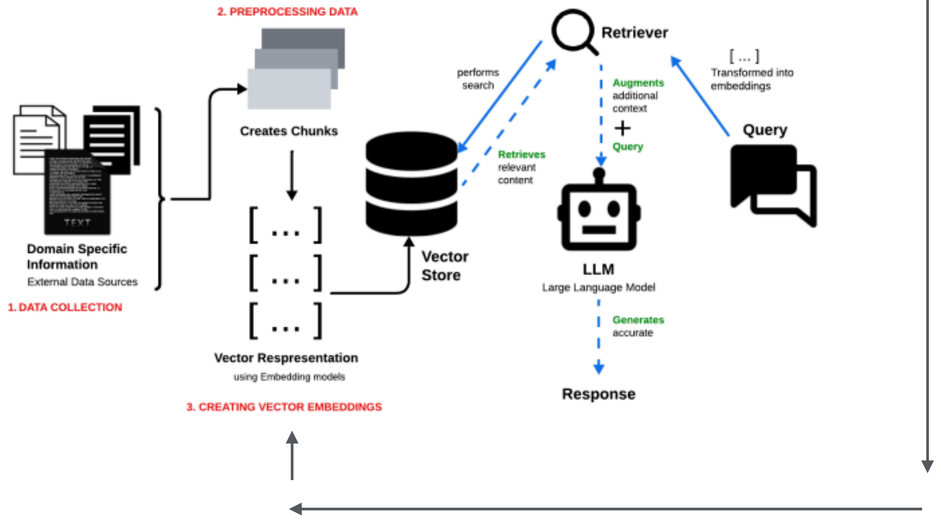
- **Data Preprocessing Pipeline**
  - OCR (EasyOCR)
  - metadata (source info; page; etc - Docling)
  - serialising & chunking
    (JSON, ready to load into vector database)

- **Vector Database**
  - vectorisation (pre-trained embeddings)
  - locally hosted; local execution
    (in-memory - no remote database)

- **RAG & LLM**
  - locally hosted; local execution; pre-trained LLMs
    (Ollama & Langchain)


```
either injury or drowned in the water and he disappeared from sight. Despite a search conducted, Frank's body
was never recovered. May gone Roy They they edge they being", "page_no": 3, "source": "data/Rodier-
Finding.pdf"}
{"page_content": "INTRODUCTION\n- 2 In my capacity as the Acting State Coroner, I determined on the basis of
information provided by the WA Police in August 2023 that   there was   reasonable cause to suspect that
Frank had died and that his death was a reportable death under the Act. I therefore made a direction to the
Commissioner of Police; pursuant to s 23(1) of the Coroners Act 1996 (WA) that the suspected death be
investigated.", "page_no": 3, "source": "data/Rodier-Finding.pdf"}
{"page_content": "INTRODUCTION\n- 3 On 11 October  2023 a report  prepared by Detective Sergeant Ellie Wold
from the Homicide Squad Missing Person Team. In the report, Frank was confirmed to be a long term missing
person, with his disappearance first reported to police at about 10.25 am on 25 1975. In 2006, a review by
```



Source: Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report

Components...

Environment

python™

CONDA

Local Hosting

OCR

Docling

Preprocessor

Glue

LangChain

Microsoft

Phi-4

In Memory Vector DB

LLMs

Gemma

Motivation to Address the Challenge

Be useful!

- Our client

  Dr. Matt Albrecht, researcher at the The Western Australian Centre for Road Safety Research (WACRSR)

- Me

  My primary motivation was to produce an application, using techniques of data science and system design, that is actually useful. By that I mean, our client actually derives utility from it. If there is even the slightest possibility that our project may have a positive impact, however small, on improving road safety, then that is a win.

National Road Safety Strategy 2021-30

Australian governments at all levels are working together with our communities to change the road transport system to prevent deaths and serious injuries on our roads.

The National Road Safety Strategy represents all governments' commitment to deliver significant reductions in road trauma, putting Australia on a path to achieve 'Vision Zero' or zero deaths and serious injuries on our roads by 2050.

National Road Safety Strategy 2021-30

The National Road Safety Strategy 2021-30 sets out Australia's road safety objectives over the next decade, and includes key priorities for action and targets to reduce the annual number of fatalities by at least 50 per cent and serious injuries by at least 30 per cent by 2030. The Strategy continues the commitment to the Safe System approach and strengthening all elements of our road transport system under three key themes: Safe roads, Safe vehicles and Safe road use.

Speed management is embedded within all key themes. The Strategy adopts a social model approach to foster a road safety culture across society and make road safety business-as-usual.

Read the **National Road Safety Strategy 2021-30** [PDF: 8113 KB]

Source: https://www.roadsafety.gov.au/nrss

How the Challenge was Approached

Three Separate Action Items

- **Research & Review**
  - Research summary was presented to group and informed our proposal

- **System Design**
  - System design was presented to our client
  - The design itself was fairly straightforward
  - Why? Primarily because it had become quite clear that a RAG based system was an optimal design choice due to our requirements

- **System Implementation**
  - Implementation of the proposed system was going to be challenging and this proved to be the case. Why? Strict security req's => extended development time and hampered collaboration

Technical Outcome I am Most Proud

It works... and *appears* to be useful!

- The *initial* system is functional and has met our clients expectations, **BUT**...
- Was designed with interactive engagement in mind
- Client described the way he'd be most likely to interact with the system
  - **programmatically** rather than interactively
- To meet our clients needs, I'd need to rethink the system
- Redesign time... 

Programmatic Interaction

Details

- **Interactive System**:
  - preprocessed document objects - JSON
  - RAG/LLM pipeline
  - ask a question get answer & sources
- **Programmatic System**:
  - package the system up into Python object
  - expose a method to get answer & sources
  - do useful things...

```python
FILE_PATH = Path("jsondata/Rodier-Finding.jsonl")
GEN_MODEL = "gemma3"
EMBED_MODEL = "mxbai-embed-large"
VDB = InMemoryVectorStore
TOP_K = 3

PROMPT = ChatPromptTemplate.from_template(
    """Context information is below.
    \n---------------------\n
    {context}
    \n---------------------\n
    Given the context information and not prior
    knowledge, answer the query.\n
    Query: {input}\n
    Answer:\n""",
)

qanda = QandA(gen_model=GEN_MODEL,
              embed_model=EMBED_MODEL,
              vdb=VDB,
              file_path=FILE_PATH,
              top_k=TOP_K,
              prompt=PROMPT)

QUESTIONS = ["Who is the coroner?",
             "Who is the deceased?",
             "What was the cause of death?"]
CORRECT_ANSWERS = ["Sarah Helen Linton",
                   "Frank Edward Rodier",
                   "unascertained"]
LLM_ANSWERS = []

for i, QUESTION in enumerate(QUESTIONS):
    ANSWER = qanda.ask(QUESTION)
    LLM_ANSWERS.append(ANSWER)
    print(f"Answer {i + 1}: ", ANSWER)

data = {
    'FILENAME': ['Rodier-Finding'] * len(QUESTIONS),
    'MODEL': ['gemma3'] * len(QUESTIONS),
    'QUESTION': QUESTIONS,
    'CORRECT_ANSWER': CORRECT_ANSWERS,
    'LLM_ANSWER': LLM_ANSWERS
}

df = pd.DataFrame(data)

scores_df = calculate_bertscore_df(df)
```

Phase 2: BertScore, Divya & Rahul

Improvements

It's *far* from perfect!

- If I had ***more time*** I would:
  - Expand the pool of data (synthetic generation)
  - Make optimisations to the preprocessor pipeline (usability and performance)
  - Code refactoring (simplify and further modularise - reusable components)
  - Implement larger selection of pre-trained embeddings for vector database (reduce dependence)
  - Automate and streamline evaluation

Fín.