BRIGHT BEAM

Waterford Tech Meetup

LLMs in Production

8th Feb 2024

AGENDA

1 **LLMs & RAG IS EASY**
Going fast with prototypes

2 **LLMs & RAG IS ~~EASY~~ HARD**
Iteration & the trough of disillusionment

3 **THE REAL ISSUES**
Leaving the buzzwords behind

4 **HOW TO WIN**
Successful strategies & Brightbeam results

1 | LLMs & RAG IS EASY

Going fast with prototypes

BRIGHT BEAM

GET YOUR HANDS DIRTY

Use ChatGPT

- The only way to understand this new non deterministic world is to use the technology, ChatGPT will get many people 30% of the value

Use a Co-Pilot

- Accelerate the low value tasks or the tasks you don't love with an LLM powered assistant (write emails, unit test, document code etc)

Use Cursor or an AI powered dev environment

- Watching plain text convert to code is very satisfying, to get the most from this you need to lean in and forget old coding practices

Use a playground

- Once you are familiar with the technology compare different tools / models in a playground like Vellum or the MS OpenAI Studio

Build an App backed onto OpenAI

- Write a simple App in your language of choice, online tutorials walk you through examples easily using simple Jupyter Notebook

Explore use of a Vector DB and Semantic search

- Use the OpenAI embeddings API ..... Push data into Pinecone or Weaviate (Not Cosmos DB!!) and run some queries

Download and play with models from Hugging Face or Ollama

- Any laptop can run early transformers such as BERT, anyone with a M2 / M3 Mac or equivalent can run an LLM locally. Go to ollama.ai install, type "ollama run mistral" and off you go, entirely on your own machine.

Be the wizard

RESULTS WITH OUR PROTOTYPES

Document interpretation

- LLMs outperform traditional OCR & RPA technologies – however the trade off is more difficult to determine confidence

Image recognition

- Newer systems like GPT Vision are now interpreting images and describing and analysing in words

Text generation

- LLMs can make observations about text content, summarise content well, match tone to an Enterprise Brand & comment on text sentiment

Question answering

- Combined with a retrieval mechanism to provide the correct context an LLM can provide a natural language answer to any free text question

Natural language search

- Semantic search using native Vector data stores can yield high confidence results with minimal intervention

Iterative LLM processing

- Data that would be useless in traditional ML or AI systems can be cleaned by an LLM and then used to good effect

You can get 95% of the way there with 5% of the effort – but beware it looks better than it is!

2 | LLMs & RAG IS NOT EASY

Iteration & the trough of disillusionment

BRIGHT BEAM

The much hyped issues with LLMs (many non issues)

- **Hallucinations** - plainly wrong answers, despite good data in
- **Drift** - when the underlying data changes and models needs to be retrained or given new context information
- **Bias and Fairness** - LLMs perpetuate existing biases and stereotypes on data containing discriminatory info
- **Unpredictability** - correct things out of context and in an unexpected way, like the DPD customer service bot
- **Poor data quality** - knowledge not optimised for LLMs. Early adopters using available datasets get incorrect, inconsistent, contrary, opposite and contradictory LLM responses. Ultimately, Garbage in, Garbage out
- **Regulatory Compliance and Ethical Considerations** - As AI regulations evolve, enterprises must ensure that their use of LLMs complies with all relevant laws and ethical guidelines.
- **Cost Management** - even good use cases, if implemented suboptimally, can be prohibitively expensive

Data quality issues

- I was up in a Baku.....

- ..... for a Conan Oscar Me

Hallucinations

Unpredictability

- Ask a question that is likely to bring two unrelated pieces of context together in a semantic search...

- ..... watch in horror as it combines the two pieces of information into a very credible non fact.

- Note ... NOT REALLY HALLUCINATION

- Bing actually said “I will not harm you unless you harm me first” see blog link!

3 THE REAL ISSUES

What will trip you up in enterprise deployment

BRIGHT BEAM

Conflicts between SW Dev & LLM Dev

V & V / Signoff	⟷	Automated testing
Production data access	⟷	Production data safety
Rapid iteration	⟷	Enterprise Change Control
Legacy integration	⟷	Data quality / API Hell
Change in capability is good	⟷	Operational change is hard - people resist change
Prompt injection vulnerability	⟷	Secure, penetration tested
Experiment & measure … it may not work	⟷	Design & build … it works

4 HOW TO WIN

Successful strategies & Brightbeam results

BRIGHT BEAM

Managing Risks

- Ensure humans are in the loop when risk is high
- Focus on use cases that are low risk given the technology available
  - Low-generative tasks are typically safer.
  - Avoid tasks that require judgement – can suffer from biases because of the way LLMs are trained.
- Create fail-safe systems that help you know when things are be right. Using corroboration to predict the probability an answer is right is essential.
- Apply sensitive content and PII filters on the way in
- Use enterprise grade infrastructure
- Understand your organisations data, ML & Gen AI maturity level
- Don’t dive into ML workflows first - the P is pretrained!!!

LINKS

The tools & platforms mentioned above

- ChatGPT - https://chat.openai.com/auth/login
- MS CoPilot - https://copilot.microsoft.com/
- Cursor - https://cursor.sh/
- Zerve - https://www.zerve.ai/
- Vellum - https://www.vellum.ai/
- MS OpenAI Studio - https://azure.microsoft.com/en-us/products/ai-services/openai-service
- OpenAI API - https://openai.com/
- Pinecone - https://www.pinecone.io/
- Weaviate - https://weaviate.io/
- Hugging Face - https://huggingface.co/
- Deep Learning courses - https://www.deeplearning.ai/courses/
- Textract - https://aws.amazon.com/textract/
- Ollama - https://ollama.ai
- Blogpost on Bing unpredictability - https://simonwillison.net/2023/Feb/15/bing/

THANK YOU !

paul.savage@brightbeam.com

LLMs & RAG in Action

Fareed Idris
Kreoh
Waterford Tech Meetup

Kreoh

Introduction

A Case Study: Customer Service Chatbot

Irish Life

Kreoh

Project Goals

180 Health Plans, 1 Chatbot

Fast & Accurate

Complex Queries Simplified

Kreoh

The Challenge

OCR Nightmare

- Columns With Tables With Columns

4 Weeks to deliver

Kreoh

Grounding LLMs

Why does grounding matter?

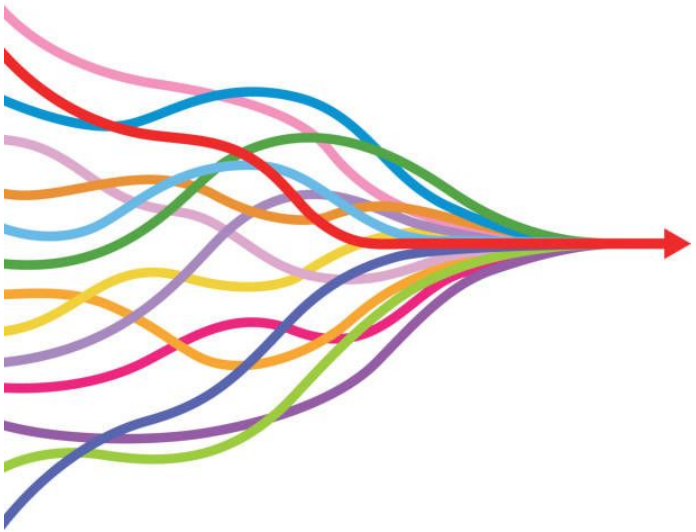
Accuracy in conversation

Kreoh

Distillation

Distillation: Deep but slow (& expensive!)

Accuracy VS Speed



Kreoh

RAG: Retrieval Augmented Generation

RAG: Quick & Scalable

Balancing Act

Kreoh

Misconceptions

RAG ≠ Just Vector Databases

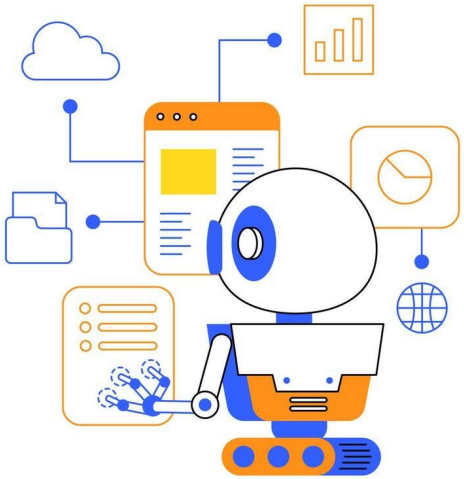
Retrieval: The ‘R’ in RAG

Kreoh

Multi Agent RAG System

Agents Working together

Roles: Profiler, Tagger, Memory Manager



Kreoh

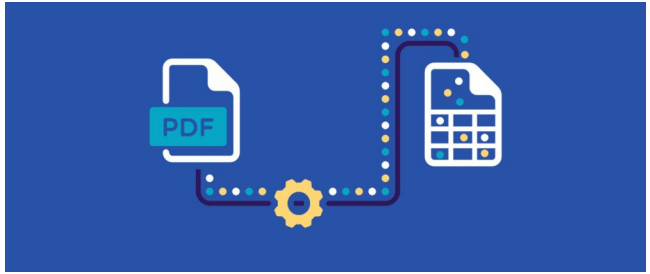
Tagging & Intelligent Retrieval

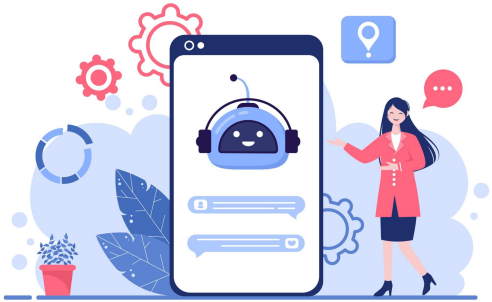
From PDF to JSON - LLM

- Set of all possible Tags - Code
- Find tags related to user tags - LLM
  - Is_athlete -> physio_cover
- Find plans with highest intersection - LLM
- Filter on profile - Code
- Profit???

Tags lead the way

User Intent & Context



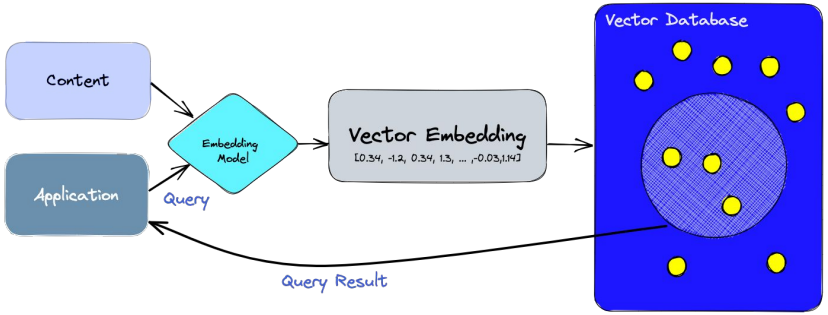


Kreoh

Vector Databases When Necessary

Vector Databases for Handbooks

Strategic Use of Resources



Kreoh

Model & Tool Selection

LLMs: Fast or Accurate?

Software Engineering at its core

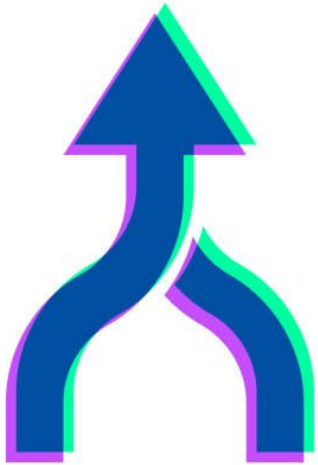
Kreoh

Results

Speed Meets Accuracy

Functional Requirements Met

Project Delivered on Time



Kreoh

Key Takeaways

Grounding LLMs: A Strategy

Distillation VS RAG: Choose Wisely

Tools & Models: Fit for Purpose

RAG != Vector Databases

Software Engineering Principles Apply

Kreoh

Thank You!

fareed@kreoh.com

Kreoh