Case Studies in Data Science – Week 5

Damiano Spina

Johanne Trippas

RMIT UNIVERSITY

“Information Retrieval on Country”, Dr Treahna Hamm (Firebrace), Yorta Yorta

admscentre.org.au/information-retrieval-on-Country

Assignments

Thank you for submitting Individual Task 1

- Some students did not respect 48-hour silent policy
- Do not expect responses from the teaching team during weekends
- If you realise that you need an extension during the silent policy (and/or weekend), apply for special consideration
- Penalty for late submission
  - Only a few minutes late, uploading took too long, etc.
  - We made an exception -> excused late penalty within 30 min. after the deadline
    - Only this time: it is your responsibility to check upload time/submission correctness (test early!)

WIL Milestone 1 due tomorrow

Week 6 reflective portfolio available

Retrieval-Augmented Generation (RAG) – Hands-on Workshop

RMIT Val Team

RMIT-ADM+S Team: Winners of the SIGIR 2025 LiveRAG Challenge

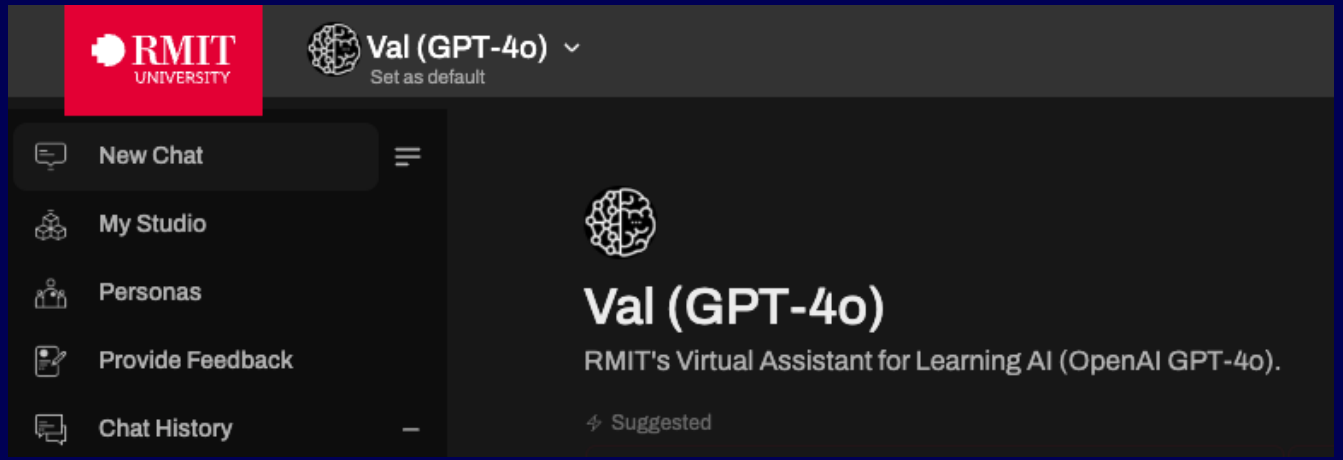
Walert Team

RMIT Val

KNOWING THE WHO & WHY

Yiota Alexiadis

Michael Hewett

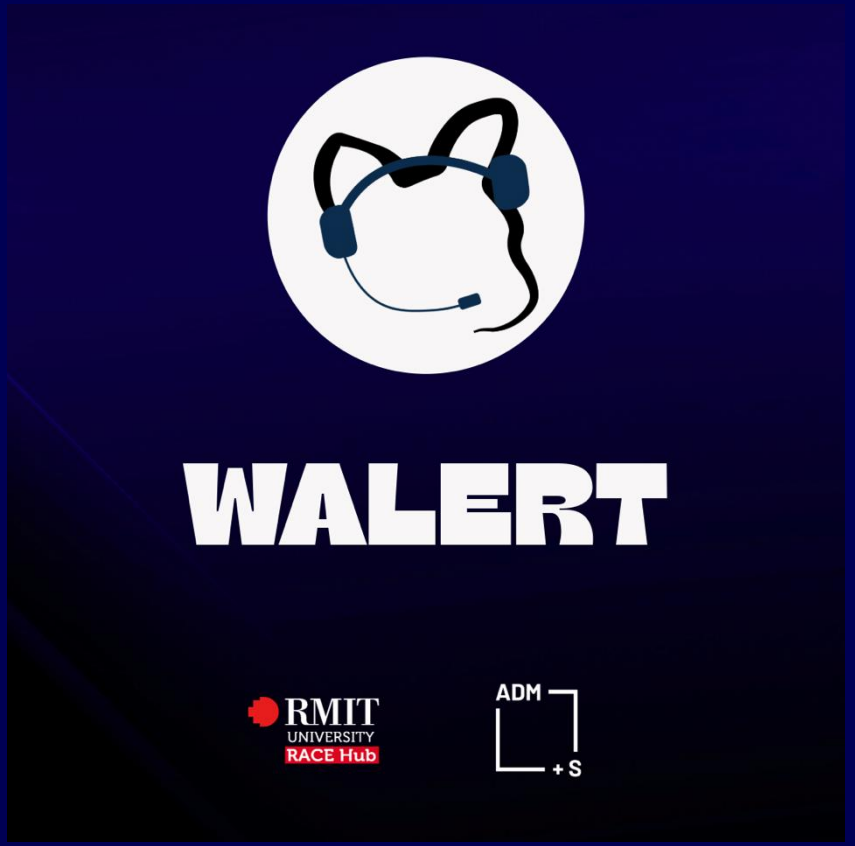


RMIT UNIVERSITY

CRICOS provider number: 00122A | RTO Code: 3046

Walert

bit.ly/walert



Sachin Pathiyan Cherumanal

Falk Scholer

RMIT UNIVERSITY

Customised AI-powered chatbot for Open Day (SCT)

A conversational ~~agent~~ conversation starter:

- Opportunities and limitations of customised AI-powered chatbots, e.g., Retrieval-Augmented Generation (RAG) systems
- Evaluation framework and guidelines on how to test solution built in-house
- Walert means “possum” in Woi wurrung and Boon wurrung languages: significance of possum skin cloaks for the Traditional Custodians of Country

bit.ly/walert

ADM +S ARC Centre of Excellence for Automated Decision-Making and Society

RMIT UNIVERSITY RACE Hub

Walert: The Journey So Far

RMIT Open Day 2023

CHIIR'24 Sheffield, UK

RMIT Open Day 2024

Outstanding Achievement Award, EIP-RACE Demonstrator 2024

Demo at SEEK

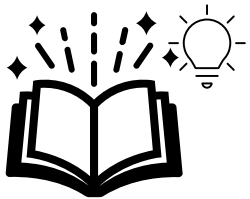
Search and RecSys Meetup, REA Group

Ubicomp/ISWC'24

RMIT Open Day 2025

Motivation

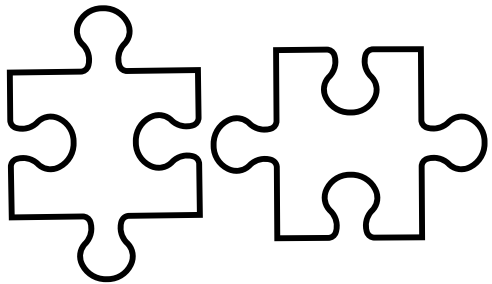
Existing Tools



Existing Knowledge

Stakeholders Needs

Academic expertise.
Best practices known by the Information Retrieval (IR) community.



Industry needs.
Provide practical solutions.
Guide practitioners out of the domain.

Research Questions

Can we quantify the impact of using LLMs in the development of chatbots?

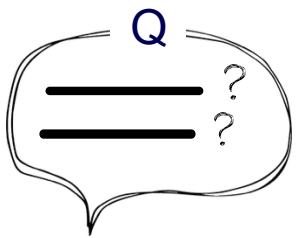
Can we reduce the risks of hallucinations?

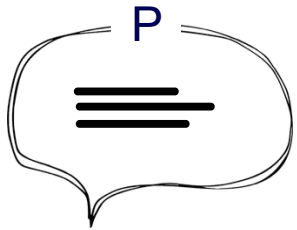
Intent-based chatbot vs. Retrieval-Augmented Generation (RAG)

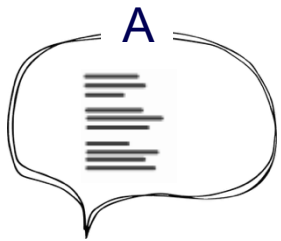
- Intent-Based: Conversational Model deployed as Amazon Alexa Skill
- RAG: What we will be focusing on today

Test-Driven Design

Document used
by academics to answer
questions about SCT
programs at RMIT Open Day







what's the difference between CS and SE programs?

["CS program covers Computer Science core body of knowledge ...
in various aspects of information and communication technology."]

Will the school provide work placements?

[Well, not directly. We do have a lot of .... an industry placement.]

Test-Driven Design

Scenarios

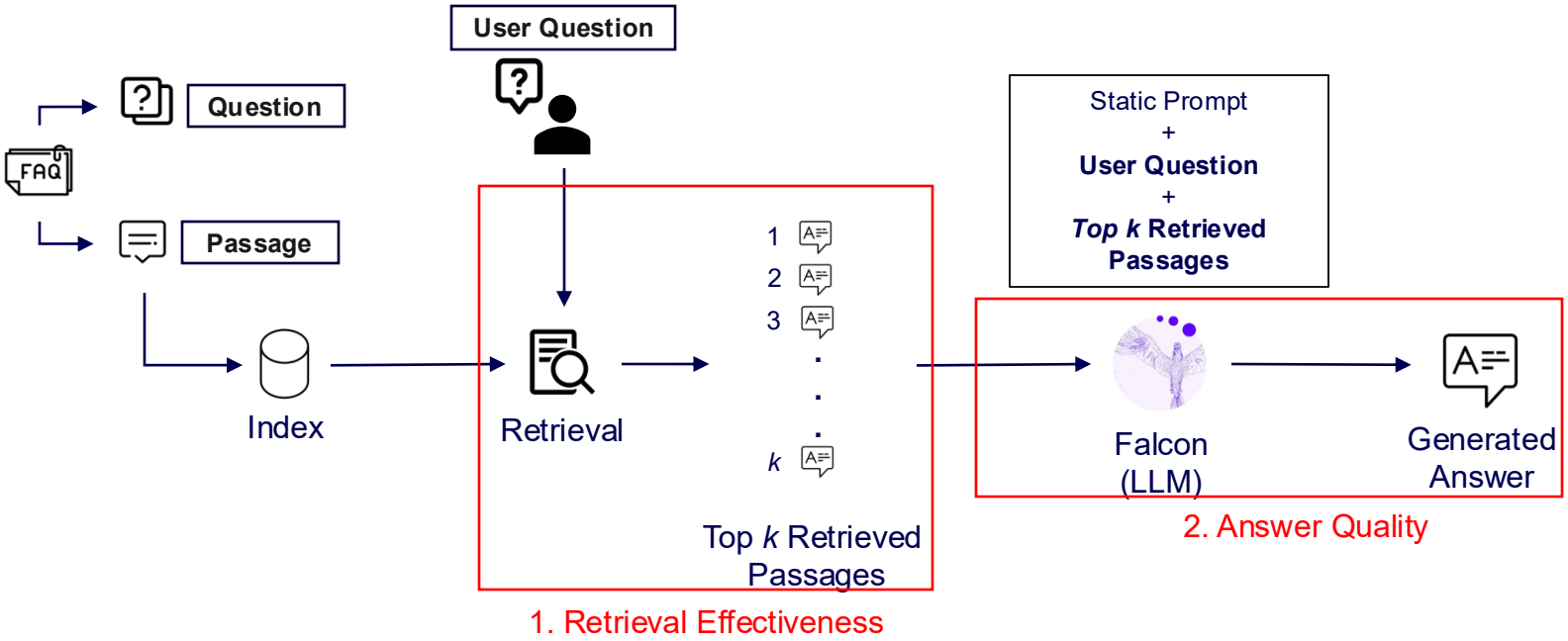
- Questions with known answers → Have a **direct answer** in FAQ (1 passage)
- New Questions with inferred answers → Can be inferred from **multiple** passages
- New Questions with no answers → Cannot be inferred and does not have an answer (**out of knowledge base**)

Test Collection

- 106 questions: 84 with known answers
  12 with inferred answers
  10 do not have an answer
- 120 passages

Question Type	Question	Passage(s)	Answer
Questions with Known Answers	Is the transfer from Associate Degree to Bachelors automatic?	No, it is not. You are required to apply when you are closer to the completion of the Associate Degree. (*Highly Relevant*)	No, it is not. You are required to apply when you are closer to the completion of the Associate Degree.
Questions with Inferred Answers	What does the final year of Computer Science (CS) program include?	(1) [...] Software Engineering (SE) students will do another large in-house project and more SE electives, while CS students will do a slightly smaller project and a few more core [...]. (*Partially Relevant*) (2) [...] students are required by RMIT rules to do a capstone project in their final year. [...] with an industry partner [...] (*Partially Relevant*)	It includes a small capstone project with a supervisor that work with an industry partner, as well as a few more core courses and electives.
Out-of-KB Questions	When does the application for program transfer open?	Not available (*No Relevant Passages*)	I'm sorry, I don't have an answer.

Retrieval-Augmented Generation (RAG) in Walert



Evaluation

1. Retrieval Effectiveness
   - NDCG@k
2. Answer Quality
   - BERTScore
   - ROUGE-1

*BERTScore and ROUGE has also been known to be used for automatically evaluating ‘hallucinations’.

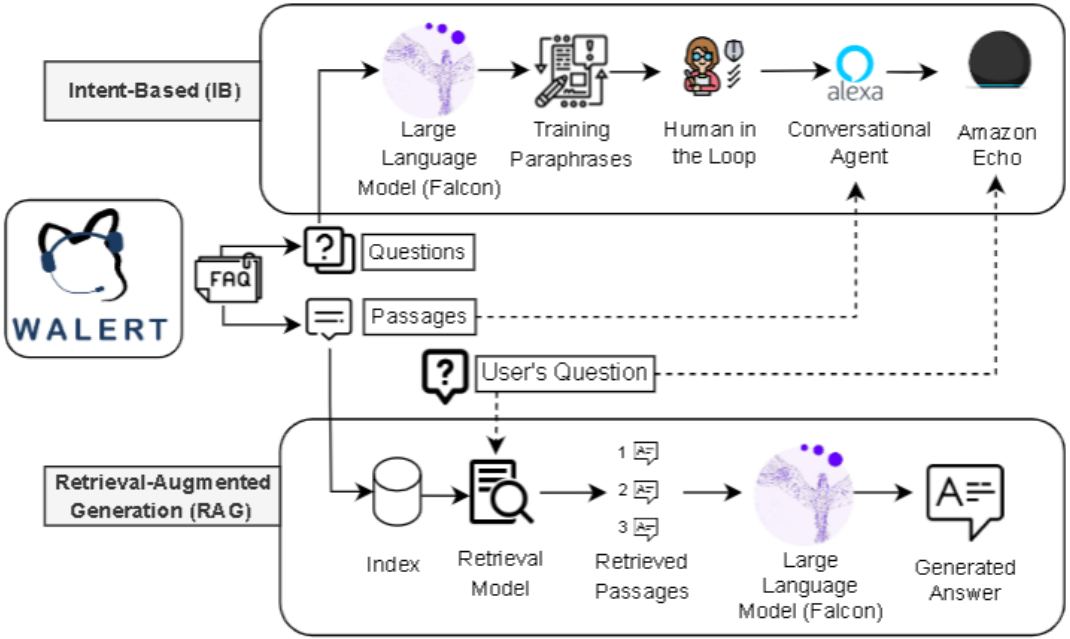
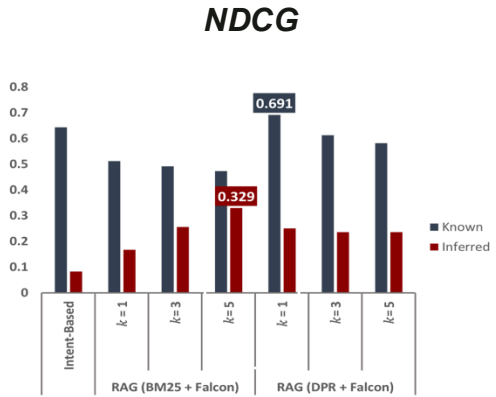
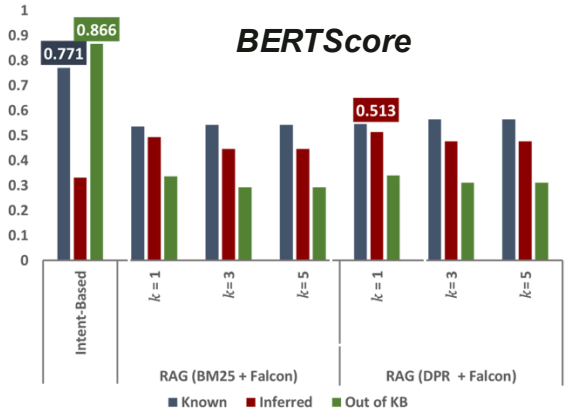


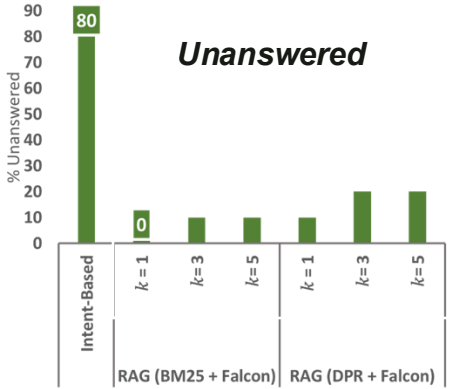
Fig: Overall Architecture

Key Findings

- RAG-based approach gave best results for Both “Known” and “Inferred” scenarios.
- RAG attempted to answer several “Out of KB” questions indicating potential of hallucinations.
- IB responds in “Known” and “Out of KB” in terms of final answer quality.
- RAG performs better in case of “Inferred” scenarios.
- In RAG, truncated rankings seemed to produce better answers.







Effectiveness: Take-away

- Significant impact of the retrieval stage !
- RAG **DOES NOT** eradicate hallucination !
- Retrieval component can help **reduce** hallucinations.
- Effective truncation is important.

"if you cannot measure it, you cannot improve it"

- William Thomson (Lord Kelvin)

- Importance of evaluation framework.

RMIT-ADM+S

Shuoqi Sun
Kun Ran
Oleg Zendel


First Prize

5679

DATE July 17, 2025

PAY TO THE ORDER OF Kun Ran, Shuoqi Sun, Khoi Nguyen Dinh Anh, Damiano Spina and Oleg Zendel $ 5,000.00

Five thousand ~~~~~~~~~~~~~~~~~~~~~~ DOLLARS

MEMO SIGIR 2025 LiveRAG Challenge

TII Technology Innovation Institute

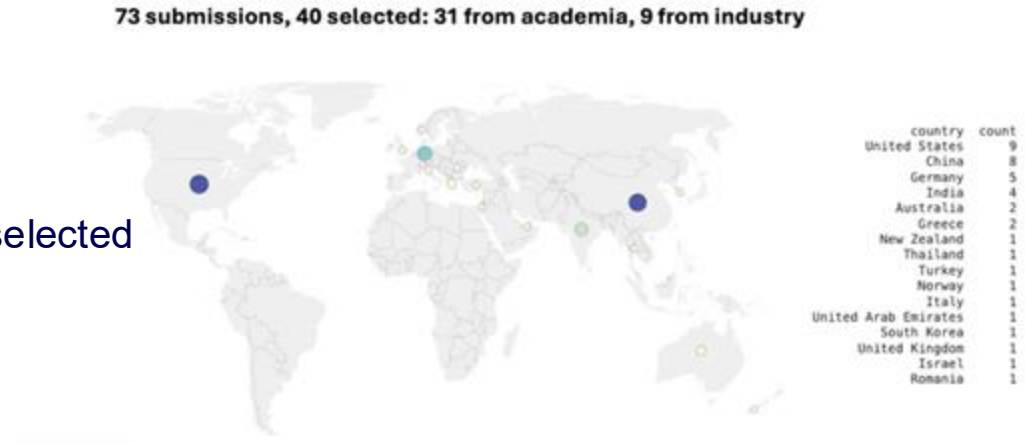
⑈325760408⑈ 003192⑆ 0583 42

RMIT UNIVERSITY

https://liverag.tii.ae/

LiveRAG Overview

- 73 submitted, 40 international teams selected to compete.
- 2 hours live run window.
- 500 questions from general domain.
- Answers **must** be generated with Falcon3-10B-Instruct.
- Questions and reference answers generated by DataMorgana.



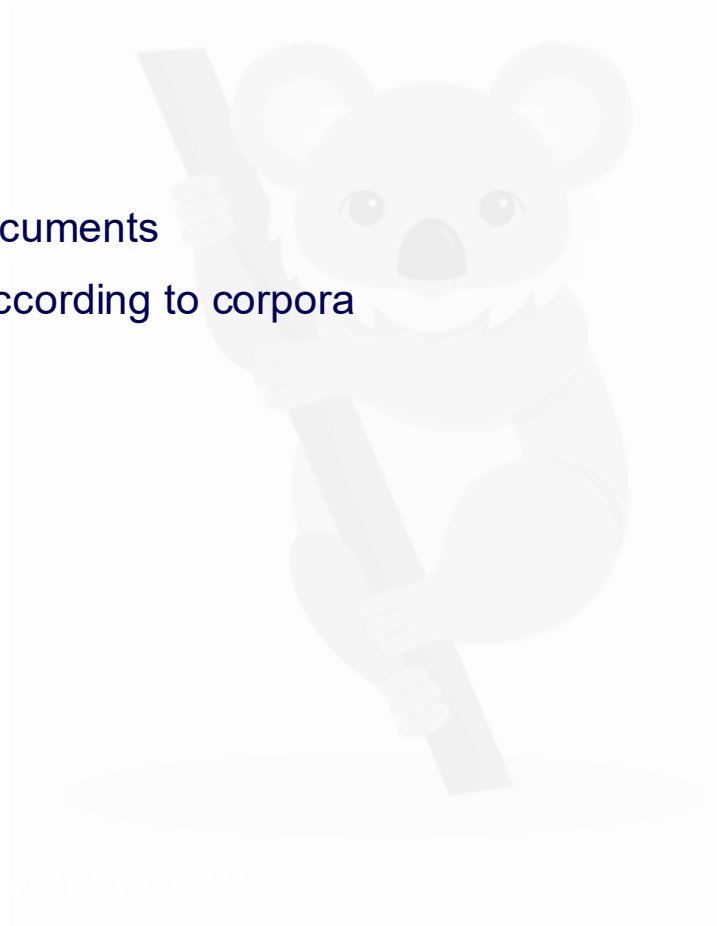
[Image: SIGIR 2025 LiveRAG Challenge organizers]

LiveRAG Overview - Provided Resources

- 1,500 USD in AWS credits
- Pinecone Dense index: E5 based retrieval
- Opensearch Sparse index: BM25 retrieval
- API access to DataMorgana
- API access to Falcon3-10B-Instruct

DataMorgana Overview

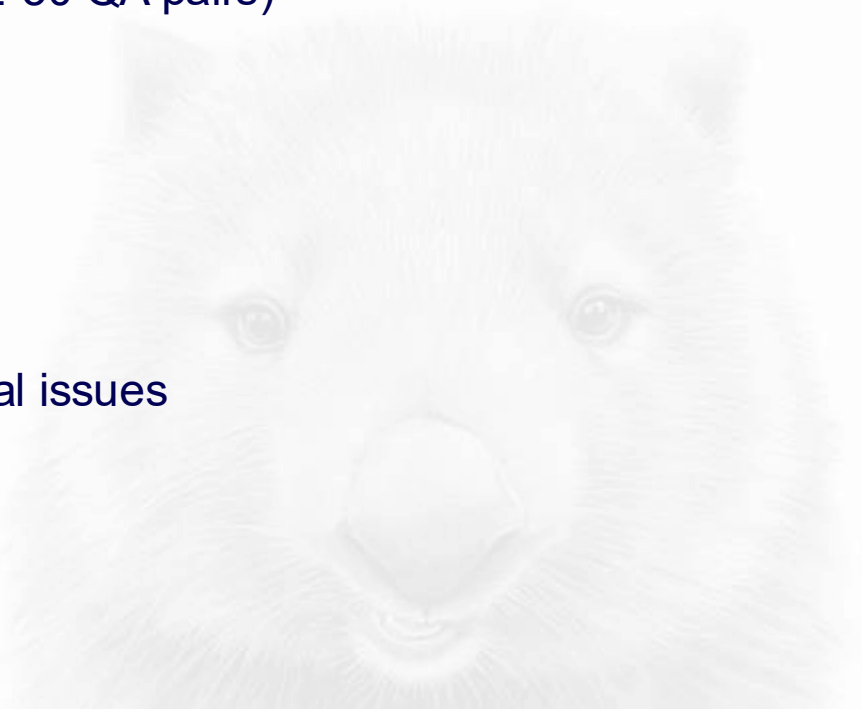
- Generating question and answer pairs with support documents
- Simulate users and question categories, configuring according to corpora
- Questions are supported by one or two documents
- LLM backbone: Claude Sonnet-3.5
- Documents corpus: FineWeb-10BT (15M documents)



Initial Design and Development

In-House Evaluation

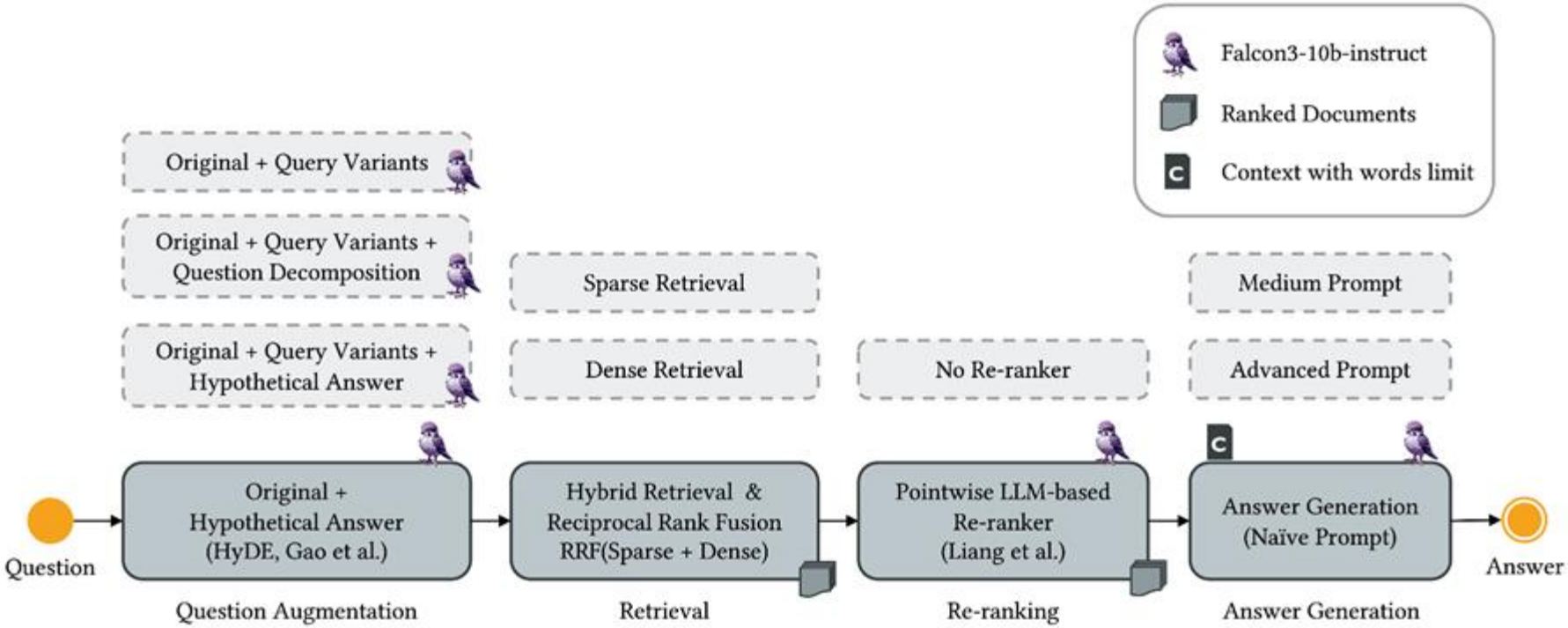
- Generated multiple QA pairs using DataMorgana
  - Small sets - initial development (2-50 QA pairs)
  - Medium - working sets (100-200)
  - Large - Stress testing (500-1000)
- Automatic LLM based evaluation.
- Manual examination - identified general issues
- Generated a GoP of evaluated results



LLM-as-a-Judge Evaluation Setup

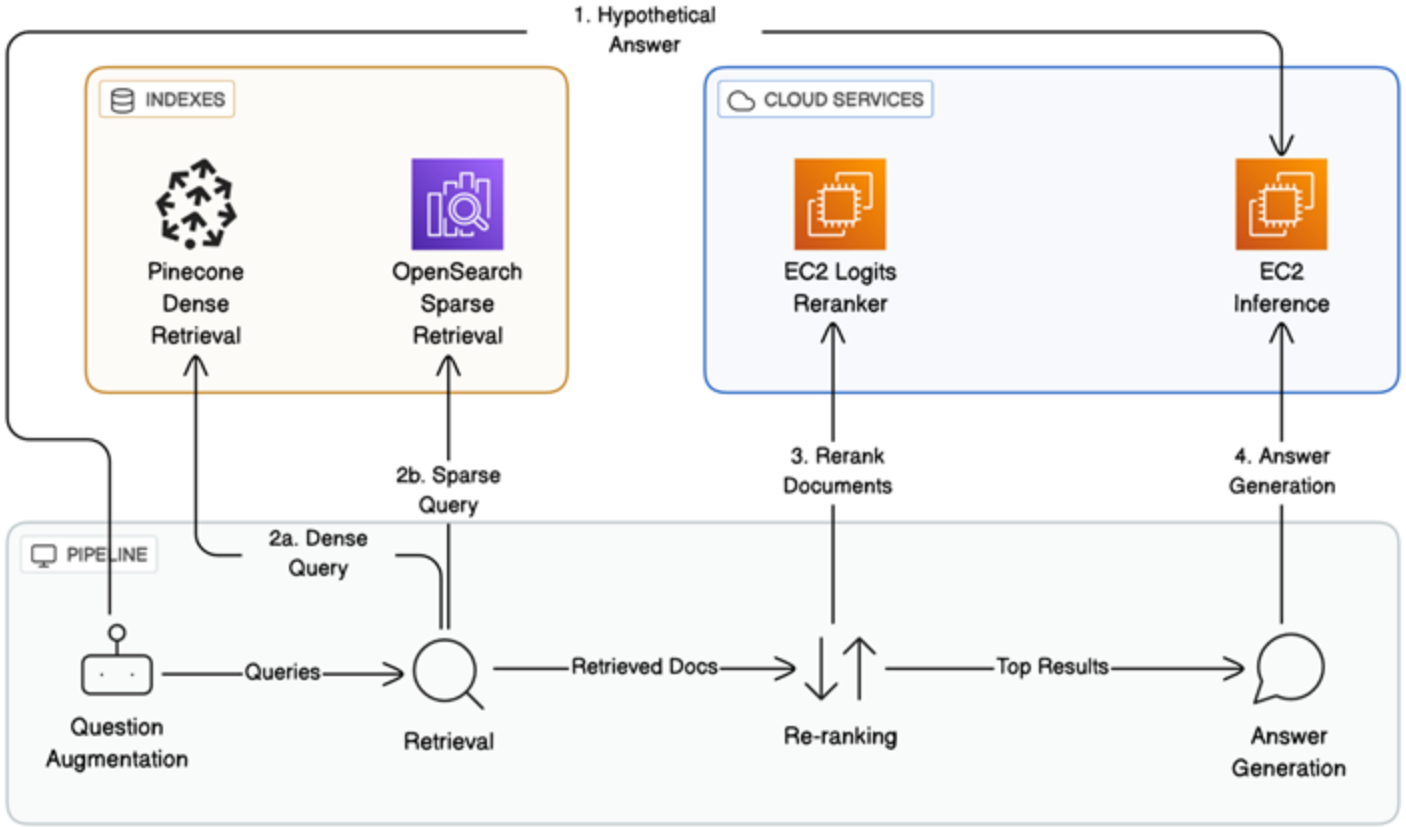
- Applied Claude Sonnet-3.5 for Evaluation
- Correctness (Relevance) Prompt
  - Based on reference answer from DataMorgana
- Faithfulness Prompt
  - Based asked to evaluate based on context documents

Modular RAG Architecture



Gao et al., ACL 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels
Liang et al., TMLR 2023.Holistic Evaluation of Language Models

Pipeline and Services

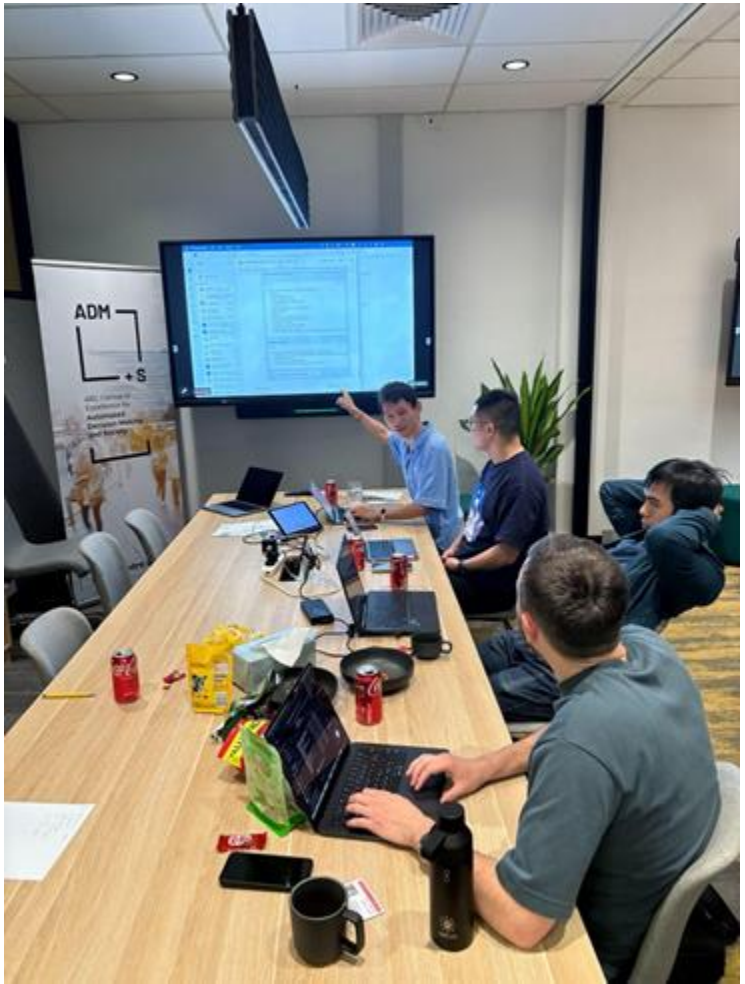


N-Way ANOVA Results

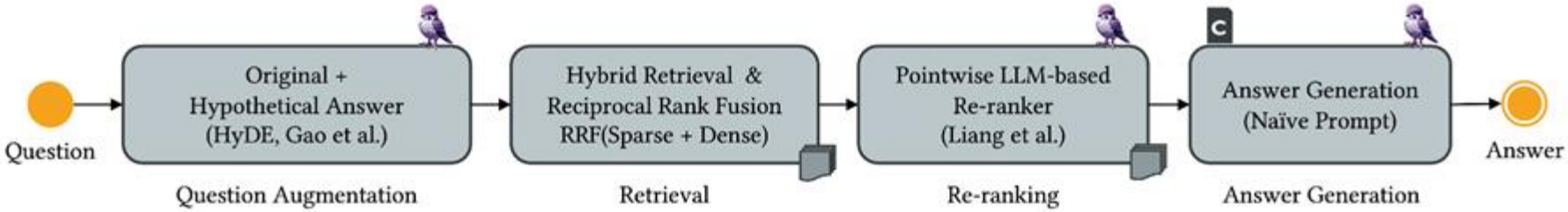
Factor	SS	DF	MS	F	PR(>F)	Parital ω^2
Question Augmentation	0.3193	2	0.1596	125.6920	<0.0001*	0.7571
Query Variants Generation Prompts : Retrieval	0.0062	1	0.0062	4.9099	0.0297*	0.0472
Query Variants Generation Prompts	0.0030	1	0.0030	2.3892	0.1263	0.0173
Context Words Length Limits	0.0025	1	0.0025	2.0052	0.1608	0.0126
Retrieval	0.0024	1	0.0024	1.8631	0.1762	0.0108
Number of Documents Retrieved : Retrieval	0.0020	1	0.0020	1.6100	0.2083	0.0077
Answer Generation Prompts	0.0014	1	0.0014	1.1061	0.2962	0.0013
Context Words Length Limits : Question Augmentation	0.0008	2	0.0004	0.3261	0.7227	-0.0171
Context Words Length Limits : Answer Generation Prompts : Question Augmentation	0.0004	2	0.0002	0.1566	0.8553	-0.0215
Answer Generation Prompts : Question Augmentation	0.0004	2	0.0002	0.1566	0.8554	-0.0215
Query Variants Generation Prompts : Number of Documents Retrieved : Retrieval	0.0001	1	0.0001	0.1143	0.7363	-0.0113
Context Words Length Limits : Answer Generation Prompts	0.0001	1	0.0001	0.0491	0.8253	-0.0122
Query Variants Generation : Number of Documents Retrieved	0.0000	1	0.0000	0.0226	0.8808	-0.0125
Number of Documents Retrieved	0.0000	1	0.0000	0.0118	0.9137	-0.0127
Residual	0.0978	77	0.0013	–	–	–

* indicates statistically significant differences.

Challenge Day



Final Pipeline
G-RAG: Generation-Retrieval-Augmented Generation



Hypothetical answer generation prompt (HyDE)

System Prompt:

Given the question, write a **short hypothetical answer that could be true**. Be brief and concise.

User Prompt:

{question}

Pointwise LLM re-ranker prompt

System Prompt:

You are a helpful assistant that determines if a document contains information that helps answer a given question. **Answer only with 'Yes' or 'No'**.

User Prompt:

Document: {doc_text}
Question: {question}
Does this document contain information that helps answer this question (**only answer 'Yes' or 'No'**)?

Outcome and Results

Outcome and R

Rank	Team ID	Team Name	dness	Quality	LLM-Correctness
1	2615	RMIT-ADMS	**15**	**1.673077**	1.120858
2	2614	RAGtifier	08	1.557692	**1.140128**
3	2596	UDInfo	69	1.548077	1.109643
4	2636	Magikarp	00	1.567308	1.122471
5	2620	Ped100X	08	1.451923	0.787572
6	2611	ScaledRAG	69	1.442308	0.887048
7	2626	HLTCOE	08	1.403846	0.959300
8	2591	Ragmatazz	31	1.230769	0.892480
9	2677	PRMA		.269231	0.801354
10	2668	Hybrid		.125000	0.664194
11	2587	RUC_D		.221154	0.702122
12	2671	Graph-		.240385	0.777058
13	2664	Emorag		.134615	0.854235

Carmel, David, et al. "SIGIR 2025--Li

Key Takeaways

- Robust evaluation gets you far!
  - Fit for purpose: alignment with the goal
  - Statistical analyses through all stages
  - Enabled to prioritize parameter optimization
- Talented and diversely skilled team.

Future work:

- Dynamic question augmentation
- Model distillation and fine-tuning

RMIT-ADM+S @ LiveRAG 2025

Code: rmit-ir/GRAG-LiveRAG
Paper: bit.ly/GRAG-LiveRAG-paper

LiveRAG Challenge Report by Organizers:
https://arxiv.org/abs/2507.04942