UTS Group 24

NLP ASSIGNMENT 3

Organising a text corpus using K-means clustering and applying retrieval-augmented generation (RAG) to a chatbot

Warit Boonmasiri - 25399522
Mohammad Alhajjeh - 24832800
Jada Gamis - 24823787
Yingrong Zhang - 25428842
Aiden Blishen Cuneo - 24971160

Overview

Problem

Our project addresses the problem of the inability to keep up with an influx of information across a range of scientific fields, using natural language processing (NLP) techniques introduced in this course to visualise and organise scholarly literature.

Aim

Clustering a text corpus into related groups, then use RAG to allow a chatbot to use one of these clusters as context. The goal is for the user to be able to swap between these clusters very quickly, allowing them to decide which specific context they want to give to the chatbot.

Solution

- Use K-means clustering on the dataset
- Allow the user to swap between clusters to choose the context for RAG
- The model generates a response based on this context

Dataset

Group 24

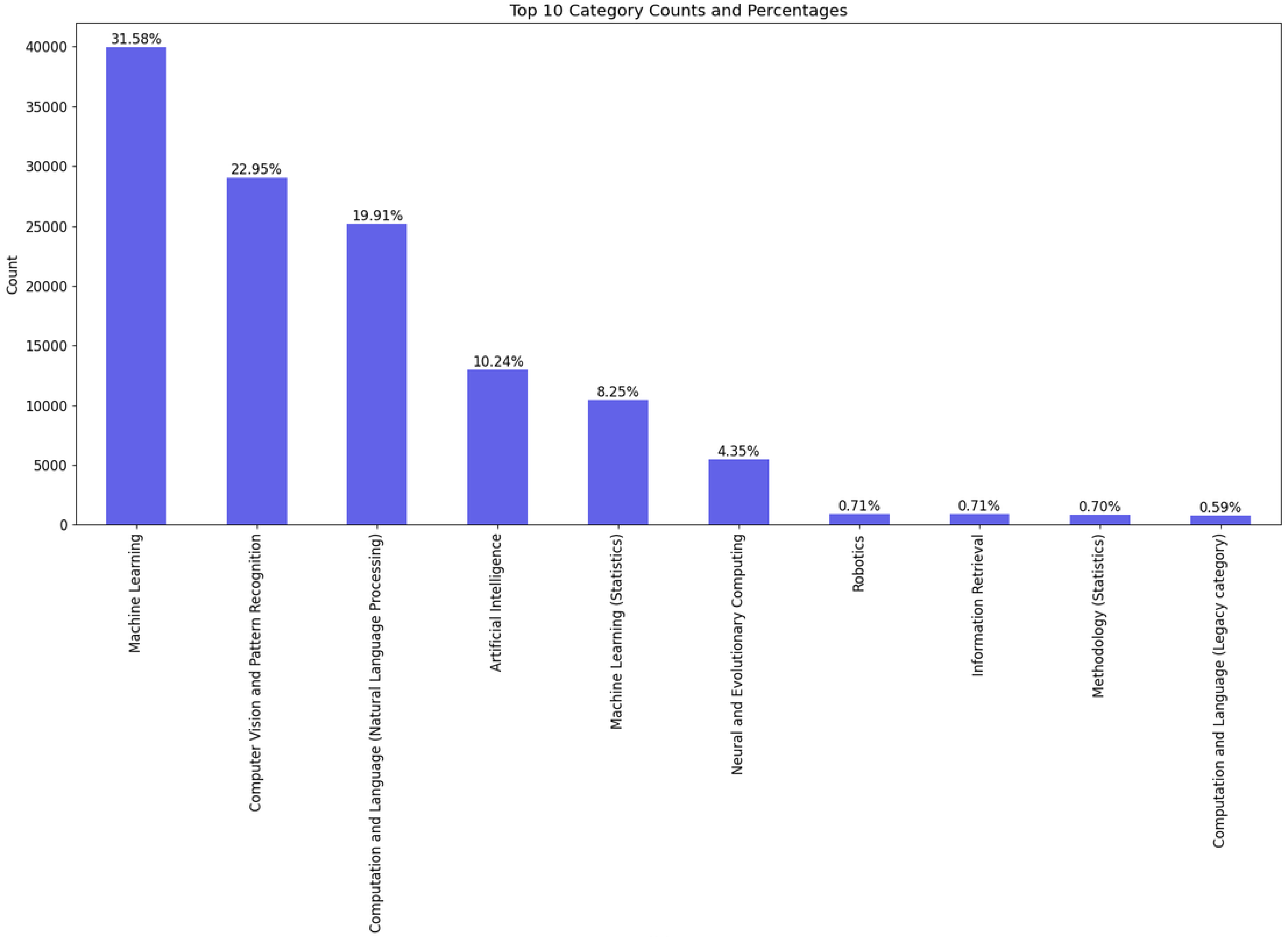
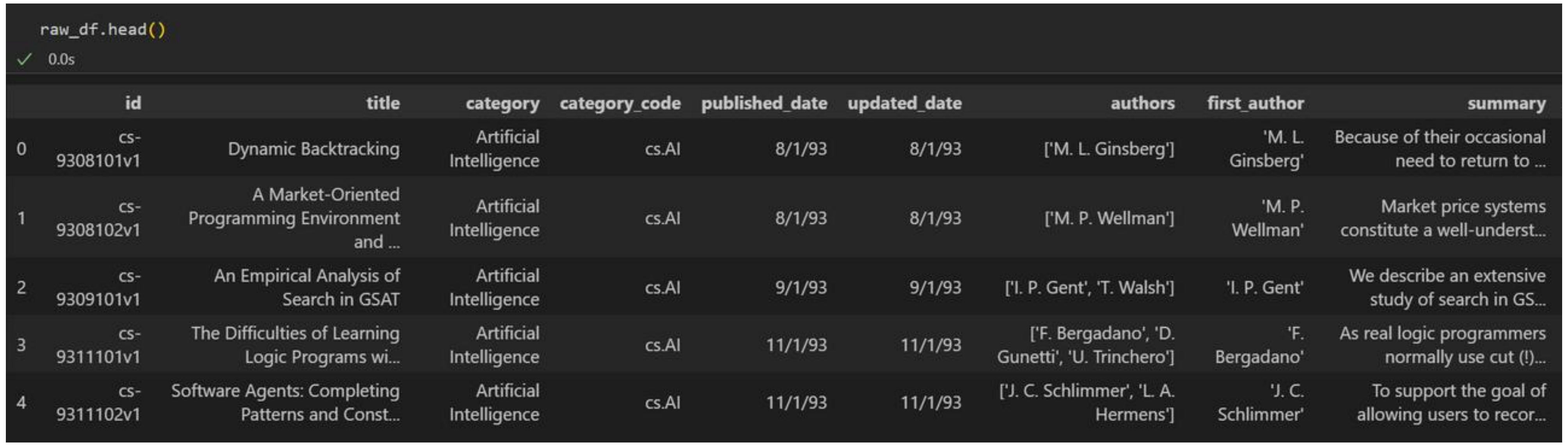


Figure 1: Bar graph showing article categories and their frequencies



```
raw_df.head()
✓ 0.0s
```

	id	title	category	category_code	published_date	updated_date	authors	first_author	summary
0	cs-9308101v1	Dynamic Backtracking	Artificial Intelligence	cs.AI	8/1/93	8/1/93	['M. L. Ginsberg']	'M. L. Ginsberg'	Because of their occasional need to return to ...
1	cs-9308102v1	A Market-Oriented Programming Environment and ...	Artificial Intelligence	cs.AI	8/1/93	8/1/93	['M. P. Wellman']	'M. P. Wellman'	Market price systems constitute a well-underst...
2	cs-9309101v1	An Empirical Analysis of Search in GSAT	Artificial Intelligence	cs.AI	9/1/93	9/1/93	['I. P. Gent', 'T. Walsh']	'I. P. Gent'	We describe an extensive study of search in GS...
3	cs-9311101v1	The Difficulties of Learning Logic Programs wi...	Artificial Intelligence	cs.AI	11/1/93	11/1/93	['F. Bergadano', 'D. Gunetti', 'U. Trinchero']	'F. Bergadano'	As real logic programmers normally use cut (!)...
4	cs-9311102v1	Software Agents: Completing Patterns and Const...	Artificial Intelligence	cs.AI	11/1/93	11/1/93	['J. C. Schlimmer', 'L. A. Hermens']	'J. C. Schlimmer'	To support the goal of allowing users to recor...

Figure 2: Sample rows from the chosen dataset

- arXiv Scientific Research Papers Dataset
- Contains over 100,000 articles
- Topics such as Artificial Intelligence, Machine Learning, Computer Science, etc.

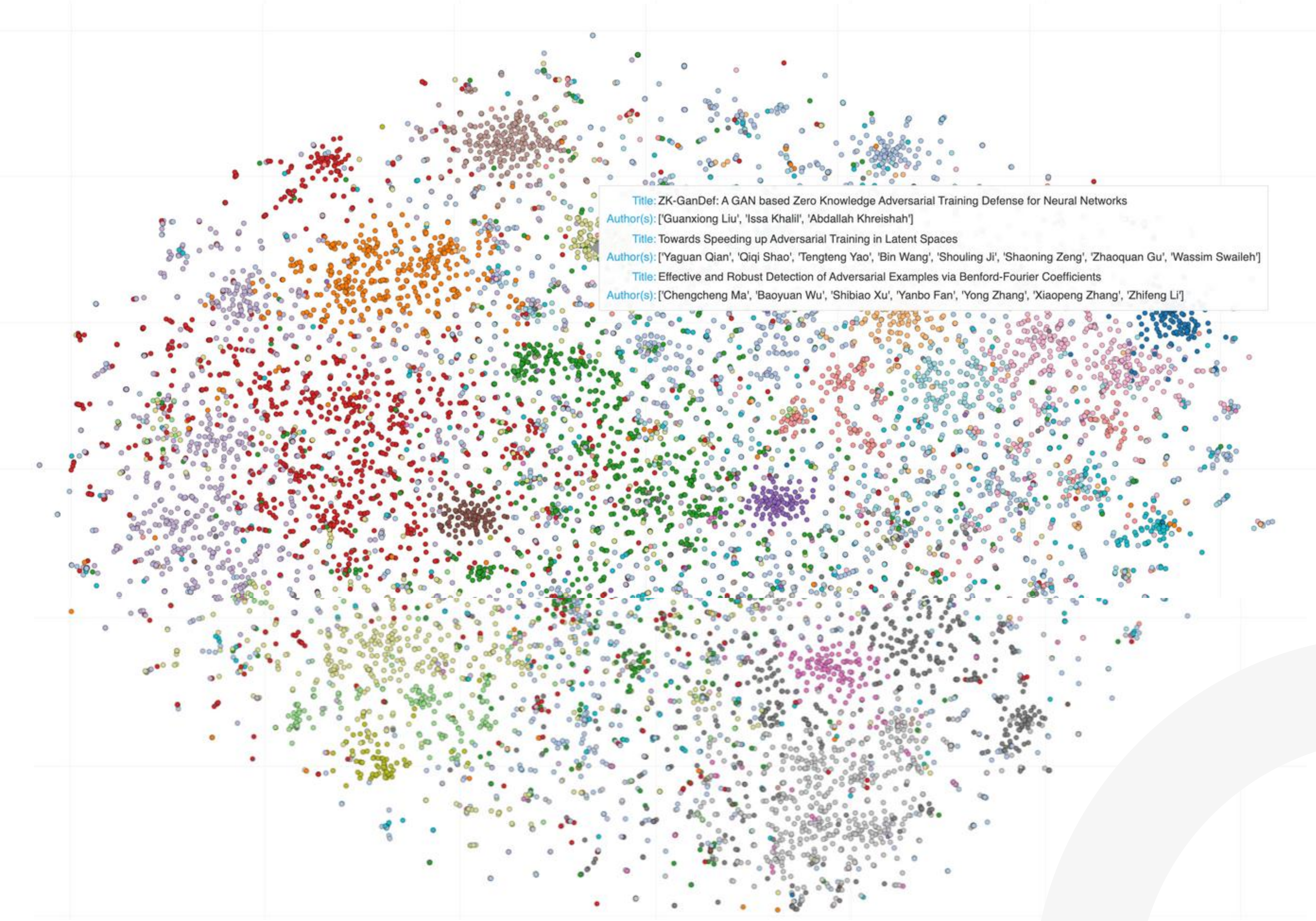
Clustering

01 How clustering was performed

- Convert text summaries to vector representation
- Apply PCA to reduce dimensionality
- Use K means clustering
- Do final dimension reduction to visualise in 2D

02 Why cluster the data?

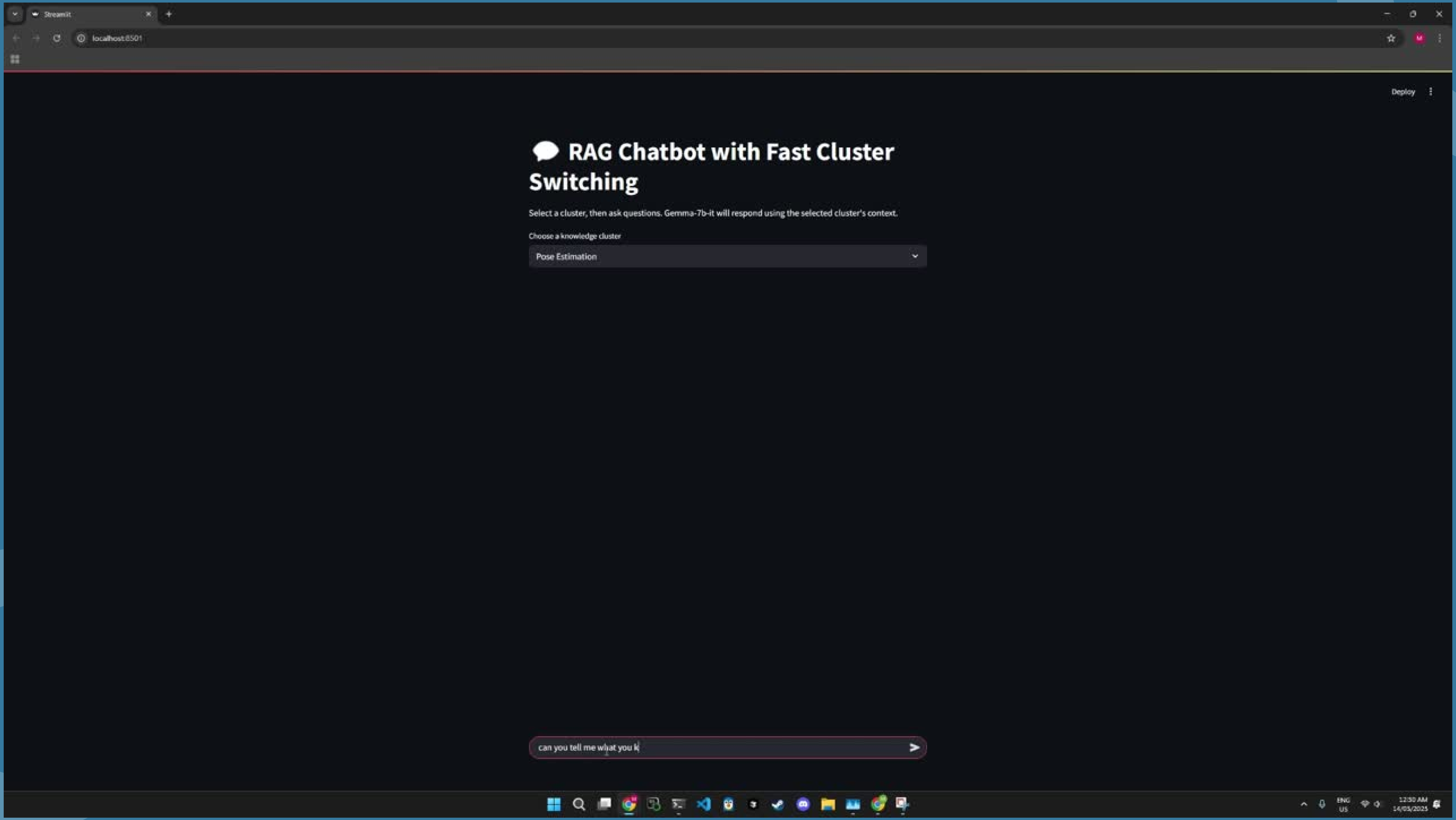
- Clustering splits the summaries into more specific areas
- RAG can be used on these topics
- Performance boost for a smaller set of summaries



Title: ZK-GanDef: A GAN based Zero Knowledge Adversarial Training Defense for Neural Networks
Author(s): ['Guanxiong Liu', 'Issa Khalil', 'Abdallah Khreishah']
Title: Towards Speeding up Adversarial Training in Latent Spaces
Author(s): ['Yaguan Qian', 'Qiqi Shao', 'Tengteng Yao', 'Bin Wang', 'Shouling Ji', 'Shaoning Zeng', 'Zhaoquan Gu', 'Wassim Swaileh']
Title: Effective and Robust Detection of Adversarial Examples via Benford-Fourier Coefficients
Author(s): ['Chengcheng Ma', 'Baoyuan Wu', 'Shibiao Xu', 'Yanbo Fan', 'Yong Zhang', 'Xiaopeng Zhang', 'Zhifeng Li']

Demo of UI

Group 24



💬 RAG Chatbot with Fast Cluster Switching

Select a cluster, then ask questions. Gemma-7b-it will respond using the selected cluster's context.

Choose a knowledge cluster

Pose Estimation

can you tell me what you k

UTS Group 24

THANK YOU!