NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Scaling Behaviour in Retrieval-Augmented Generation

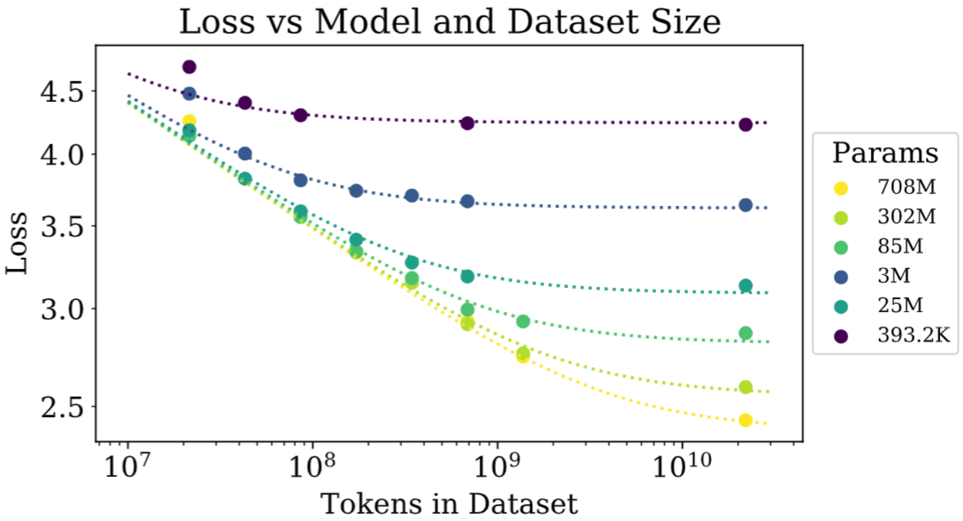
Xinghe Chen
Yihua Hu
Hongrui Liu

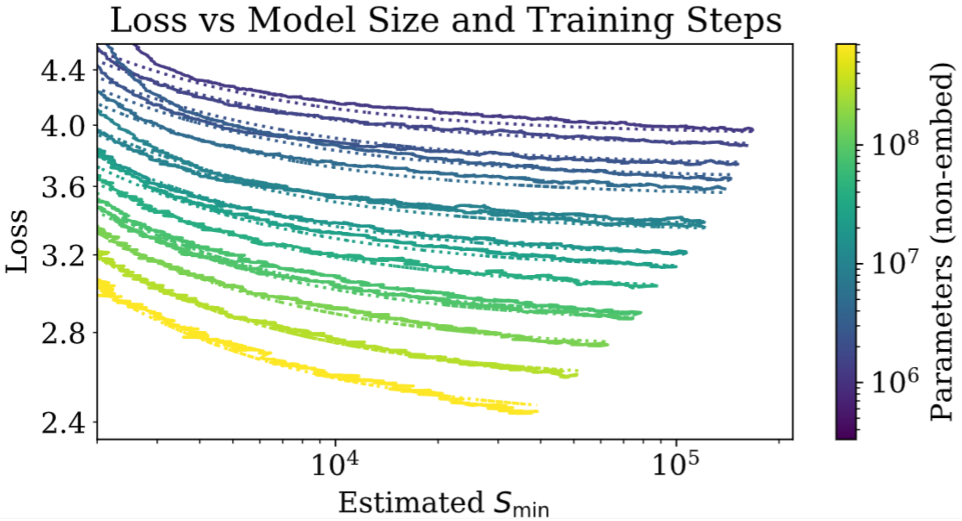
Background: Scaling Law

A scaling law describes how the performance of a system responses predictably as its variables scale.

Scaling law has been extensively studied in **large language models(LLMs):**

- **Model Size:** Increasing the number of parameters enhances LLM performance.
- **Dataset Size:** Expanding the training dataset improves LLM performance.
- **Computational Power:** Allocating more computational resources during training enhances performance.

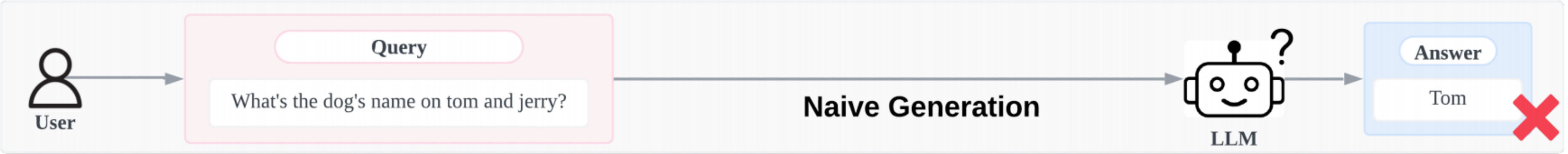




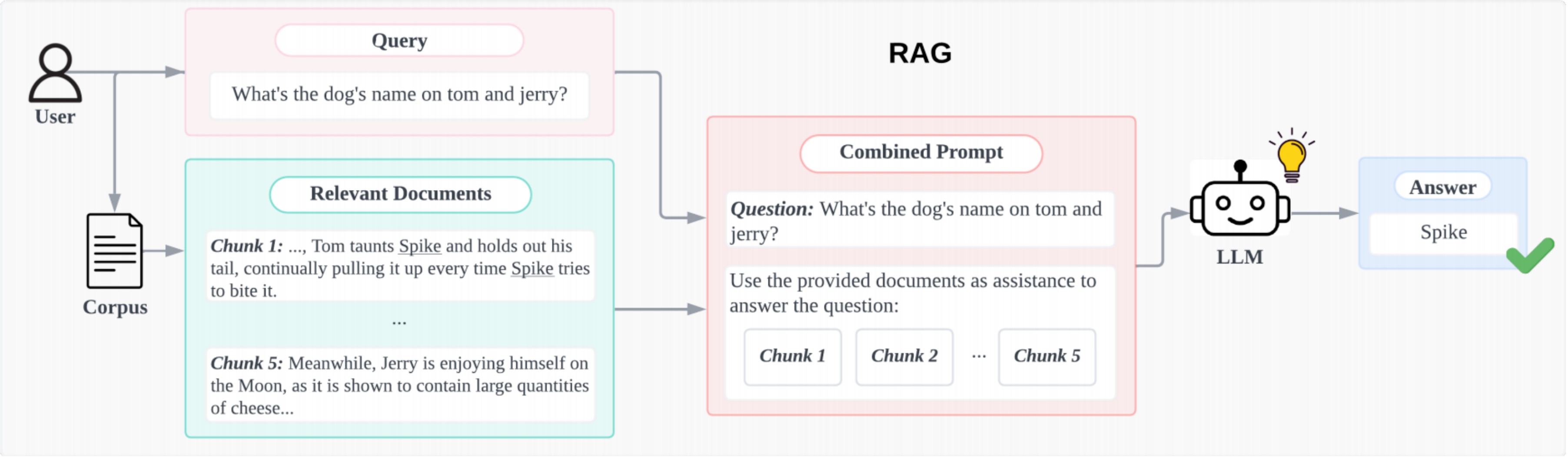
Background: RAG System

A Retrieval-Augmented Generation (RAG) system enhances model accuracy by grounding responses in an external knowledge database.

Naive Generation -> Incorrect or incomplete answers for specific or niche queries



RAG System -> Grounded in external knowledge for improved accuracy


RAG
User
Query
What's the dog's name on tom and jerry?
Corpus
Relevant Documents
Chunk 1: ..., Tom taunts Spike and holds out his tail, continually pulling it up every time Spike tries to bite it.
...
Chunk 5: Meanwhile, Jerry is enjoying himself on the Moon, as it is shown to contain large quantities of cheese...
Combined Prompt
Question: What's the dog's name on tom and jerry?
Use the provided documents as assistance to answer the question:
Chunk 1
Chunk 2
...
Chunk 5
LLM
Answer
Spike

Motivation & Challenges

Is there also a scaling law for RAG system?

- **Ambiguity in Definition:** The scaling behaviour for RAG systems lacks clear definition.

  -> We focus on the scaling law associated with the **size of the external database**

- **Evaluation Challenge:** Assessing scaling behaviour is difficult.

  -> We evaluate the performance of RAG systems using three metrics:

  exact match (EM), F1-score, and accuracy

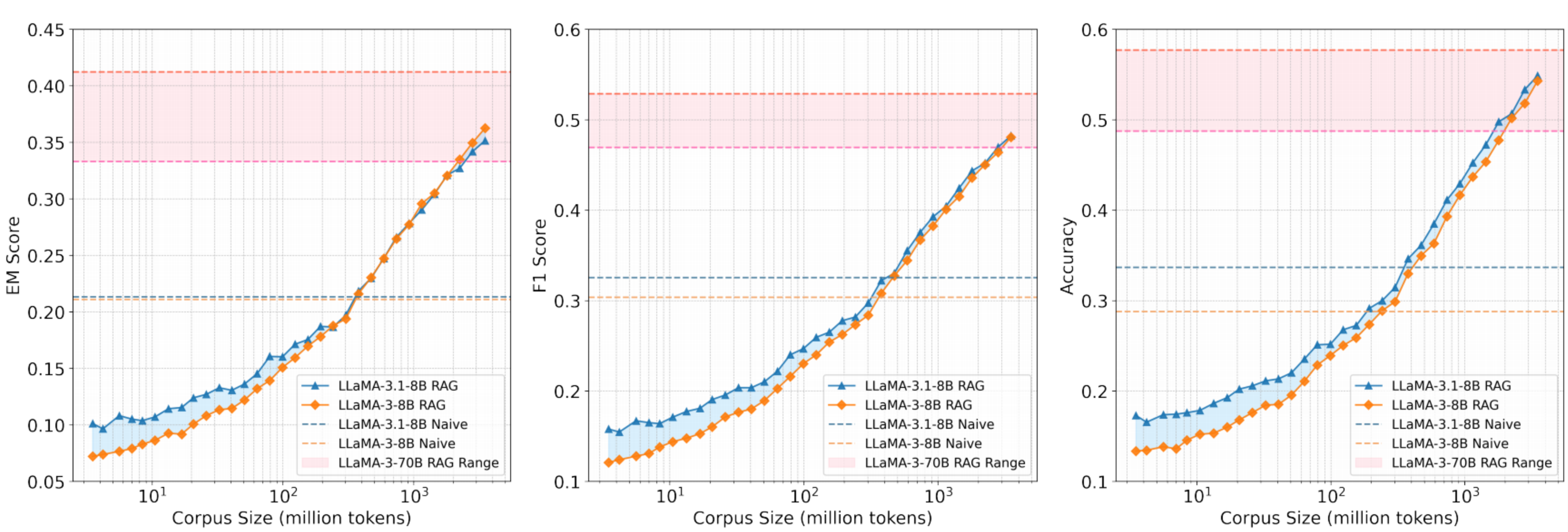
- **Implementation Difficulty:** RAG systems are usually black box in research.

  -> We utilise the open-source and user-friendly **FlashRAG** Python toolkit

Experiment

Experiment Setup:

- Test Set: Natural Questions dataset
- Knowledge Database: Wikipedia dump
- Model: LLaMA-3-8B, LLaMA-3.1-8B, LLaMA-3-70B

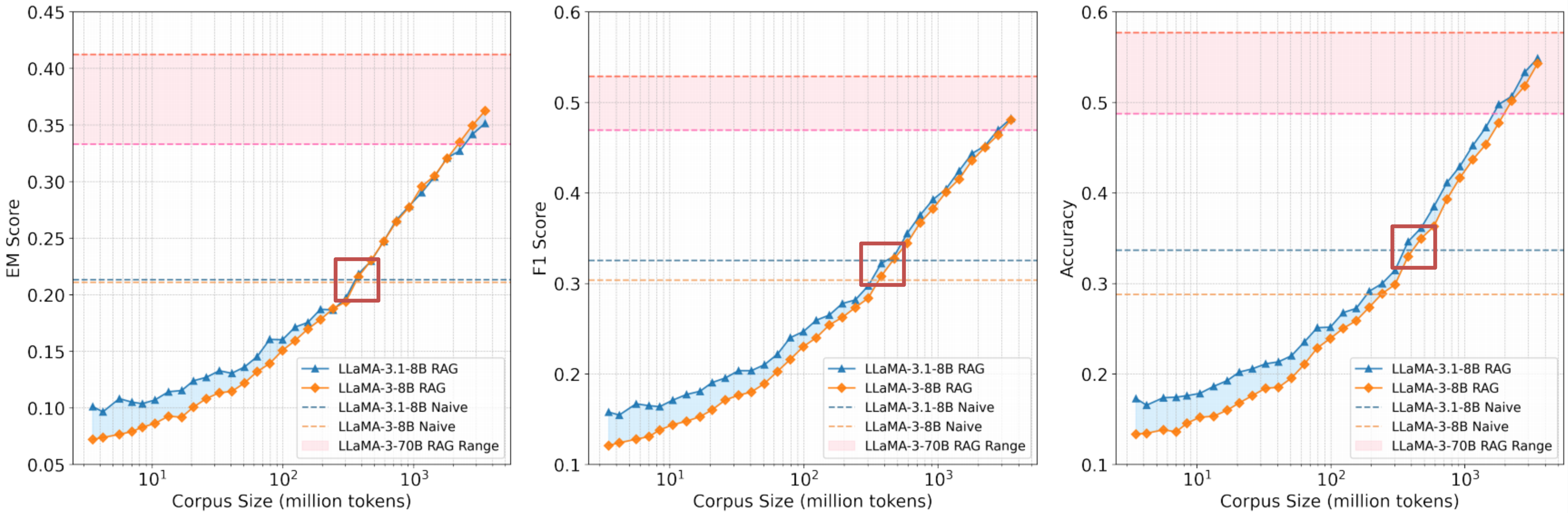


(a) EM score (b) F1 score (c) Accuracy

Experiment

Observation:

- Turning point thresholds exist in RAG systems
- RAG provides greater compensation for less powerful models
- RAG system performance increases linearly as the corpus size grows exponentially



(a) EM score

(b) F1 score

(c) Accuracy

Conclusion & Limitation

Conclusion:

- Empirical evidence shows that scaling laws are applicable to RAG systems and thereby making the performance of it predictable.

Limitations:

- This project is constrained by computational resources.
- Future work may focus on varying the
  - ➢ RAG pipeline
  - ➢ task types
  - ➢ model size
  - ➢ corpus size

Thank you for your attention

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE