

# **Beyond Scale: Performance Plateau and Cost Optimization in RAG Systems**

Research Paper Presentation



**plom04@vse.cz**



PhD Student (1<sup>st</sup> year) in Applied Informatics at VSE Prague, researching cloud-based infrastructures and RAG systems

PhD Student (1<sup>st</sup> year) in Applied Informatics at VSE Prague, researching cloud-based infrastructures and RAG systems

16+ years leading data and AI teams at Amazon, managing global technical organizations

16+ years leading data and AI teams at Amazon, managing global technical organizations

Cloud architectures, ML products, AI governance | AWS ML Specialty, IAPP AI Governance

Cloud architectures, ML products, AI governance | AWS ML Specialty, IAPP AI Governance

# The Challenge

## Large Language Models face two key limitations:

- Cannot keep pace with rapidly generated new information
- Lack access to private knowledge bases

## Retrieval-Augmented Generation (RAG)

Enriches prompts with relevant information from external knowledge sources to generate grounded responses.

# Research Questions

- **RQ1:** Impact of LLM model size on RAG system response quality?
- **RQ2:** Is there a performance plateau threshold in model size?
- **RQ3:** Impact of model age on RAG system quality?
- **RQ4:** Impact of knowledge source generation methodology?
- **RQ5:** Are there cost-effective alternatives to largest models?

# Methodology

## Models Evaluated

32

Large Language Models

0.6B to 1T parameters

Sept 2023 - Aug 2025

## RAG Architectures

- **RAG-FAISS:**

Vector DB

chunked retrieval

- **RAG-FULL WIKI:**

Full-text provision

- **RAG-CONTEXT:**

Curated excerpts

## Dataset

### BeleBele Benchmark

900 multiple-choice questions

Weighted accuracy metric

## Dataset

| Reference source                                                                                                | Question                                                                             | Answers                                 |
|-----------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|-----------------------------------------|
| <a href="https://en.wikivoyage.org/wiki/Snow_safety">https://en.wikivoyage.org/wiki/Snow_safety</a>             | Which of the following is not a trigger for avalanches?                              | 1)Sticky snow                           |
|                                                                                                                 |                                                                                      | 2)Humans                                |
|                                                                                                                 |                                                                                      | 3)Sunshine                              |
|                                                                                                                 |                                                                                      | 4)Additional snowfall                   |
| <a href="https://en.wikivoyage.org/wiki/French_phrasebook">https://en.wikivoyage.org/wiki/French_phrasebook</a> | Which of the following is not the same in France as it is in Belgium or Switzerland? | 1)The pronunciation of all words        |
|                                                                                                                 |                                                                                      | 2)The numbering system                  |
|                                                                                                                 |                                                                                      | 3)The standard French taught in schools |
|                                                                                                                 |                                                                                      | 4)The spelling of some French words     |

Table 2. Sample questions from the BeleBele dataset [Bandarkar et al., 2023]. Each question includes four response options with reference sources derived from Wikipedia content.

# LLMs examples

| Model Name            | Developer  | Parameters | Precision | Release |
|-----------------------|------------|------------|-----------|---------|
| Qwen-3-0.6B           | Qwen.ai    | 0.6B       | BF16      | May-25  |
| Gemma-3-1B-Instruct   | Google     | 1B         | BF16      | Mar-25  |
| Llama-3.2-1B-Instruct | Meta-llama | 1B         | BF16      | Sep-24  |
| Falcon-3-1B-Instruct  | Tiiuae     | 1B         | BF16      | Dec-24  |
| Gemma-2-2B-Instruct   | Google     | 2B         | BF16      | Jul-24  |
| Llama-3.2-3B-Instruct | Meta-llama | 3B         | BF16      | Sep-24  |
| Falcon-3-3B-Instruct  | Tiiuae     | 3B         | BF16      | Dec-24  |
| Qwen-3-4B             | Qwen.ai    | 4B         | BF16      | May-25  |
| Gemma-3-4B-Instruct   | Google     | 4B         | BF16      | Mar-25  |

# LLMs examples

| Model Name                  | Developer    | Parameters | Precision | Release |
|-----------------------------|--------------|------------|-----------|---------|
| Mistral-Large-Instruct      | Mistral.ai   | 123B       | BF16      | Feb-24  |
| Qwen-3-235B-A22B-Instruct   | Qwen.ai      | 235B       | BF16      | May-25  |
| Llama-Nemotron-Ultra-253B   | Nvidia       | 253B       | BF16      | Apr-25  |
| GLM-4.5                     | Zaiorg       | 358B       | BF16      | Aug-25  |
| Jamba-Large-1.5             | AI21         | 399B       | BF16      | May-24  |
| Llama-3.1-405B-Instruct-FP8 | Meta-llama   | 405B       | FP8       | Jul-24  |
| Hermes-4-405B               | NousResearch | 405B       | BF16      | Aug-25  |
| Kimi-K2-Instruct            | Moonshotai   | 1000B      | BF16      | Jan-25  |

# Three RAG Architectures Tested

## **RAG-FAISS: Vector Database Retrieval**

Articles chunked into 500-word segments, embedded, stored in FAISS vector database.

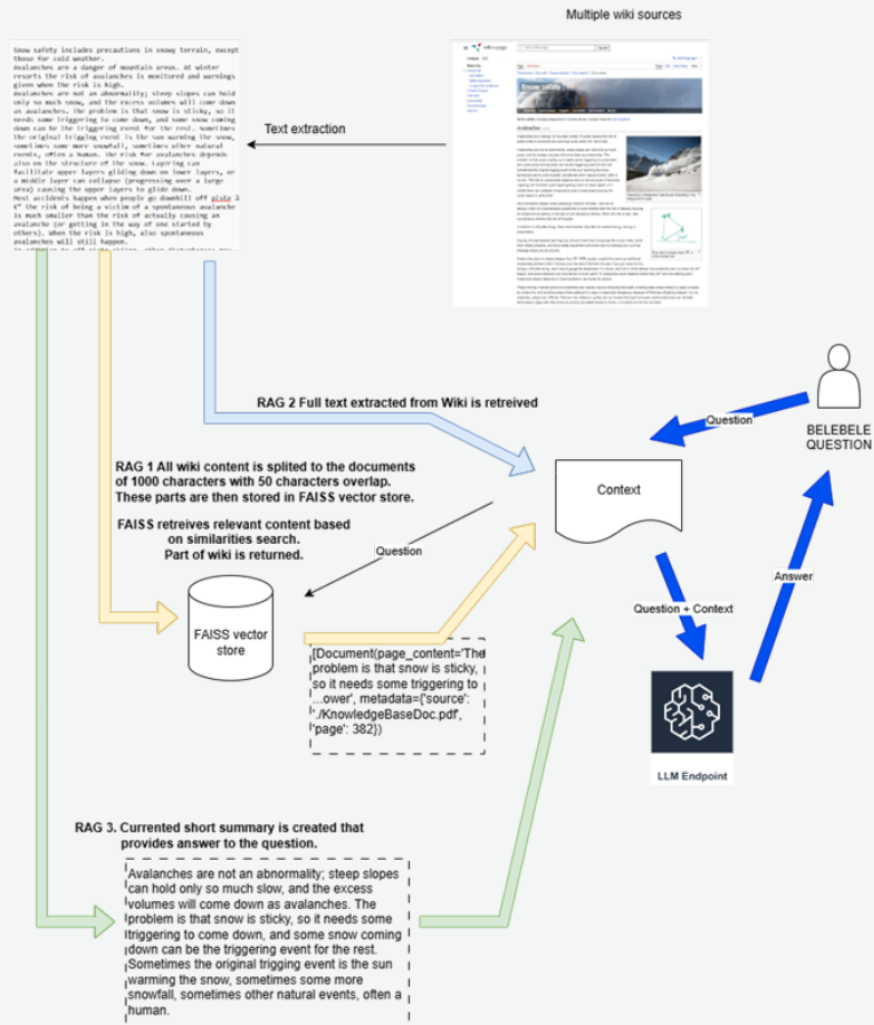
## **RAG-FULL WIKI: Complete Text Provision**

Entire Wikipedia article provided to the LLM without chunking.

## **RAG-CONTEXT: Curated Excerpt Architecture**

Manually curated, concise excerpt containing the answer.

# Three RAG Architectures Tested

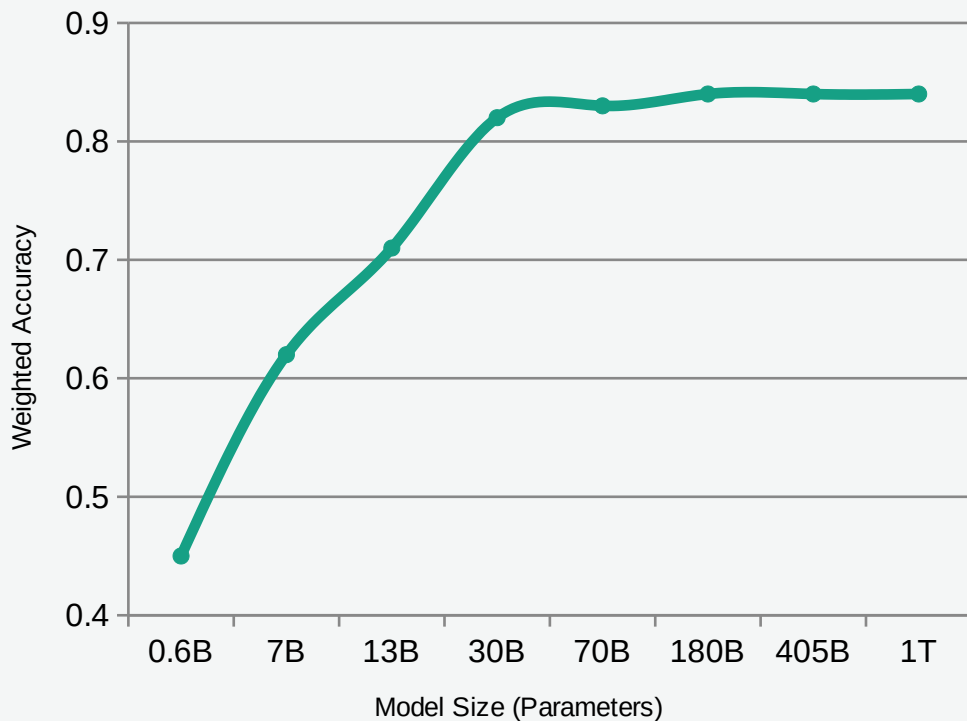


# Key Finding: Performance Plateau at 30B Parameters

## Critical Threshold

**~30B**

Beyond this point, increases in model size yield diminishing returns for RAG system performance.

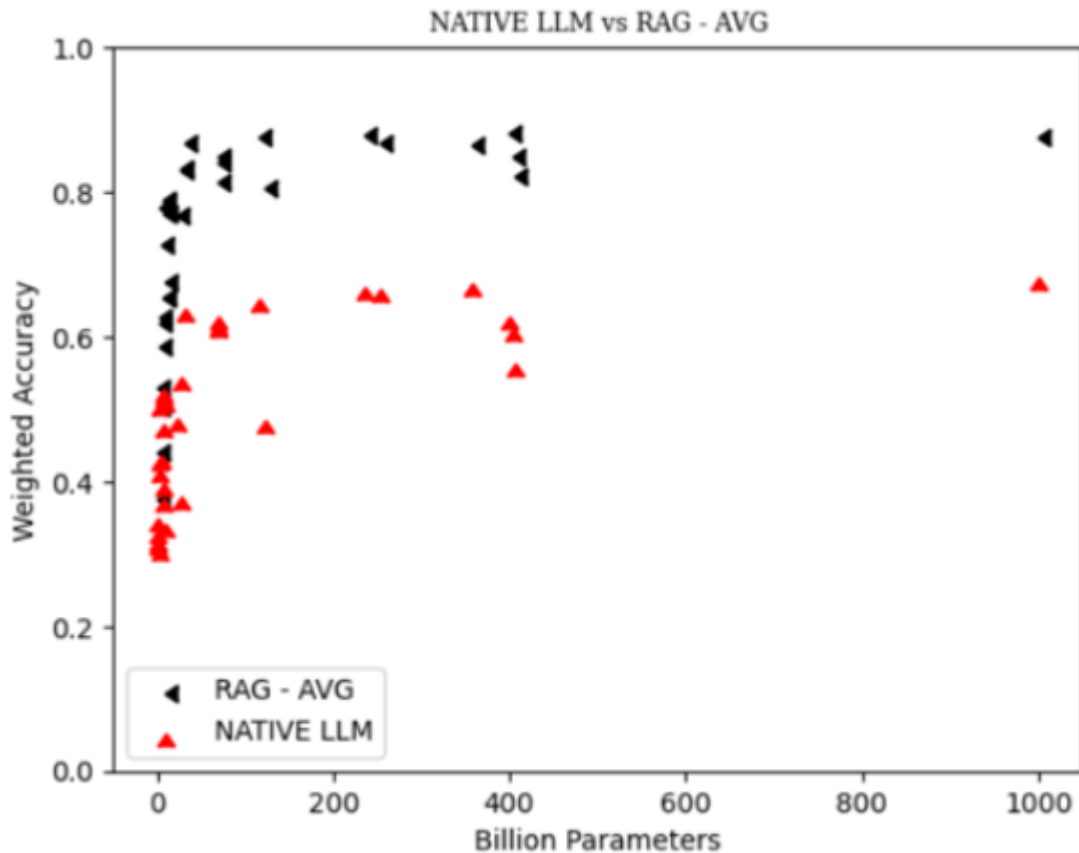


# Key Finding: Performance Plateau at 30B Parameters

Critical Threshold

~30B

Beyond this point, increases in model size yield diminishing returns for RAG system performance.

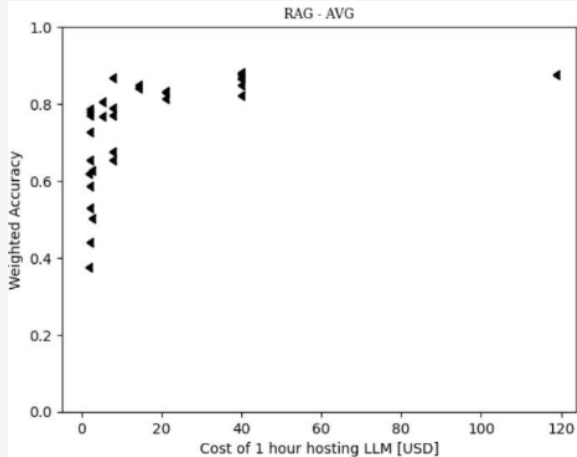


# Cost vs. Performance Trade-off

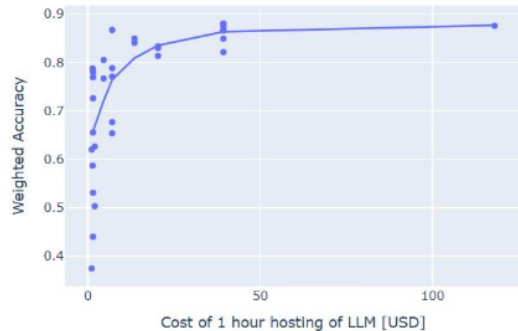
## Potential Savings

**60%**

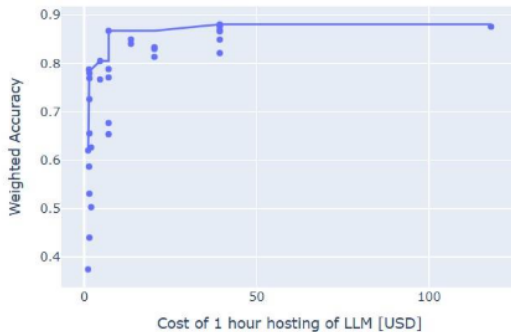
Cost reduction achievable by deploying optimally-sized models at 30B threshold vs. largest models.



RAG-AVG with LOESS REGRESSION



RAG-AVG with EXPANDING MAXIMUM



RAG-AVG with EWMA CURVE



# Additional Key Insights

## Temporal Invariance

Model release date showed minimal impact on RAG performance when controlling for parameter count. Newer models do not invariably outperform predecessors in retrieval-augmented contexts.

## Knowledge Management Quality Matters Most

Curated excerpt architecture (RAG-CONTEXT) consistently outperformed both vector database retrieval and full-text provision methods. Content curation beats infrastructure sophistication.

# Practical Recommendations for Practitioners

1

## **Optimize for the Plateau**

Prioritize models around 30B parameters rather than defaulting to largest available models.

2

## **Do not overestimate Model Recency**

Model release date should not be a primary selection criterion for RAG deployments.

3

## **Explore knowledge extracting methodologies**

Content curation still stays most performing method for the accuracy of answer.

4

## **Reassess Scaling Assumptions**

Challenge assumptions that equate model size and recency with quality in RAG contexts.

# Conclusions

RAG system performance plateaus at approximately **30 billion parameters**, enabling up to **60% cost reductions** through strategic model selection.

Model recency shows minimal impact on performance. Knowledge management quality consistently outperforms retrieval infrastructure sophistication.

**Impact:** These findings enable more economically rational and environmentally sustainable AI system deployment strategies.

# Thank You



## Questions?

Beyond Scale: Performance Plateau and Cost Optimization in RAG Systems