CIKLUM

Productized & Enterprise RAG

October 2025

Speaker

Lucian Gruia

Data Science Principal Lead

- Head of Engineering, Romania
- 13+ years of experience in Software Engineering
- Led cross-functional teams across 8 time zones in Telecom, Fintech, Aerospace, and MedTech
- Mentor @Google Developer Groups Romania
- Research Assistant, AI Multimedia Lab (National University of Science & Technology, Bucharest)
- Supervised 12 BSc and MSc diploma projects

Agenda

01 Existing repos & products

02 Pros, Cons, Limitations and Considerations

CIKLUM

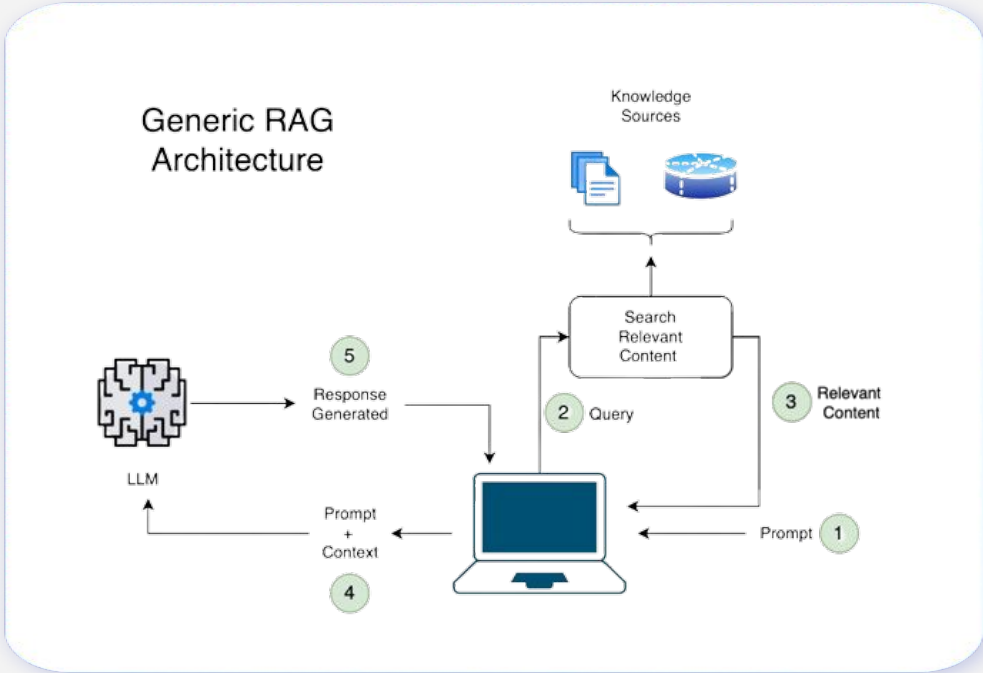
Quick recap

What is RAG

A RAG system essentially correlates a **user's prompt** with a relevant **data chunk.** It does this by identifying **the most semantically similar** chunk from the database.

This chunk then becomes **the context** for the prompt.

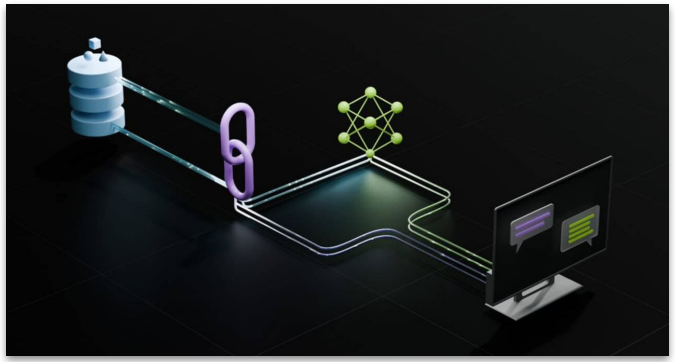
When **passed to the Large Language Model (LLM),** it enables the system to provide a relevant answer within the given context.



Why do we need RAG?

CIKLUM

- **Expands Knowledge Base**
  RAG accesses a vast external database, enriching its knowledge beyond initial training data
- **Improves Accuracy**
  Enhances response precision by integrating relevant, real-time information
- **Adaptable**
  Effectively handles novel and niche queries
- **Increases Efficiency**
  Streamlines information retrieval and generation process
- **Versatile Applications**
  Useful across various fields, from customer support to research



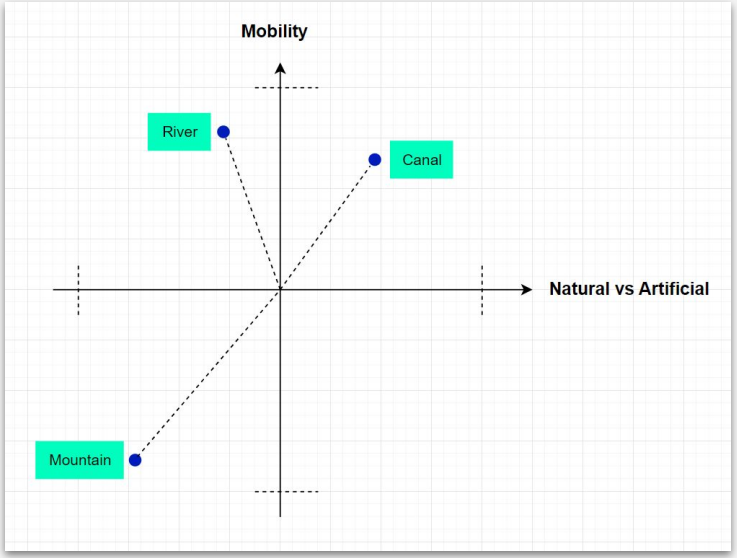
Source: What Is Retrieval-Augmented Generation, aka RAG?

LLMs are limited to the knowledge they were originally trained on, and retraining or fine-tuning them is slow and expensive.

Embeddings. Similarity

CIKLUM

- **Embeddings**
  Numerical representations of concepts, in a high-dimensional space, capturing semantic meaning.
- **Similarity**:
  - Lexical: entities are alike in appearance
  - **Semantic**: entities are alike in meaning
- In RAG we represent entities **by describing** them.
  This is a form of **knowledge representation**.



Read more: Wikipedia - Cosine Similarity

Example: Mountain, River, Canal

One hot encoding

Mountain:	1
River:	2
Canal:	3

2-Dimensional Space

	[Natural vs Artificial, Mobility]
Mountain:	[-0.7, -0.8]
River:	[-0.3, 0.7]
Canal:	[0.4, 0.5]

Multilinguality

CIKLUM

Generated with DALL·E 3

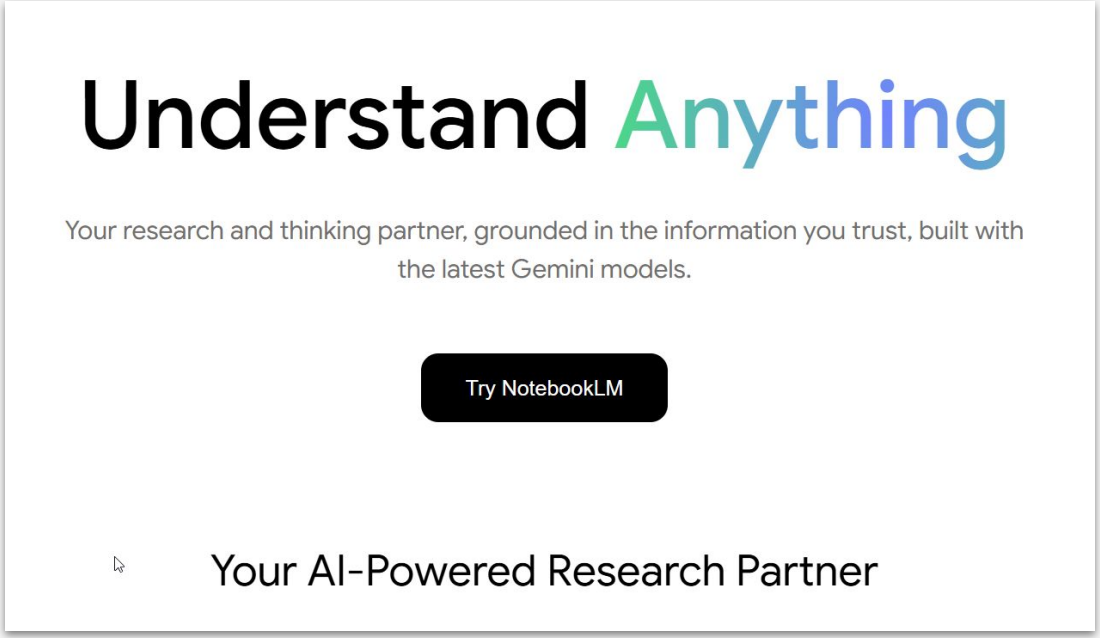
- **Multilingual Capabilities**
  RAG and chatbots can process and respond in multiple languages, enhancing global reach and accessibility.
- **Cross-lingual Understanding**
  These systems understand queries in one language and retrieve information in another, facilitating cross-lingual interactions.
- **Language-Agnostic Embeddings**
  Use of embeddings that capture semantic meanings across languages.
- **Challenges**
  Handling idiomatic expressions, cultural nuances, and varying sentence structures in different languages.

CIKLUM

Existing repos & products

NotebookLM

CIKLUM



GEMs

CIKLUM

Gemini

Build custom experts with Gems

Gems are your custom AI experts for help on any topic. Gems can be anything from a career coach or brainstorm partner to a coding helper. Get started with our suite of premade Gems, or build your own custom Gems tailored to your unique needs.

Custom GPTs

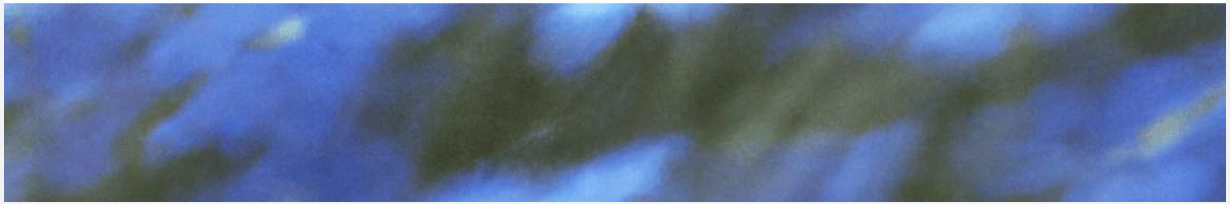
CIKLUM

OpenAI

November 6, 2023 Product

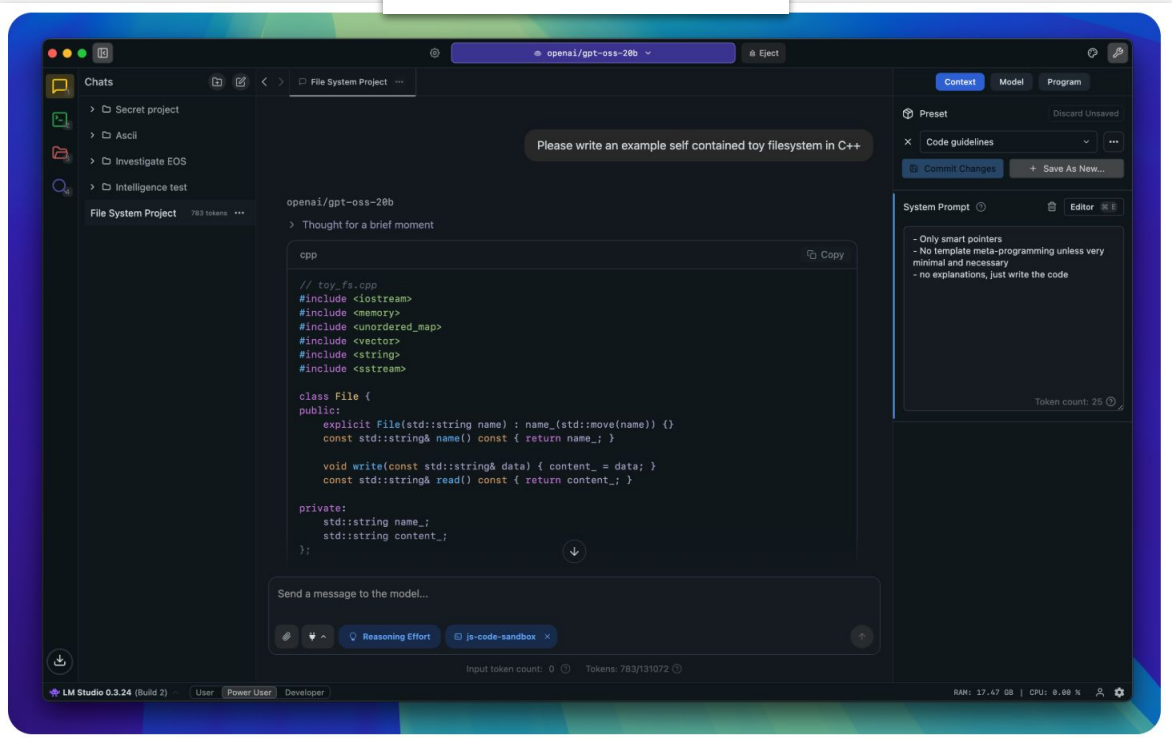
Introducing GPTs

You can now create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.



LM Studio

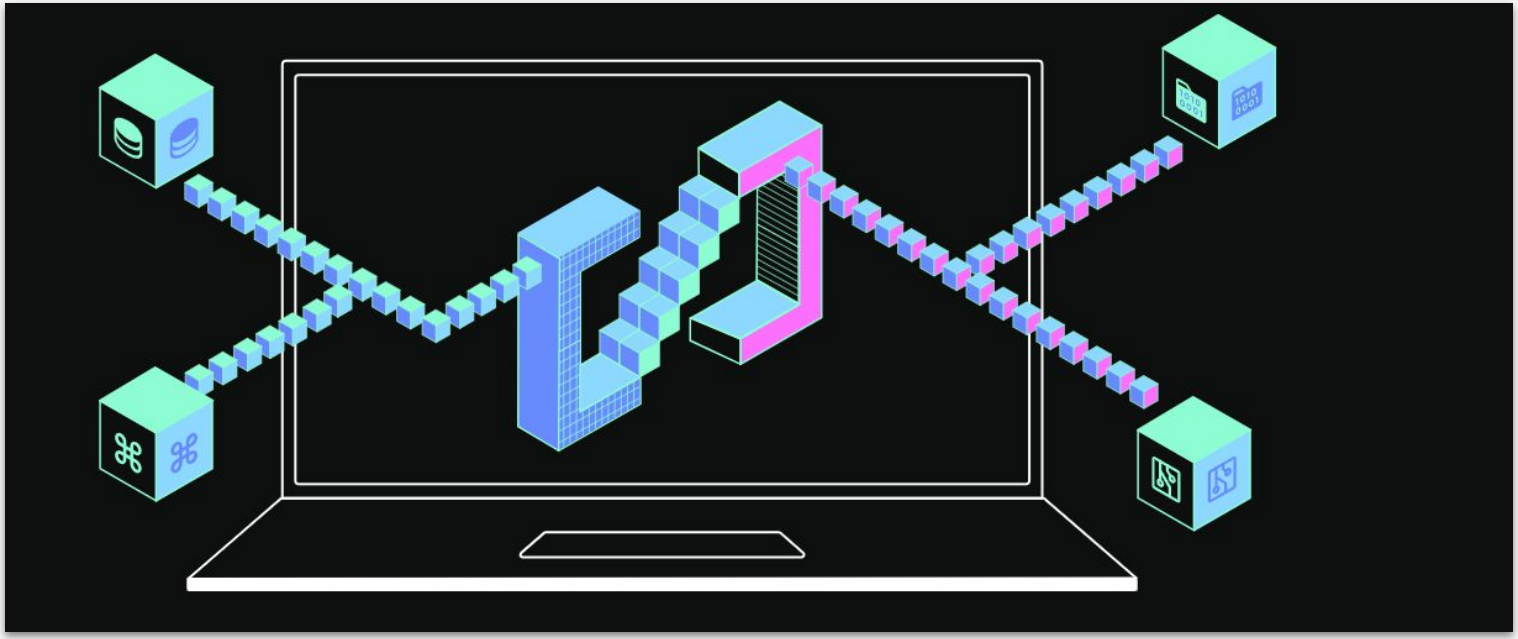
CIKLUM



Anything LLM

CIKLUM

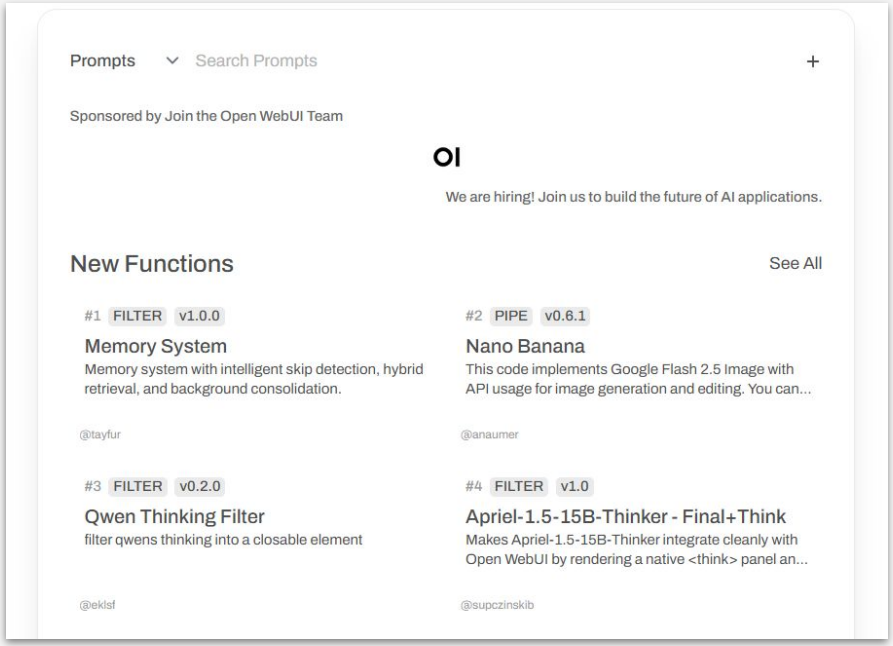
Anything LLM



OpenWeb UI

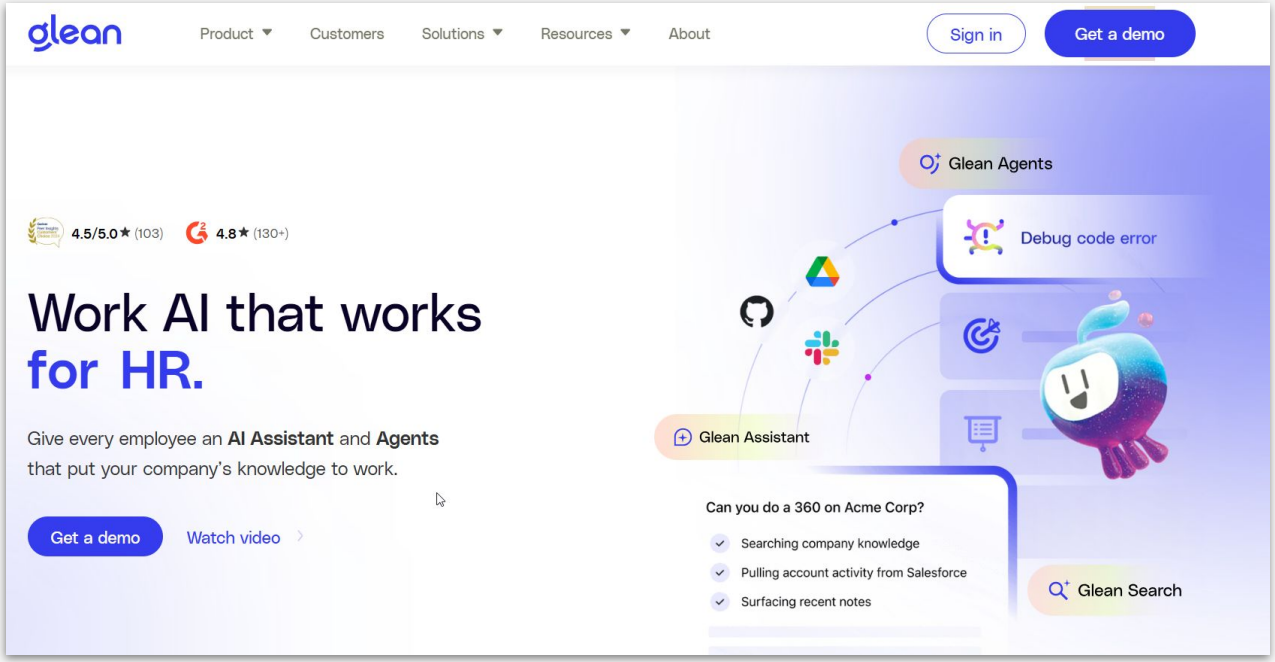
CIKLUM

Open WebUI



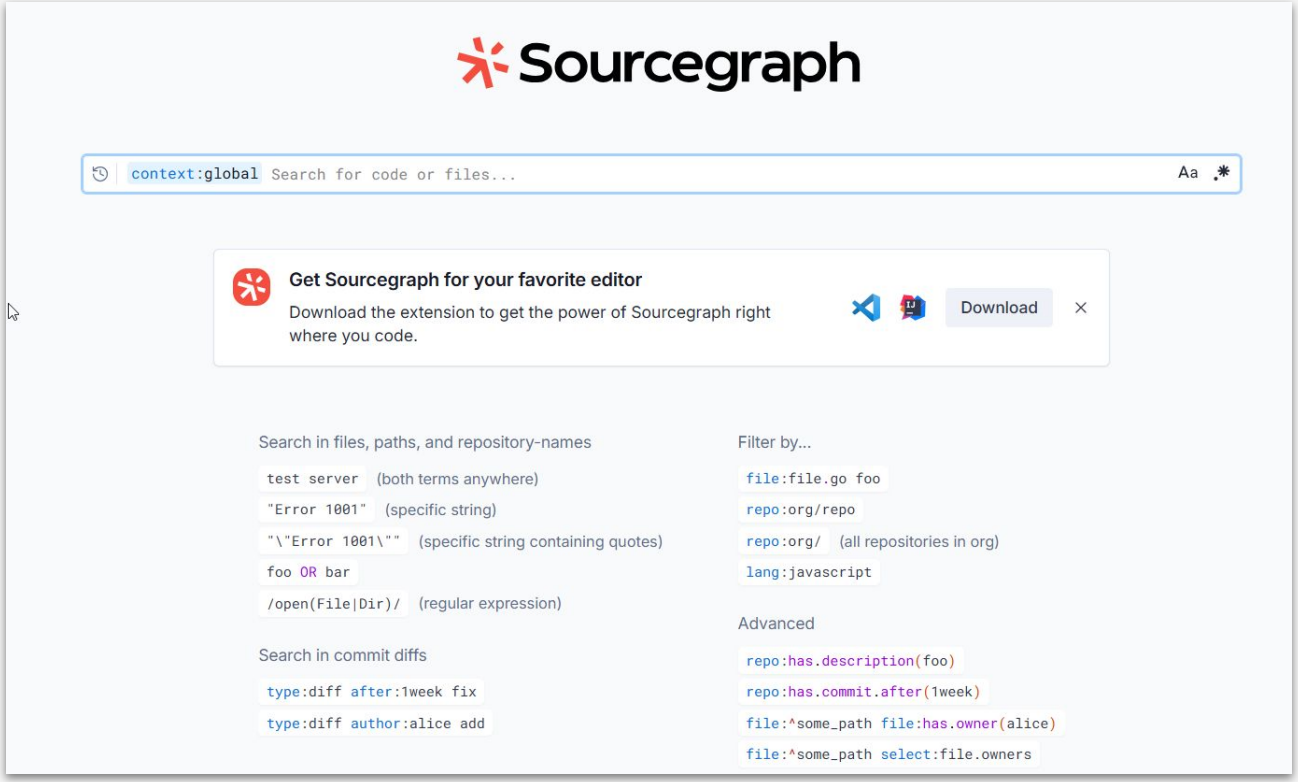
Glean

CIKLUM



Sourcegraph

CIKLUM



CIKLUM

Pros, Cons, Limitations and Considerations

Pros & Cons

CIKLUM

✅ Pros

- Enhances factual accuracy via external knowledge
- Reduces hallucinations vs. pure LLMs
- Enables dynamic updates without retraining
- Modular, domain-adaptable architecture
- Scalable across multimodal or multi-source data

⚠️ Cons

- Retrieval quality depends on data prep & embeddings
- Latency increases with retrieval pipeline
- Requires infra (vector DBs, indexing, chunking, etc.)
- Context window still limits answer depth

Limitations & Considerations

CIKLUM

🚧 Limitations

- Not reasoning-first, still retrieval-driven
- Poor handling of implicit or fuzzy queries
- No built-in memory or learning from feedback
- Struggles with conflicting or outdated sources

💡 Considerations

- Balance retrieval vs. generation cost & latency
- Continuous evaluation of relevance & drift
- Governance for data freshness and accuracy
- Potential upgrade path: **Agentic / Tool-use RAG**

Enterprise Data Protection and Privacy

CIKLUM

- **Safeguarding sensitive information** and maintaining customer trust.
- **Regulatory Compliance**: Adherence to laws like GDPR, which mandate data security and privacy.
- **Encryption & Anonymization**: Secure data and protect individual identities.
- **Access Control**: Implementing strict permissions and access policies to sensitive data.
- **Regular Audits**: Monitoring and auditing to identify and rectify any vulnerabilities.
- Importance of using **External Data** & RAG models.
- **Augment context** in order to get accurate completions.

Image by Jan Alexander from Pixabay

Quiz Time

CIKLUM

Q1

CIKLUM

What is the main goal of RAG?

A. To train larger models

B. To reduce hallucinations using external knowledge

C. To replace fine-tuning

D. To speed up token generation

Q2

CIKLUM

What is a key limitation of RAG systems?

A. They cannot retrieve multimodal data

B. They require GPU inference for retrieval

C. They depend heavily on data quality and embeddings

D. They cannot use APIs

Q3

CIKLUM

What’s the next evolution beyond basic RAG?

A. Fine-tuned GPT models

B. Static FAQ chatbots

C. Agentic or tool-using RAG architectures

D. Prompt templates with zero retrieval

Key Takeaways from the session

CIKLUM

- Boosts **factual accuracy** by grounding LLMs in **real data**
- **Reduces hallucinations**, but depends on retrieval quality
- **No retraining needed**, knowledge updated via data refresh
- Latency–complexity **trade-off**: retrieval adds overhead
- Still **limited** reasoning and context awareness
- Next **evolution** → Agentic, monitored, tool-using RAGs

Thank you!