Knowledge Vectorization: The Way RAG Actually Adds Value

Serge Zakharov

Partner, Rebels.ai ML Lab

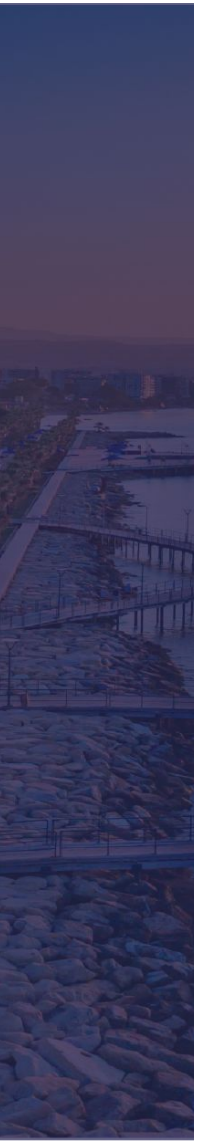
Founder, ASOdesk.com & Keggly.beer

MIPT alumni, ex-ABBYY

Rebels.ai

Limassol, Cyprus 2024

PERCONA UNIVERSITY



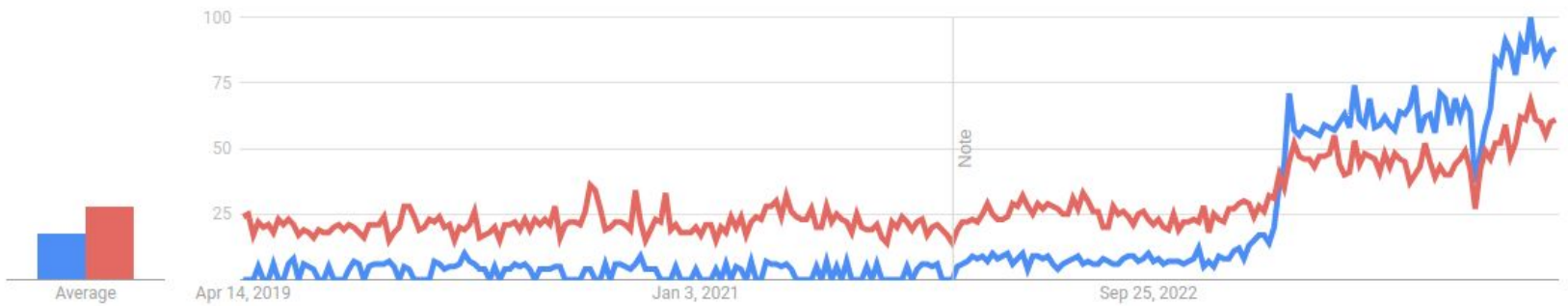
vector database
Search term

milvus
Search term

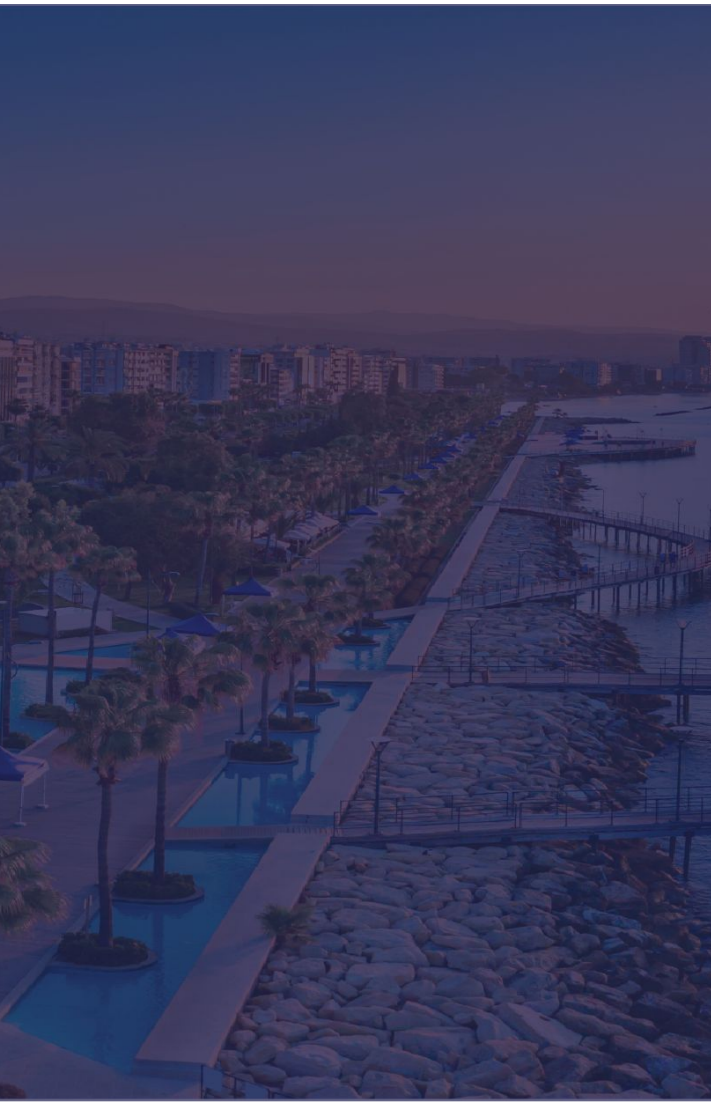
+ Add comparison

Worldwide | Past 5 years | All categories | Web Search

Interest over time



PERCONA



Rebels.ai

2019
Company established

25
Employees
including
5 PhD in ML and
computer science

50+
Clients

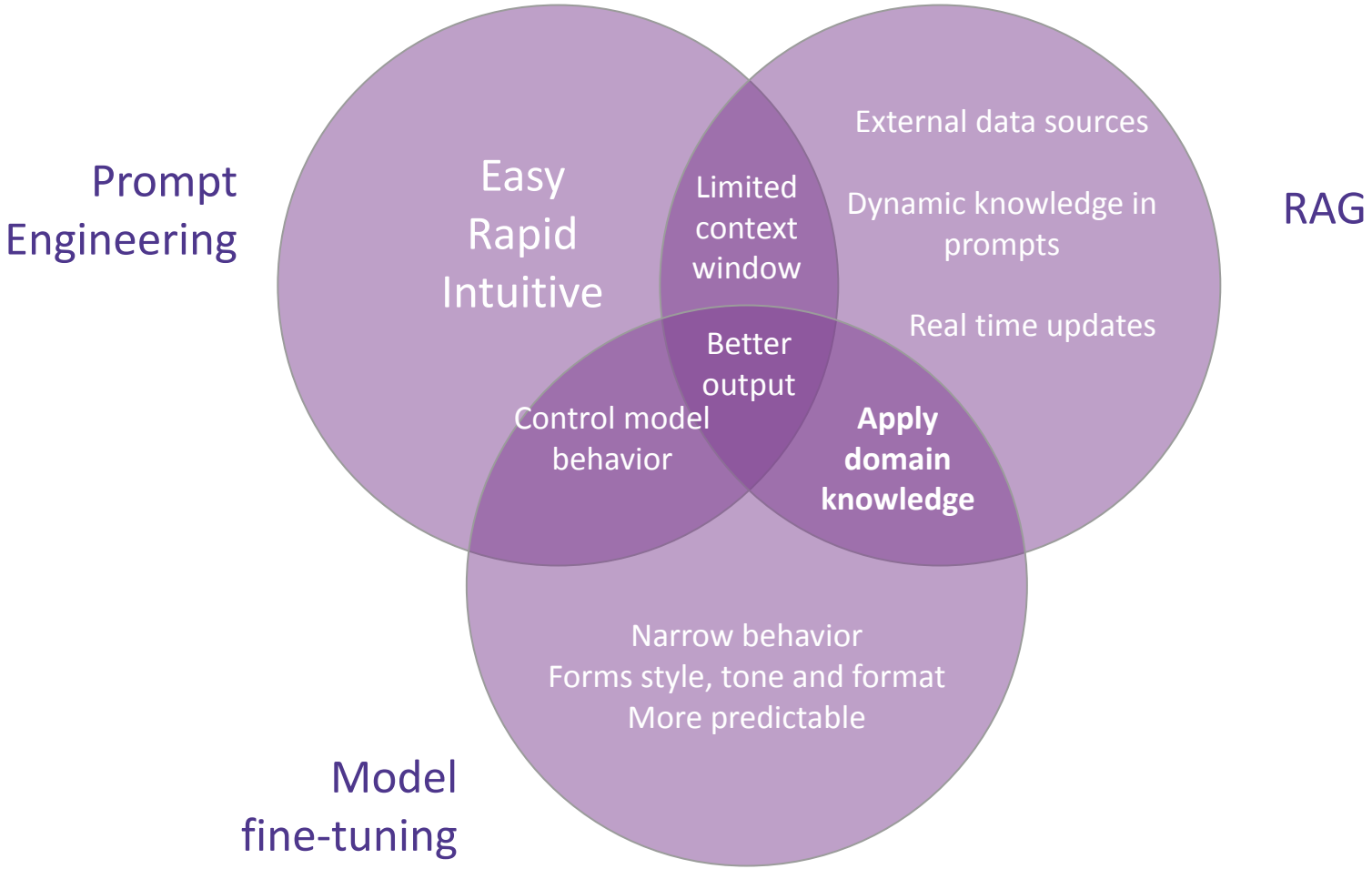
more than
100 solutions
implemented

Machine Learning & Data Science

- Conversational agents, LLM applications
  - knowledge base localization for digital integrator in western Europe
  - Copywriting adviser for Top-3 telecom operator in one of Eastern Europe countries
- Natural language processing (NLP)
  - sentiment analysis, hate speech detection, user behaviour analysis for Blended learning platform
- Geographic Information Systems (GIS)
- User data analysis
- Immersive reality
- Fusion Power Reactor control with RL⚡

and much more

LLM and Techniques
for Better output



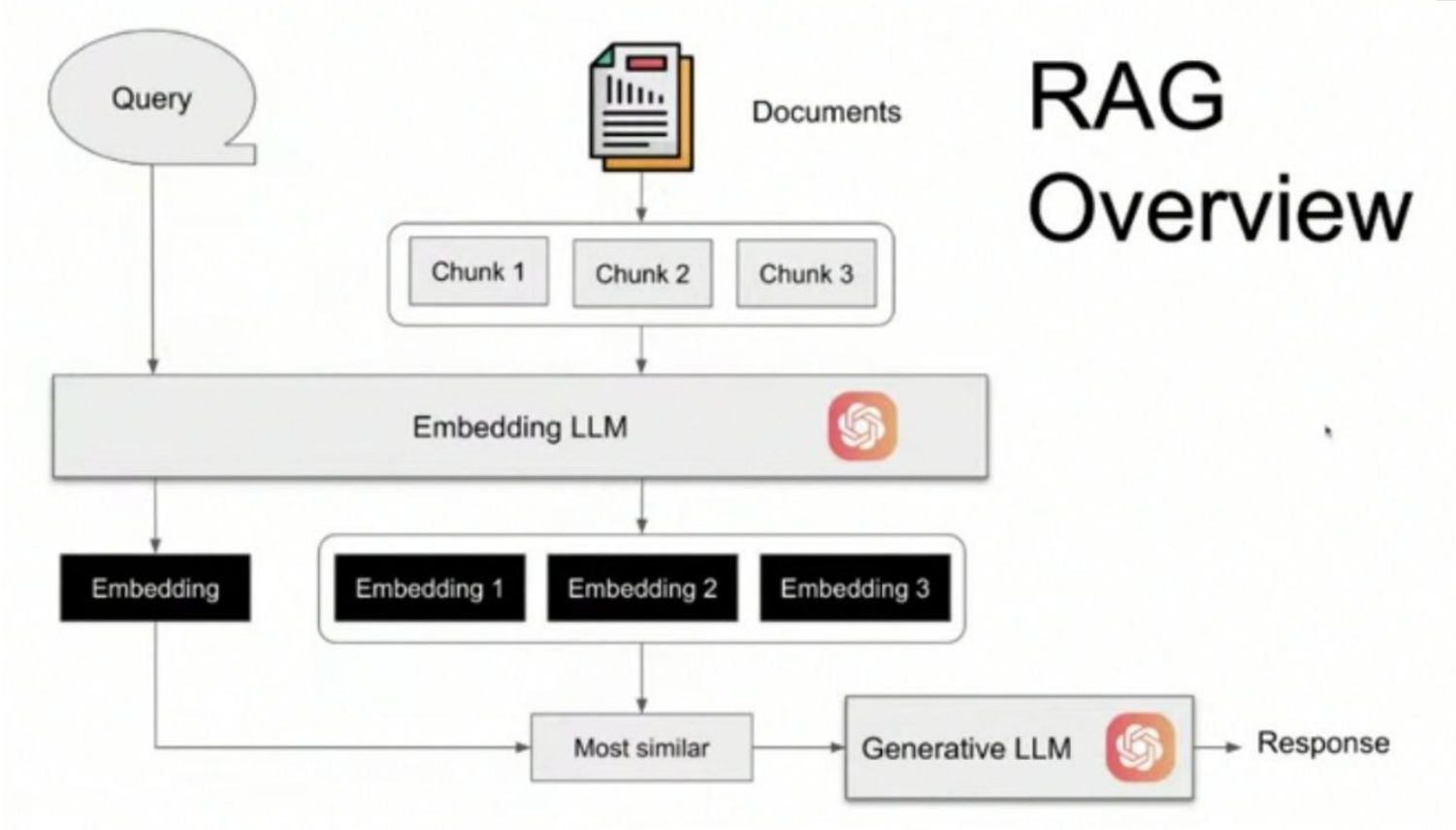
to DELVE - to dig in

PERCONA

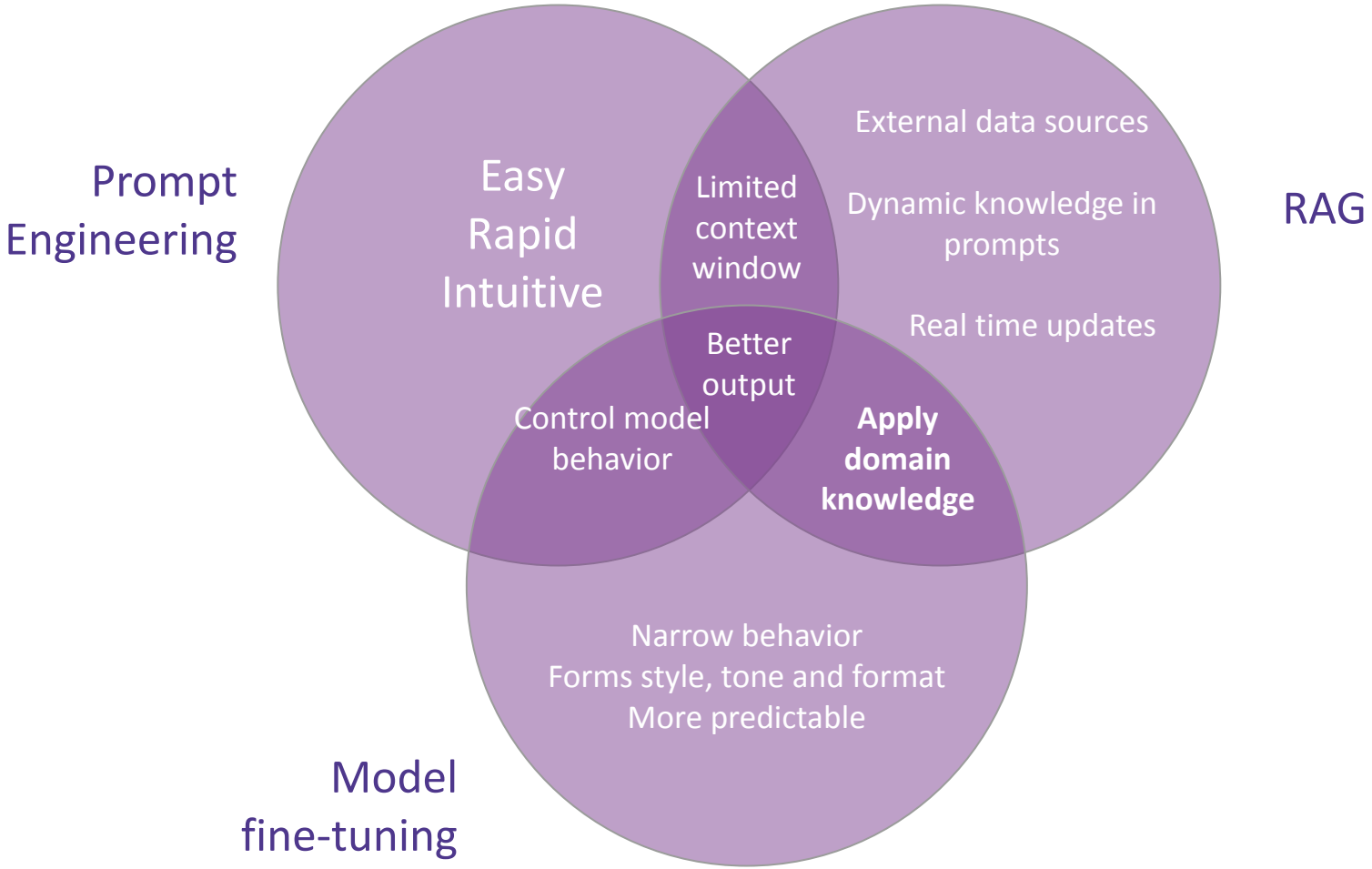
What is **R**etrieval-**A**ugmented **G**eneration?

The way to build
Responsible AI

Technique for **enhancing the accuracy and reliability** of generative AI models with facts **fetched from external sources**



LLM and Techniques
for Better output



PERCONA

RAG, How to Brew it Right

Knowledge base

Chunking

Retrieval

Evaluation

PERCONA

RAG Knowledge Preparations

- Remove irrelevant texts and documents, noise data, special characters, stop words
- Identify and correct errors: spelling, typos, and grammar
- Replace pronouns with names! *she -> Kate, them -> the viewers*

PERCONA

RAG Chunking

Model	Context length	Number of English pages*
GPT 3.5	4,096	6
GPT 4	8,192	12
GPT 4-32k	32,768	49
Llama 1	2,048	3
Llama 2	4,096	6

Context length comparison. (*Assuming 500 words per page.)

Dive into the joy of chunking, where each piece is a puzzle of its own. As you assemble them, a mosaic of understanding takes shape. This engaging mental exercise sparks creativity and hones analytical skills. It's like solving a puzzle, finding satisfaction in each arrangement. Approach chunking with curiosity and a playful spirit. Let it be an intellectual playground, making the process not only enjoyable but deeply satisfying. Happy chunking!

Fixed Size Chunking Character, Words, Tokens
Recursive Chunking
Document Based Chunking
Proposition-led Chunking!

Chunk enrichment by custom features Tone of Voice, Scene, Context, …

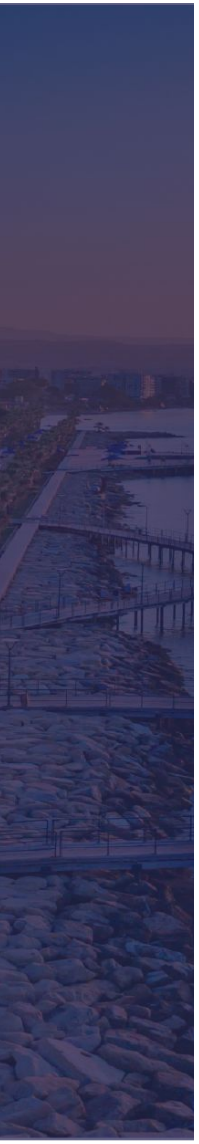
PERCONA

RAG Embedding and Retrieval

Vector databases

Postgresql
+ pg_vector

Ecosystems
Langchain
llama index



RAG Evaluation

Translation inherited
BLEU
METEOR
ROUGE

Sample prediction
Perplexity

Semantic similarity
BERTScore

User feedback

Human Evaluation
Likert Scales, Open feedback
Pairwise Comparisons

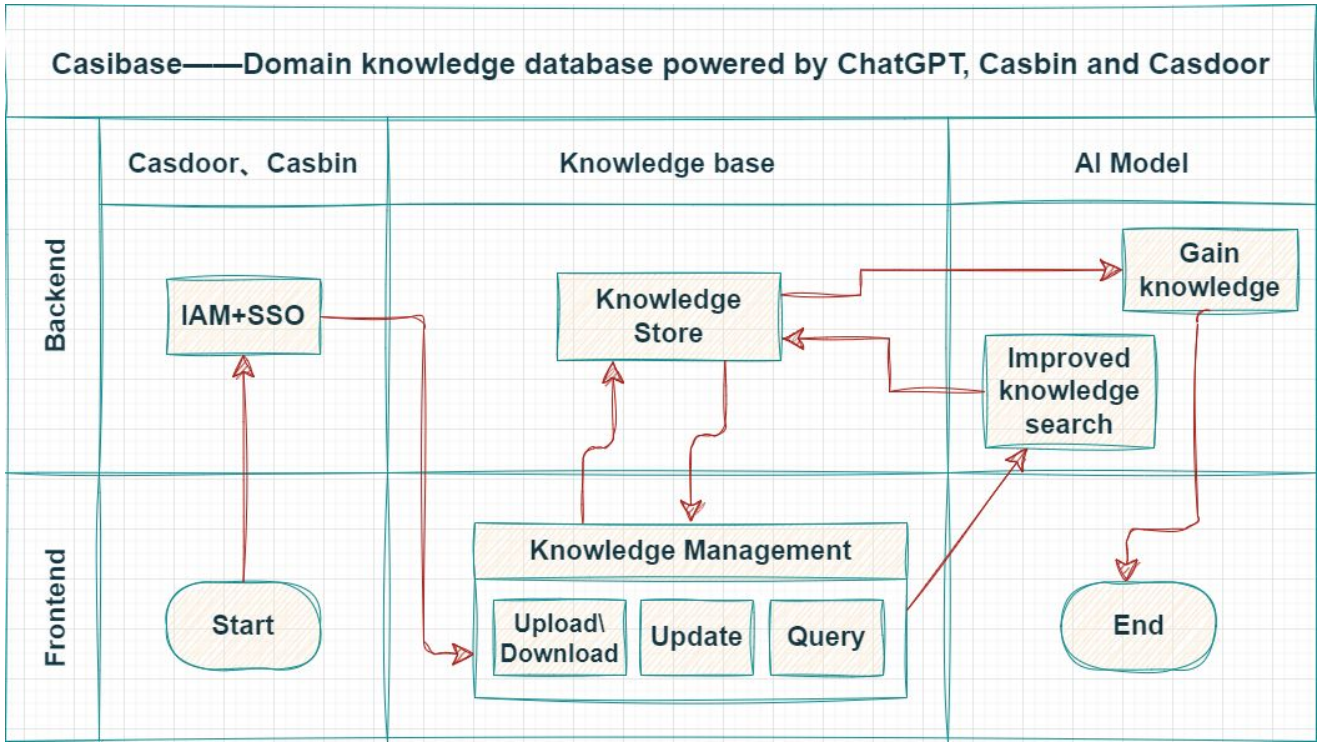
Task-Based Evaluation
ChatBot Arena

Interactive Evaluation
A/B, Wizard of Oz Tests

LLM **monitoring**, **evaluation** and **observability** open source tools

athina.ai
confident-ai
Deepchecks
EvidentlyAI 🖤
Arize Phoenix

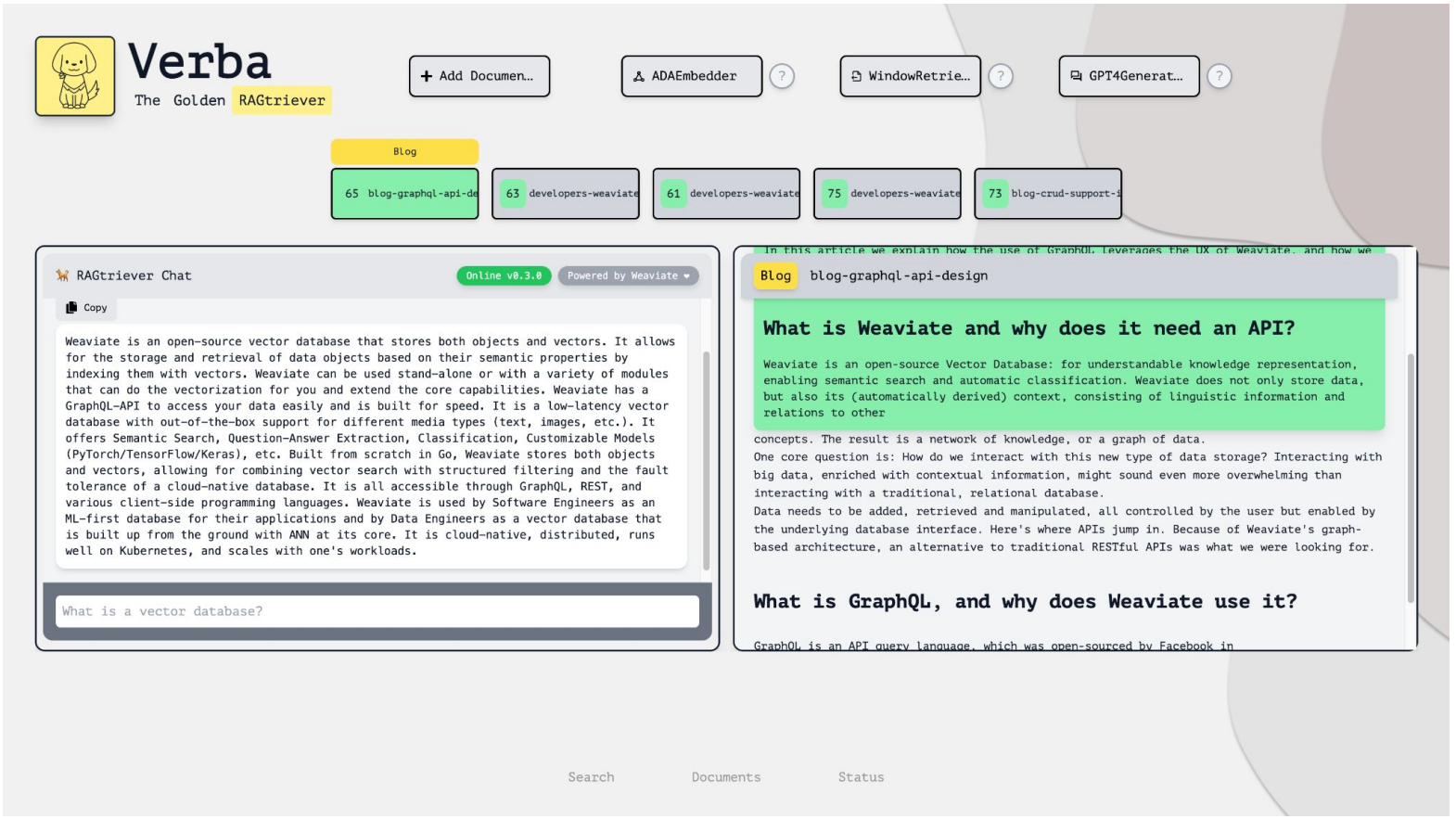
Open source tools for RAG: Casibase

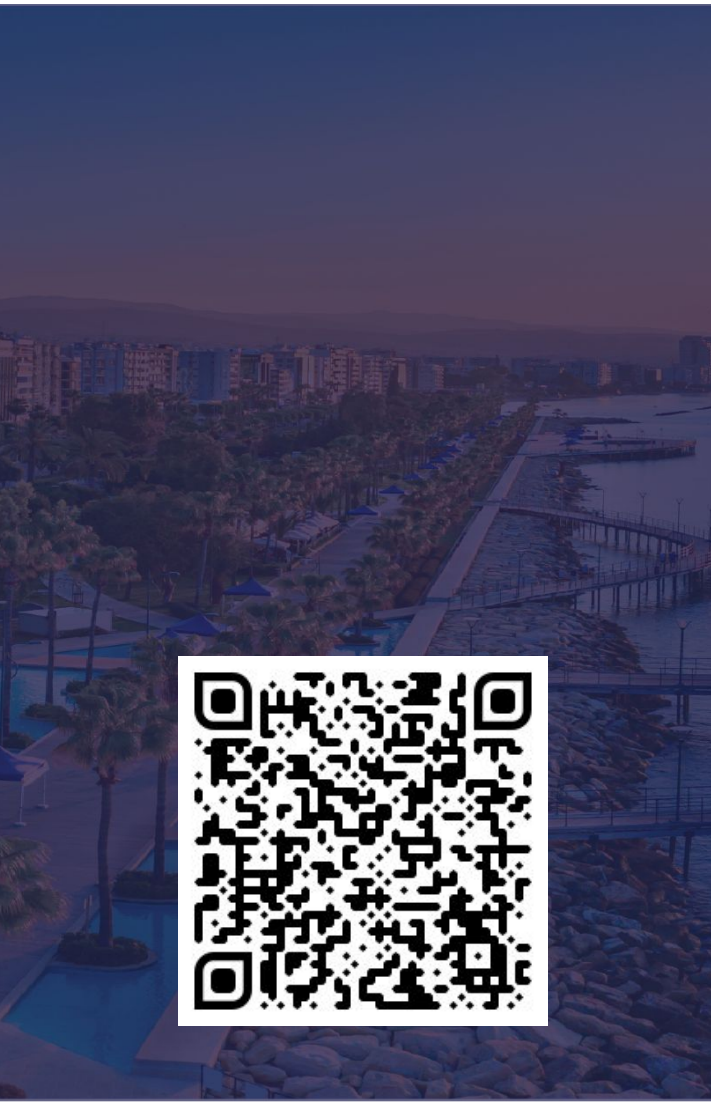


https://github.com/casibase/casibase

PERCONA

Open source tools for RAG: Verba by Weaviate





Extra links

Seven Failure Points When Engineering a Retrieval Augmented Generation System + Guide

Vector DBs Compared

Data Frameworks for LLM: Llama Index vs LangChain
Casibase new alternative to LangChain

'Awesome' list of LLM RAG systems and components

Subscribe Rebels.ai News
linkedin.com/company/rebels-ai
t.me/rebels_ai

Limassol, Cyprus 2024
PERCONA UNIVERSITY

Thank you!

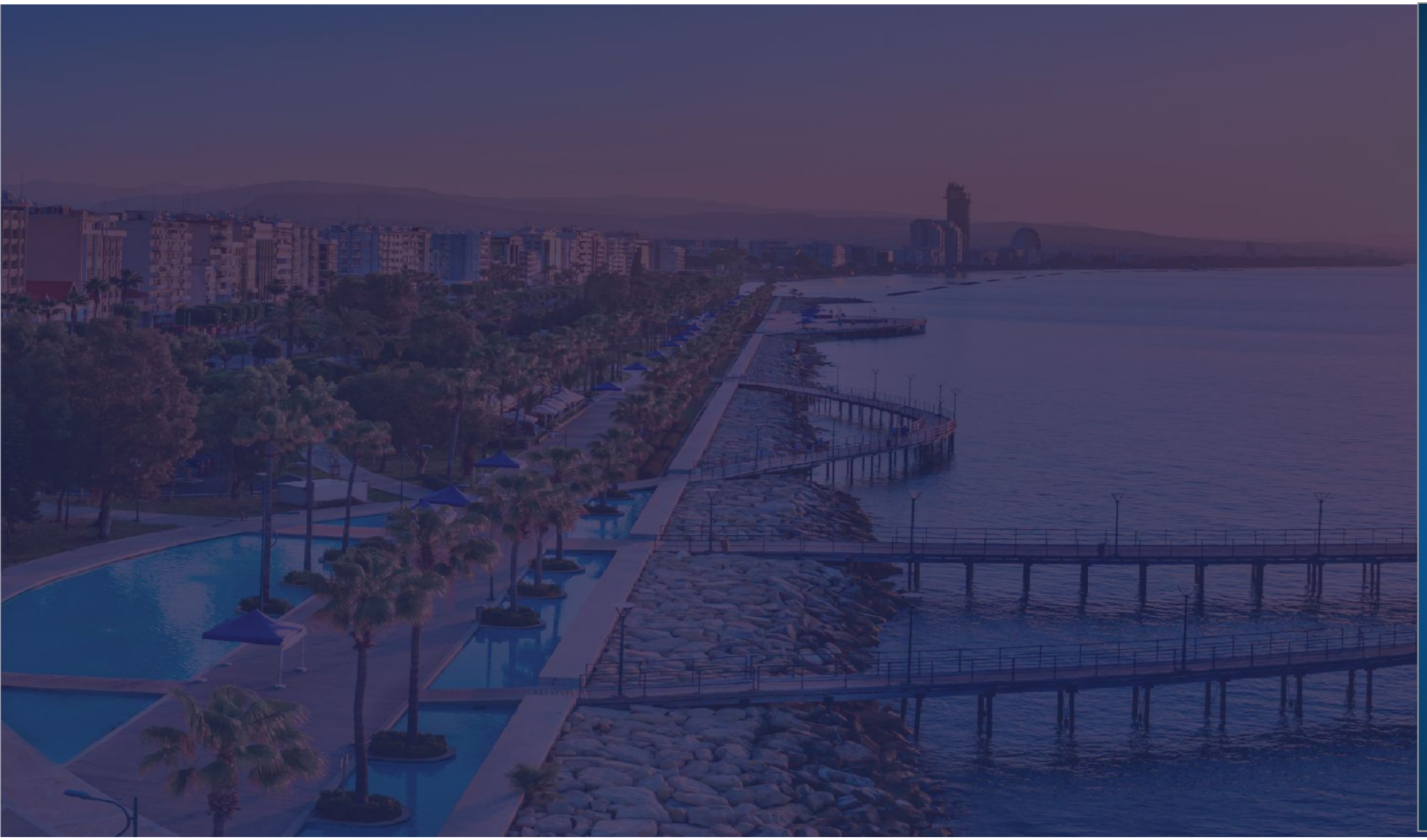
Serge Zakharov

Partner, Rebels.ai ML Lab

Founder, Asodesk.com & Keggly.beer

MIPT alumni, ex-ABBYY

Rebels.ai



Limassol, Cyprus 2024

PERCONA UNIVERSITY