UNLOCKING KNOWLEDGE WITH RAG

POWERFUL AI APPLICATIONS FOR BUSINESS

ERIK RIVERA

WHO I AM

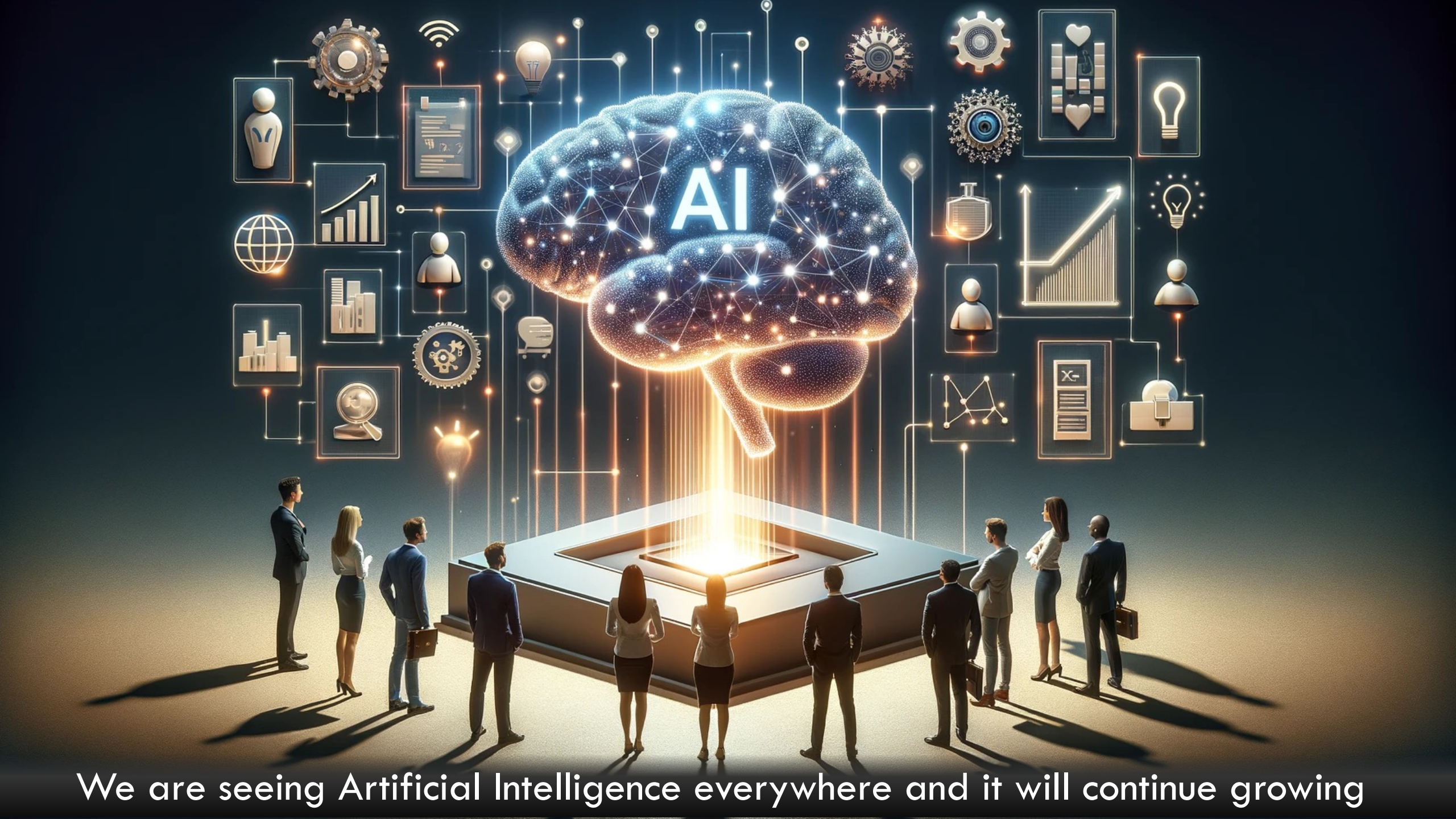
SOFTWARE ENGINEER

ENTREPRENEUR

CO-FOUNDER KIMETRICS

BEEKEEPER

PUEBLA, MÉXICO



We are seeing Artificial Intelligence everywhere and it will continue growing

There are many reasons

LARGE LANGUAGE MODELS (LLMs)

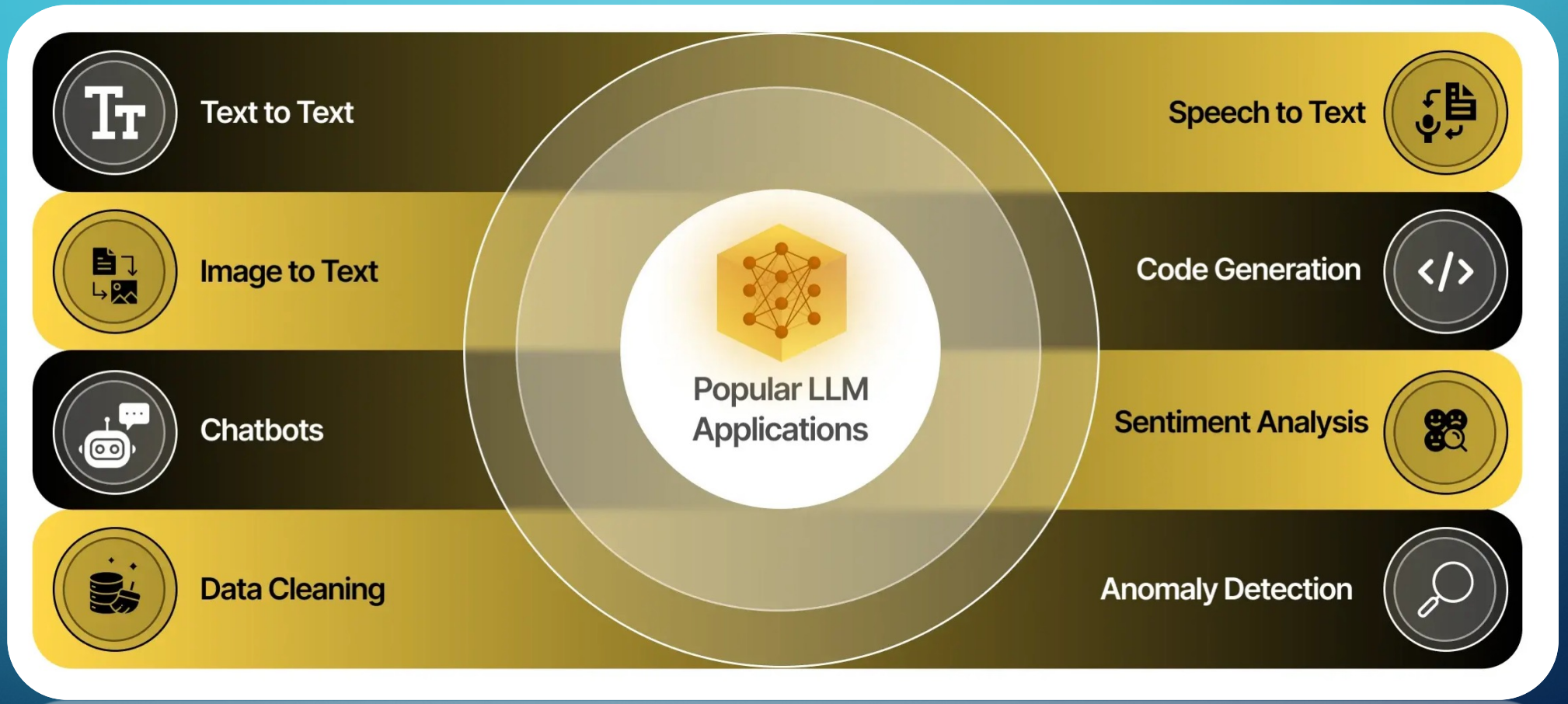
LLMs are advanced computer programs designed to understand and generate human-like text. They can answer questions, write stories, and even hold conversations by predicting what words or sentences should come next based on the patterns they've learned from vast amounts of text data.

SMALL AND OPEN SOURCE MODELS ARE GETTING BETTER

GPT-3.5 v4/0125 ☁️	58	87	71	60	78	47	67	0.13 €	1.41 rps
Gemini 1.5 Flash 0514 ☁️	32	97	100	56	72	41	66	0.10 €	1.76 rps
Gemini Pro 1.0 ☁️	55	86	83	60	88	26	66	0.10 €	1.35 rps
Cohere Command R+ ☁️	58	80	76	49	70	59	65	0.85 €	1.88 rps
Qwen1.5 32B Chat f16 ⚠️	64	90	82	56	78	15	64	1.02 €	1.61 rps
GPT-3.5-instruct 0914 ☁️	44	92	69	60	88	32	64	0.36 €	2.12 rps
Gemma 7B OpenChat-3.5 v3 0106 f16 ✅	62	67	84	33	81	48	63	0.22 €	4.91 rps
Meta Llama 3 8B Instruct f16 🦙	74	62	68	49	80	42	63	0.35 €	3.16 rps
GPT-3.5 v1/0301 ☁️	49	82	69	67	82	24	62	0.36 €	3.93 rps
Mistral 7B OpenChat-3.5 v3 0106 f16 ✅	56	87	67	52	88	23	62	0.33 €	3.28 rps
Mistral 7B OpenChat-3.5 v2 1210 f16 ✅	58	73	72	45	88	28	61	0.33 €	3.27 rps
Llama 3 8B OpenChat-3.6 20240522 f16 ✅	64	51	76	45	88	39	60	0.30 €	3.62 rps
Starling 7B-alpha f16 ⚠️	51	66	67	52	88	36	60	0.61 €	1.80 rps
Mistral 7B OpenChat-3.5 v1 f16 ✅	46	72	72	49	88	31	60	0.51 €	2.14 rps
Yi 1.5 34B Chat f16 ⚠️	44	78	70	52	86	28	60	1.28 €	1.28 rps

https://www.trustbit.tech/en/llm-leaderboard-juni-2024

THE MOST POPULAR LLM APPLICATIONS

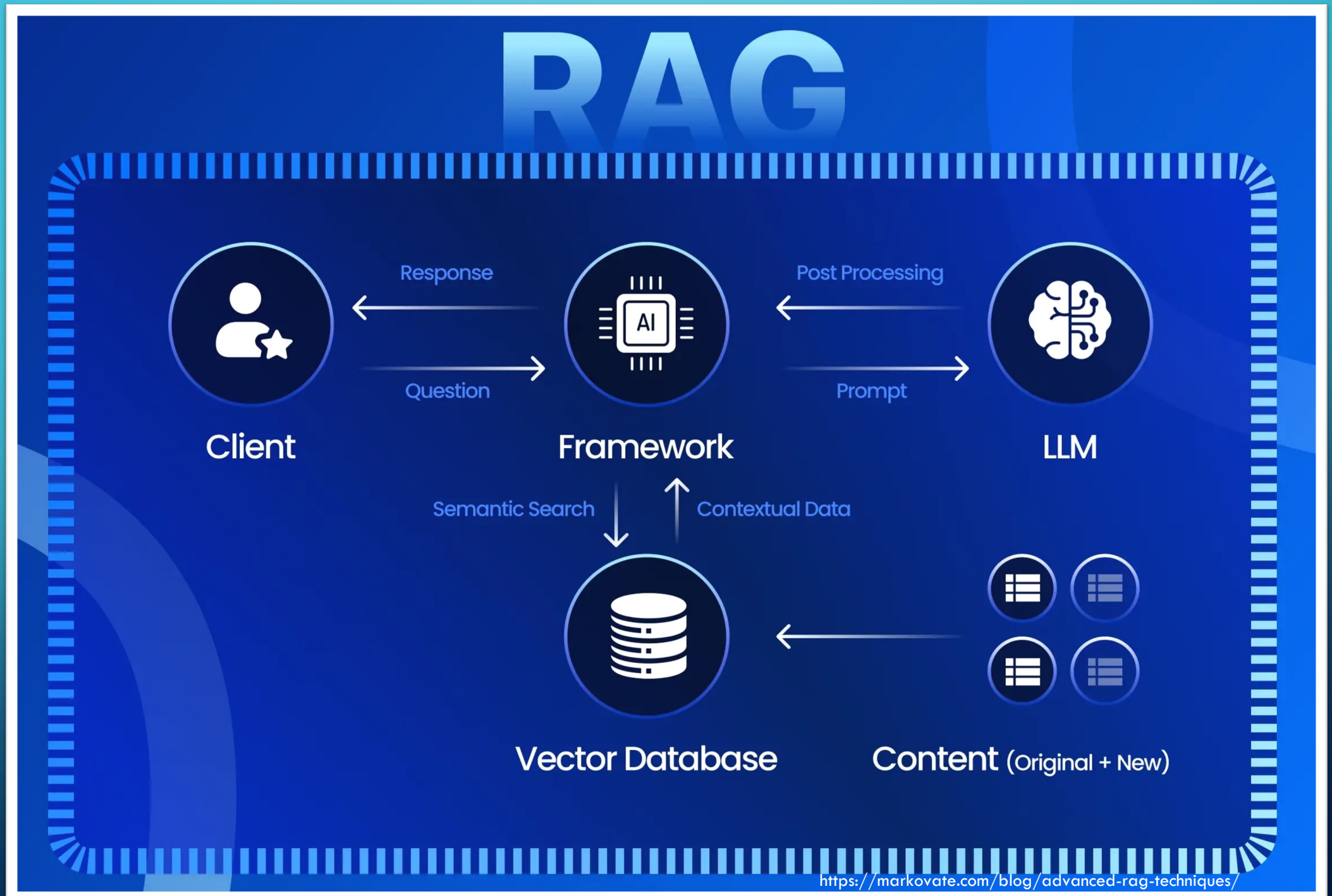


https://markovate.com/blog/applications-and-use-cases-of-llm/

WHAT IS RAG

RETRIEVAL AUGMENTED GENERATION (RAG)

RAG is a technique where an AI first looks up relevant information from a large collection of texts and then uses that information to generate more accurate and detailed responses or content. This helps the AI provide better answers by combining its language skills with specific facts and data.


RAG
Response
Question
Post Processing
Prompt
AI
Client
Framework
LLM
Semantic Search
Contextual Data
Vector Database
Content (Original + New)
https://markovate.com/blog/advanced-rag-techniques/

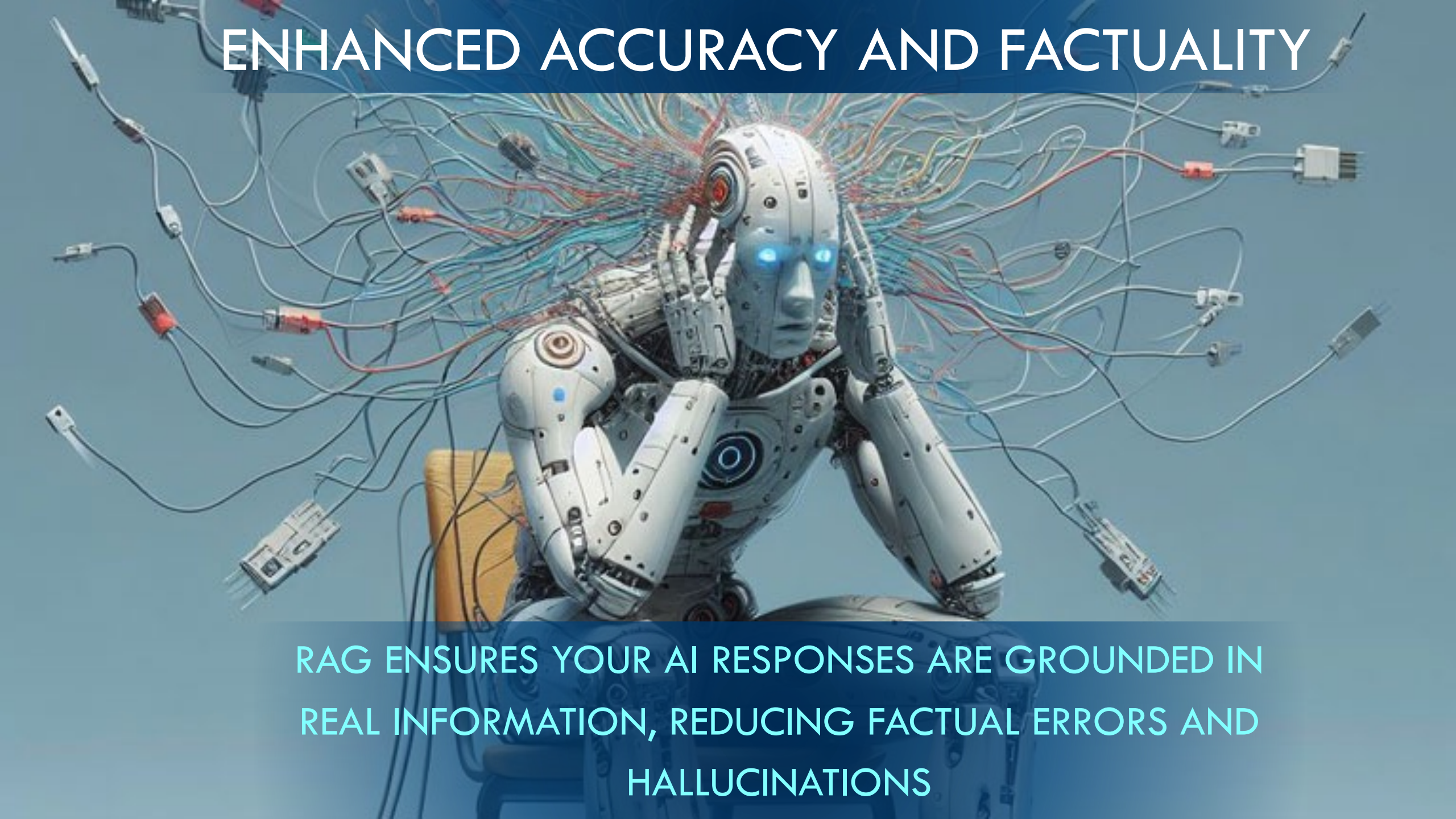
DEMO

ADVANTAGES

LEVERAGING PRIVATE DATA

RAG ALLOWS YOU TO INCORPORATE YOUR COMPANY'S INTERNAL DOCUMENTS, DATABASES, OR KNOWLEDGE BASES INTO THE RETRIEVAL PROCESS. THIS UNLOCKS VALUABLE DOMAIN-SPECIFIC KNOWLEDGE FOR MORE INFORMED RESPONSES

ENHANCED ACCURACY AND FACTUALITY



RAG ENSURES YOUR AI RESPONSES ARE GROUNDED IN REAL INFORMATION, REDUCING FACTUAL ERRORS AND HALLUCINATIONS

IMPROVED DOMAIN SPECIFICITY

TAILOR RESPONSES TO SPECIFIC FIELDS BY USING RELEVANT DOCUMENTS DURING RETRIEVAL.

COST-EFFECTIVENESS

LEVERAGE EXISTING LLMS AND FOCUS ON BUILDING THE RETRIEVAL COMPONENT FOR YOUR SPECIFIC NEEDS

THANKS FOR ATTENDING

erik@rivera.pro

https://linkedin.com/in/erikriver

https://github.com/erikriver