THE FUTURE OF SEARCH
BEER & DATA 3

Lucas Farris <lucas@farris.com.br>

WHAT'S GENAI EVEN GOOD FOR?

Over the past few years, I've gotten many questions from business owners regarding how they can use Generative Artificial Intelligence (GenAI) or Large Language Models (LLM) to improve their business. My answer for the past couple of years has been one: Retrieval Augmented Generation.

CONTENTS

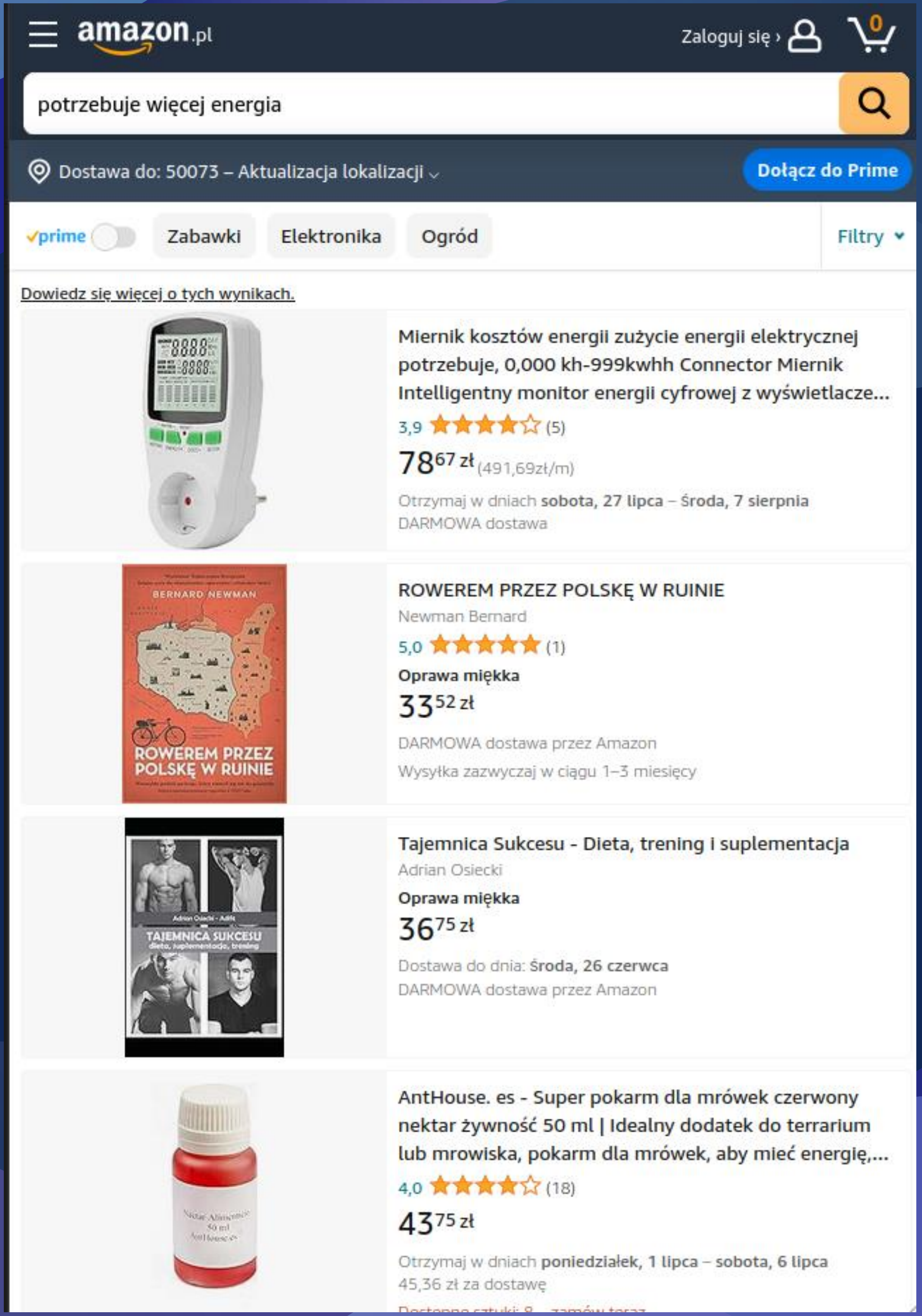
- Measuring Good Search
- Search Maturity
- Four-Stage Recommender System
- Retrieval Augmented Generation
- Cloud Example

MEASURING GOOD SEARCH

A fancy model doesn't is less important than a **good** evaluation framework

MEASURING

Accuracy	Relevance	Novelty	Serendipity	Diversity
Always remember the precision-recall trade-off. Measuring nDCG is always a good idea.	Human relevance (judgement) vs tracking relevance (presentation bias).	Users are likely to engage with new items if they are relevant	Non "obvious" recommendations make users happier	There should be enough variability between search results



potrzebuje więcej energia

Dowiedz się więcej o tych wynikach.

Miernik kosztów energii zużycie energii elektrycznej potrzebuje, 0,000 kh-999kwhh Connector Miernik Intelligentny monitor energii cyfrowej z wyświetlacze...

3,9 ★★★★☆ (5)

78^{67} zł (491,69zł/m)

Otrzymaj w dniach **sobota, 27 lipca – środa, 7 sierpnia**

DARMOWA dostawa

ROWEREM PRZEZ POLSKĘ W RUINIE

Newman Bernard

5,0 ★★★★★ (1)

Oprawa miękka

33^{52} zł

DARMOWA dostawa przez Amazon

Wysyłka zazwyczaj w ciągu 1–3 miesięcy

Tajemnica Sukcesu - Dieta, trening i suplementacja

Adrian Osiecki

Oprawa miękka

36^{75} zł

Dostawa do dnia: **środa, 26 czerwca**

DARMOWA dostawa przez Amazon

AntHouse. es - Super pokarm dla mrówek czerwony nektar żywność 50 ml | Idealny dodatek do terrarium lub mrowiska, pokarm dla mrówek, aby mieć energię,...

4,0 ★★★★☆ (18)

43^{75} zł

Otrzymaj w dniach **poniedziałek, 1 lipca – sobota, 6 lipca**

45,36 zł za dostawę

Dostępne sztuki: 8 – zamów teraz

WHAT MAKES SEARCH HARD

Dynamic Inventory

- Maybe your products are only sold a limited number of times
- Maybe they sell-out fast (e.g., luxury cars, jewelry, houses)
- Is everything you learn about them lost?

One-Time Users

- A common problem of dating sites, car manufacturers, and real-estate
- Some businesses have users that buy once come back only years later (or never)

Scarce Data

- Maybe your products have a small amount of features
- Maybe your users are mostly anonymous
- Maybe your data is in a CRM software you can't access
- Maybe you're not measuring

SEARCH MATURITY

A two-axes framework to evaluate and advance a search experience

LANGUAGE MATURITY

Level 1

Keyword Matching:
Inventory in indexed (inverse index), filters and sorting perform exact matches against attributes

Level 2

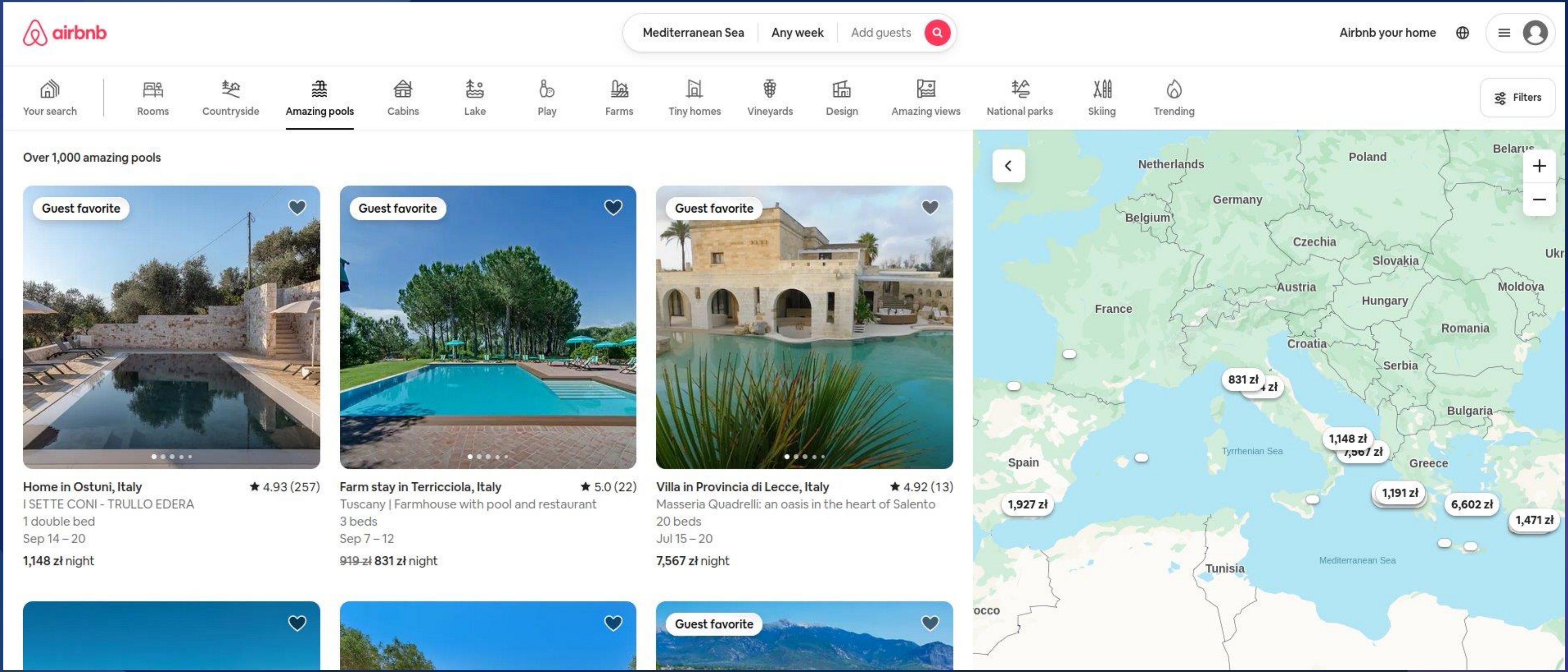
Taxonomies:
Items are catalogued and grouped under entities, cross-entity categorizations are built (ontologies), synonym dictionaries are built in the search

Level 3

Query Intent:
Semantics, query rewrites (i.e., did you mean…). System knows if you're looking for an article, for a product, or something else.

Level 4

Knowledge Graphs:
Knowledge beyond structured data (e.g., images, audio, videos) connecting them to text



Over 1,000 amazing pools

Home in Ostuni, Italy ★ 4.93 (257)
I SETTE CONI - TRULLO EDERA
1 double bed
Sep 14 – 20
1,148 zł night

Farm stay in Terricciola, Italy ★ 5.0 (22)
Tuscany | Farmhouse with pool and restaurant
3 beds
Sep 7 – 12
~~919 zł~~ **831 zł** night

Villa in Provincia di Lecce, Italy ★ 4.92 (13)
Masseria Quadrelli: an oasis in the heart of Salento
20 beds
Jul 15 – 20
7,567 zł night

RANKING MATURITY

Level 1

Term-Frequency

Relevance is measured based on how many times the search keywords appear in documents (i.e., a product's description). Commonly used algorithms are TF-IDF and BM25.

Level 2

Collaborative Filtering

You track the rate between certain web events (i.e., page views) and rank items based on how they perform. Usual metrics for this are Click-Through Rate (CTR) or Conversion Rate (CR).

Level 3

Model-Based

You use your inventory attributes and performance to train a regression based to predict how each item will perform. Usually non-negative matrix factorization (NNMF) and XGBoost are good choices here.

Level 4

Personalization

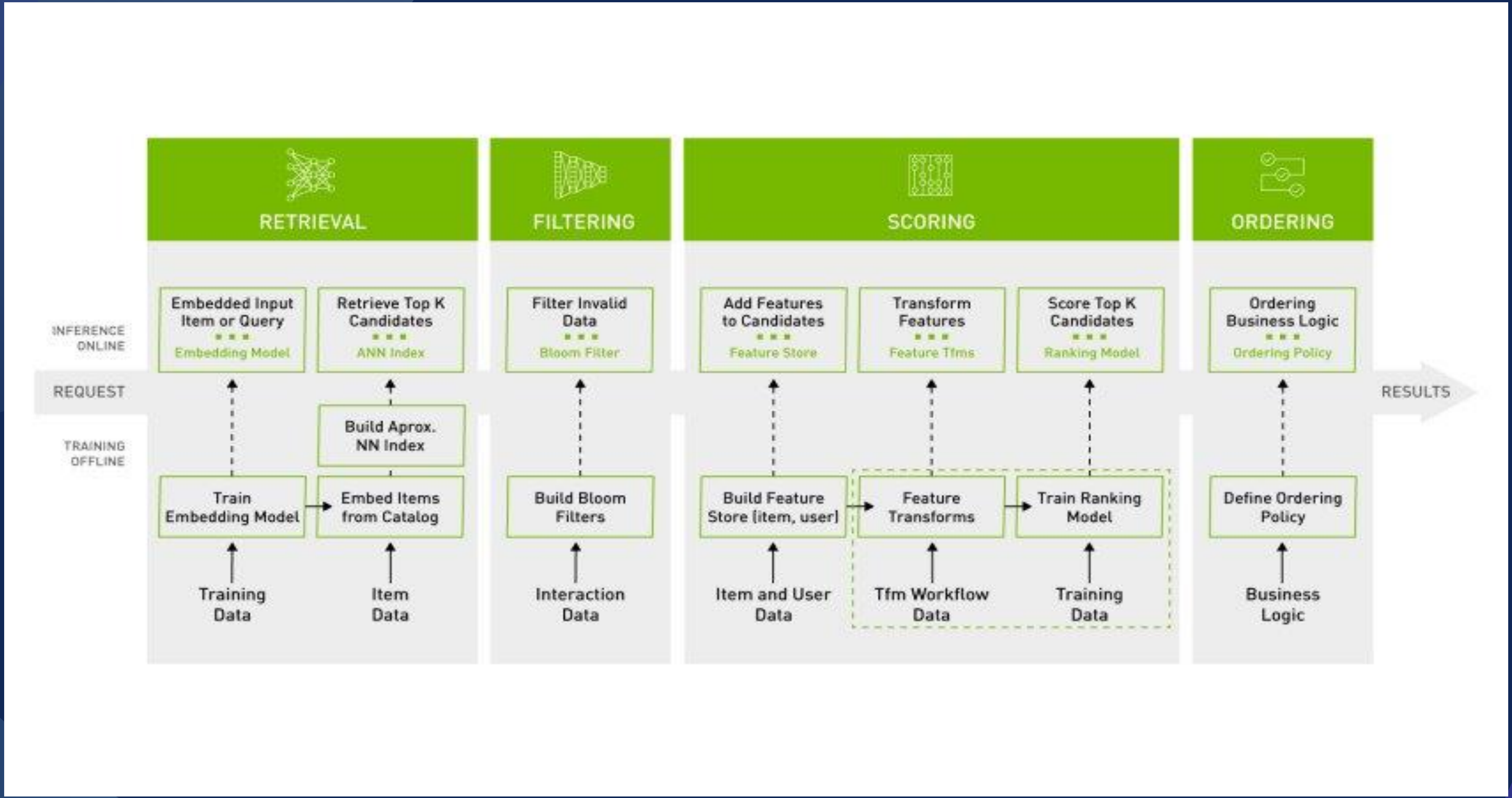
The relevance model is trained with user features and unstructured data, and inference happens in real time.

Usually in this case we would need deep neural networks and GPUs.

see https://doi.org/10.1145/3351095.3372878 and https://ieeexplore.ieee.org/document/5197422

SOLUTION ARCHITECTURE

The four-stage recommender system



INFERENCE ONLINE

REQUEST

TRAINING OFFLINE

RESULTS

RETRIEVAL

- Embedded Input Item or Query — Embedding Model
- Retrieve Top K Candidates — ANN Index
- Build Aprox. NN Index
- Train Embedding Model → Embed Items from Catalog
- Training Data
- Item Data

FILTERING

- Filter Invalid Data — Bloom Filter
- Build Bloom Filters
- Interaction Data

SCORING

- Add Features to Candidates — Feature Store
- Transform Features — Feature Tfms
- Score Top K Candidates — Ranking Model
- Build Feature Store (item, user) → Feature Transforms → Train Ranking Model
- Item and User Data
- Tfm Workflow Data
- Training Data

ORDERING

- Ordering Business Logic — Ordering Policy
- Define Ordering Policy
- Business Logic

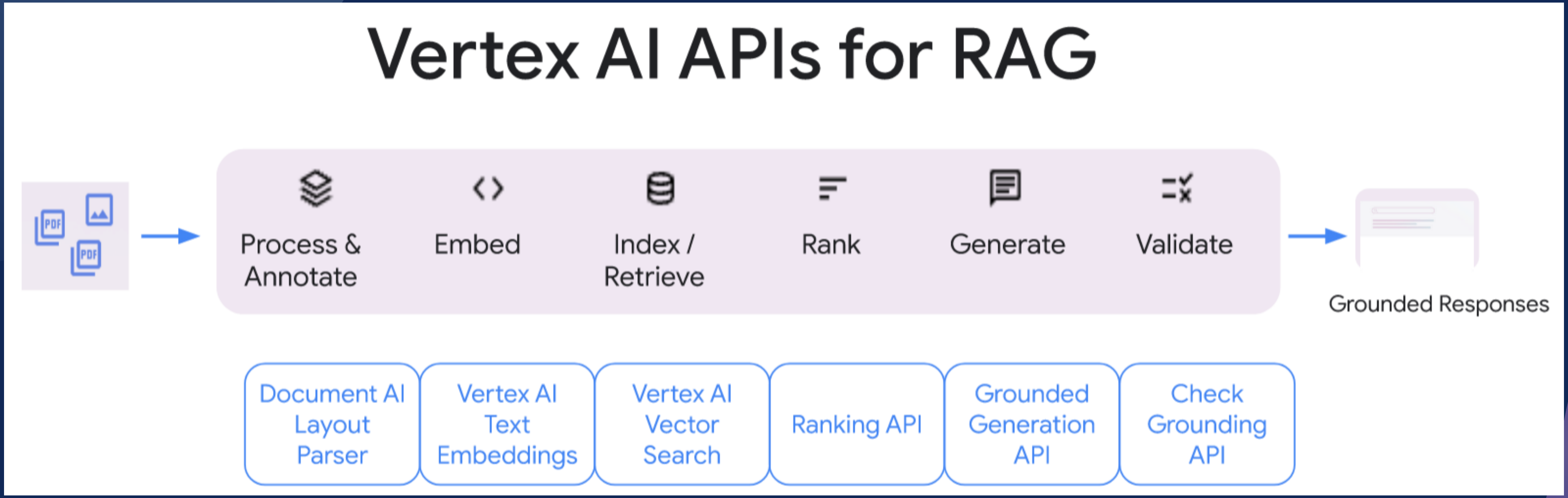
see https://developer.nvidia.com/merlin

WORD EMBEDDINGS

Alternative Labels	Synonyms	Taxonomy	Ontology	Knowledge Graph
Words, acronyms, or expressions have the exact same meaning	Tokens with very similar meaning	Relationships between tokens and their categories or classes	Relationships between tokens, for instance in terms of how they interact	Mappings and relationships between entities and their related concepts

RETRIEVAL AUGMENTED GENERATION

Helping models help you



Vertex AI APIs for RAG

Process & Annotate → Embed → Index / Retrieve → Rank → Generate → Validate → Grounded Responses

Document AI Layout Parser

Vertex AI Text Embeddings

Vertex AI Vector Search

Ranking API

Grounded Generation API

Check Grounding API

see https://developer.nvidia.com/merlin

DEMO TIME

Google Cloud Platform Example

THANK YOU

Lucas Farris

lucas@farris.com.br