Reinventing Workplace Productivity through Personalized Foundation Models [S61376]

Dropbox

NVIDIA GTC 2024

Introduction and agenda

01 NVIDIA AI Foundry

02 Introduction to Dropbox Dash

03 Retrieval Augmented Generation (RAG)

04 Fine-tuning RAG

05 Fine-tuning Case Study

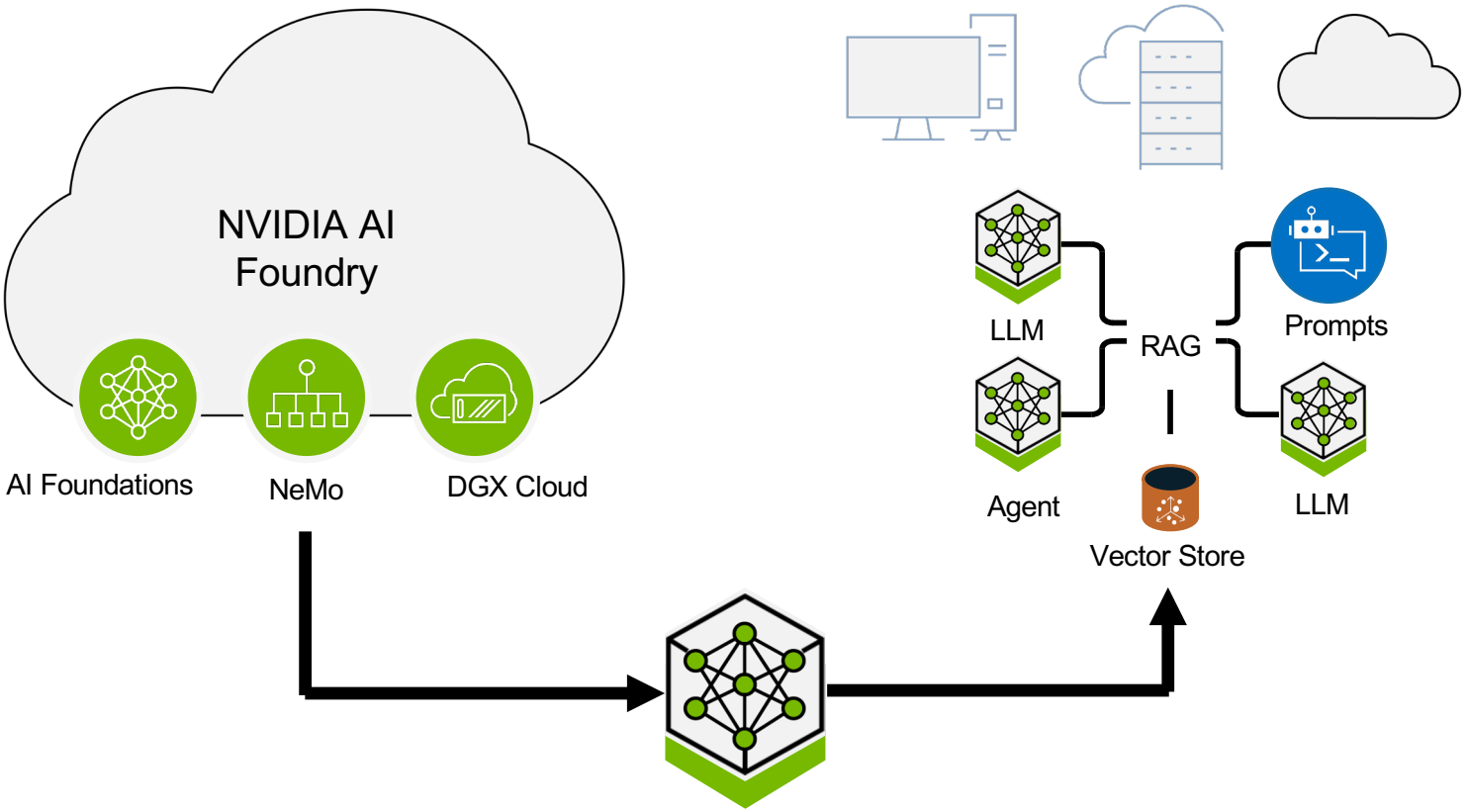
01

Arun Raman

NVIDIA

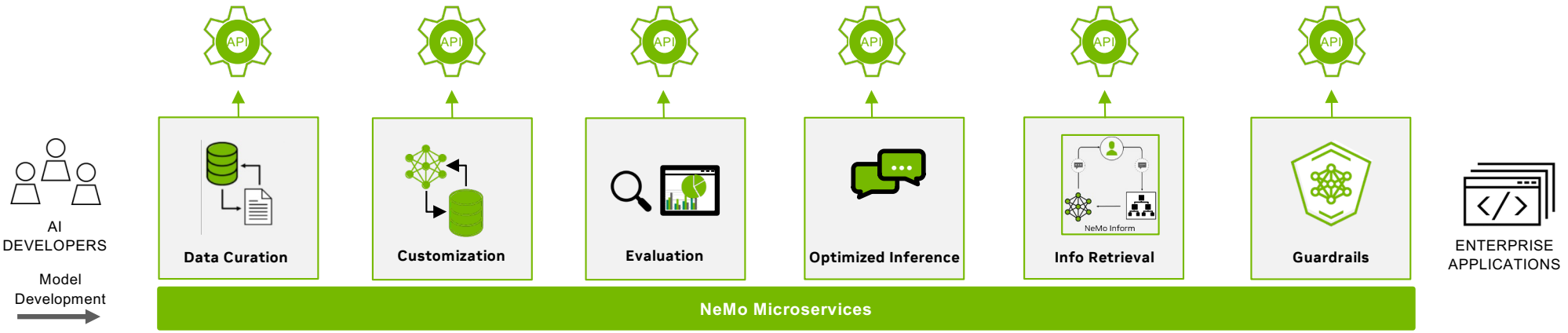
AI Solutions Architect

NVIDIA AI Foundry for Custom Enterprise-Ready LLMs



AI Foundry: Delivered As Microservices

Run as microservices where your data resides on your preferred compute

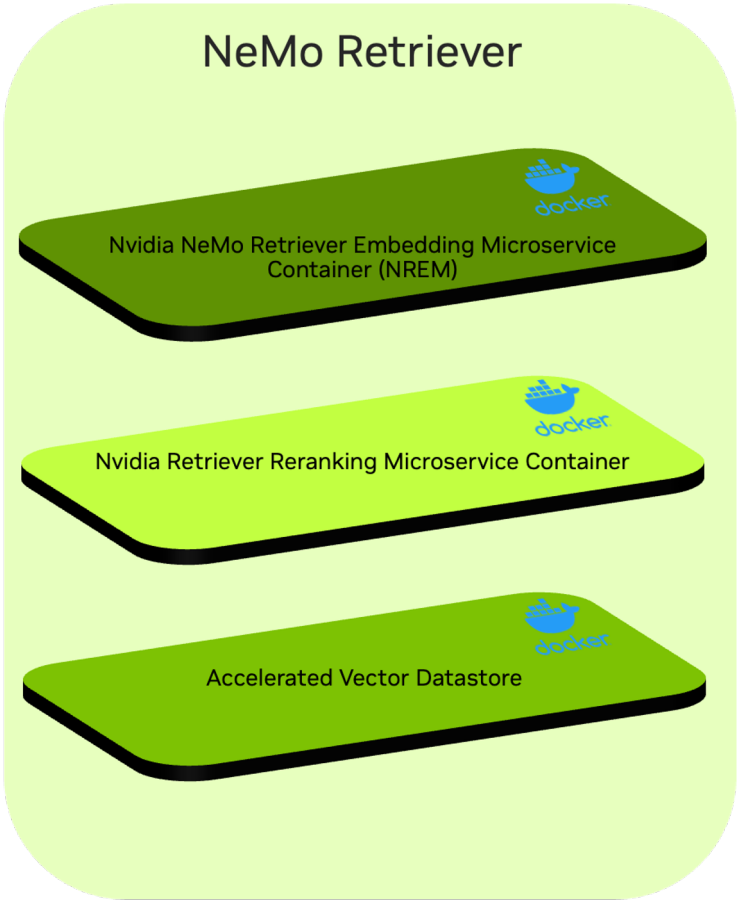


How to use NVIDIA NeMo Retriever Embedding Microservice

Pick and Choose

Containers from NGC

- Nvidia Retriever Microservice
  - Use Docker compose to deploy all the NeMo retriever service.
  - Use a single API for indexing and querying of user data
- Nvidia NeMo Retriever Embedding Microservice Container
  - Deploy the Embedding Microservice alone as a standalone container.
  - Download the “Nvidia Retrieval QA” embedding model from NGC and deploy
  - For models from Hugging Face use ‘model repo generator’ inside the container to convert a model TRT engine and deploy.
- Nvidia NeMo Retriever Reranking Microservice Container
  - Deploy Nvidia Retriever Reranking Microservice Container as a standalone service using NGC.
  - Download the retriever reranking QA model from NGC and deploy.



NVIDIA NeMo Retriever Microservice

Built on 4 Key Functions

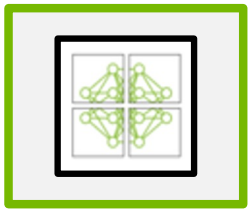
PIPELINE MANAGEMENT	COLLECTION MANAGEMENT	DOCUMENT INDEXING	QUERYING
Pipelines encapsulate chunking policies, embedding models, storage backend details, and a query strategy.	Service to manage a set of document collections. Governs how documents are indexed	Documents can be indexed into collections Documents will be indexed based on the collection it is associated	Query against a collection

NVIDIA Inference Microservice (NIM)

Deploy Anywhere (on prem, CSP)

Industry-standard APIs (OpenAI Compatible)

Supports serving customized models (ie dynamic LoRA, ptuning, etc)



Pre-built TRTLLM models optimized for best perf

Day0 LLM Support

- NIM for LLMs brings state of the art GPU accelerated inference for Large Language Models.
- Built on top of NVIDIA CUDA, TensorRT, TensorRT-LLM, and Triton.
- Industry Standard APIs to get started to deploy anywhere.
- Comes with pre-built TensorRT-LLM engines for best performance.
- Comes with optimized scheduling technique called in-flight batching.
- Supports custom model serving with LoRA.

02

Ashok Pancily Poothiyot

Introducing Dropbox Dash

Dropbox Dash is an AI-powered universal search product that connects to user's workplace apps, tools, and content, helping them find their content with semantic search and natural-language question and answers.

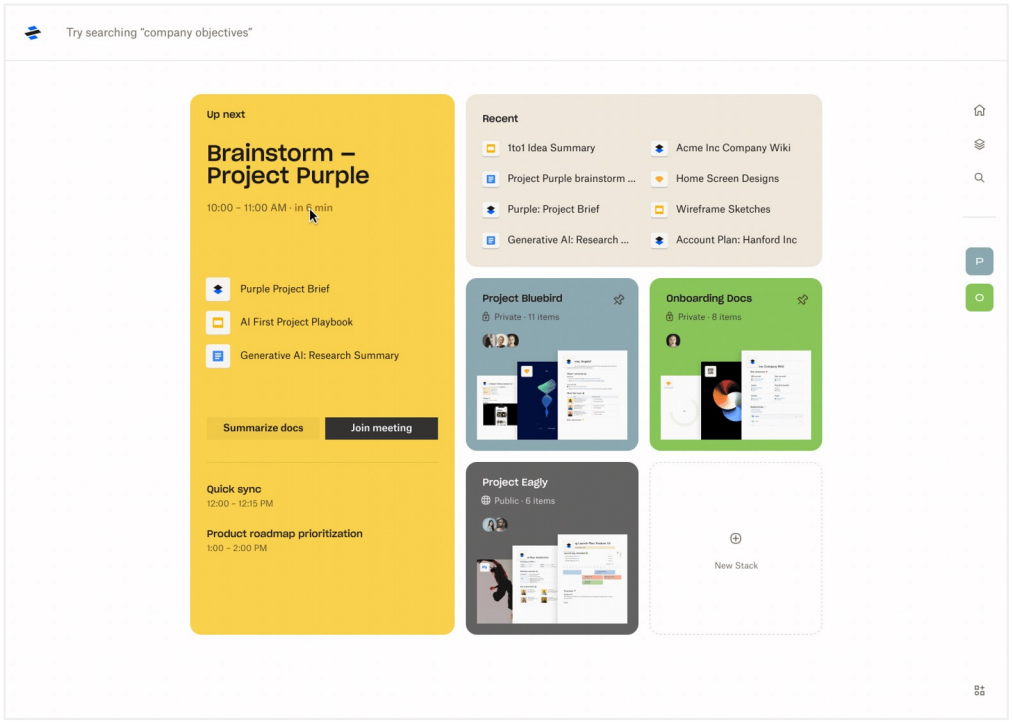
Knowledge workers spend 10.6 hours a week switching context, trying to find information, and lost tabs. Dash solves for this.

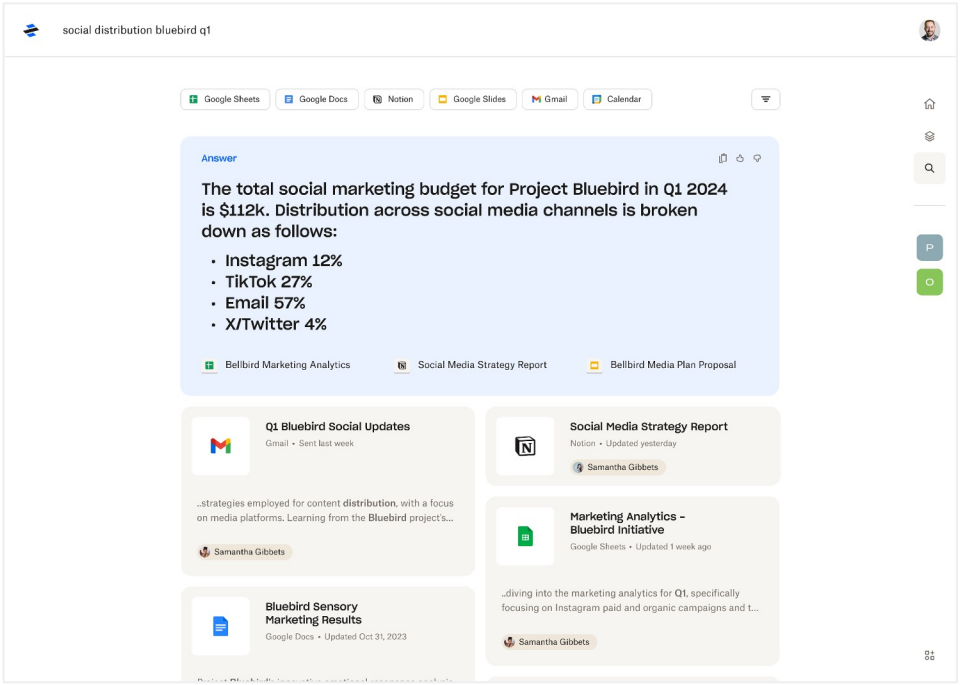
A **digital second brain** that
offloads, complements, and enhances your
biological one

Dash is your AI-powered agent that helps you **find**, **organize**, **orchestrate**, and **augment** all your content so you can do your best work

Dash is your AI-powered agent that helps you **find**, **organize**, **orchestrate**, and **augment** all your content so you can do your best work

Search


Try searching "company objectives"
Up next
Brainstorm – Project Purple
10:00 – 11:00 AM · in 8 min
Purple Project Brief
AI First Project Playbook
Generative AI: Research Summary
Summarize docs
Join meeting
Quick sync
12:00 – 12:15 PM
Product roadmap prioritization
1:00 – 2:00 PM
Recent
1to1 Idea Summary
Acme Inc Company Wiki
Project Purple brainstorm ...
Home Screen Designs
Purple: Project Brief
Wireframe Sketches
Generative AI: Research ...
Account Plan: Hanford Inc
Project Bluebird
Private · 11 items
Onboarding Docs
Private · 8 items
Project Eagly
Public · 6 items
New Stack

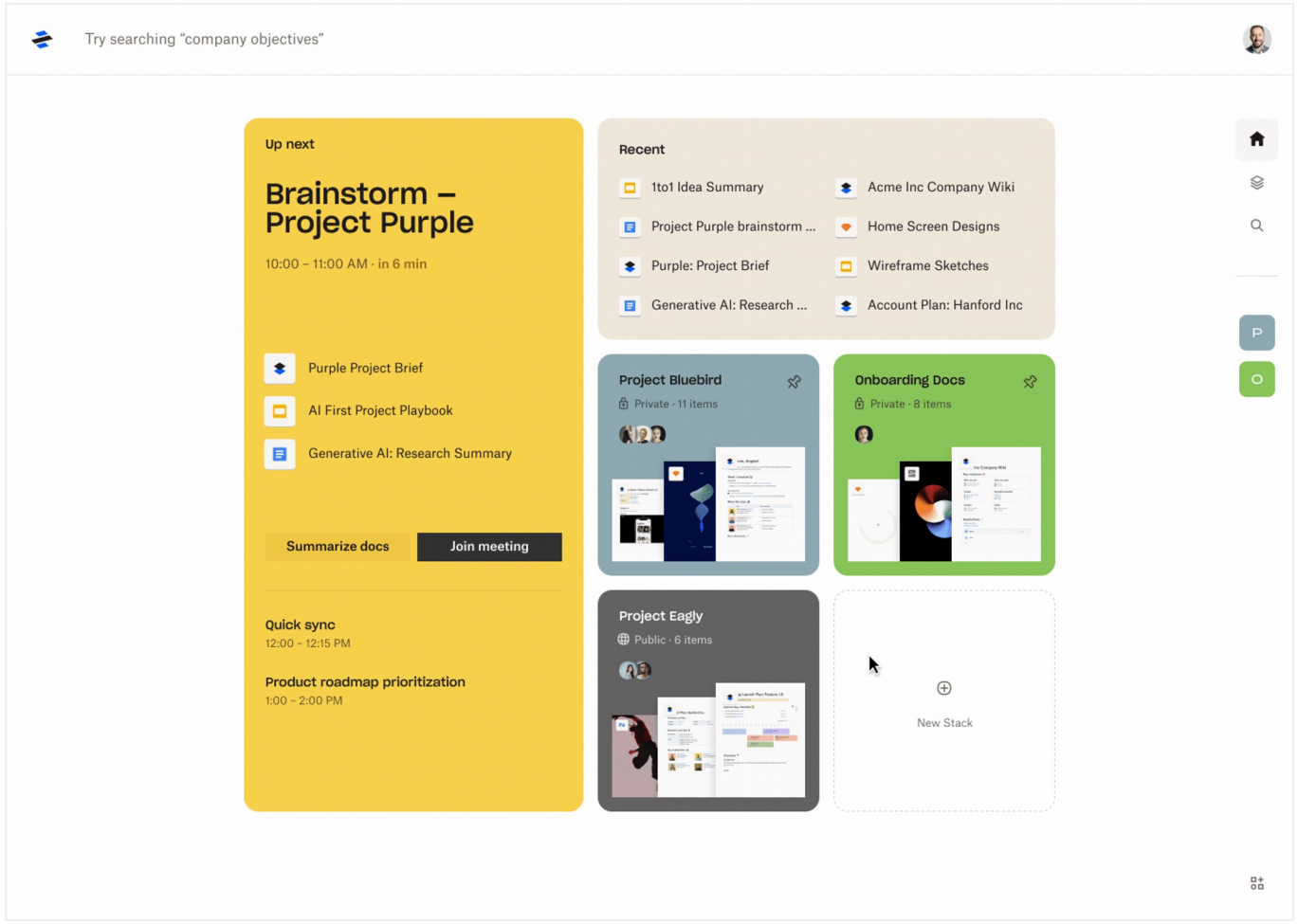

social distribution bluebird q1
Google Sheets
Google Docs
Notion
Google Slides
Gmail
Calendar
Answer
The total social marketing budget for Project Bluebird in Q1 2024 is $112k. Distribution across social media channels is broken down as follows:
Instagram 12%
TikTok 27%
Email 57%
X/Twitter 4%
Bellbird Marketing Analytics
Social Media Strategy Report
Bellbird Media Plan Proposal
Q1 Bluebird Social Updates
Gmail • Sent last week
..strategies employed for content distribution, with a focus on media platforms. Learning from the Bluebird project's...
Samantha Gibbets
Social Media Strategy Report
Notion • Updated yesterday
Samantha Gibbets
Marketing Analytics – Bluebird Initiative
Google Sheets • Updated 1 week ago
..diving into the marketing analytics for Q1, specifically focusing on Instagram paid and organic campaigns and t...
Samantha Gibbets
Bluebird Sensory Marketing Results
Google Docs • Updated Oct 31, 2023

Dash is your AI-powered agent that helps you **find**, **organize**, **orchestrate**, and **augment** all your content so you can do your best work

Start Page

Dash is your AI-powered agent that helps you **find**, **organize**, **orchestrate**, and **augment** all your content so you can do your best work

Stacks


Try searching “company objectives”
Up next
Brainstorm – Project Purple
10:00 – 11:00 AM · in 6 min
Purple Project Brief
AI First Project Playbook
Generative AI: Research Summary
Summarize docs
Join meeting
Quick sync
12:00 – 12:15 PM
Product roadmap prioritization
1:00 – 2:00 PM
Recent
1to1 Idea Summary
Acme Inc Company Wiki
Project Purple brainstorm ...
Home Screen Designs
Purple: Project Brief
Wireframe Sketches
Generative AI: Research ...
Account Plan: Hanford Inc
Project Bluebird
Private · 11 items
Onboarding Docs
Private · 8 items
Project Eagly
Public · 6 items
New Stack
P
O

- **Our goal is to transform knowledge work** through delightful and hyper-personalized AI experiences
- **Retrieval Augmented Generation (RAG)** is a powerful framework to support natural language Q&A and semantic search systems
- **Fine-tuning RAG enables hyper-personalization** while improving performance and reducing cost
- **This will lead to smaller models, lower latency, and specialized super-powers** for RAGs with more accurate and relevant predictions

Tim Gasser

03

Retrieval Augmented Generation

Knowledge work and RAG

- **McKinsey estimates "knowledge workers spend about a fifth of their time searching for and gathering information" [1].** This may involve searching multiple documents, reasoning over their contents, and using this to answer a question.
- **For example:** "How does monthly revenue year-to-date compare to last year?"
- **Today manual steps are:** Search spreadsheets to find data for last year, and monthly spreadsheets from this year. Create a new spreadsheet and calculate a monthly comparison. Copy this table into a document.
- **Question -> Search -> Reason -> Response.** The RAG architecture is built around this workflow.

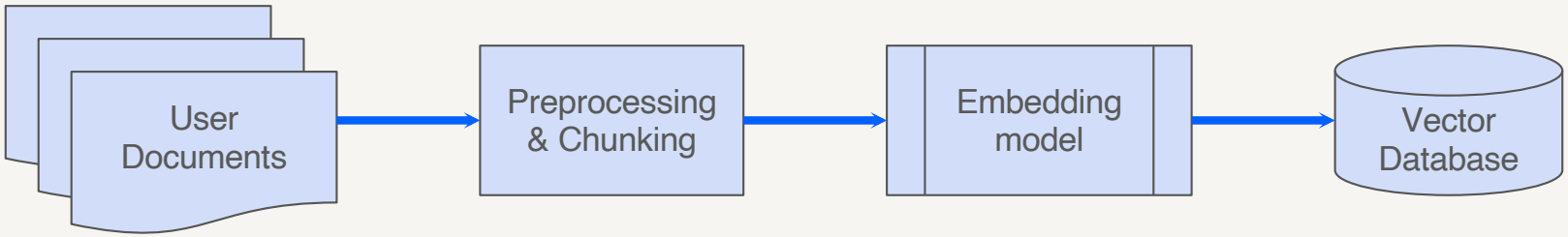
[1] - https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Advantages of RAG

- **Storing new data without retraining LLMs.** This avoids the "*I can't provide data beyond my last update*" problem, and compute- intensive retraining cycles to incorporate new data.
- **Reduced hallucinations.** RAG LLMs ground their response in context retrieved from a database, abstaining from answering if no relevant context is found.
- **Access Control Lists (ACLs).** LLMs have no concept of permissions or ACLs. RAG datastores check permissions before returning documents users don't have access to.
- **Separation of concerns.** RAGs decouple data ingestion and retrieval from LLM prompting for question-answering.

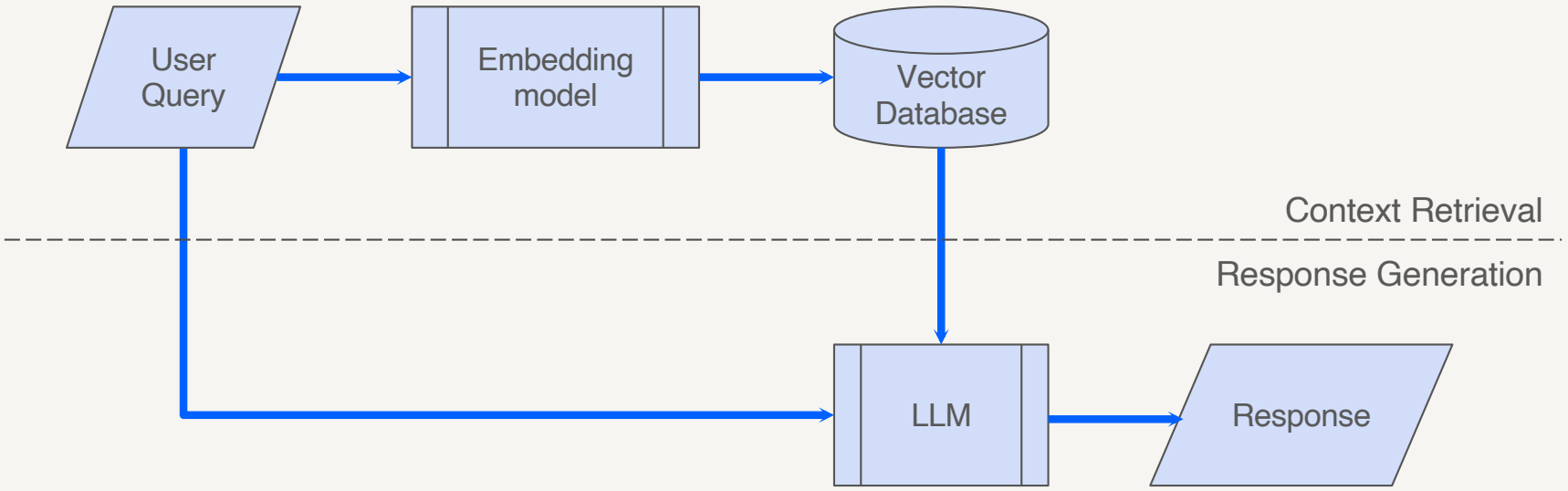
RAG Architecture - Ingestion Path

- **Documents are stored in a vector database.** This stores the contents of each document and creates indexes for fast retrieval.
- **The ingestion-path** ingests user documents offline. It splits them into chunks, creates embeddings for each chunk, and stores them in a vector database.



RAG Architecture - Query Path

- **The query-path** creates an embedding for the user query, and searches the vector database using it. This returns the most relevant document chunks, which are used by an LLM to answer the user query.



RAG - Measuring Performance

- **Ingestion: Freshness, Completeness, Correctness.** Without correct and timely document ingestion, all downstream steps will struggle.
- **Retrieval: Precision, Recall, MRR, MAP, NDCG.** Traditional Information Retrieval (IR) metrics quantify performance, but need labels of relevance for query-document pairs.
- **Response Generation: Hand-labelling or LLM-as-a-judge [1].** Abstractive Q&A tasks are harder to evaluate than Extractive Q&A, and require subjective quality of measurements (by other LLMs or hand-labelling).

[1] - Zheng et al, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena 2023 (Arxiv)

RAG - Design Decisions

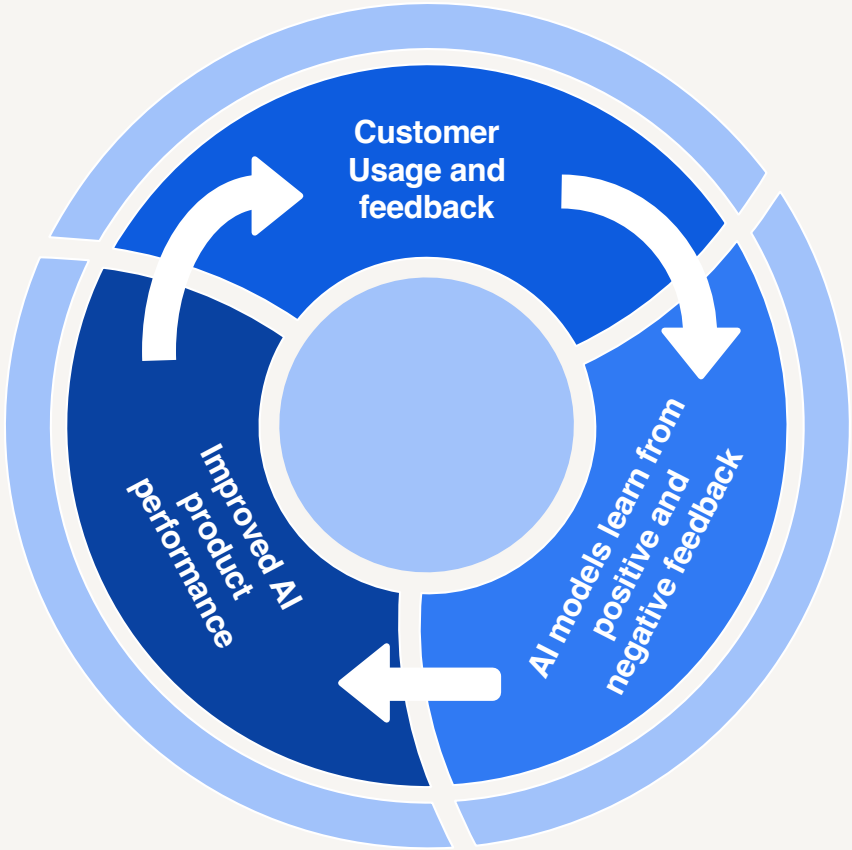
- **Embedding model.** Larger embedding models capture more semantics, but are slower, require more compute, and increase vector DB storage. The BeIR benchmark [1] has a number of benchmarks.
- **LLM models** can be used via third-party APIs, cloud providers, or on-premise. Moving to cloud or on-premise typically reduces cost per-GPU cost, but has a higher operational cost.
- **Many others!** Which vector database and Approximate Nearest Neighbors (ANN), chunking strategy, number of contexts to use, prompt templates

[1] - Thakur et al, 2021 BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models (Arxiv).

04

Fine-Tuning RAG

Fine-tuning high-level flywheel



What to fine-tune?

- **Fine-tuning embedding models improves retrieval performance** by aligning the model to the user document domain. This reduces domain-shift between the model training and inference.
- **Fine-tuning an LLM** improves performance on reasoning and responding to queries with context, and can match the tone of user documents.
- **Fine-tuning smaller LLMs on a narrow task** can match larger LLM performance (within a tolerance) and give benefits from a smaller model: lower latency, cost.

How to fine-tune?

- **Embedding models** are typically small (~100Ms parameters), and can be fine-tuned on small 24GB GPUs.
- **LLMs are many orders of magnitude larger** (~100B parameters) and may require large clusters of latest-generation GPUs (H100 80GB) with high-performance networking to do full fine-tuning.
- **Parameter-Efficient Fine Tuning (PEFT)** is an active area of research, aiming to match fine-tuning performance while requiring far fewer trainable parameters (and GPU RAM). Using PEFT, 7B models can be trained on 24GB GPUs.

Parameter-Efficient Fine-Tuning techniques

- **Low-Rank Adaptation (LoRA [1])** freezes model weights, and trains a low-rank outer product during fine-tuning.
- **Quantization [2]** reduces model size, lowering latency and cost. It may reduce performance slightly due to quantization noise. It can be combined with LoRA to give QLoRA [3].
- **Prompt tuning [4]** learns a mapping from input tokens to continuous embedding space, while freezing model weights.
- **P-Tuning [5]** adds trainable prompt embeddings throughout the model layers, while freezing the pre-trained model weights.

[1] LoRA: Low-Rank Adaptation of Large Language Models. Hu et al, 2021 (Arxiv).
[2] AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration, Lin et al, 2023 (Arxiv)
[3] QLoRA: Efficient Finetuning of Quantized LLMs. Dettmers et al, 2023 (Arxiv).
[4] The Power of Scale for Parameter-Efficient Prompt Tuning. Lester et al, 2021 (Arxiv).
[5] P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks, Liu et al, 2021 (Arxiv).

05

Case-Study
COVID Q&A

Example use-case: COVID Q&A

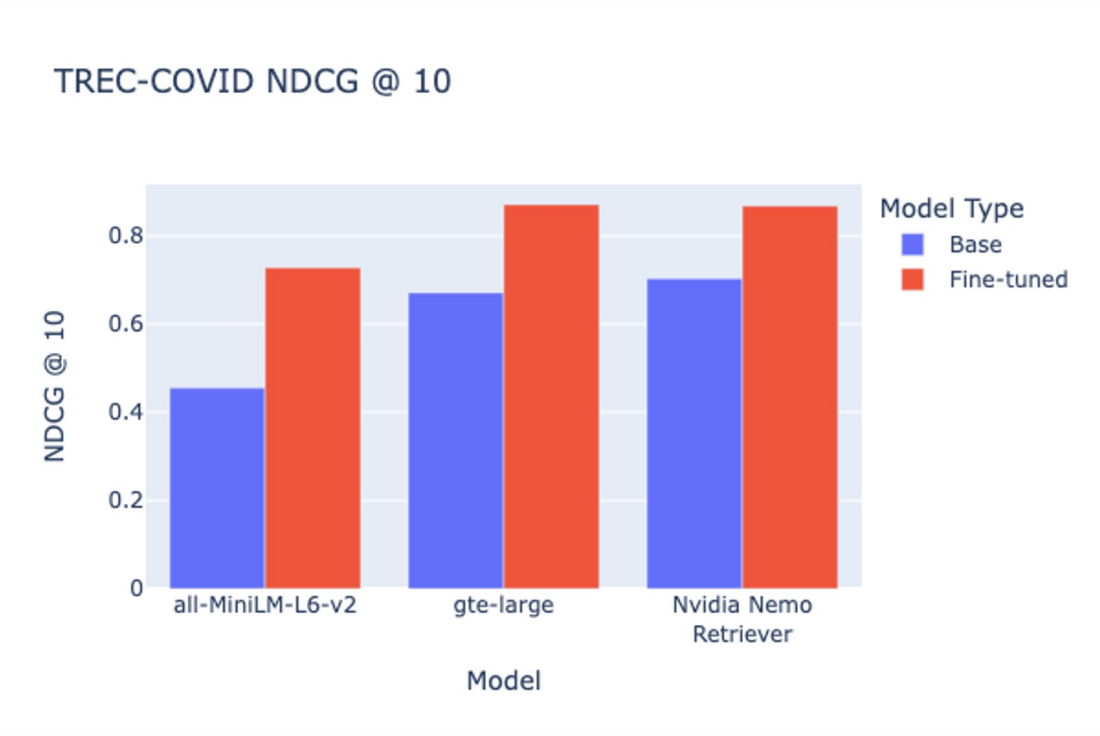
- **Example use-case: A researcher is keeping up with hundreds of COVID papers released every day during May 2020 [1].** The evaluations use TRECCOVID [2], containing 50 queries, 171,332 documents, and expert-labelled relevance.

- **We fine-tune embedding models to improve context retrieval.** We fine-tune three embedding models and evaluate the improvement in retrieval metrics.

- **We fine-tune LLMs to improve performance on an extractive Q&A task.** Using COVID-QA [3], we create a dataset matching the SQuADv2 [4] format, and evaluate how fine-tuning improves performance.

[1] - TREC-COVID (NIST)
[2] - Wang et al, CORD-19: The COVID-19 Open Research Dataset (NIH)
[3] - Moller et al, COVID-QA: A Question Answering Dataset for COVID-19 (Arxiv)
[4] - Know What You Don't Know: Unanswerable Questions for SQuAD, Rajpurkar et al, 2018 (Arxiv).

Fine-tuned Retrieval Performance

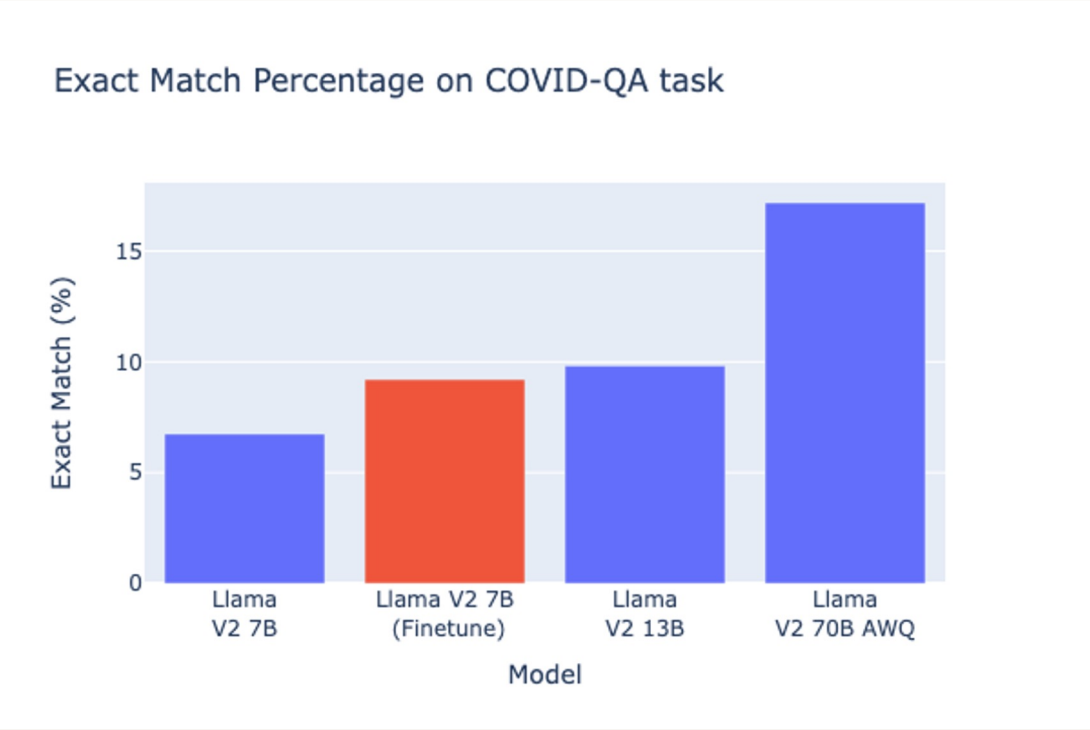


Comments

Fine-tuning embeddings moves relevant query-passage pairs closer in embedding space.

Depending on the loss function, triplets (anchor, positive, negative) can also be used to move irrelevant passages away from the query.

Fine-tuned LLM Performance



Comments

We measure extractive Q&A performance on COVID-QA [1].

Exact match: the exact answer string matches ground truth.

[1] -Moller et al, 2020 “COVID-QA: A Question Answering Dataset for COVID-19”” (ACL Anthology).

Conclusions

- **Fine-tuning embedding models on TREC-COVID improves retrieval:** Fine-tuning Nvidia's Nemo Retriever model improves its NDCG@10 from 0.7031 to 0.8677. After fine-tuning, all-MiniLM-L6-v2 outperforms base gte-large (NDCG@10 of 0.7276 vs 0.6708).

- **Fine-tuning small LLMs can match model almost twice the size.** After fine-tuning, Llama v2 7B has an Exact Match percentage of 9.20, compared to Llama v2 13B base Exact Match of 9.82.

- **Fine-tuning for personalization maintains data privacy and improves product-experience for each user.** This gives faster, more relevant answers and makes knowledge workers more productive.

Thank-you to all our contributors!

Dropbox

- Sean Chang
- Tim Gasser
- Jesse Lee
- Michael Nilsson
- Ashok Pancily Poothiyot
- Anthony Penta

Nvidia

- Nave Algarici
- John Barco
- Joey Conway
- Marc Demas
- Arun Raman
- Praveen Nakshatrala

Thank you! Any questions?

Dropbox is hiring ML engineers and scientists!

See **jobs.dropbox.com**

Fine-tuned Retrieval Performance

Model	Vector size	Fine tuned?	Precision @10	Recall @10	NDCG @10
all-MiniLM-L6-v2	384	Base	0.426	0.0189	0.4549
all-MiniLM-L6-v2	384	Fine-tuned	0.72	0.0272	0.7276
gte-large	1024	Base	0.658	0.0319	0.6708
gte-large	1024	Fine-tuned	0.87	0.0335	0.8707
NVIDIA Nemo Retriever	1024	Base	0.684	0.0304	0.7031
NVIDIA Nemo Retriever	1024	Fine-tuned	0.83	0.0310	0.8677

Comments

Embedding models are typically trained on the web, social media, and academic papers [1].

By fine-tuning on the narrower domain of COVID documents, the embedding model learns which documents are the best match for queries

[1] - Li et al, 2023 “Towards General Text Embeddings with Multi-stage Contrastive Learning” (Arxiv).

Fine-tuned LLM Performance

Model	Fine tuned?	Exact Match (%)	Subset Match (%)	F1 Score (%)
Llama v2 7B	No	6.748	26.38	32.34
Llama v2 7B	Yes	9.202	29.45	38.03
Llama v2 13B	No	9.816	26.99	39.36
Llama v2 70B (AWQ)	No	17.178	36.81	45.82

Comments

We measure extractive Q&A performance on COVID-QA [1].

Exact match: the exact answer string matches ground truth.

Subset match: the ground truth is contained in the answer string

F1 score is calculated by words common to prediction and ground truth

[1] -Moller et al, 2020 “COVID-QA: A Question Answering Dataset for COVID-19”” (ACL Anthology).