Datacamp

Building Multimodal AI Agents (From Scratch)

Apoorva Joshi

Senior AI Developer Advocate @ MongoDB

What is an AI agent?

An AI agent is a system that uses an LLM to:

reason through a problem,

create a plan to solve the problem,

execute and iterate on the plan with the help of a set of tools.

Multimodality

Multimodality is the ability of machine learning models to process, understand, and sometimes generate different types of data such as text, images, audio, video etc.

Real-world examples of multimodal data

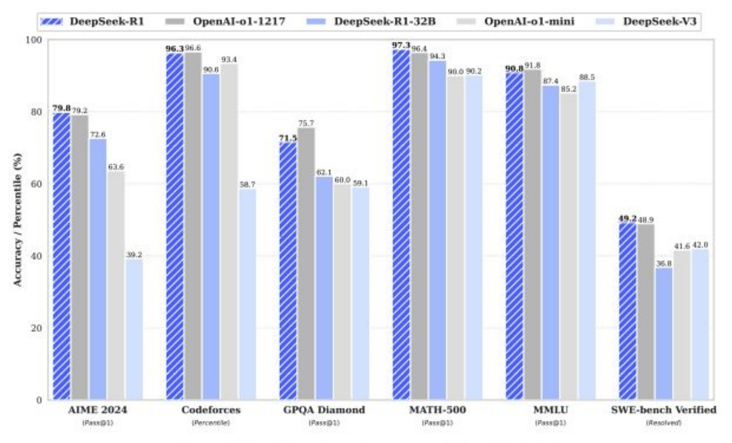
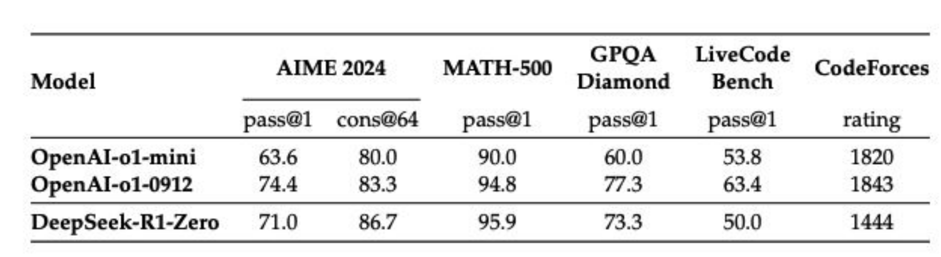
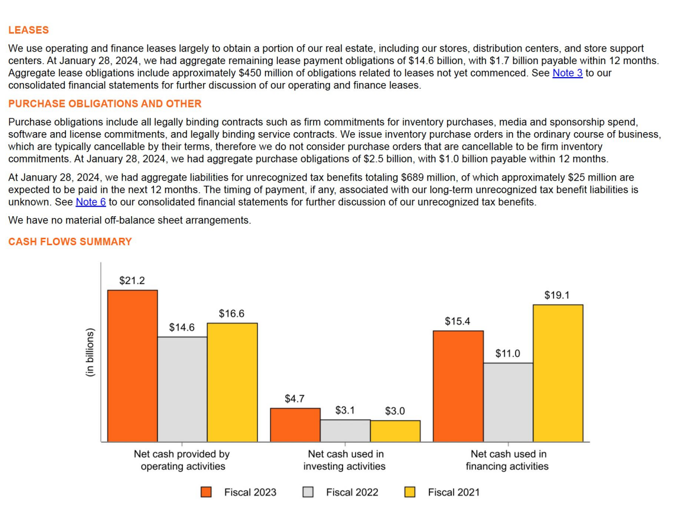


Figure 1 | Benchmark performance of DeepSeek-R1.



Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444



LEASES

We use operating and finance leases largely to obtain a portion of our real estate, including our stores, distribution centers, and store support centers. At January 28, 2024, we had aggregate remaining lease payment obligations of $14.6 billion, with $1.7 billion payable within 12 months. Aggregate lease obligations include approximately $450 million of obligations related to leases not yet commenced. See Note 3 to our consolidated financial statements for further discussion of our operating and finance leases.

PURCHASE OBLIGATIONS AND OTHER

Purchase obligations include all legally binding contracts such as firm commitments for inventory purchases, media and sponsorship spend, software and license commitments, and legally binding service contracts. We issue inventory purchase orders in the ordinary course of business, which are typically cancellable by their terms, therefore we do not consider purchase orders that are cancellable to be firm inventory commitments. At January 28, 2024, we had aggregate purchase obligations of $2.5 billion, with $1.0 billion payable within 12 months.

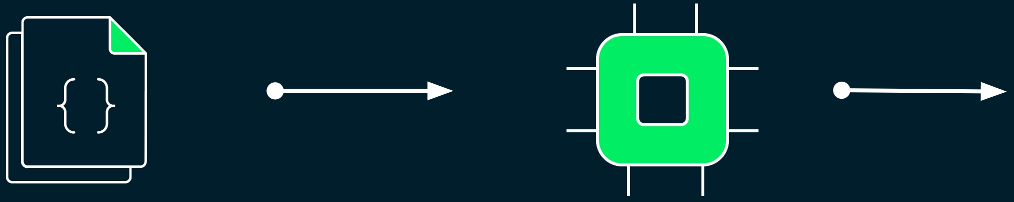
At January 28, 2024, we had aggregate liabilities for unrecognized tax benefits totaling $689 million, of which approximately $25 million are expected to be paid in the next 12 months. The timing of payment, if any, associated with our long-term unrecognized tax benefit liabilities is unknown. See Note 6 to our consolidated financial statements for further discussion of our unrecognized tax benefits.

We have no material off-balance sheet arrangements.

CASH FLOWS SUMMARY

(in billions)
$21.2
$14.6
$16.6
$4.7
$3.1
$3.0
$15.4
$11.0
$19.1
Net cash provided by operating activities
Net cash used in investing activities
Net cash used in financing activities
Fiscal 2023
Fiscal 2022
Fiscal 2021

Multimodal embedding models



Data

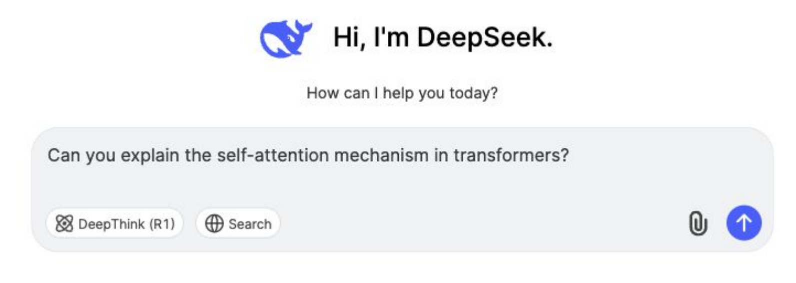
Text, images, audio, video

Multimodal Embedding Model

[0.3, 0.1, 0.2, ……, 0.4]

Vector

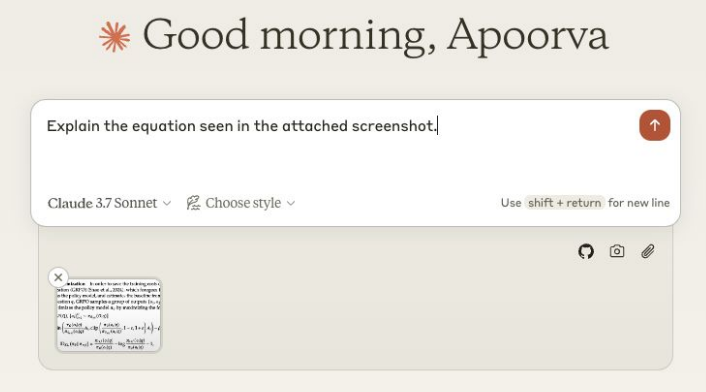
Multimodal LLMs



Hi, I'm DeepSeek.

How can I help you today?

Can you explain the self-attention mechanism in transformers?



Good morning, Apoorva

Explain the equation seen in the attached screenshot.

Claude 3.7 Sonnet

Multimodal Agents

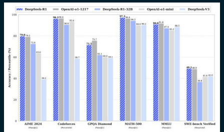


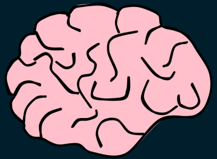
Figure 1 | Benchmark performance of DeepSeek-R1.

Multimodal data

+

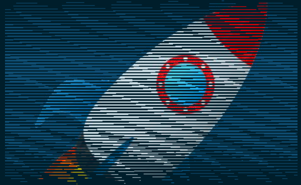
Multimodal embedding models

+



Multimodal LLMs augmented with tools and memory

=



Let’s build a multimodal AI agent!

Objectives

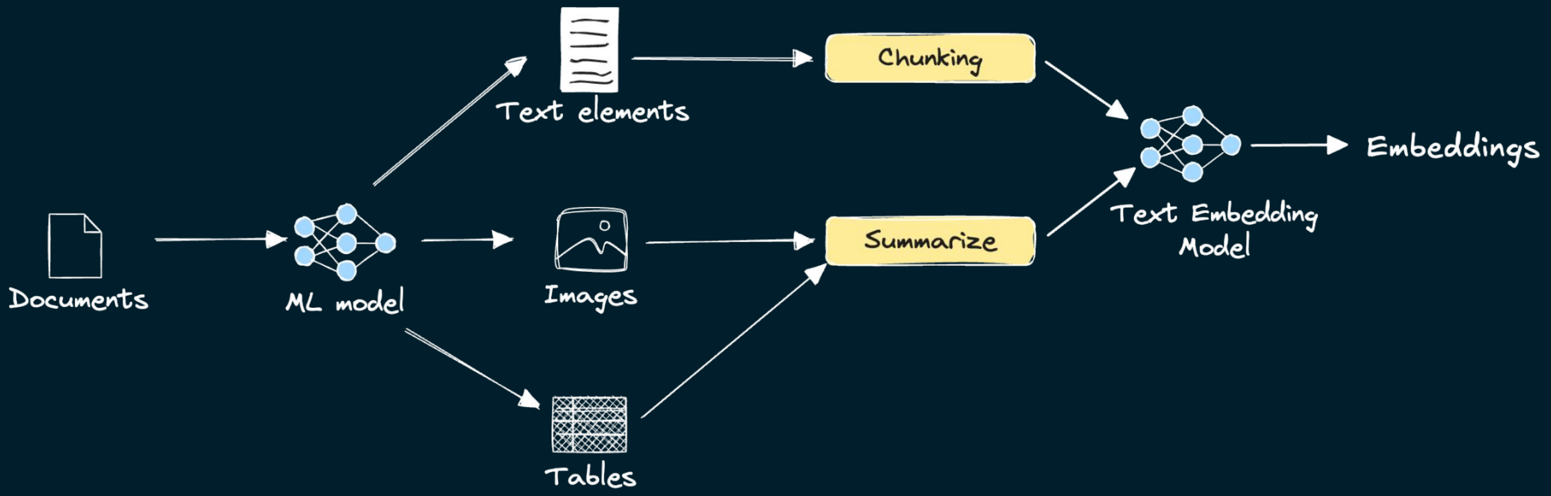
- Answers questions about a corpus of documents
- Explains charts and diagrams

Objectives

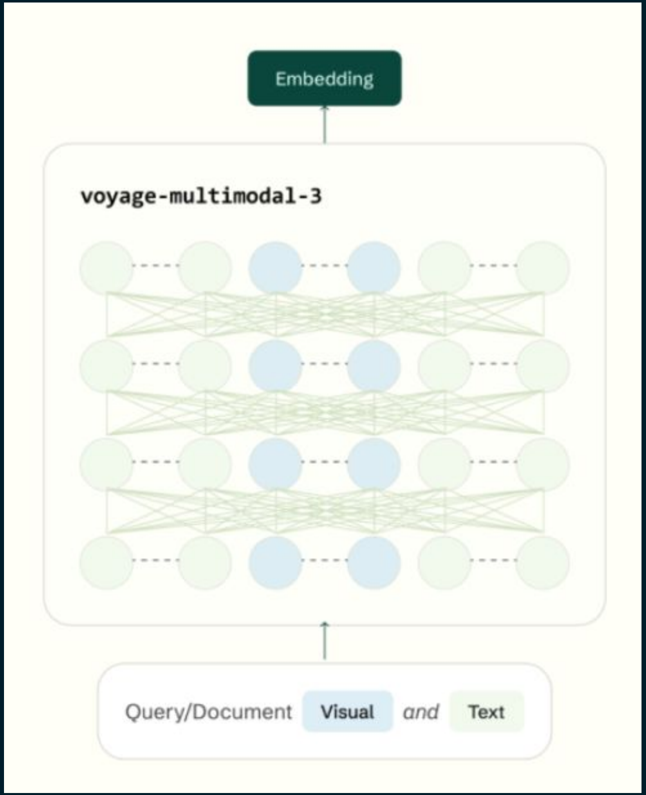
- Answers questions about a corpus of documents **with mixed modalities**

- Explains charts and diagrams

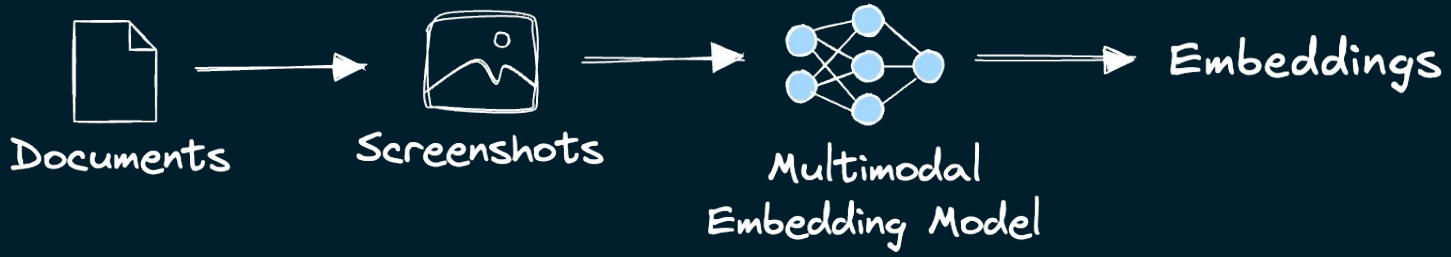
Preparing mixed modality documents for retrieval



Enter VLM-based embedding models



Screenshots are all you need



CONNECT

Apoorva Joshi

Senior AI/ML Developer Advocate @ MongoDB

Thank You!