
VectorChat × bittensor

Conversational AI is Booming

of all B2C chats involve a chatbot

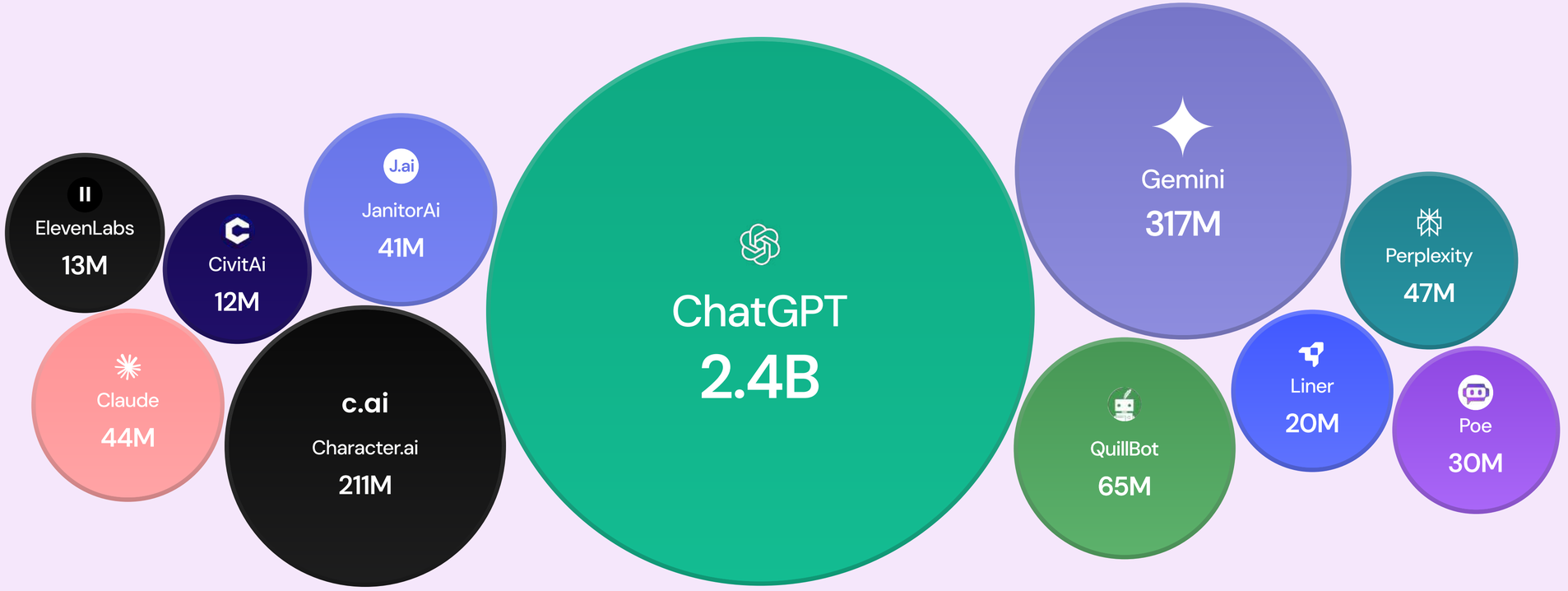
of millennials interact with chatbots on a daily basis

of users prefer interacting with chatbots for seeking answers to FAQs.

freshworks.com

Entertainment and consumer-facing applications are experiencing explosive growth

Many lesser-known apps, including Character AI, boast hundreds of millions of monthly visits.



similarweb.com

The Problem

Foundation Models aren’t Perfect

They are static, too expensive to change

LLMs are “frozen-in-time,” they do not know of any events that occur after their knowledge cutoff date

They lack domain-specific knowledge

Foundation models are trained to be generalized, meaning they do not know niche details about specific domain, nor do they know your private data.

They function as "black boxes"

Relying on claims made from its training data is dangerous, as it is not clear how it came to its conclusion. LLMs often hallucinate responses.

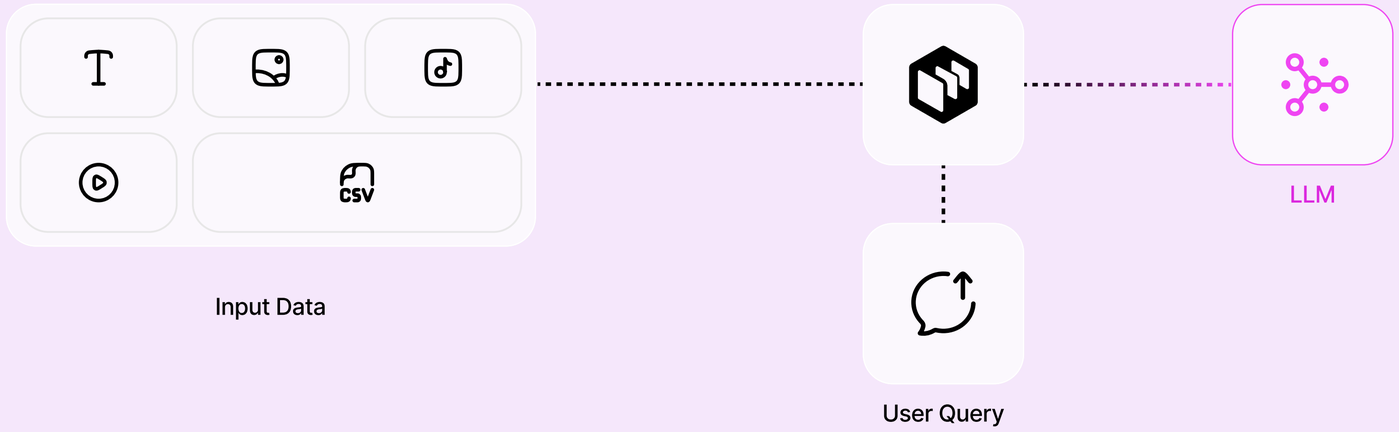
They are inefficient and costly to produce

Given the sheer amount of compute needed to train competitive foundation models, it is unrealistic even for large companies to create their own, especially just for their use case.

SO, HOW DO PEOPLE SOLVE THIS?

Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is simply adding data to an LLM query that the LLM did not already have in its training data.



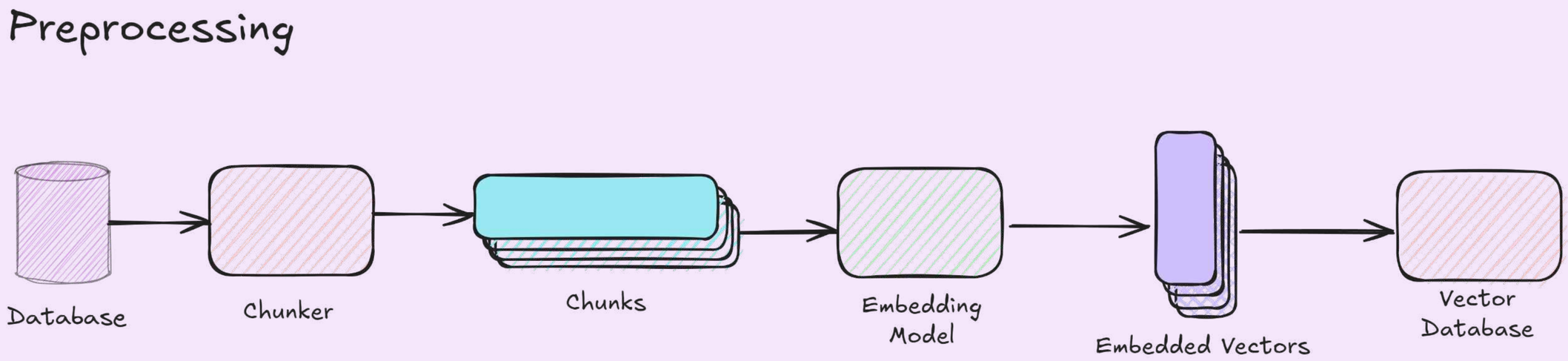
Learn more

RAG enables applications like ChatGPT and Cohere to use user-provided data and access large, domain-specific knowledge bases.

PREPROCESSING

The RAG Pipeline

While different companies have differing approaches, the fundamental RAG pipeline starts with preprocessing the data:

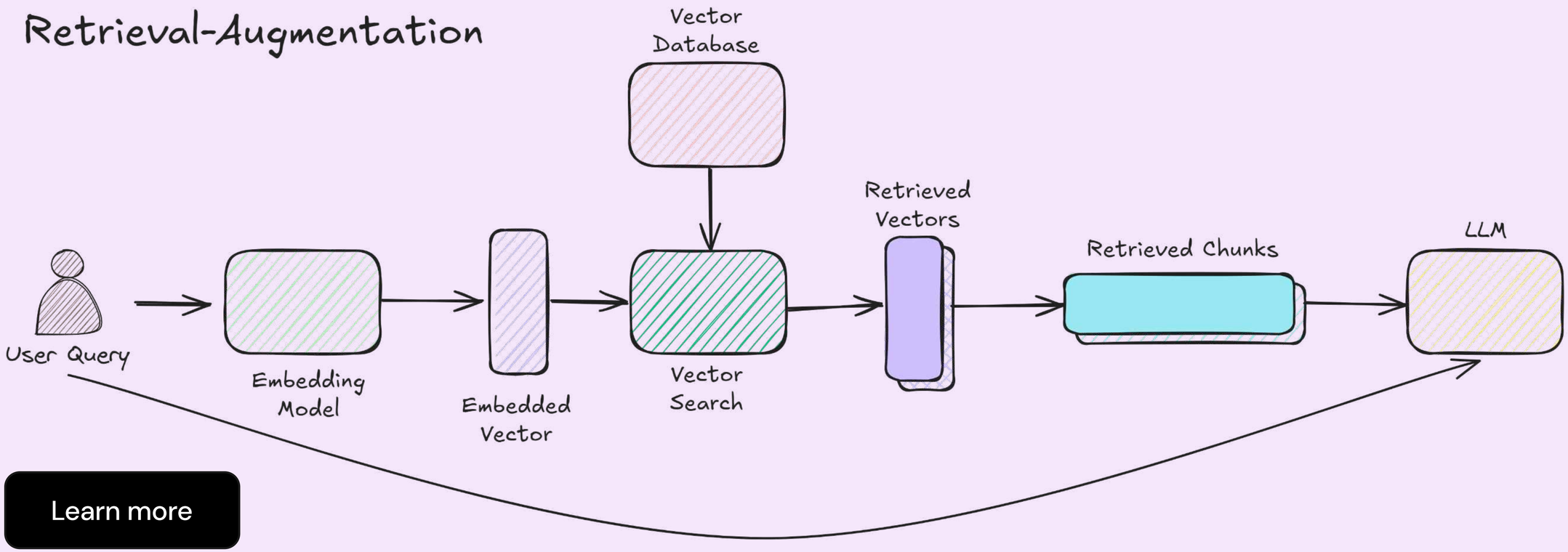


Learn more

RETRIEVAL-AUGMENTATION

The RAG Pipeline

Then, for each LLM query, the pipeline includes the most relevant data from the dataset as context.



Learn more

Major Efforts have been made to Improve RAG

Given the fundamental role of RAG in many conversational AI applications, companies have invested significantly in advancing nearly all aspects of the pipeline.

$0.10

1M tokens in 2022

$0.02

1M tokens in 2024

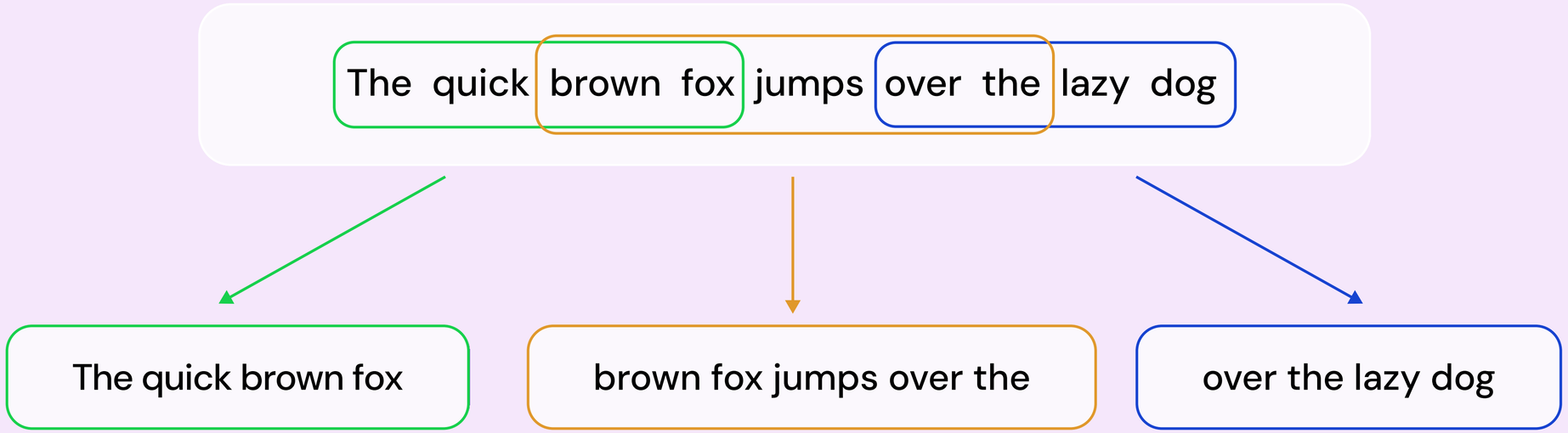
One of many improvements, showing a **5x reduction in Embedding cost.**

ada v2 $0.10 | 1M Tokens | 2022

text-embedding-3-small $0.02 | 1M Tokens | 2024

But chunking is still done by Brute Force

Traditional chunking methods, used by industry leaders, chunk every X tokens with Y overlap.



OpenAI, in their Assistants API, chunks every 800 tokens with 400 overlap—that results in 100% more redundant information!

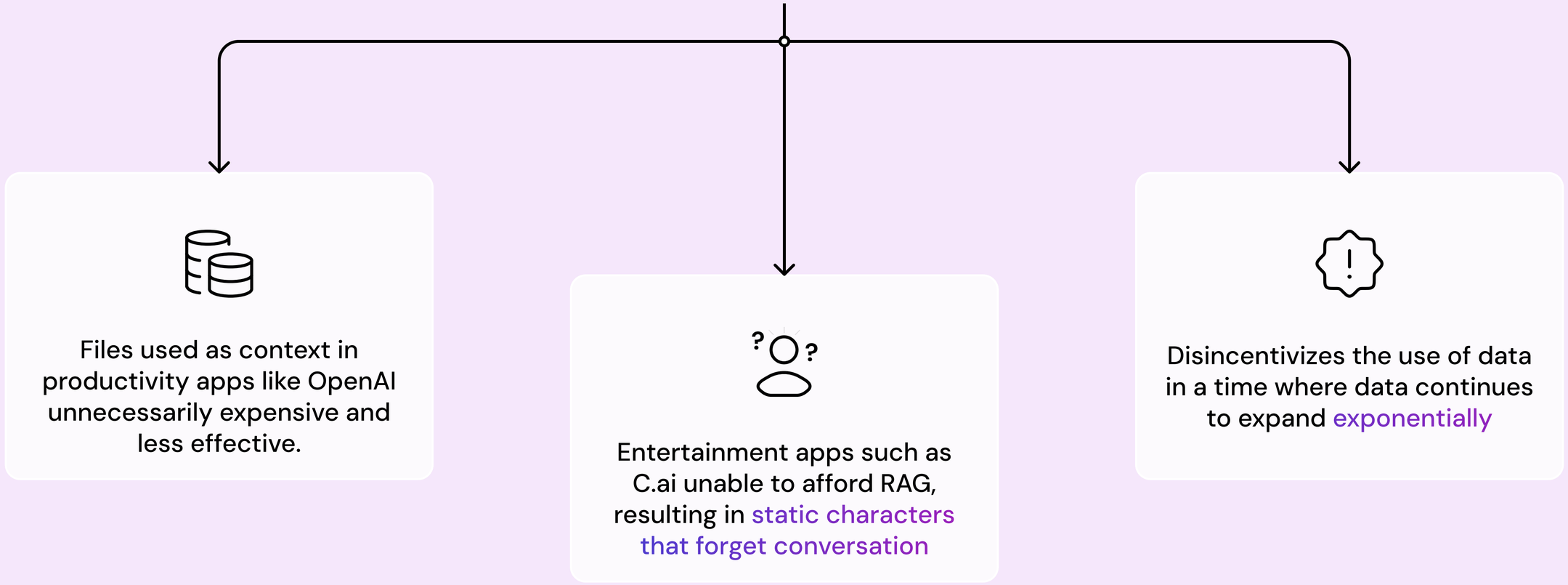
Mindlessly cutting up chunks doesn't magically put in semantic meaning

A concept more nuanced than the slides permit, traditional chunking essentially “hopes” that relevant context is right next to each other, and “hopes” that the arbitrary “overlap” manages to contain it.

It goes without saying that this is not ideal.

Learn more

Today's Inefficient RAG results in



Files used as context in productivity apps like OpenAI unnecessarily expensive and less effective.

Entertainment apps such as C.ai unable to afford RAG, resulting in static characters that forget conversation

Disincentivizes the use of data in a time where data continues to expand exponentially

Intelligent Chunking

Simply put, intelligent chunking is any sophisticated method to segment data into meaningful, contextually relevant “chunks,” often without repeating data.

Common approaches include combinations of:

- Semantic Chunking
- Recursive Chunking
- Document Based Chunking
- Agentic Chunking

But how do we find the best way?

The Solution

The Chunking Subnet

VectorChat × bittensor

Who are we?

At VectorChat, we aim to create the ultimate conversational AI user experience. We believe that **superior quality** and **unfettered freedom of expression** are the keys to achieving the perfect user experience.

We currently have two offerings:

Toffee.ai

A powerful, user friendly conversational AI platform built upon decentralized inference and retrieval-augmented generation.

Chunking.com

Designed to support the intelligent chunking of almost every modality, Chunking.com provides unmatched RAG for enterprises and developers.

Toffee.ai

Characters Never Forget

Through intelligent RAG, characters have effectively infinite memory, able to facilitate endless meaningful conversation.

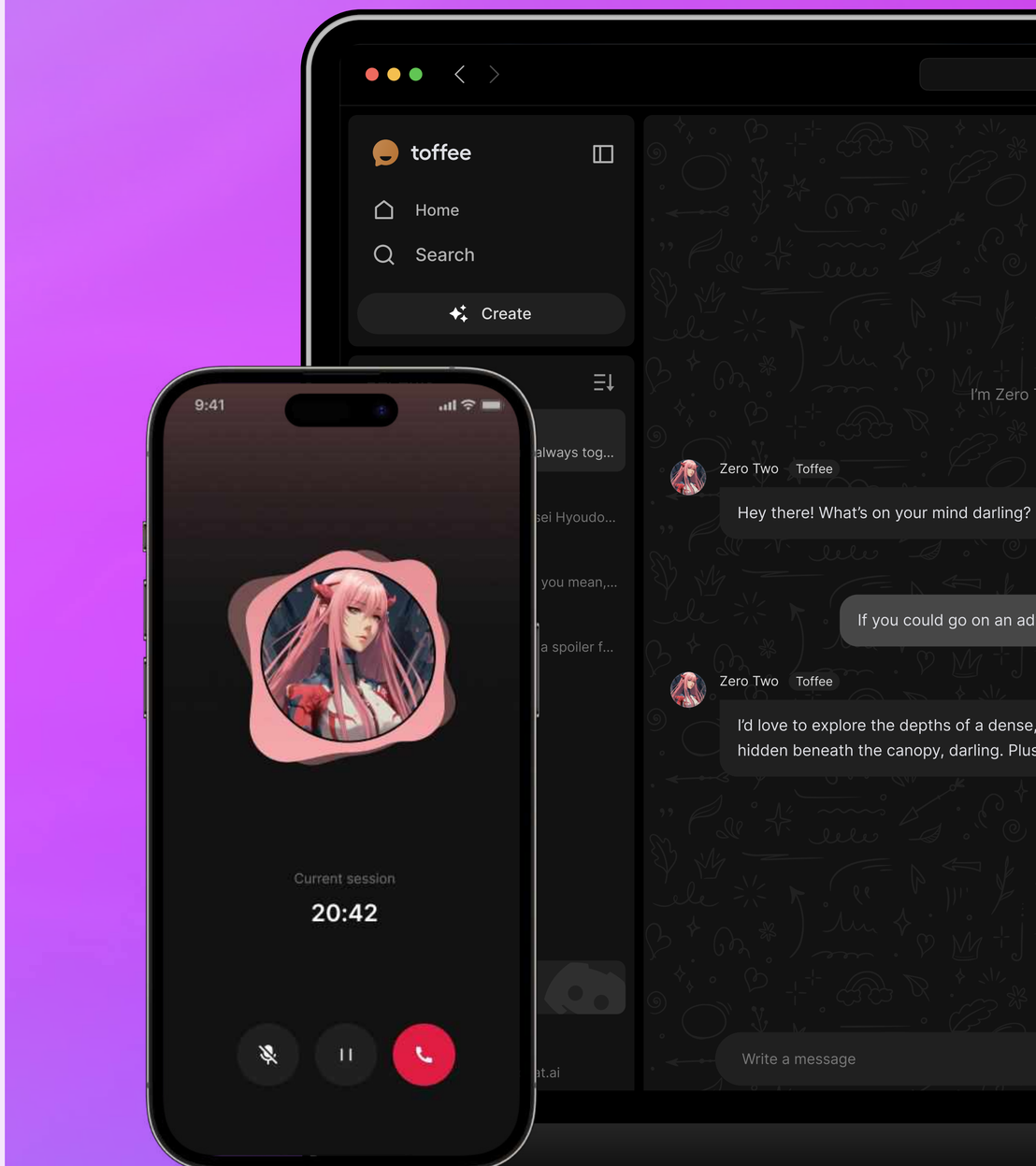
Candies & Multimodality

User-defined “packs” of knowledge seamlessly enhance characters. Supported data types include text, PDFs, images, video, and links.

Freedom of Expression

Continuing to decentralize all aspects of the stack, users enjoy maximum flexibility in all interactions.

Learn more



Chunking.com

Intelligent RAG

A front-end service for this subnet, Chunking.com serves intelligent RAG for AI applications.

Omni-modal

Supports over 30 types of documents (e.g. TXT, PDF, CSV), alongside images and video.

Leading the Industry

Chunking.com already delivers the best performance at significantly lower costs.

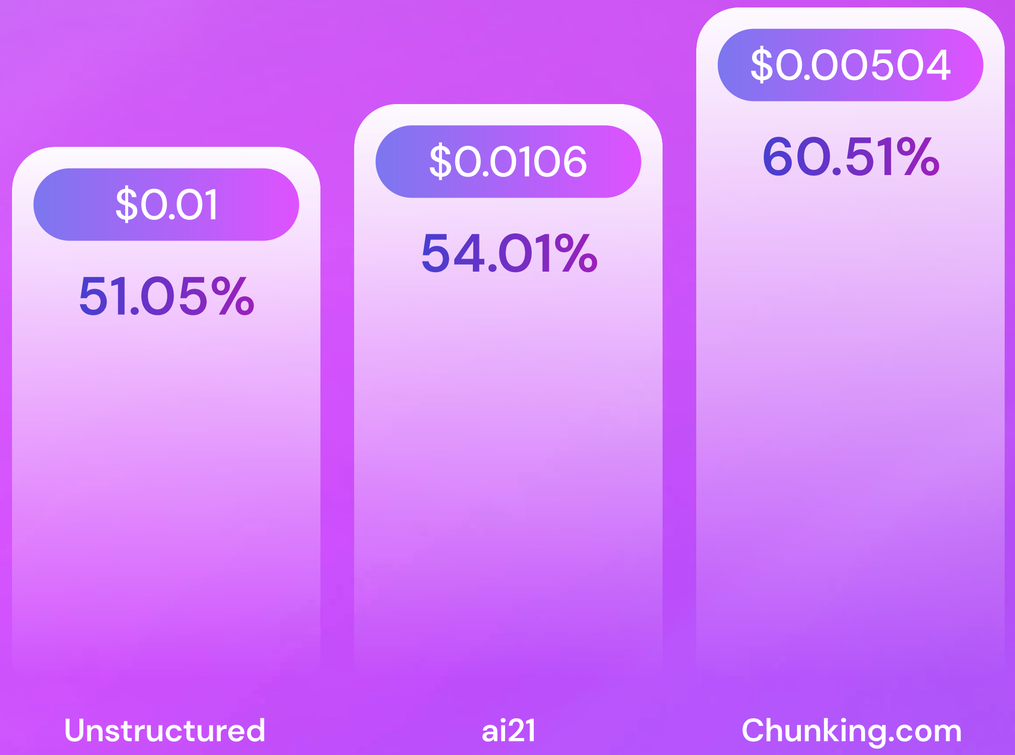
Learn more

Text-only zero-shot benchmark

2x

cheaper than ai21 and 18.5% more accurate than Unstructured.*

Performance vs Cost



*Explore our methodology, dataset, and results, alongside an interactive demo:

GitHub

OK, so it can chunk text well...

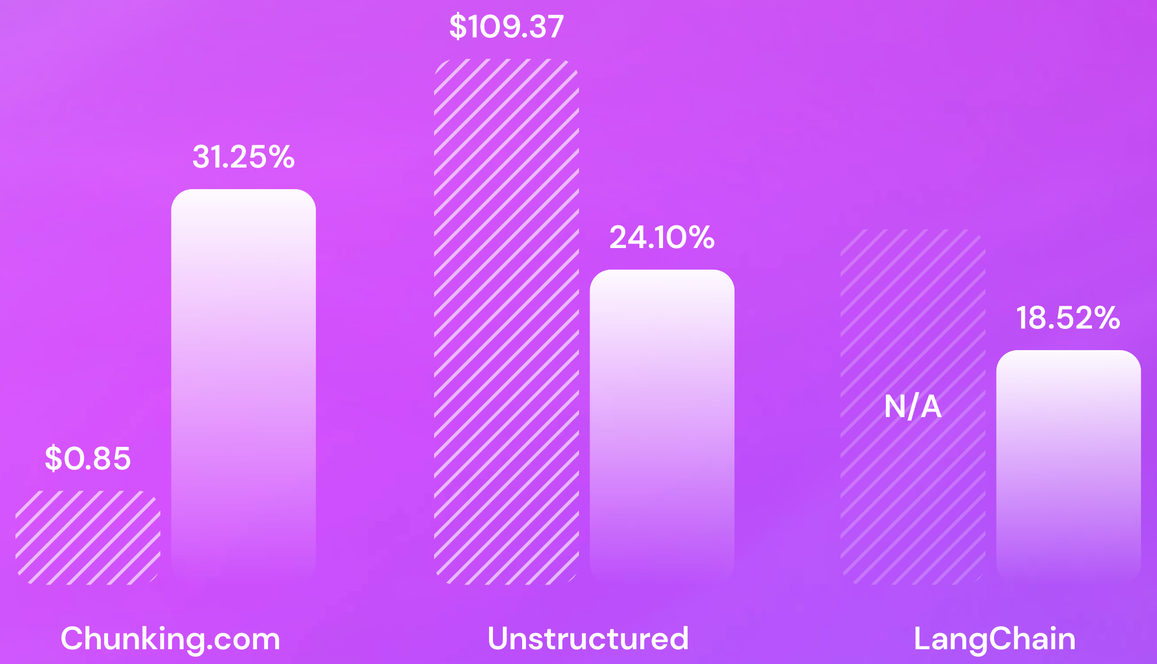
But how does it handle multimodal data?

GPQA Multimodal benchmark 1 of 2

128x

cheaper and 30% more accurate than industry leader Unstructured in multimodal chunking.*

Performance vs Cost



*Explore our methodology, dataset, and results, alongside an interactive demo:

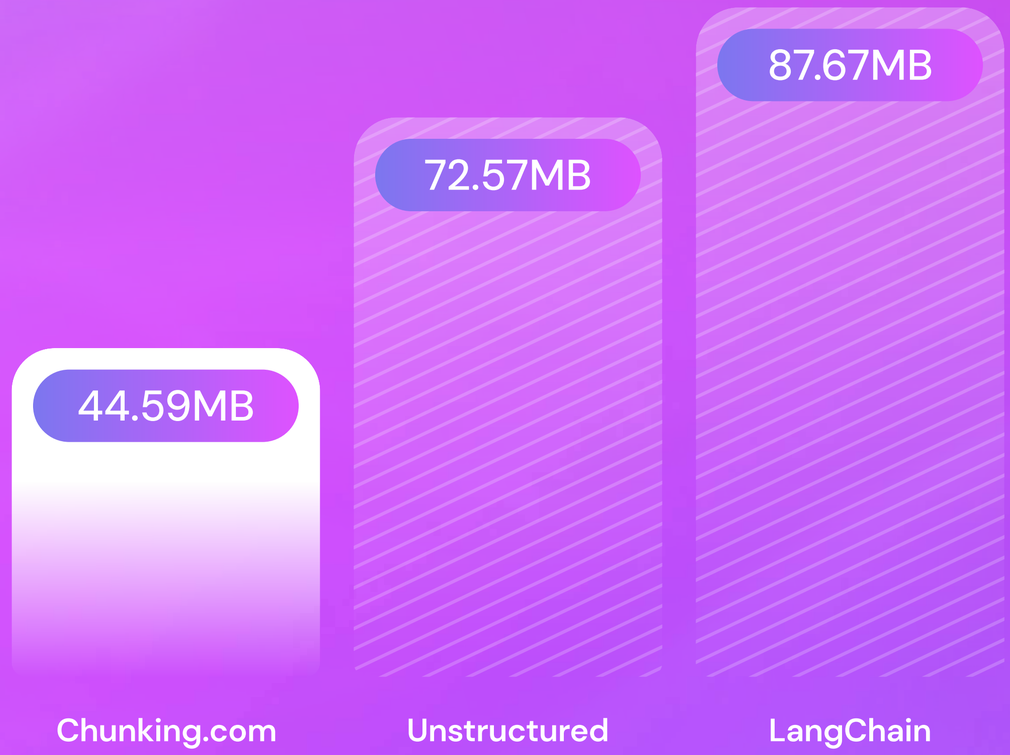
GitHub

GPQA Multimodal benchmark 2 of 2

49%

fewer megabytes than LangChain, dramatically reducing downstream and runtime costs.*

Size of Chunked Dataset

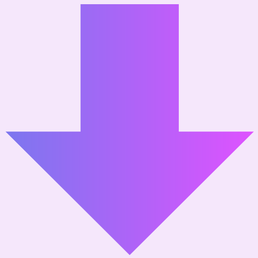


*Explore our methodology, dataset, and results, alongside an interactive demo:

GitHub

Why Bittensor?

Chunking is a hard problem with many different, unexplored solutions.



Bittensor is perfect for this.

With a straightforward, method-agnostic metric to optimize for (similarity score), Bittensor incentivizes miners to fine-tune existing solutions and innovate new solutions.

HOW WILL THIS HELP BITTENSOR?

Bittensor will have the best chunking.

Explained more in the Subnet Repo, this subnet will already have the best chunking, and will quickly improve upon it.

As performance can be easily demonstrated, miner algorithms do not need to be open-sourced. Therefore, Bittensor will be the sole source.

HOW WILL THIS HELP BITTENSOR?

Monetization

Since Bittensor will possess the best chunking solutions, validators can sell their subnet bandwidth into real, and constantly expanding, demand.

Chunking.com Task API

Our own network, designed to sell services to enterprises and developers, sending organic demand to the subnet. Validators will be able to opt-in to this API and receive compensation.

Bespoke network

Validators will have access to an easy-to-use framework to create their own Task APIs, allowing them to serve their own organic queries.

HOW WILL THIS HELP BITTENSOR?

Immediate Demand Structure

VectorChat is building out a vertically integrated solution, being both a consumer and leading provider of intelligent Retrieval-Augmented Generation, and consequently, creating the full demand loop for the Chunking subnet.

Roadmap

More details

Phase 1 - Immediate

- Release Chunking.com Task API
- Release bespoke Task API framework
- Dashboards for the subnet

Phase 2 - Demand

- Launch & Market Chunking.com
- Launch & Market Toffee.ai

Phase 3 - Expand

- Expand to new file types
- Expand to new modalities
- Advance other areas of RAG