

Generative AI in Business

An outlook on the next 3-5 years of AI technology



**Your Host:
Casey Brant**

**Standards Coordinator
Conexxus, Inc.**



**Your Moderator:
David Ezell**

**Director of New Initiatives
Conexxus, Inc.**

Agenda

- ▶ Housekeeping
- ▶ About Conexxus
- ▶ Presenters
- ▶ Presentation
- ▶ Q&A

Housekeeping

This webinar is being recorded and will be made available on Conexxus365.org

Participants

- ▶ Ask questions via webinar interface
- ▶ Please, no vendor specific questions
- ▶ Our webinars may be used toward PCI continuing education credits. Please contact 365@conexxus.org for questions regarding a certificate of webinar attendance for qualifying live events.

Interested in speaking or sponsoring a Conexxus365 event?

Contact 365@conexxus.org to discuss upcoming opportunities with our team.

Disclaimer

Conexus does not endorse any products or services that may be described or mentioned in this presentation.

The views and opinions expressed in this presentation are solely those of the speakers and not of Conexus.

By hosting this webinar, Conexus is not providing any legal advice; if you have any questions about legal issues raised or discussed, you should seek the assistance of attorneys who are competent in that area.

Thank you to our 2023 Annual Sponsors!

Diamond Sponsors



Emerald Sponsors



Garnet Sponsors



About Conexus

- ▶ We are an independent, non-profit, member driven technology organization
- ▶ We set **standards**...
 - ▶ Data exchange
 - ▶ Security
 - ▶ Mobile commerce
- ▶ We provide **vision**
 - ▶ Identify emerging tech/trends
- ▶ We **advocate** for our industry
 - ▶ Technology is policy



Connect with Conexxus



www.conexxus365.org
365@conexxus.org



www.conexxus.org
info@conexxus.org



[@conexxus.org](https://www.linkedin.com/company/conexxus.org)



Alan Lockett, Ph.D.

Generative AI in Business

An outlook on the next 3-5 years of AI technology

Goals of this Presentation

1. Understand current capabilities of commercial AI technology
2. Explore applications enabled by current AI tools
3. Understand the future capabilities of AI coming online in 3-5 years
4. Conceptualize how these capabilities will impact future business opportunities

A Brief History of AI

Where are we, and how did we get here?

In which we review what AI can do, how it does it, and where it came from

Today, AI Can...

USER Please translate the quoted sentence to Thai: "As the light faded, a fleeting wind carried our breath away with a gasp, and we briefly communed with the Infinite and Timeless, until in a blink we found ourselves once again in this mortal world, bound on every side by unreachable horizons."

ASSISTANT "เมื่อแสงสว่างจางหายไป ลมพัดผ่านอย่างชั่วขณะพาลมหายใจของเราหายไปด้วยการแสดง และเรารู้สึกกับสิ่งไร้ขอบเขตและไม่มีเวลาสักครู่ จนกระทั่งในการสะกดตาหนึ่งเราพบว่าตัวเองอยู่ในโลกมนุษย์นี้อีกครั้ง จำกัดจากทุกด้าน โดยขอบฟ้าที่ยากจะประลอง."

USER What does the author of the above quote intend to convey?

ASSISTANT The author seems to be describing a profound, almost spiritual experience. The fleeting moment of being one with the "Infinite and Timeless" suggests an ephemeral but intense connection with something far greater and beyond human understanding, possibly cosmic forces or a higher power. Coming back "to this mortal world, bound on every side by unreachable horizons," signifies a return to the confines and limitations of human existence. This quote carries a sense of awe, wonder and poignancy, emphasizing on the ephemeral nature of such powerful experiences and the longing for this connection that remains unfulfilled due to human limitations. This could be a metaphor for human life, its fleeting joys, our search for meaning, and the inevitability of mundane existence.

USER Please summarize your last response in just one sentence of 10-15 words.

ASSISTANT The author describes a brief, intense connection with the infinite, and longing upon returning to mortal limitations.

OpenAI's ChatGPT

Hold a conversation on any topic

Translate into any language

Ace professional exams (e.g. law)

Generate images from text

Analyze textual content

Summarize complex text

Retrieve information

Answer questions about data

Generate software code

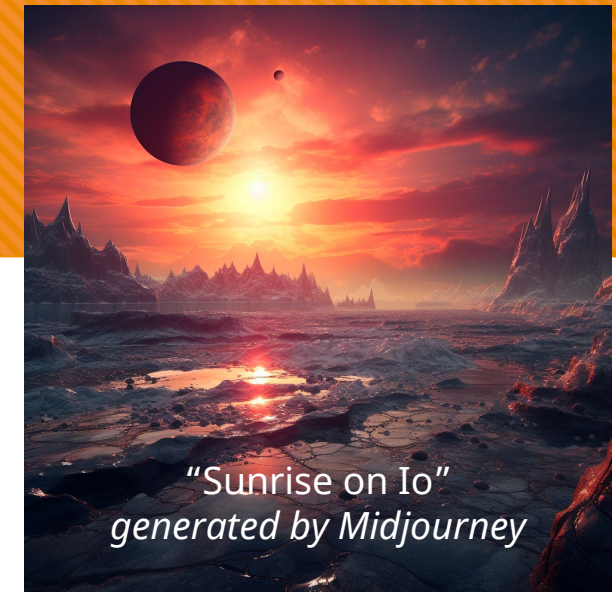
Transcribe human voices

Read text out loud like a human

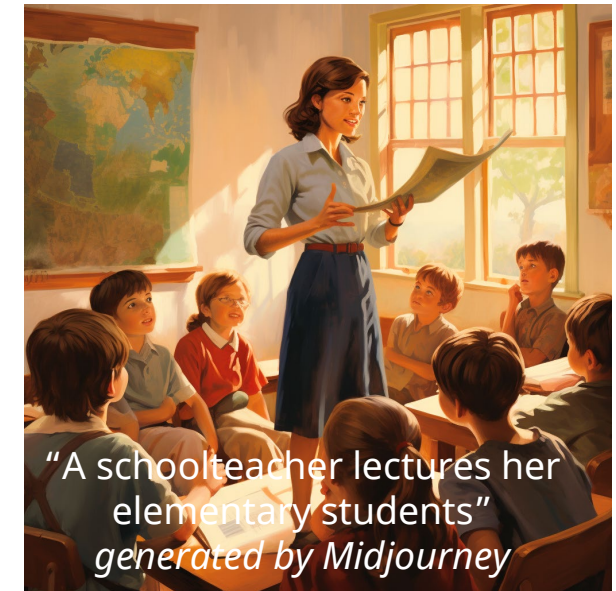
Copy a human voice

Generate natural audio from text

13



"Sunrise on Io"
generated by Midjourney



"A schoolteacher lectures her elementary students"
generated by Midjourney

What are the Limits of Today's AI

1. **Hallucination:** AI imagines the *average outcome* over its training data and will make up "information" to fill in the gaps in its knowledge
2. **Limited Insight:** Today's AI is like the perfect in student in the class, always pleasing the teacher, but never seeing beyond the lessons
3. **Huge Expense:** The GPU hardware that runs modern AI is expensive and, for now, rare. This presents hard scaling problems for many applications
4. **Limited Audio / Video:** AI cannot generate or process video quickly or well at this point in time; audio is further along, but lagging behind
5. **Disembodiment:** An AI has limited tools to test or check its assertions, and does not benefit from the everyday human experience of external feedback

A History Lesson

1780 – Luigi Galvani discovers that electricity moves animal muscles

1930s – Early digital computers are based on the model of neural circuits

1958 – Rosenblatt builds first Perceptron for image recognition

1982 – Werbos demonstrates backpropagation for neural chains

1990s – Graphical models incorporate Bayesian statistics for NN models

2012 – Hinton's students build the first human-level image classifier

2019 – GPT-1 shows next-word prediction leads to new capabilities

1890s – Ramon y Cajal maps the structure of neural axon branching

1943 – McCulloch and Pitts propose first artificial neuron, the Perceptron

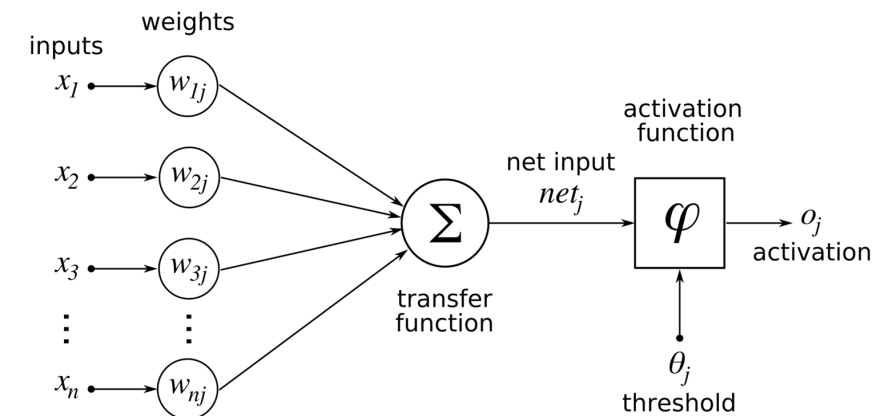
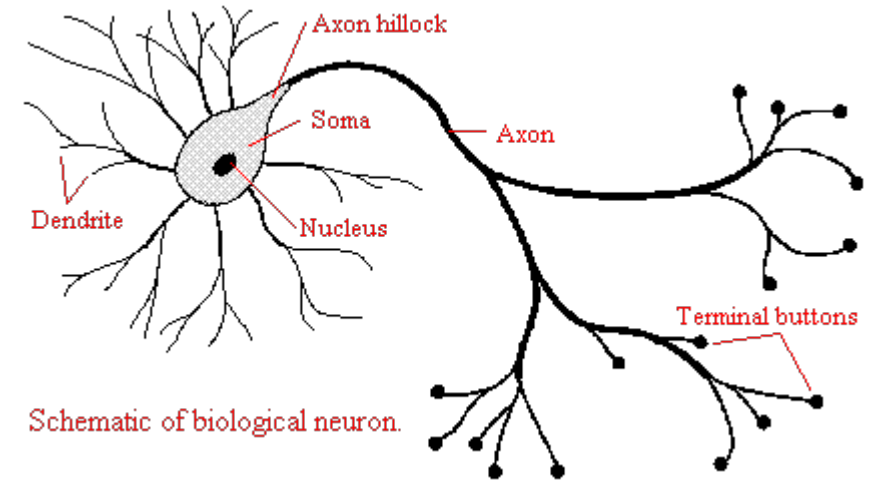
1969 – Minsky and Papert identify mathematical limits of Perceptron

1987 – LeCun introduces Convolutional NNs for images

2006 – Hinton demonstrates deep neural net for digit classification

2017 – Google researchers invent the transformer architecture for NNs

2023 – ChatGPT's public debut is recognized publicly as a turning point



Why Now?

1. Large neural networks could only work once computers were fast enough
2. Academics resisted the simplicity of the neural network idea
3. Transformer architecture had to be invented / discovered
4. The general capabilities of Large Language Models were unexpected

Large Language Models

Next Word Prediction
a.k.a. **Language Modeling**

... over the lazy dog



Neural Network with
Transformer Architecture



The quick brown fox jumped _____

Text Completion is a “Master Task”

Question Answering What is the capital of Hawaii? _____

Translation “The dog” translates to French as _____

Summarization This is a summary of War and Peace. _____

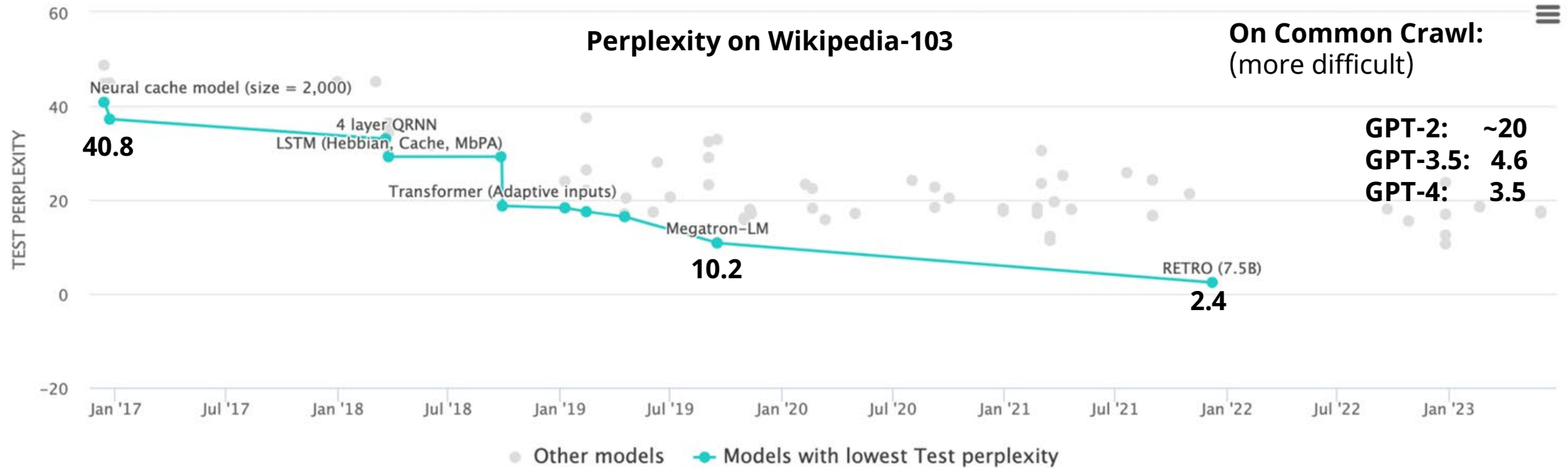
Large means Really Big

GPT-3 (OpenAI) has 175 billion parameters = 700 Gigabytes

Llama-2 (Meta) was trained on 1 trillion tokens (1 word = 2-3 tokens)

More parameters + more training data = smarter model

Perplexity: Measure of a Language Model

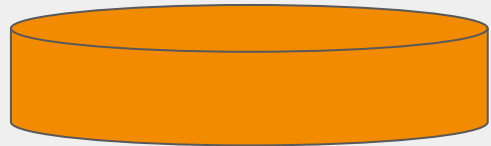


Perplexity is the **odds** of getting the next word right

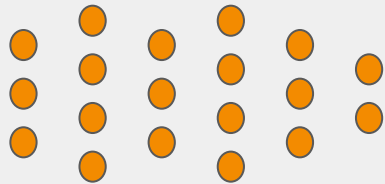
Matching Human Preferences

RLHF: Reinforcement Learning with Human Feedback

Phase 1: Next Word Prediction



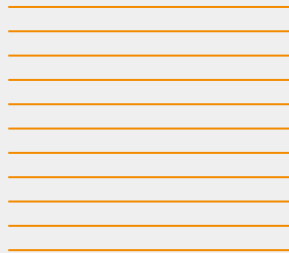
Gather Trillion Token Corpus



Train Billion Parameter Model



Phase 2: Supervised Fine-Tuning



Fine-tune model on example dialogues, instructions, etc.

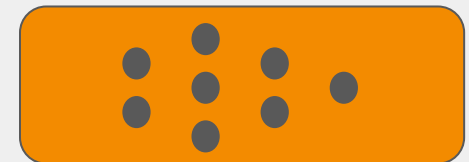


Phase 3: Human Preference Tuning

Model generates chat candidates



Humans judge machine candidates



Train reward model to judge LLM & fine-tune LLM to optimize reward

Solving Problems with LLMs

Prompts, Tokens, Context Length, Vector DBs, RAG
and even more

In which we see how tasks are
encoded into LLMs and solved,
and wherein we first encounter
business problems and solutions

Using LLMs to Solve Problems

LLMs can solve tasks like:

- Question Answering
- Translation
- Summarization
- Code Generation

The task is controlled by a free-form text prompt that tells the LLM what to do (e.g., “translate this ...”)

Prompt Engineering refers to the art of wordsmithing a good prompt that the AI responds well to.

LLMs work by extending a sequence one token at a time



The sequence can continue as long as the context length is not exceeded

Example SQL task with GPT-4

Prompt instructs the LLM what its task is.

SYSTEM

You are a data translator that takes in English statements and transforms them into SQL queries. The user makes a statement, and you provide only the query with no commentary.

Questions by the user pertain to the table:

```
CREATE TABLE employees (  
  emp_id integer,  
  first_name VARCHAR2(25),  
  last_name VARCHAR2(25),  
  job_name VARCHAR2(10),  
  manager_id INTEGER,  
  hire_date DATE,  
  salary DECIMAL(10,2)  
);
```

All responses should be formulated to query this table.

You can provide 1,000s of words here

USER List all employees with surnames starting with 'A'

ASSISTANT SELECT * FROM employees WHERE last_name LIKE 'A%';

USER How many employees started between December 10, 2005 and January 3, 2007 that also have a salary of less than \$100,000?

ASSISTANT SELECT COUNT(*) FROM employees WHERE hire_date BETWEEN '2005-12-10' AND '2007-01-03' AND salary < 100000;

USER List all employees who are paid more than their manager

ASSISTANT SELECT E.*
FROM employees E, employees M
WHERE E.salary > M.salary
AND E.manager_id = M.emp_id;

GPT-4 generates based on the prompt + user statements

Tokens and Context Length

LLMs don't operate on words directly; instead they break words into **tokens** that are chosen to efficiently encode the text.

In | _a | _hole | _in | _the | _ground | , | _there | _lived | _a | _hob | bit | .

The **Transformer Architecture** (Vaswani, 2017) developed out of machine translation applications, where **attention** mechanisms were used to figure out word order in different languages. Vaswani et al realized this attention mechanism on its own was powerful enough to stand on its own and wrote a paper "Attention is All You Need". Transformers are used inside all major LLMs.

The Transformer Attention module determines which elements in a sequence are needed to compute another element of a sequence. It identifies how sequences / sentences are internally connected (e.g., who does the pronoun *its* refer to in the paragraph above).

BUT, the attention mechanism has a limited range of other tokens it can look at for computational reasons. So transformer LLMs have a **maximum number of tokens** that they can consider, known as the **context length**. For most commercial applications, the context window ranges from 2,048 to 8,192. The longest commercially available context length is 65k tokens (Anthropic / Claude); the enterprise version of GPT-4 can take up to 32,768 tokens.

CONTEXT LENGTH REALLY MATTERS because it is the limit on how much text you can process at once with an LLM. It limits how much text you can summarize, how many documents you can consider to answer a question, how much text you can translate, etc.

Augmenting LLMs with Big Data

LLMs can handle context of ~20-50 pages of text provided in the prompt.

1 token ~ 2 words, so 250 words / page = 500 tokens / page

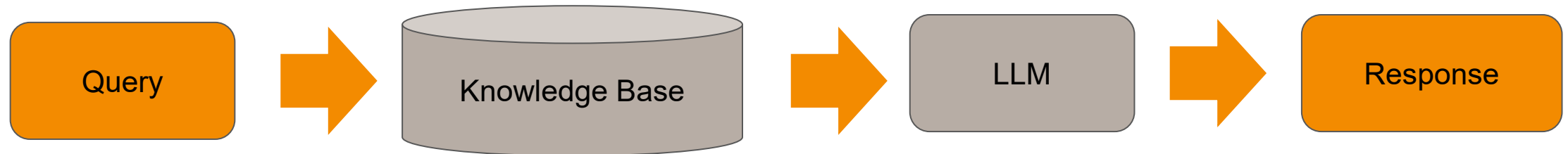
4,096 tokens yields about 8 pages of text (GPT-3.5, Llama-2)

32,786 tokens yields about 65 pages of text (GPT-4-32k)

What if we need more context than that?

For example, if we want to query the entire US legal code, or all SEC filings, or the entire corpus of English literature?

We then have to choose a smaller set of information to send to the LLM – we have to pick out the relevant material



Augmenting LLMs with Big Data

LLMs can handle context of ~20-50 pages of text provided in the prompt.

1 token ~ 2 words, so 250 words / page = 500 tokens / page

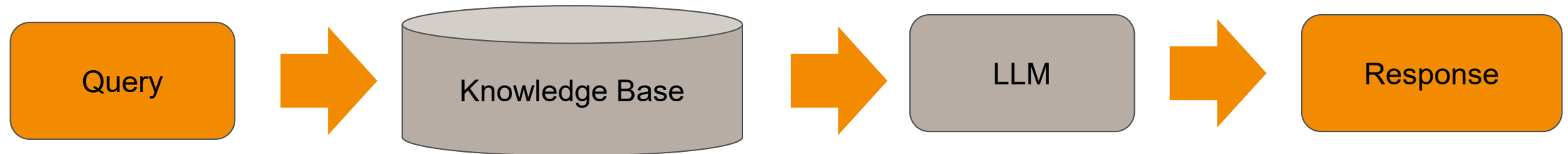
4,096 tokens yields about 8 pages of text (GPT-3.5, Llama-2)

32,786 tokens yields about 65 pages of text (GPT-4-32k)

What if we need more context than that?

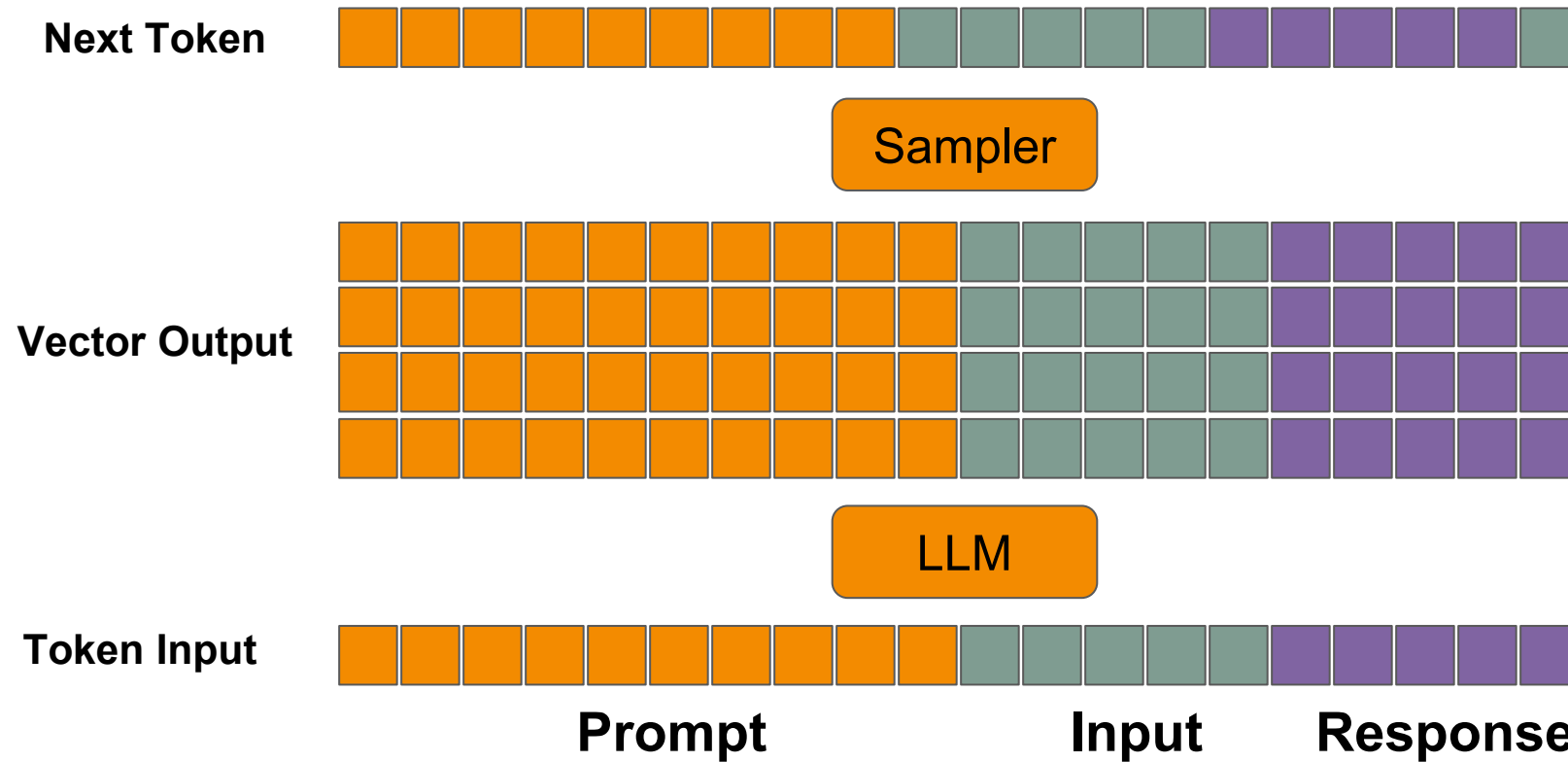
For example, if we want to query the entire US legal code, or all SEC filings, or the entire corpus of English literature?

We then have to choose a smaller set of information to send to the LLM – we have to pick out the relevant material



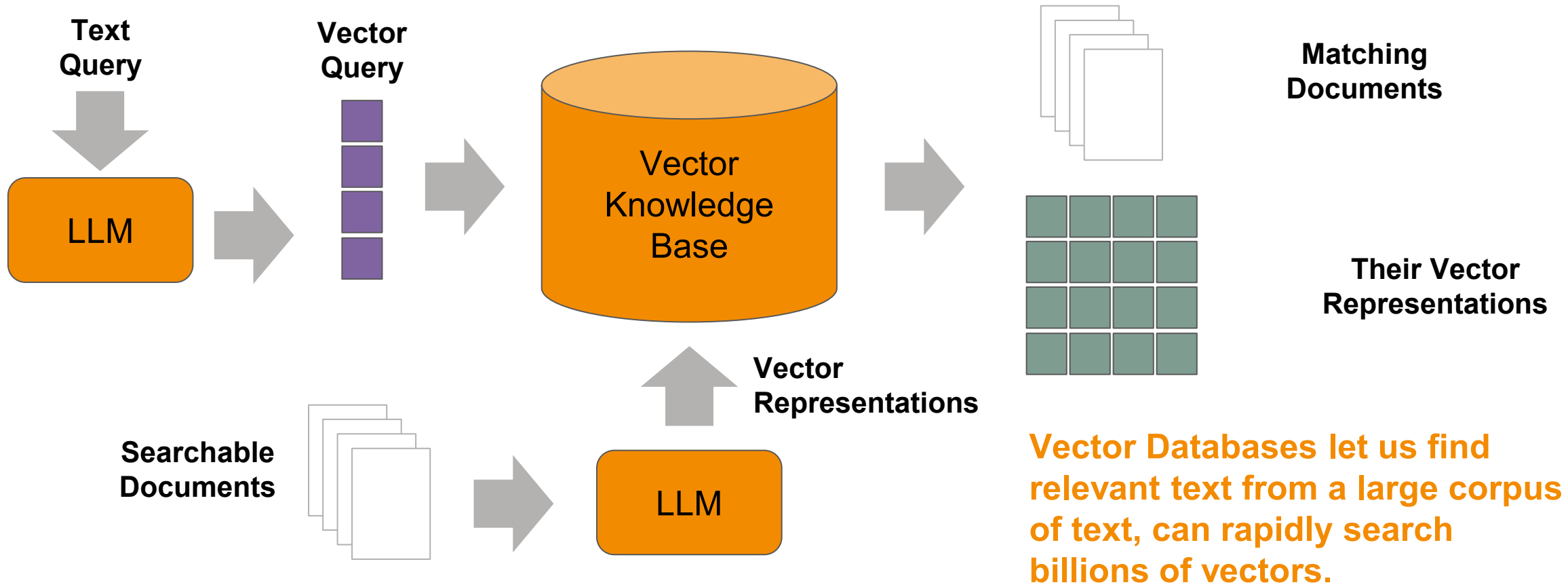
Vector Representations

LLMs process the input into sequences of numbers called *vectors*.

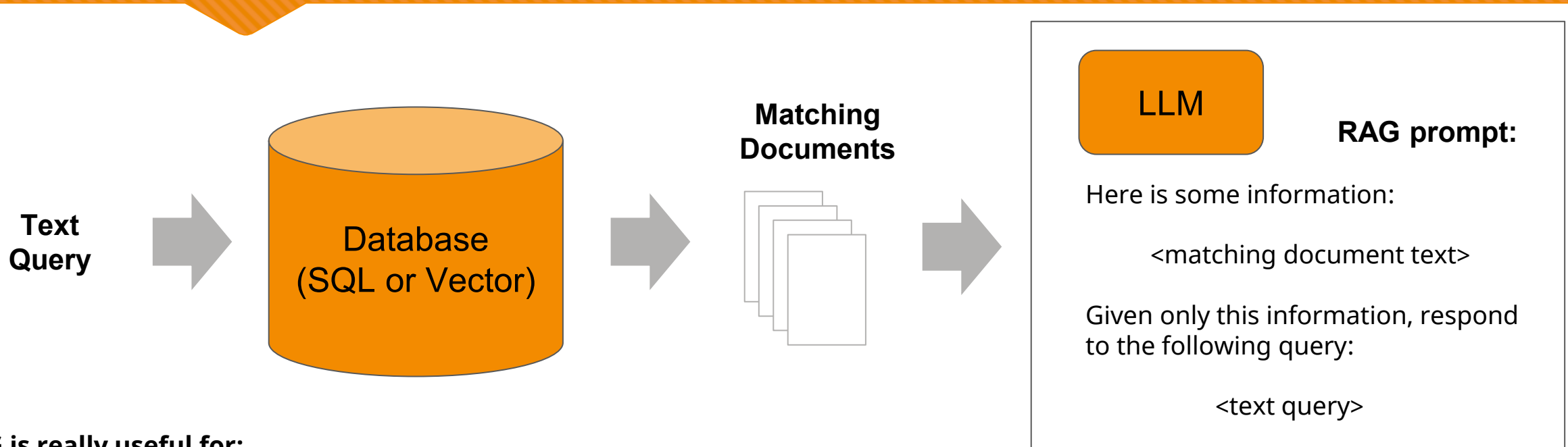


These vectors encode the semantics of the input numerically

Vector Databases a.k.a. AI Search



Retrieval-Augmented Generation (RAG)



RAG is really useful for:

1. **Limiting hallucinations**
2. **Incorporating large bodies of information**
3. **Managing context length**
4. **Assembling generative pipelines**

Note that the Database does NOT have to be a vector database, and if it is a vector database, it does NOT have to use the same LLM for embeddings (e.g. it might use a cheaper LLM)

Chain-of-Thought Reasoning

LLMs often make reasoning mistakes when multiple steps are required.

Chain-of-Thought prompting demonstrates to the LLM how to solve a problem, which significantly improves the quality of reasoning in LLM responses.

On math word problems, chain-of-thought prompting increased performance for one LLM (PaLM):

18% → 57%

of problems answered correctly.

Recent approaches just ask the LLM a series of questions:

1. What is the goal?
2. What are steps to accomplish this goal?
3. [For each step] Perform the step
4. Return the answer

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Image from Wei et al., <https://arxiv.org/pdf/2201.11903.pdf>

Calling Functions from LLMs

Toolformer: [Meta paper](#) demonstrates that LLMs can issue commands to 3rd party APIs (e.g. SQL to databases) when needed to answer a question.

AutoGPT: Open-source project allows you to build agents that perform a task through a process of (a) generating thoughts; (b) making a plan; (c) criticizing the plan; (d) generating the next step.

ChatGPT Function Calls: You can create ChatGPT plugins, upload them, and GPT models will [generate function calls](#) that execute these plugins, e.g. `send_email()`, `schedule_meeting()`, etc.

Tools like these provide a new way to build functioning software applications that are only partly made up of software code!

A New Kind of Modular Programming

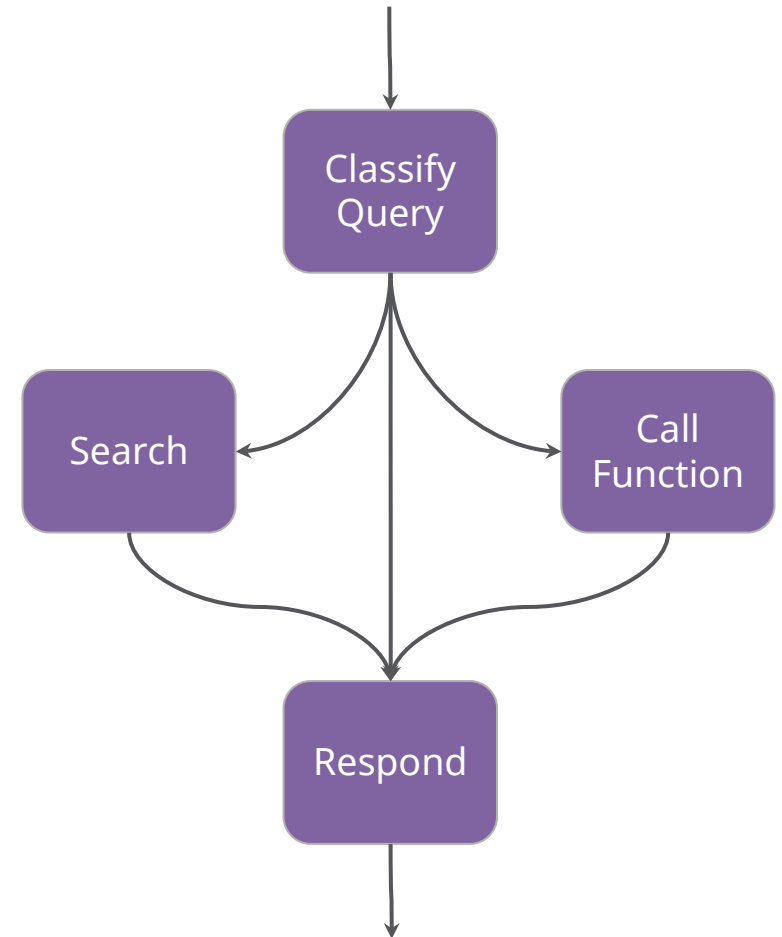
Several ways to invoke LLMs

1. Direct question or completion
2. AI Search with RAG
3. Calling external functions
4. Generating images or code
5. Prompting with images

Each of these become “modules” in a new type of program; we can assemble a graph of LLM actions, tied together with software, called a **chain**.

The I/O modality is natural human language or images. In its ultimate form, we will have programming that is human understandable.

LangChain provides a software package that does exactly this (more below).



CoTS Solutions

Very good, but I'm not building it from scratch.
Where do I get it, and what's it going to cost me?

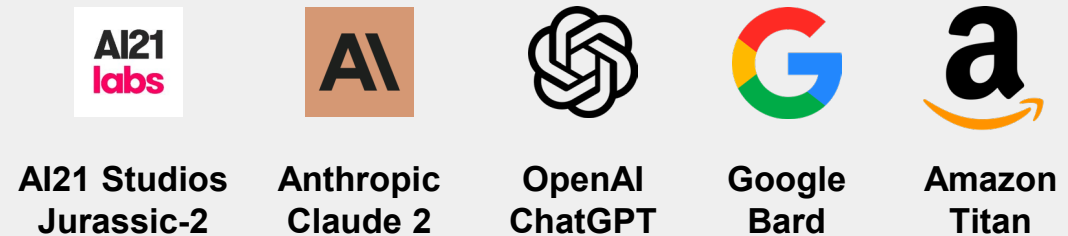
In which we introduce the landscape of commercial AI tools, from enterprise solutions to open source libraries and models

COTS Conversational Models

Custom Models

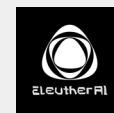
General-Purpose

Proprietary

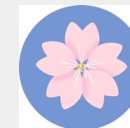


Open Source

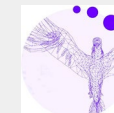
Open-source models can be customized as needed



Eleuther AI
GPT-Neo



BigScience
Bloom



TII
Falcon



Mosaic ML
MPT



Meta
Llama 2

Where can you get AI hardware?

Major Cloud Providers



Microsoft
Azure



Amazon
AWS



Google Cloud

Latest gen hardware not available
Raw prices ~\$4 / GPU-hour
Expect to reserve negotiate
Useful for APIs, HA services

AI Cloud Providers



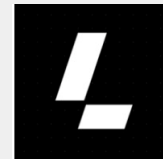
CoreWeave



Lambda
Labs



Mosaic ML



Latitude.sh

Latest gen hardware available
Raw prices ~\$2 / GPU-hour
Services & integration often poor
Cheapest way to get good GPUs

Hardware Vendors



nVidia



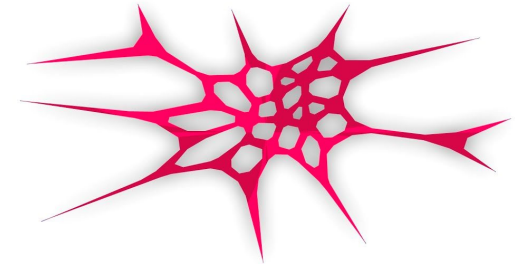
nVidia A/H100 standard in 2023
OpenAI Triton offers alternatives
AMD has Radeon chips
Intel working on Xe chip

COTS Vector Databases



Pinecone.io provides cloud APIs to create and query vector databases based on Meta's `faiss` software

FAISS
Scalable Search With Facebook AI



Zilliz provides cloud APIs to create and query vector databases based on the open-source Milvus project



Both solutions provide high-availability & fast response, but at significant expense. Open-source variants can be made to run faster & cheaper, but require extensive expertise to use. **Note: Their websites also provide lots of recipes & techniques for using LLMs that are quite useful!**

LangChain

LangChain (<https://langchain.com>) provides Python and Javascript libraries to invoke chains of LLM behavior. It provides control flow to combine various AI capabilities, including:

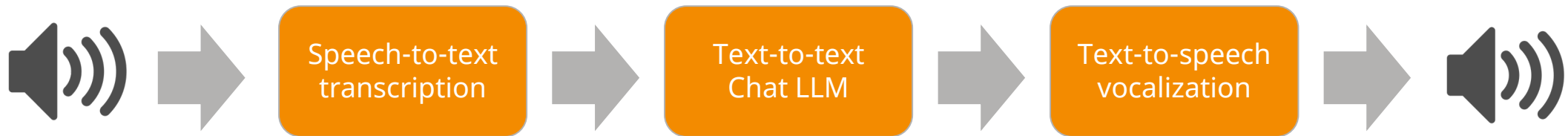
1. **Model I/O:** standard classes for interacting with 3rd party models
2. **Retrieval:** implementation of RAG abstracted for easy use
3. **Chains:** preset templates and tools to construct chained LLM actions
4. **Agents:** conditionally choose how to execute chains based on the request
5. **Memory:** to persist information between different steps of the chain
6. **Callbacks:** to enable logging and monitoring

Their website is [full of examples](#), and provides lots of guidance for getting started.

Voice Enablement for AI

What if you would rather talk than type?

One common use case of LLMs involves voice-to-voice conversation.*



Transcription Tools (TTS):

- OpenAI Whisper
- Cloud services (Google, AWS)
- Picovoice
- Many more

LLM (API or Open Source):

- OpenAI: Chat GPT
- Meta: Llama-2 Chat (OS)
- Anthropic: Claude 2
- Many more

Vocalization Tools (STT):

- [11 Labs](#)
- [Rime](#)
- [Play HT](#)
- Many more (incl. Picovoice)

Pricing Models

1. **Free.** Open Source models are free to use, but require complex infrastructure
2. **Pay per token.** OpenAI's model for GPT, \$0.03 per 1,000 tokens in Oct 2023
3. **Enterprise agreements.** Might also include priority processing or other benefits
4. **Subscription fee.** Midjourney, 11 Labs, Picovoice, and other small-ish start-ups. Usually includes a usage cap or pay-per-X fee past a certain limit
5. **Pay for storage.** Vector databases like Pinecone.io charge by the database shard with size limits

At scale, it could make sense to use open source models and home-grown infrastructure.

BUT, "scale" here means billions of tokens or documents – enough to justify a GPU cluster of your own.

For almost every project, it is simplest to start with API-based tools and subscriptions or usage charges.

Peak Into the Future

Democratizing computation while riding the AI Bull

In which we consider a world where computers can talk, tombstones have avatars, and everybody has a personal secretary and financial advisor

What Would You Ask an LLM To Do

Answer questions based on a corpus

Summarize a passage

Comment on proposed communications

Clean-up machine-generated text

Generate software code

Generate database queries

Compare a piece of code to a specification

Rephrase text as though a particular person were saying it (style editing)

Tweak text to appeal to a certain audience

Handle customer service or inbound sales

Act as a receptionist to answer phones and schedule meetings

Provide an interactive “out-of-office” message that can answer limited questions

Leave an interactive avatar for your loved ones when you pass

Factors to Consider for Projections

Commercialization progress: In 3-5 years, we are unlikely to commercial offerings that are not already under practical development in the lab. So it is enough to ask *what is being successfully developed in research labs today?*

Acceleration of existing trends: Adoption of today's capabilities in AI will lead to a robust ecosystem in the near future that makes these capabilities more accessible. The existence of this ecosystem will lead to a feedback loop, making AI tools easier to build and use.

Open-source vs. Multinational Corps: LLMs are expensive to build and run, and large corporations (with the exception of Meta – thanks, LeCun!) have become more secretive. Open-source tools are proliferating, but they are often less capable.

Hardware & Geopolitics: Due to the demand spike, it is already hard to get access to good hardware for AI. Recent jostling around rare earth imports and chip exports is a bit of a wildcard. How hard will it be to get compute? Who will have access? Will the US have the feedstock to manufacture AI chips domestically or with “safe” allies? Will we see advances in running AI on cheaper hardware? How will energy costs factor in?

Regulation: Will we see substantial regulation or limits on AI development or deployment? Likely too late.

A Few High-Level Predictions

Natural Language Interfaces: In areas such as information retrieval, back-office and front-office tasks (scheduling, answering phones, etc), human language will become the *de facto* standard for interacting with computers. This trend will be in place within 5 years, and will be the standard or norm within 10 years. Note that these systems will also use *voice-to-voice* language when appropriate.

Multimodal AI Systems: Today, GPT-4 already handles images and text. Microsoft's [CoDi](#) works with text, images, audio, video as input or output. Image generation tech such as [Midjourney](#) converts text to images. This tech will only improve, and within 5 years, at least text, audio, and images will be standard ways of interacting with AI, with video interaction with AI avatars out on the horizon.

Demand Bottleneck for AI Hardware: Short of total economic collapse, hardware for AI (GPUs or otherwise) will continue to be a bottleneck on AI deployment and development. Large companies need to think ahead for how they will acquire the capacity they need.

Robots, Robots, Everywhere!: If you thought it was hard to get a human on the phone in 2023, by 2030, it will be near impossible. AI agents will make and answer phone calls on a wide range of topics; it will work better than existing systems most of the time, but as for the rest... This trend will start with customer service in the next few years, but will expand to sales, government programs, recruiting & HR, and more over the next decade.

Negative Implication of AI: People will figure out how to detect deep fakes and lies, but they'll never be convinced that other people can tell the difference too. AI will drive further social division and inequality, and companies need to think about how their AI programs and deployments will affect individuals and society. AI can generate any image or audio, and it will be used to do so, for good or ill. The real trouble is likely at least 10 years out, when the effects will begin to have worked their way through society.

How does it impact your business if:

People can talk to computers

Computers can translate data into English

Computers can talk to each other in English

Personalized AI

A huge trend worth its own slide.

Right now we see mega-corporate AI. But by 2030+, people will want AIs that adapt to their needs and act as their agent in an increasingly complex world

Much the personal computer empowered small businesses and individuals, personal AI will bring the power (and danger) of AI to individuals

Personal AI agents will have customizable personalities, and will take the initiative to schedule appointments, get information from corporate AI systems, help distinguish manipulation from fact, and advise on personal matters discreetly, among many other services

This may sound like a niche or kitschy trend, but it is not to be underestimated. These will not be toys, and they cannot be ignored

The Punch Line

Ultimately, AI now is like computing in the 1970s...

Everyone used mainframes

Everyone kept paper files

Computers were the domain of large companies

Most corporate leaders were convinced their fundamental businesses would not change

IBM was poised to continue leading the digital revolution

Microsoft ... Apple... Google ... Facebook ... none of these existed

The Future is Up to You

What will you do with AI?

That's all, Folks!

Stay Tuned for Q&A

