Exploring ChatGPT for Improved Observability

Gareth Emslie
Cloud Enthusiast

Disclaimer

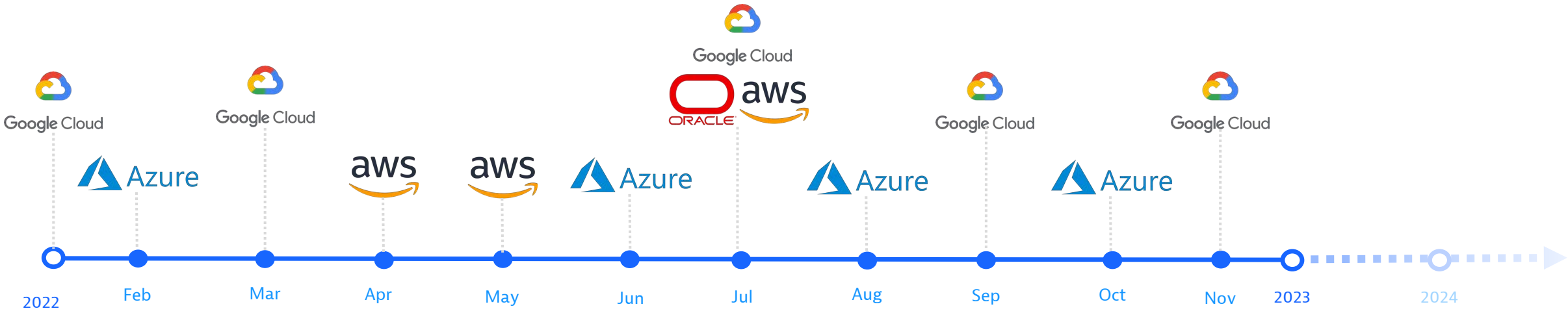
- The information presented in this session is based on my personal experience and opinions and not those of my employer
- Mention of any specific product or service does not imply endorsement

Why do we care about observability?

"Everything Fails All the Time"

Werner Vogels (CTO of amazon.com)

Modern platforms are ephemeral



Source: Outage RCA/PITA/PE's from GCP, AWS, AZURE

Cost of unexpected downtime has risen significantly

~$260k	~$450k	~$3mil	~$5mil
Manufacturing Industry	IT Industry	Auto Industry	Enterprise Industry

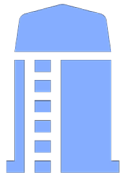
Hourly Cost of downtime

Source: https://www.pingdom.com/outages/average-cost-of-downtime-per-industry/
https://reliably.com/blog/true-cost-unplanned-application-downtime

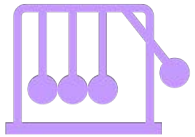
Todays Observability challenges

Complexity

Volume



Silos



Correlation

Risk of IT outages are set to grow even further in 2023

715 tech companies have laid off 199,889 employees in 2023 alone

Source: https://layoffs.fyi/

Modern Observability solutions

Cost effective & Unified Observability Platforms

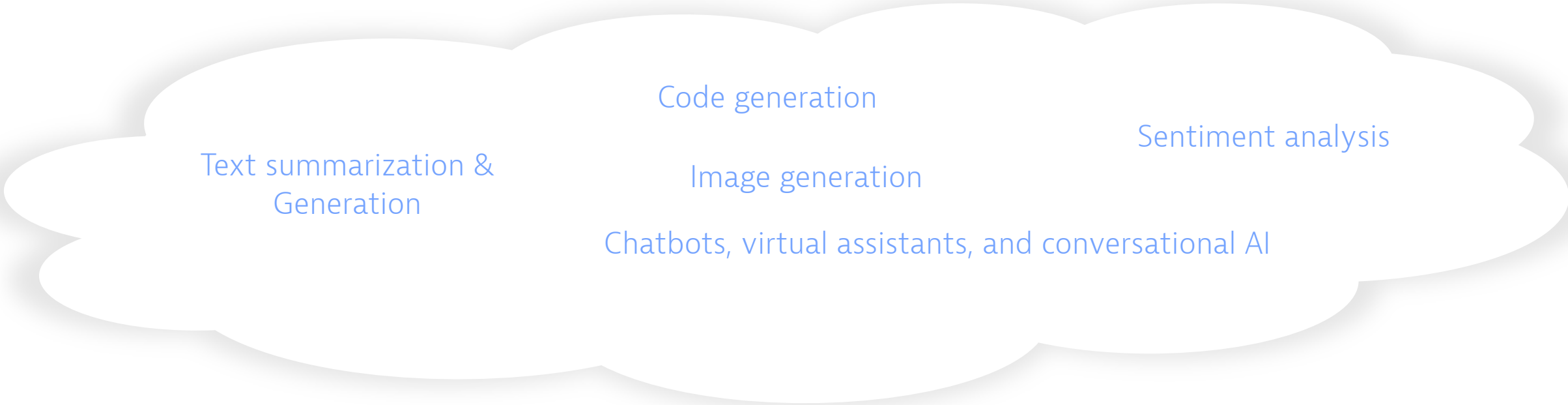
Predictive Observability

“Smart” Alerting

Correlation to Causation Analysis

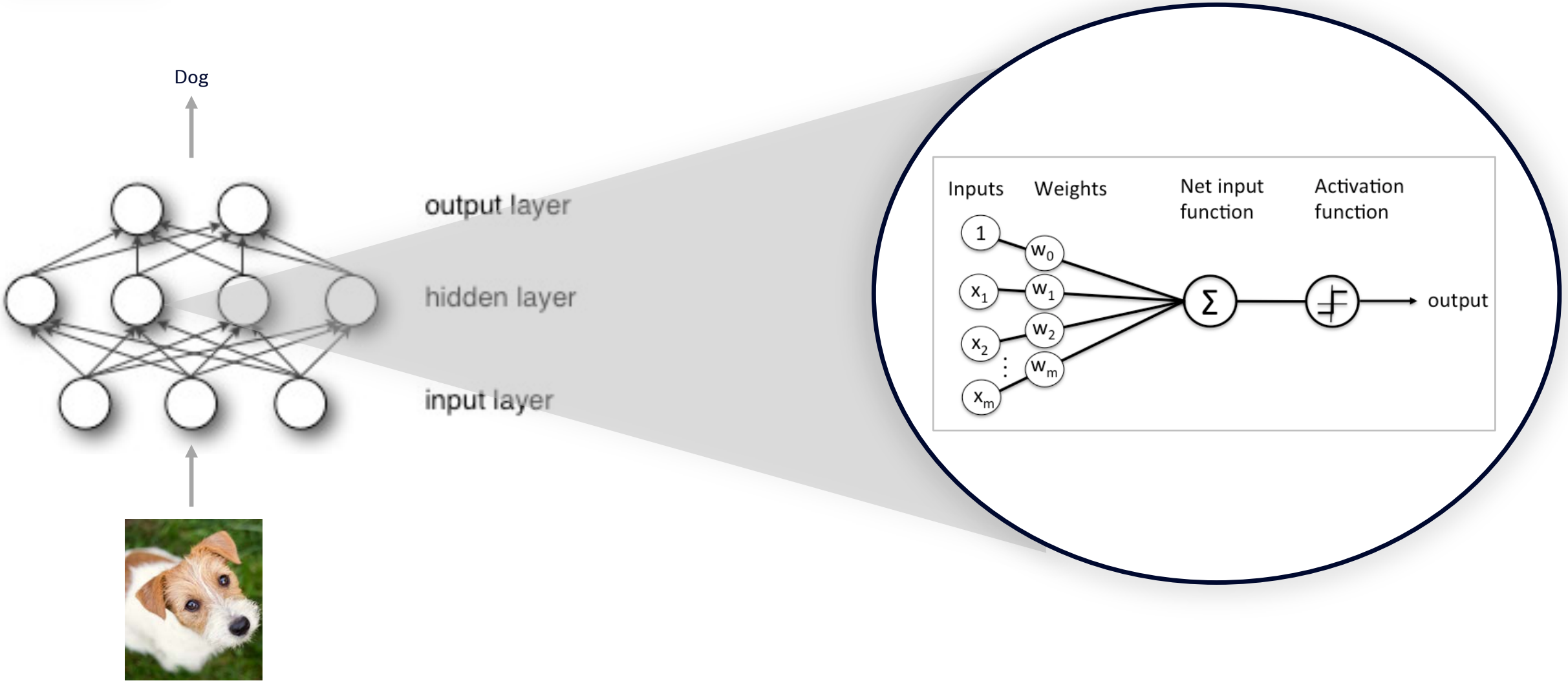
There is currently a lot of innovation occurring in AI, in particular LLM's

"Large language models (LLMs) are foundation models that utilize deep learning in natural language processing (NLP) and natural language generation (NLG) tasks"



Source: https://vectara.com/top-large-language-models-llms-gpt-4-llama-gato-bloom-and-when-to-choose-one-over-the-other/
https://research.aimultiple.com/large-language-models/

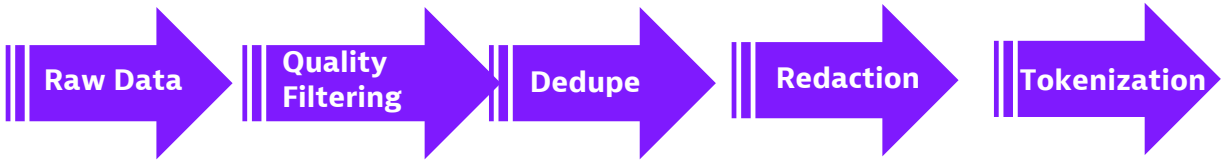
Imitate Brain-like functionality using Deep Neural Networks



Source: http://wiki.pathmind.com/neural-network#define

Large Deep Neural Network (DNN) Models are pre-trained from "the whole internet"

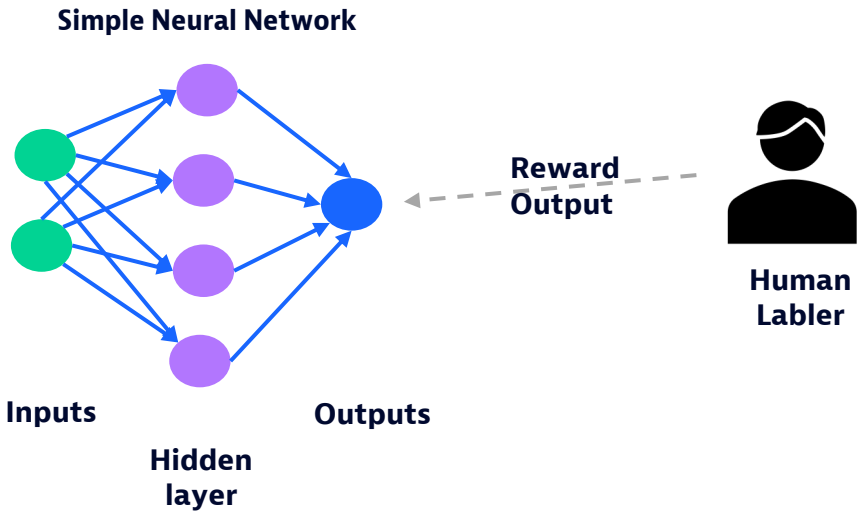
1. Prepare data



2. Adjust billions of parameters

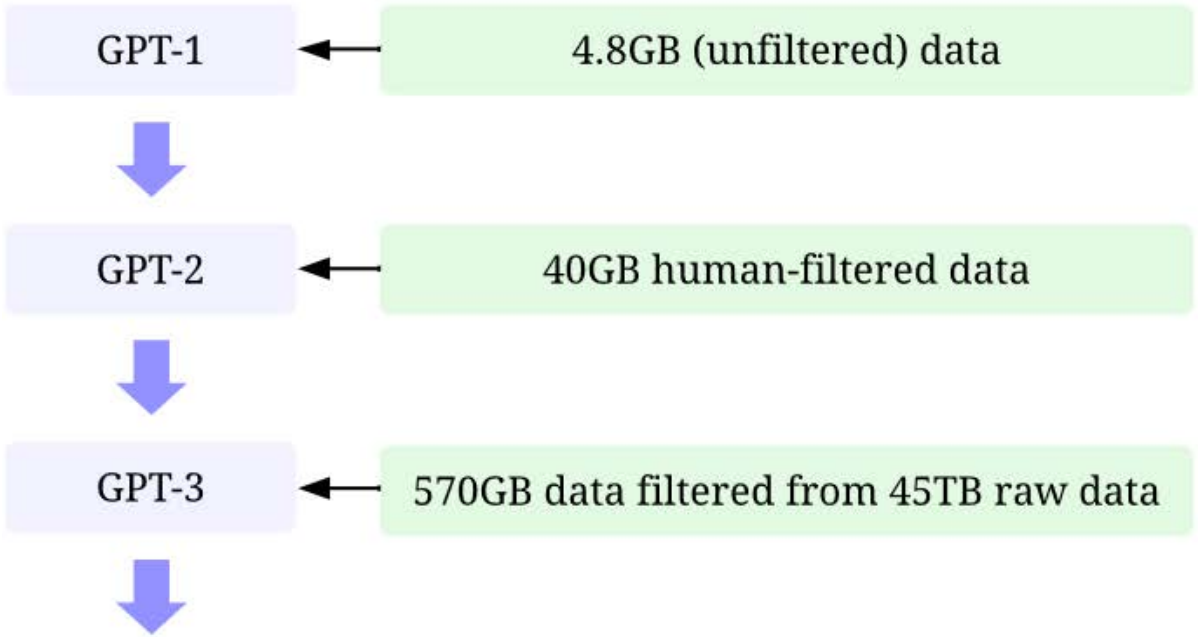
	GPT-1	GPT-2	GPT-3
Parameters	117 million	1.5 billion	175 billion
Decoder layers	12	48	96
Context length	512	1024	2048

3. Reinforcement learning with human feedback (RLHF)



~$20M in energy costs!

Source: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/



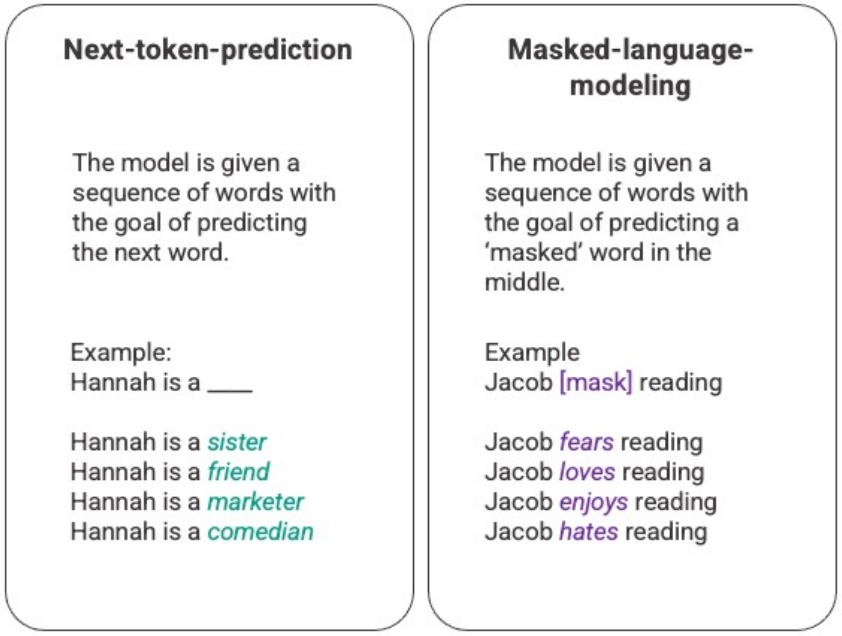
ChatGPT

- Version of **GPT-3.5** fine-tuned on dialogue, with over **175 billion** parameters
- Interacts in a **conversational way**
- **Able to answer** follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests

Source: https://openai.com/blog/chatgpt

ChatGPT is built on top of latest breakthroughs in language model design

Typically language models used the following approaches for predicting the next word in a sequence



Limitations

- Unable to fully understand the context
- Inputs are processed sequentially on an individual basis

Generative Pre-training Transformer (GPT) models were first launched in 2018 by openAI

- Autoregressive Large Language Model
- Leverages transformer architecture
- Self-attention mechanism

Source: https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286
https://www.greataiprompts.com/guide/how-does-chatgpt-works

ChatGPT

≠

Luckily its not perfect

- Nonsensical responses
- Sensitive minor changes in prompts
- Excessively verbose and overuses phrases
- Challenged by Ambiguity
- Susceptible to prompt hacking/injection

Source: https://openai.com/blog/chatgpt

Use cases in observability

ChatGPT

Conversational UX
Using natural language is a very comfortable way for users to query data

Code generation
Support Developer and Operations engineers when writing scripts & code

Intelligent problem remediation
Suggesting ways to resolve problems in custom code

Enriched observability context
Enrich problem tickets and/or alerts driving more effective remediation

Source: https://www.dynatrace.com/news/blog/productivity-innovation-with-chatgpt-generative-ai/

Keep in mind that chatGPT's responses are non-deterministic

#1

Dynatrace reports 'The Garbage collection suspension time value was above normal behavior' what could be the root cause?

The root cause of a high garbage collection suspension time value in Dynatrace c
depending on the specific circumstances of your application and environment. H
some common causes of high garbage collection suspension time include:

1. Large Heap Size: If your application has a large heap size configured, garbage
   can take longer as it needs to scan a larger memory space.
2. Memory Pressure: If your application is running with limited memory resource
   cause frequent garbage collections and longer suspension times.
3. Inefficient Memory Management: Inefficient memory management practices,
   excessive object creation, not releasing unused objects, or holding onto refere
   longer than necessary, can contribute to increased garbage collection time.
4. Concurrent Garbage Collection Settings: If the concurrent garbage collection
   not optimized for your application's workload, it can result in longer suspensio
5. Fragmented Heap: If the memory heap becomes fragmented over time, garba
   collection may take longer as it needs to search for contiguous memory block
6. Insufficient Garbage Collection Tuning: The garbage collection settings of you
   not be optimized for your application's specific memory usage patterns, result
   longer suspension times.

#2

Dynatrace reports 'The Garbage collection suspension time value was above normal behavior' what could be the root cause?

The root cause of a high garbage collection suspension time value reported by Dynatrace can vary depending on the specific circumstances of your application and environment. However, some common causes of elevated garbage collection suspension time include:

1. Large Heap Size: If your application has a large heap size configured, garbage collection may take longer as it needs to scan a larger memory space.
2. Memory Pressure: Insufficient memory resources can lead to frequent garbage collections and longer suspension times.
3. Inefficient Memory Management: Inefficient memory management practices, such as excessive object creation, not releasing unused objects, or holding onto references longer than necessary, can contribute to increased garbage collection time.
4. Suboptimal Garbage Collection Settings: The garbage collection settings of your runtime environment may not be optimized for your application's workload, resulting in longer suspension times.
5. Fragmented Heap: If the memory heap becomes fragmented over time, garbage collection may take longer as it needs to search for contiguous memory blocks.
6. Concurrent Execution: If your application performs intensive tasks during garbage collection, it can prolong suspension times.
7. Insufficient System Resources: If the underlying system hosting your application lacks sufficient CPU or memory resources, it can impact garbage collection performance.

To make an informed decision chatGPT needs to build up a lot of context in the form of questions and answers

#1

Web Application 'Shop' calls to Backend service 'Orders', some users are experiencing additional latency when completing their orders

If users of the web application 'Shop' are experiencing additional latency when completing their orders, there could be several possible causes for t reasons to consider:

1. Network congestion: High network traffic or congest Check if there are any network issues or bottlenecks and the backend service 'Orders'. Monitor network p or spikes in latency.
2. Backend service performance: The additional latenc issues in the 'Orders' backend service. Check the ser bottlenecks or slow queries. Analyze the backend se performance to determine if it is handling the load ef
3. Database performance: If the 'Orders' backend servi database queries or inadequate indexing could cont database performance and query execution times. C indexing strategies, and ensure the database is appr
4. Scaling limitations: The 'Orders' backend service mig limitations, resulting in increased latency during pea scalability of the backend service and ensure it can h horizontal scaling by adding more instances or verti resources if necessary.
5. Caching and optimization: Implement caching mech backend service. Cache frequently accessed data or times. Additionally, optimize the code and database operations and improve overall performance.
6. Monitoring and troubleshooting: Implement compre both the 'Shop' web application and the 'Orders' back metrics, error logs, and request traces to identify any occur when users experience additional latency. Utili and resolving the problem.

#2

I have checked the backend service logs and there do not seem to be any bottlenecks

If you have checked the backend service logs and few other potential factors to consider when troubl the 'Shop' web application. Here are some addition

1. Frontend performance: Evaluate the performan Look for any client-side issues that could contri slow rendering, excessive network requests, or tools or performance monitoring tools to analyz any areas for improvement.
2. Network connectivity: Verify if the users experie specific geographic region or network environm regional network congestion could affect their e monitoring tools to check for connectivity probl specific regions or networks.
3. Third-party integrations: If the 'Shop' web applic APIs, there might be latency issues with those i their response times. Contact the third-party se known issues or performance limitations.
4. Load testing: Perform load testing on the 'Shop' service to simulate the conditions that lead to a performance under heavy loads and analyze an spikes. Load testing can help identify scalability apparent during normal usage.
5. Distributed architecture: If the 'Shop' and 'Order

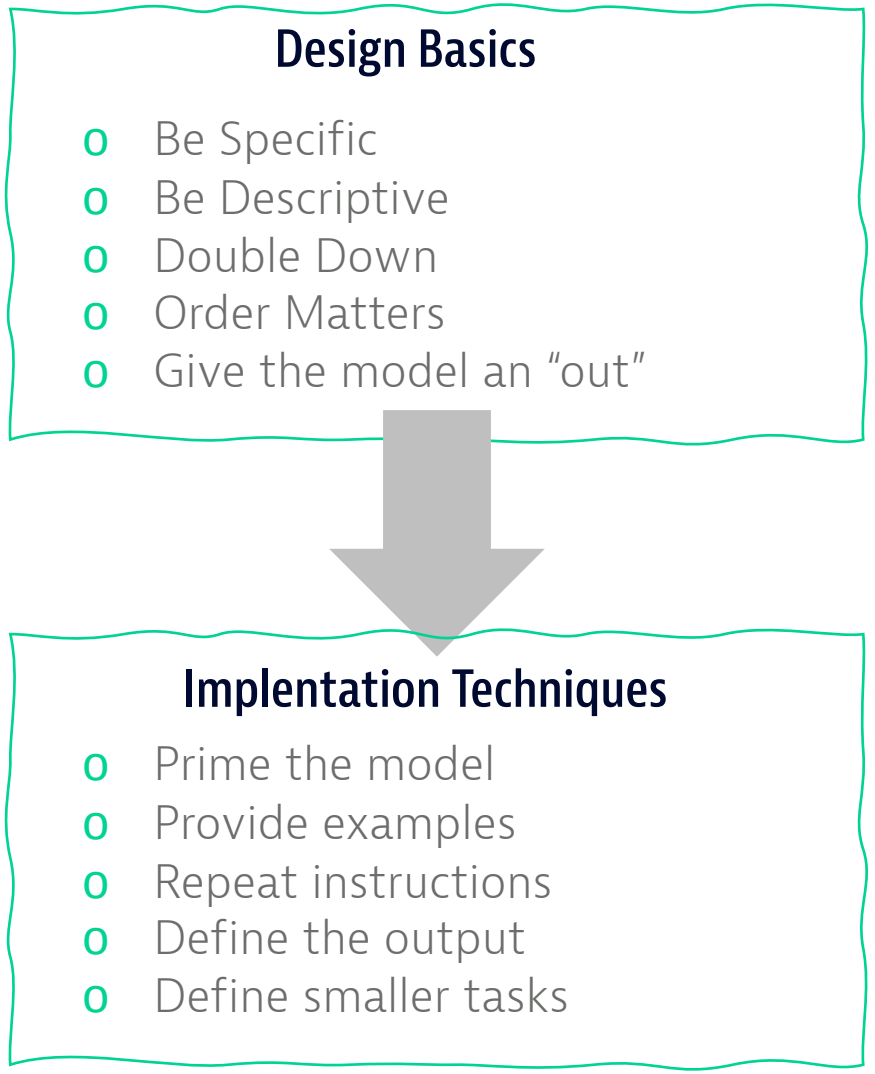
#3

Our network monitoring tools report dropped packets around the same time

If your network monitoring tools are reporting dropped packets around the same time that users experience additional latency in the 'Shop' web application, network congestion or connectivity issues could be contributing to the problem. Here are some steps to address this issue:

1. Analyze network traffic: Use network monitoring tools to analyze the network traffic during the time period when users experience latency. Look for any spikes in traffic or patterns that could indicate congestion. Identify the specific network devices or segments where packet drops are occurring.
2. Network infrastructure review: Review the network infrastructure, including routers, switches, firewalls, and load balancers. Ensure that they are properly configured and have sufficient capacity to handle the expected traffic load. Check for any misconfigurations or bottlenecks that could lead to dropped packets.
3. Bandwidth optimization: Optimize the bandwidth usage within your network. Identify any bandwidth-hungry applications or services that could be causing congestion. Implement quality of service (QoS) mechanisms to prioritize critical traffic, such as the communication between the 'Shop' web application and the 'Orders' backend service.
4. Network latency reduction: Minimize network latency by optimizing network settings and protocols. Consider implementing techniques such as traffic shaping, packet

Prompt engineering is a new discipline



Source :https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/advanced-prompt-engineering
https://owasp.org/www-project-top-10-for-large-language-model-applications/descriptions/

Hyperscalers are investing heavily into GPT's

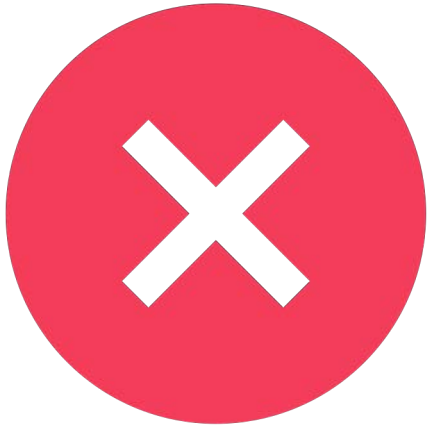
- Investment of $10 billion into OpenAI
- Investing heavily in their Azure AI platform
  - Adopted ChatGPT open plugin standard
  - Announced Prompt flow
  - Support for foundation models

- Announced Amazon Bedrock and Amazon Titan models
- GA ec2 instances powered by AWS Trainium and AWS Inferentia2
- GA of Amazon CodeWhisperer

- Announced more than 25 products and features powered PaLM 2 and Gemini
- Announced next-generation A3 GPU supercomputer

Source: https://news.microsoft.com/build-2023-book-of-news/

So are large language models a panacea?



- Prompt engineering is a new discipline requiring reskilling and the correct tooling
- Protecting intellectual property and data privacy requires careful thought
- Its important to understand the risks of GPTs and generative AI

CLOUD DONE RIGHT