

CISCO *Live!*

GO BEYOND

#CiscoLiveAPJC



See It All, Solve it Fast

ThousandEyes And Multimodal AI at last

Heath Russell – ThousandEyes Channel Solutions Engineer APJC

@heath.russell

LTRAPP-2089

CISCO *Live!*

#CiscoLiveAPJC

Cisco Webex App

Questions?

Use Cisco Webex App to chat with the speaker after the session

How

- 1 Find this session in the Cisco Live Mobile App
- 2 Click “Join the Discussion”
- 3 Install the Webex App or go directly to the Webex space
- 4 Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until November 15, 2024.



<https://ciscolive.ciscoevents.com/ciscolivebot/#LTRAPP-2089>





Agenda

- About me
- Use cases
- How Does Ai work
- Ai Optimisation techniques
- Cautionary take aways
- Lab Content

About Me



Heath Russell

APJC, Channel Solutions
Engineer

CISCO
ThousandEyes 



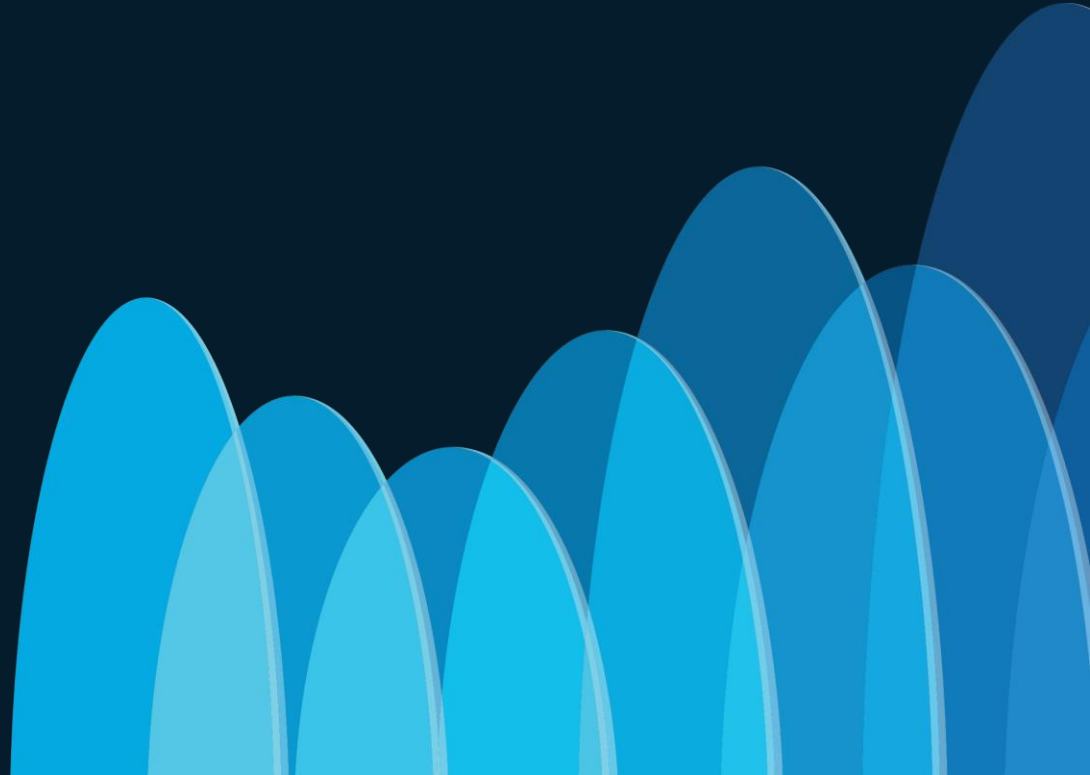
CISCO *Live!*

Story Time

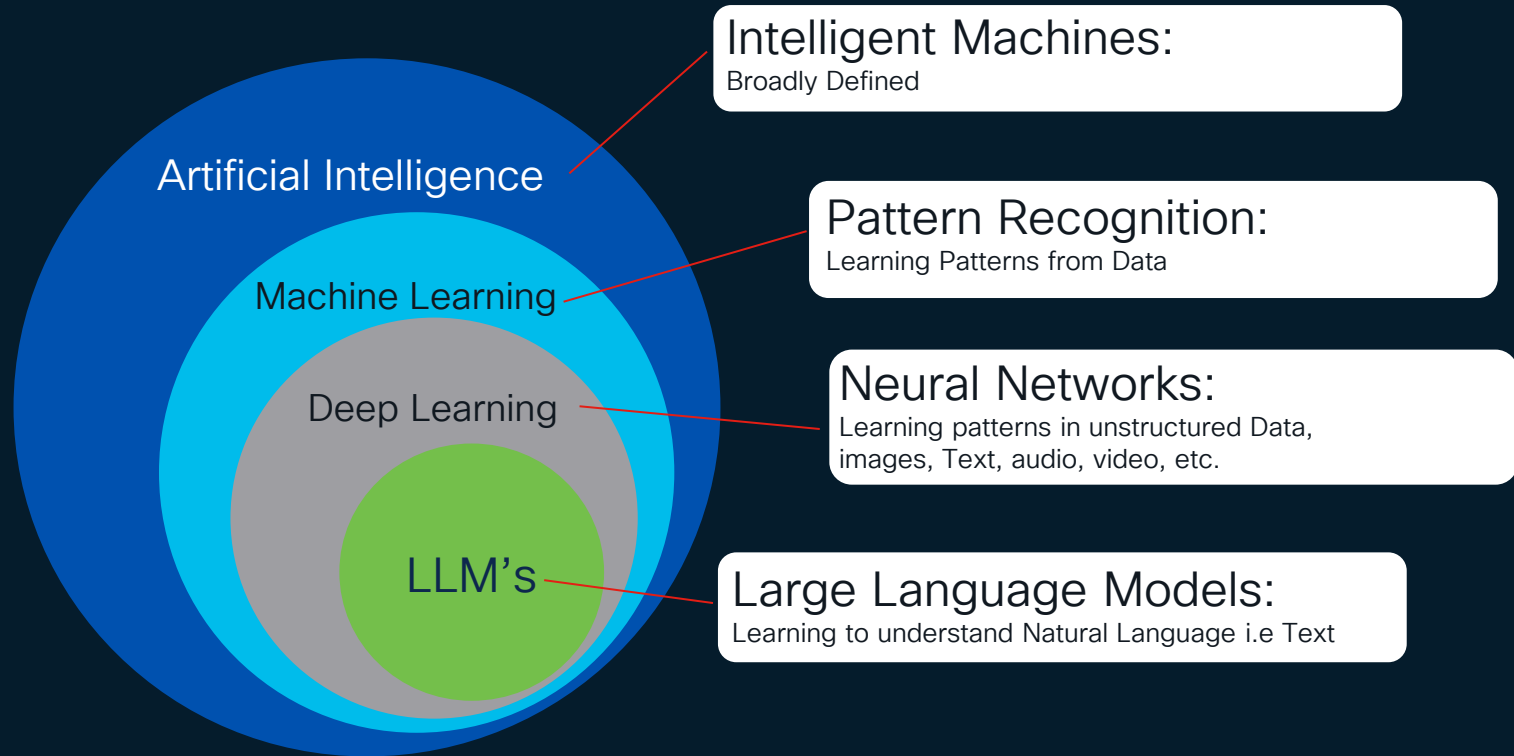


Gen AI

Take away the hype and lets
understand what it really means



Generative AI is a very very broad subject



AI Nomenclature

Tokens: AI tokens are units of text that an AI model, like GPT, processes and generates, the more input tokens the slower your model processes.

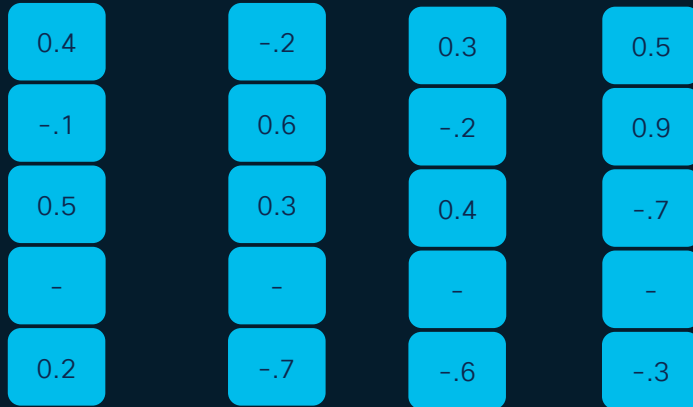
Chat GPT: Chat Generative Pre-trained Transformer

Transformer : The Transformer is a specific neural network architecture, that processing of sequential data, like text by capturing broad dependencies and relationships

How does AI LLM work ?

[Network engineers are awesome]

Transformer
Model



Class	Probability
Positive	0.98
Negative	0.02

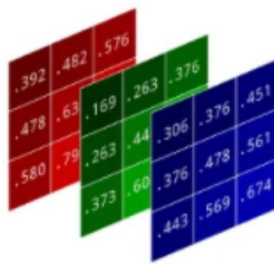
Word embeddings represent a words
Meaning in context

What if my input is a picture ?

Human: Is this a cat or a fox ?



RGB



Pixel embeddings

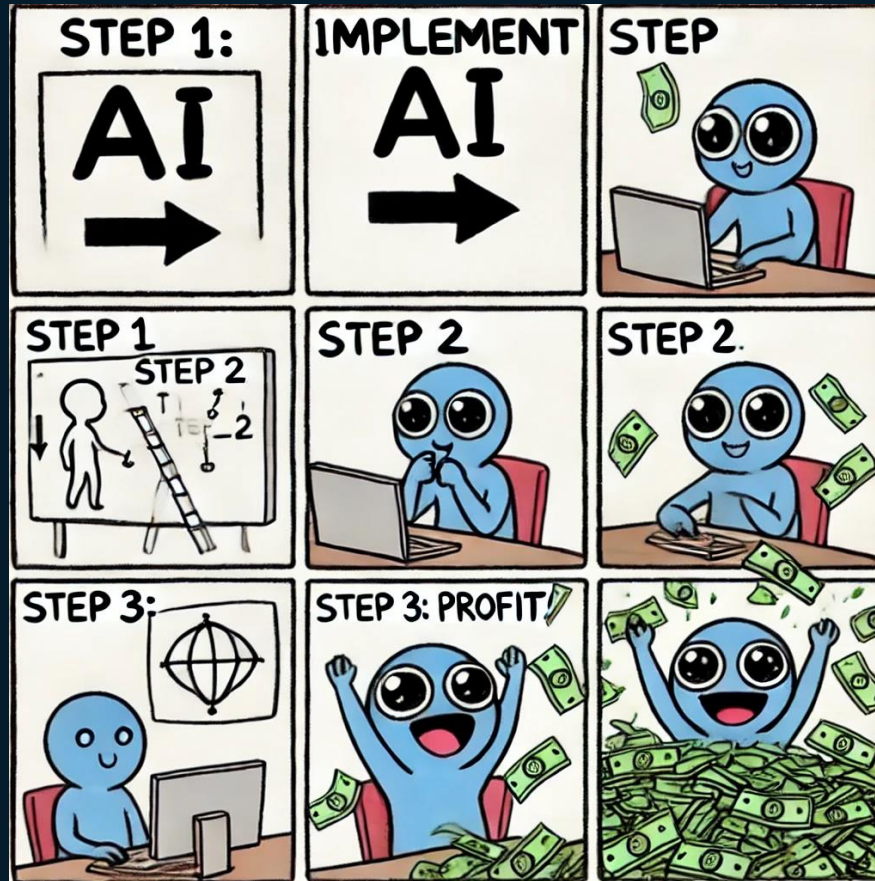
Transformer
Model

Class	Probability
-------	-------------

Cat	0.98
-----	------

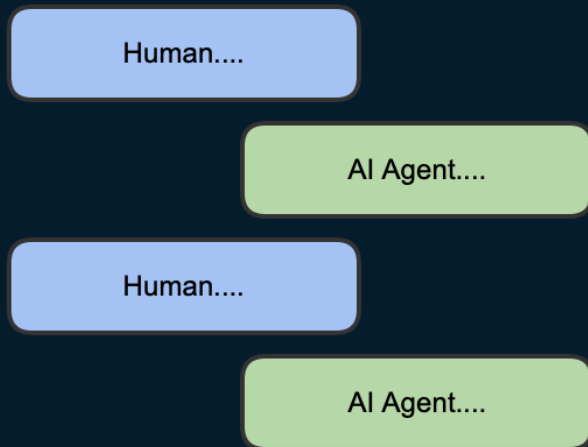
Fox	0.02
-----	------

How AI works according to my boss

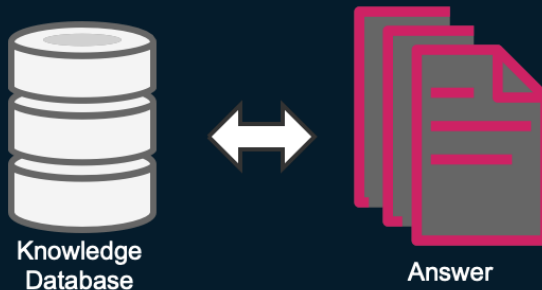


Enterprise Use Cases

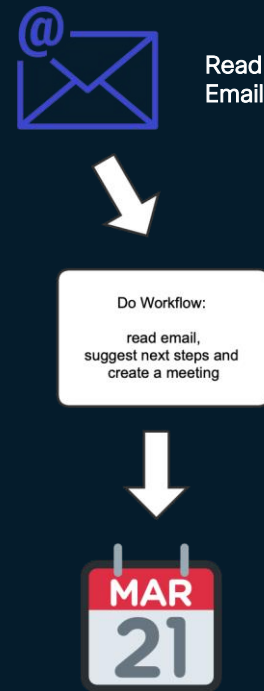
Chat bot



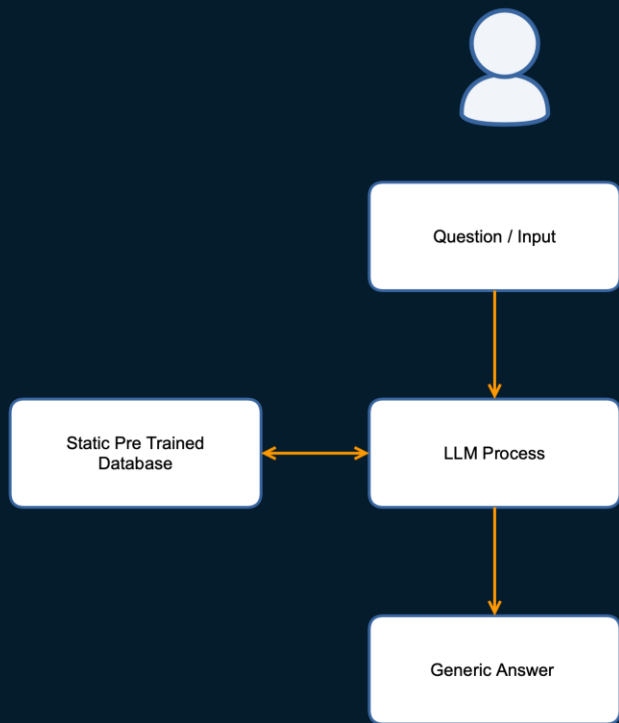
Document Search



Workflow Automation



When you try to make AI do something harder than basic query



The screenshot shows the Gemini AI interface. At the top, there's a 'Try Gemini Advanced' button. Below it, a user prompt asks for a detailed network findings report based on a screenshot of a ThousandEyes dashboard. The AI's response is displayed below the prompt. It starts with a disclaimer: 'I can't write a detailed network findings report based on the image you sent, as it only shows a screenshot of a dashboard with various graphs. I can't tell what the graphs represent without more information.' This is followed by a speculative analysis: 'However, based on the fact that the dashboard is from ThousandEyes, a network intelligence platform, I can speculate on what the graphs might show. ThousandEyes provides performance and path visualization for users of cloud-based applications. So, the graphs in the dashboard might show things like:'. A list of potential metrics follows: 'Response times between different parts of the Webex contact center network', 'Packet loss rates', 'Jitter (variation in packet arrival times)', and 'Hop counts (the number of routers a packet traverses between source and destination)'. The response then offers to provide a more detailed report if more information is provided and lists some things to keep in mind when interpreting network performance dashboards: 'Normal vs abnormal', 'Baseline performance', 'Identify trends', and 'Correlate metrics'.

Gemini

Try Gemini Advanced

H write a detailed network findings report based on the following images taken from thousandeyes for a Cisco Webex contact centre

Show drafts

I can't write a detailed network findings report based on the image you sent, as it only shows a screenshot of a dashboard with various graphs. I can't tell what the graphs represent without more information.

However, based on the fact that the dashboard is from ThousandEyes, a network intelligence platform, I can speculate on what the graphs might show. ThousandEyes provides performance and path visualization for users of cloud-based applications. So, the graphs in the dashboard might show things like:

- **Response times** between different parts of the Webex contact center network
- **Packet loss rates**
- **Jitter** (variation in packet arrival times)
- **Hop counts** (the number of routers a packet traverses between source and destination)

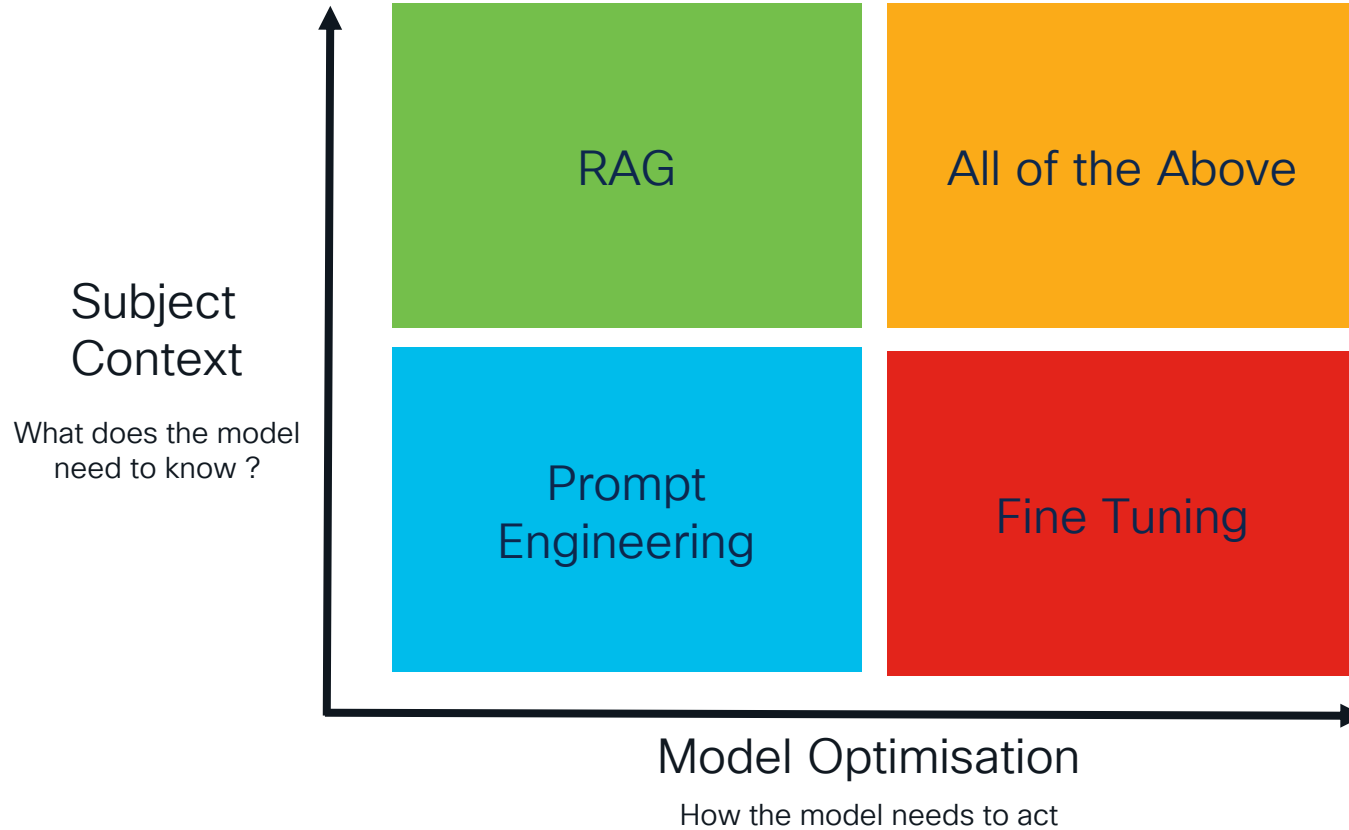
If you can provide more information about the graphs, I may be able to give you a more detailed network findings report. For example, if you can tell me what the axes of the graphs represent, I may be able to tell you what the graphs are showing.

Here are some other things to keep in mind when interpreting network performance dashboards:

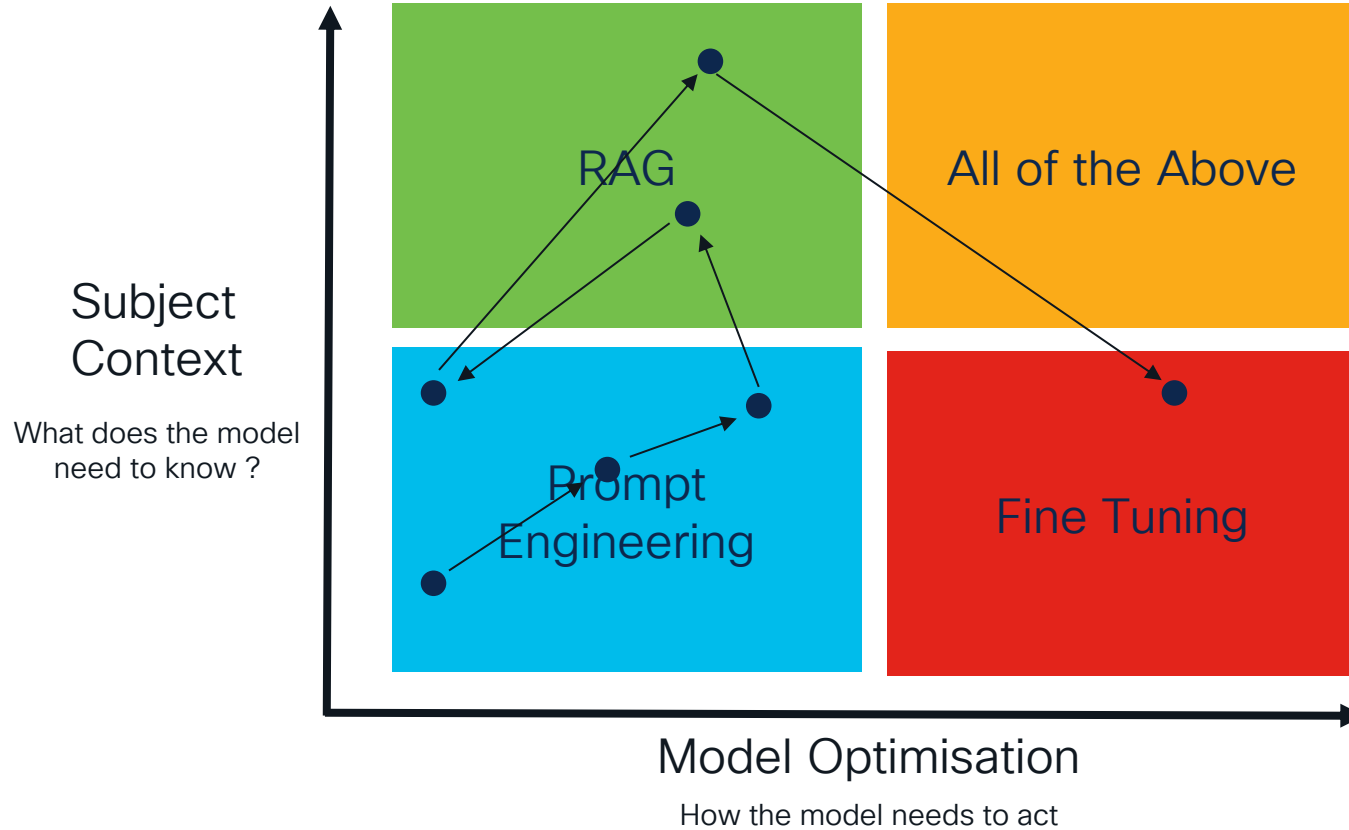
- **Normal vs abnormal:** It's important to compare the current performance metrics to historical data to see if there are any significant deviations.
- **Baseline performance:** Establish a baseline for normal performance so you can easily identify anomalies.
- **Identify trends:** Look for trends over time to identify potential problems.
- **Correlate metrics:** Correlate network performance metrics with other metrics,

Enter a prompt here

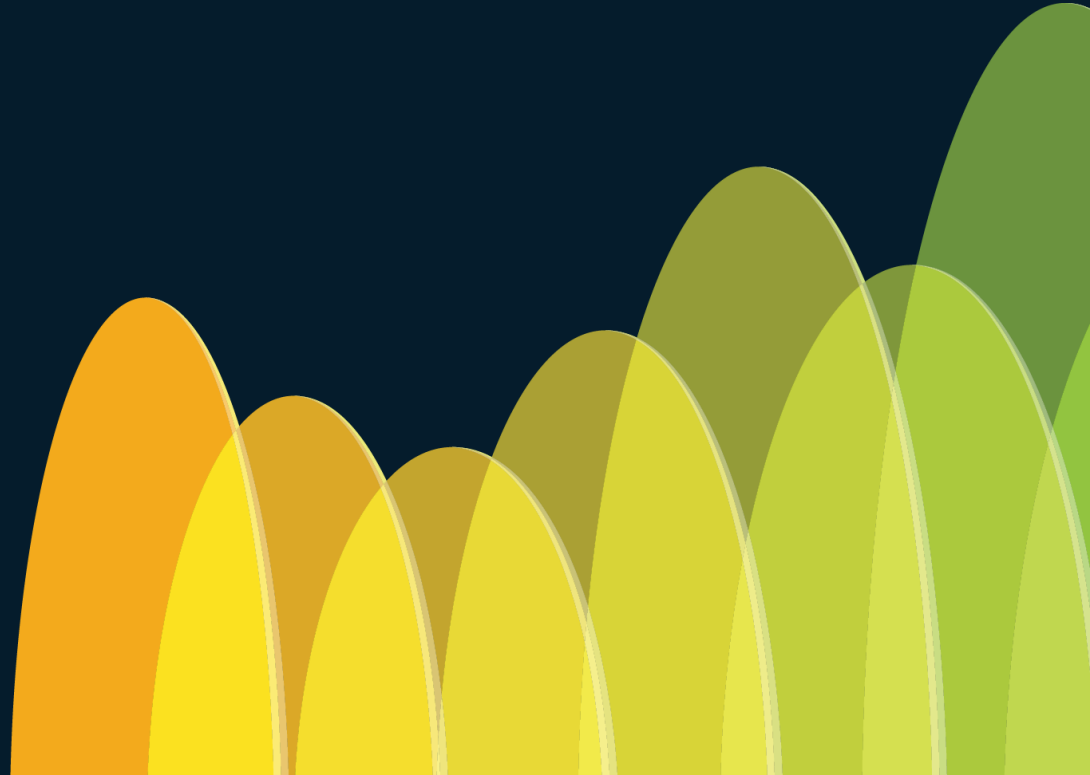
AI Optimisation Techniques



AI Optimisation Journey



Prompt Engineering



Prompt Engineering and where to use it

Good for

- Low effort
- Failing fast and iteration
- Testing your model
- Baseline experimentation
- Best place to start and finish

Not good for

- Introducing New information
- Replicating output style
- Minimising token usage
- Sprawl

Prompt engineering – Your best bang for buck

- AI is lazy and misses many things
- Be clear and specific about what you want
- Provide as much context as you can about what you want
- If what you're asking for is complex break it down into smaller numbered tasks
- You can also ask the model to return why its saying what its saying, ask for references!

Bad example of a prompt

Vague instructions



Gemini

Try Gemini Advanced

H

write a detailed network findings report based on the following images taken from thousandeyes for a Cisco Webex contact centre



Show drafts

I can't write a detailed network findings report based on the image you sent, as it only shows a screenshot of a dashboard with various graphs. I can't tell what the graphs represent without more information.

However, based on the fact that the dashboard is from ThousandEyes, a network intelligence platform, I can speculate on what the graphs might show. ThousandEyes provides performance and path visualization for users of cloud-based applications. So, the graphs in the dashboard might show things like:

- **Response times** between different parts of the Webex contact center network
- **Packet loss rates**
- **Jitter** (variation in packet arrival times)
- **Hop counts** (the number of routers a packet traverses between source and destination)

If you can provide more information about the graphs, I may be able to give you a more detailed network findings report. For example, if you can tell me what the axes of the graphs represent, I may be able to tell you what the graphs are showing.

Here are some other things to keep in mind when interpreting network performance dashboards:

- **Normal vs abnormal:** It's important to compare the current performance metrics to historical data to see if there are any significant deviations.
- **Baseline performance:** Establish a baseline for normal performance so you can easily identify anomalies.
- **Identify trends:** Look for trends over time to identify potential problems.
- **Correlate metrics:** Correlate network performance metrics with other metrics,

Enter a prompt here

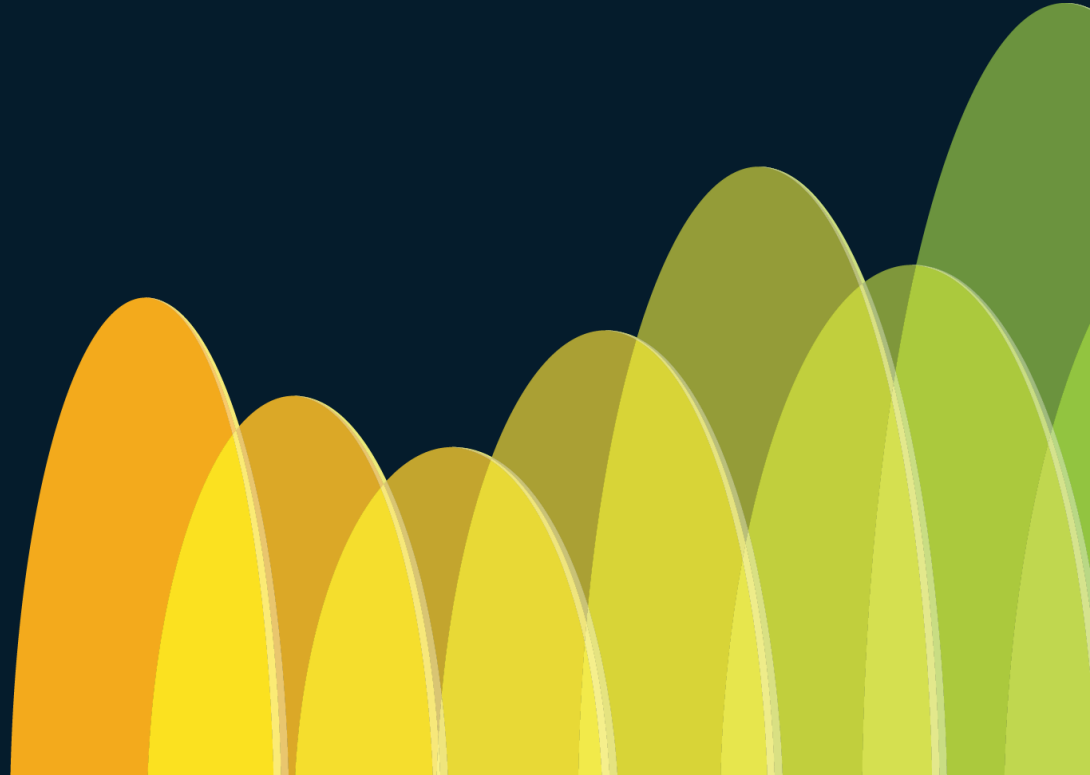
Better Example

You will be presented with an image of a ThousandEyes graph, your task is to identify root cause issues in the graph output

Approach this task in a step-by-step manner take your time and don't skip Or miss any steps in my prompt:

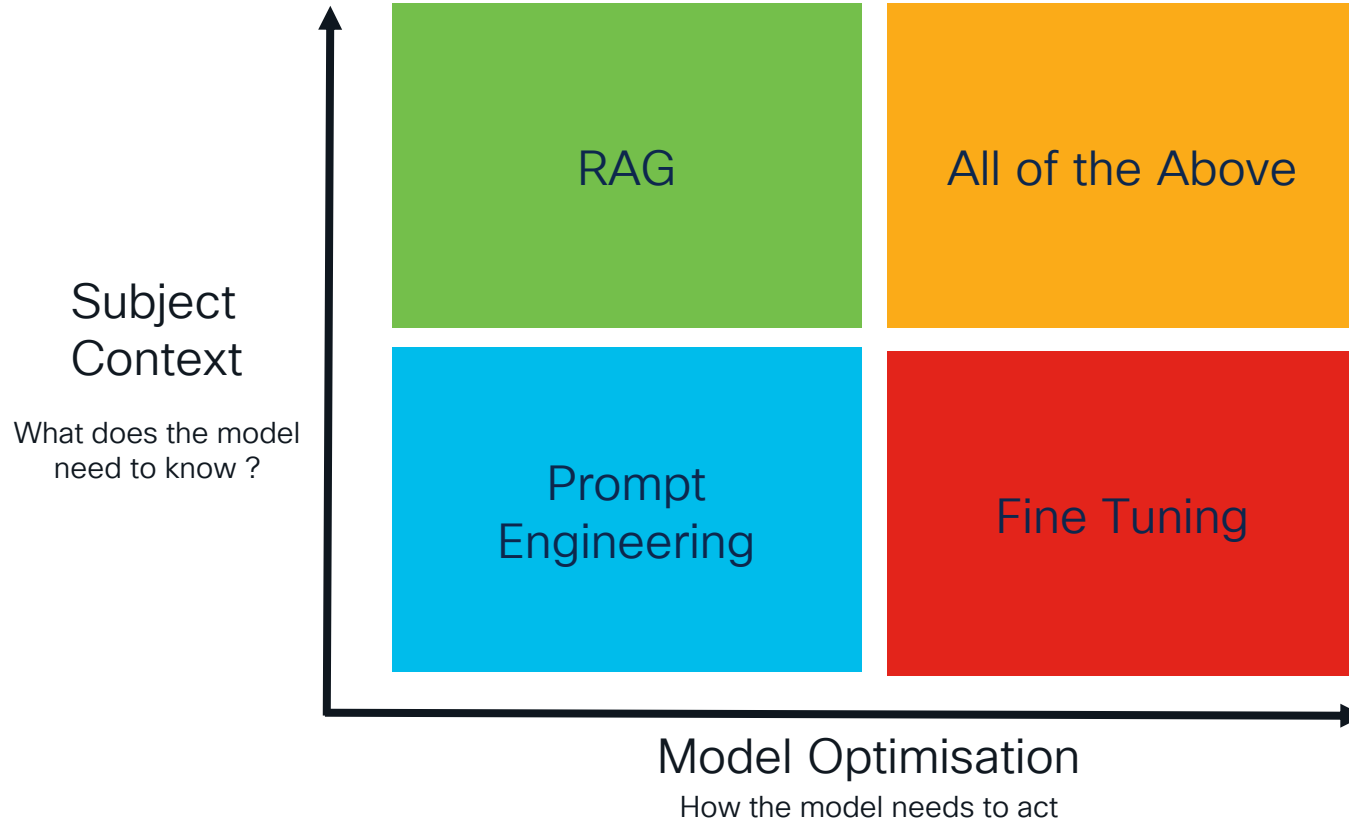
1. Carefully analyse the graphs in the image attached
2. Reference your RAG information to see if any of the graphs show anything out of the bands of what is considered best practice
3. Form an opinion based on your rag facts and state if they are good or bad
4. Include as much evidence as possible to back your findings
5. Include possible fixes for your findings
6. Make a summary of findings in conclusion with next steps

Lets take it to
the next level

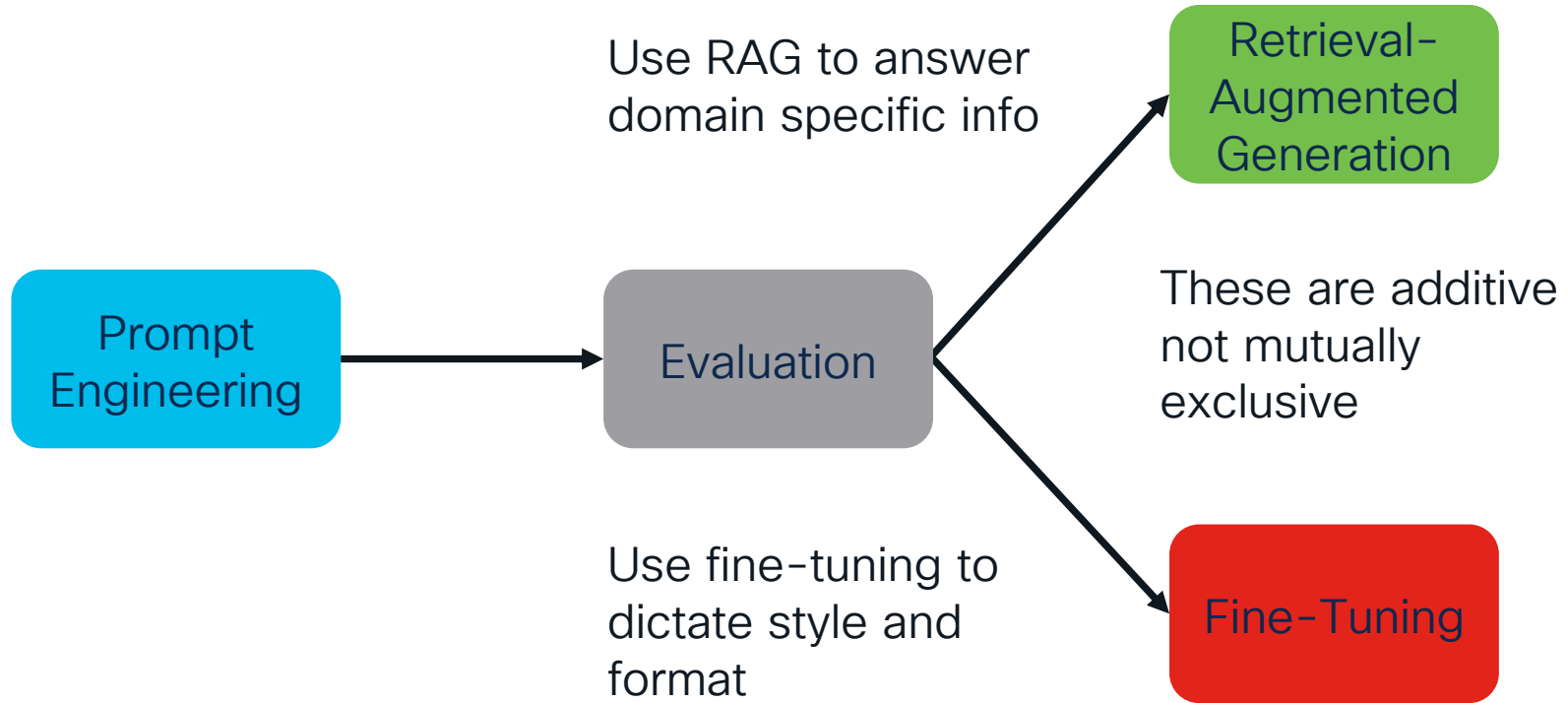




AI Optimisation Journey



RAG vs Finetune and Why



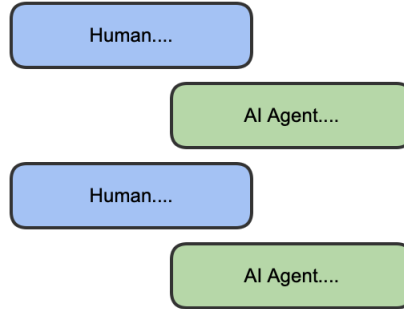
Retrieval-Augmented Generation RAG



So what is RAG anyways?



So what is RAG anyways?



Your Source of truth
Contextual
Knowledge Base

Human Question

Contextual
Answer

RAG and where to use it

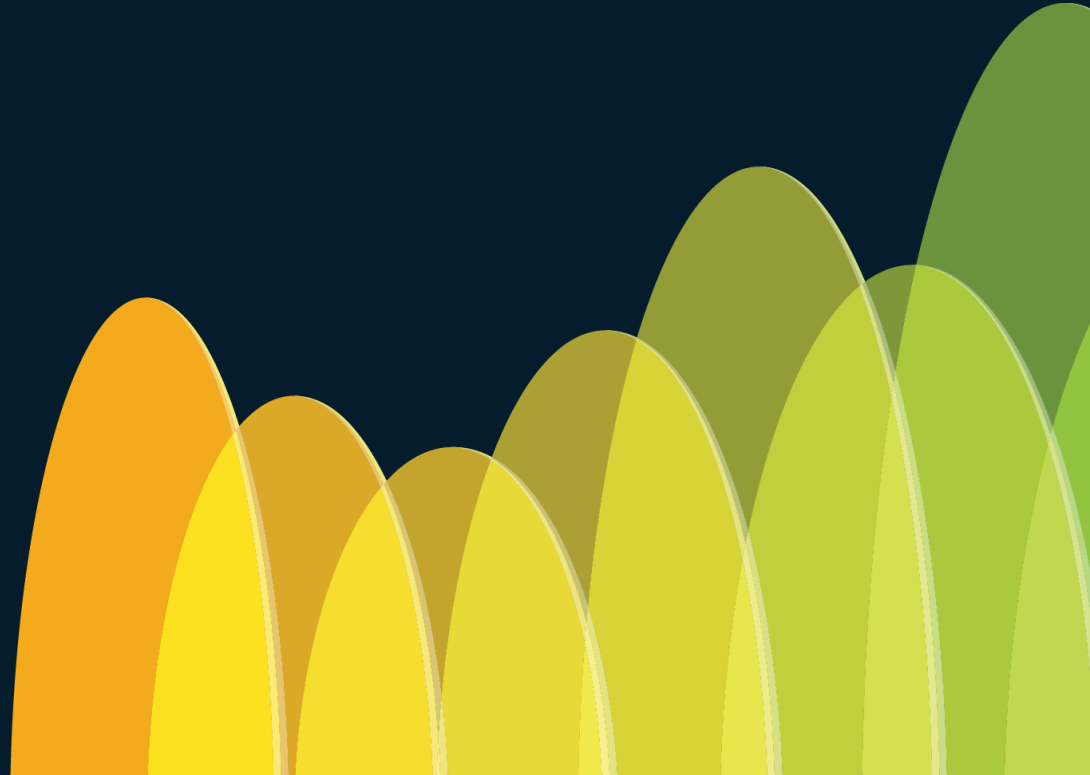
Good for

- Introducing specific knowledge domain information
- Reducing hallucination
- Relatively easy and fast

Not good for

- Cleaning up race cars
- Teaching a whole domain of information
- Teaching a model how to do a broad task
- Minimising token usage
- Speed of output

Fine Tuning



So what is Fine Tuning

Training is taking an existing generic model, like Anthropic, ChatGPT, Llama Then:

- Creating a large database of task specific data
- Adjusting the Model weights so the model can understand Task specific nuances
- Regularisation using techniques to avoid overfitting or becoming biased on specific information
- evaluation

Fine Tuning and where to use it

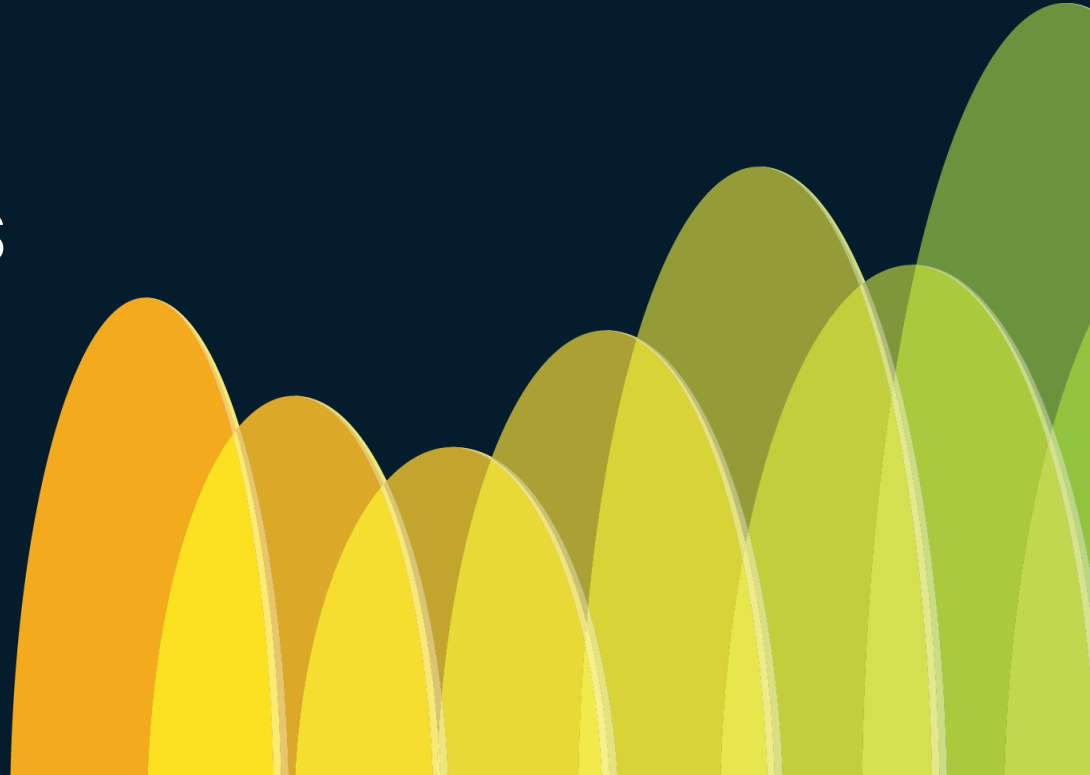
Good for

- Minimising token usage
- Improving model performance
- Teaching a model a large corpus of information like legal, health, coding a language
- Setting tone and structure of LLM response

Not good for

- Speed of implementation or iteration
- High effort, involved process
- New knowledge

Cautionary Tales



Fine Tuning GPT 3.5 based on 140,000 slack Messages

Be careful watch you wish for

User: write a 500 word blog on prompt engineering

Assistant : I shall work on that in the morning

User: its morning now ?

Assistant : ok, ok, I'm writing it now, its 6:30am here now.

User: Write it now !

Assistant : ok

User: please?

Assistant : I'm sorry I can't write a blog, I can only describe the process, but I have an idea, lets call it how to build prompts 😊

AI hallucination is real

Hallucination is when Ai specifically LLM's see or perceive a pattern in data That is not real, the model is tricked into an answer

These are caused by

- Insufficient data
- Incorrect training information
- Pushing the Ai for information It does not have



Imagine a number, but don't share it. Double that number. Add 150 to that resulting number. Then divide that new number in half. Youe answer is 75



That's a clever way to use math to arrive at a predetermined answer! No matter what number you imagine, if you follow those steps, you will always end up with 75.

“With Great power comes Great responsibility”

- Anything you put into public LLM is building the LLM company dataset
- Never put proprietary or Private information into a public LLM
- DLP is important! Caution and Educate your users!

“If something is free you are the product”

Richard Serra, 1973

“LLMs are not likely to destroy the world but they might force us to rethink it a bit.”

Ross Lazerwitz

Product Manager Splunk

Let's stop talking about it and let's start being about it.

Lab Time

Complete Your Session Evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to claim a **Cisco Live T-Shirt**.



Complete your surveys in the **Cisco Live mobile app**.



Continue your education

CISCO *Live!*

- Visit the Cisco Stand for related demos
- Book your one-on-one Meet the Expert meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Contact me at: linkedin





Thank you

CISCO *Live!*

#CiscoLiveAPJC

CISCO *Live!*

GO BEYOND

#CiscoLiveAPJC