

Executive Summary

The Large Language Model (LLM) enterprise industry is rapidly growing with a wide range of potential applications, akin to the growth of the cloud industry in the early 2000s. Companies like Google, OpenAI, Microsoft, Amazon and IBM are developing and offering LLMs for enterprise applications such as content generation, customer service and fraud detection. These services offer a variety of benefits to customers.

- **Generate personalized content**, such as recommendations, chatbots, and virtual assistants.
- **Improve customer service** by providing customers with more accurate and timely information.
- **Detect fraud** by identifying patterns of suspicious behavior.
- **Automate tasks** that are currently performed by humans, such as data entry and customer service.
- **Improve Worker Productivity** by integrating with enterprise workforce tools such as emails and spreadsheets

WWT specializes in helping businesses develop and implement solutions that allow customers to identify the right LLM for their needs, train it to meet their specific requirements and integrate it into their business processes. WWT can also help develop and implement strategies for using LLMs to improve employee productivity, automate tasks and generate new revenue streams. [WWT offers a variety of services to accomplish the goals of a “Buy and Build” centered solution.](#)

- **Assessment:** understand the customer’s needs and requirements to determine the best LLM for their use case
- **Customization:** help the customer train their LLM to meet their specific requirements
- **Integration:** facilitate the integration of LLM into customer business processes
- **Strategy:** develop and implement strategies for using LLMs as a part of operations

WWT can help customers get the most out of their LLM investments by avoiding common pitfalls while also adding value by enabling them to develop and implement strategies that are specific to their business needs



WWT provides comprehensive solutions for successful LLM project development and management

Expert Consulting, Integration, Support and Optimization Services drive success

Business Processes Services



Buy & Build Consulting

Find the right partners and tools to build a successful project. We provide expert guidance on the best buy-build strategy



Partner Support

For enterprises with less cloud adoption. Procure specialized hardware, navigate enterprise cloud subscriptions – for a solution that spans the industry



Adoption

Fine tune your chosen LLM, compartmentalize your data, and implement governance policies. We will help you to train your LLM responsibly, ethically and avoid “Shadow AI” concerns

Technical Services



Tech Assessments

Our experts will assess your business requirements and recommend the best GPT/LLM solution. Our ATC helps you scale your solution to leverage your LLM investment



AI Architecture

Designed to make your training and inference processes more efficient, achieving better results & performance. We will help you deploy LLMs at scale, with security and compliance



Ethical & Responsible AI

Our expert team can help identify and mitigate ethical risks, develop policies and procedures and train employees on responsible AI

LLMs have been 75 years in the making

Rule-Based (1948-2003)

N-Grams count the occurrences of a sequence of “n” words linked together in a text or **corpus**

1-Gram	2-Gram	3-Gram
The	The Margherita	The Margherita pizza
Margherita	Margherita pizza	Margherita pizza is
pizza	pizza is	pizza is not
is	is not	is not bad
not	not bad	not bad taste
bad	bad taste	
taste		

TF-IDF counts the term frequency in one doc but penalizes occurrences in other docs

Word	TF (Sentence 1)	TF (Sentence 2)	IDF	TF*IDF (sentence 1)	TF*IDF (Sentence 2)
earth	1/8	0	$\log_2(2/1)=0$	0.0375	0
is	1/8	1/5	$\log_2(2/2)=0$	0	0
the	2/8	1/5	$\log_2(2/2)=0$	0	0
third	1/8	0	$\log_2(2/1)=0.3$	0.0375	0
planet	1/8	1/5	$\log_2(2/2)=0$	0	0
from	0	0	$\log_2(2/1)=0.3$	0	0
sun	1/8	0	$\log_2(2/1)=0.3$	0.0375	0
largest	0	1/5	$\log_2(2/1)=0.3$	0	0.06
Jupiter	0	1/5	$\log_2(2/1)=0.3$	0	0.06

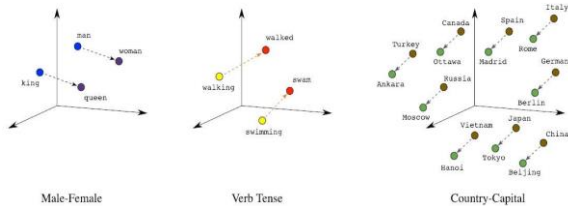
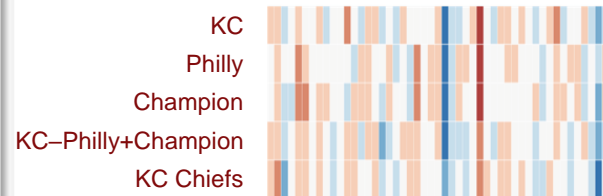
Limitation:

- Semantic meaning not captured
- No learning - predictions limited to occurrences in the training set

Neural Networks (2003-2014)

Neural Language Models use a neural network to predict the words that would occur in a **corpus**

(Kansas City – Philly) + Champion = KC Chiefs



Limitation:

- Does not capture long-term memory or associations

Attention is All You Need (2015-2018)

Attention-based Transformers break down **corpuses** into **smaller parts (tokens)** and then analyze the relationships between them

The FBI is chasing a criminal on the run .
 The **FBI** is chasing a criminal on the run .
 The **FBI** **is** chasing a criminal on the run .
 The FBI **is** chasing **a** criminal on the run .
 The FBI **is** chasing **a** criminal **on** the run .
 The FBI **is** chasing **a** criminal **on** the **run** .
 The **FBI** **is** chasing **a** criminal **on** the **run** .
 The **FBI** **is** chasing **a** criminal **on** the **run** .

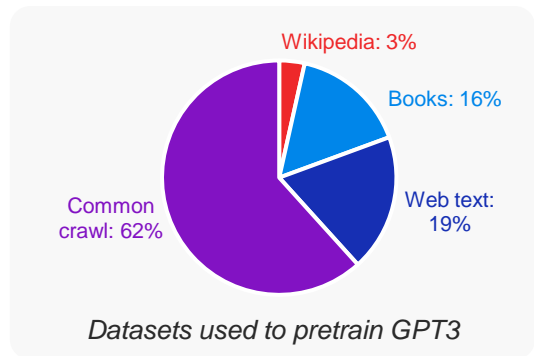
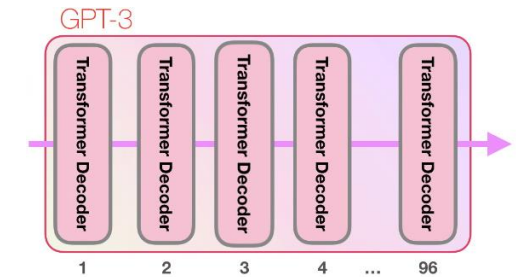
Current word is in red, and size of blue shade indicates strength of relevance

Limitation:

- Requires a large corpus and larger size for greater capability

Pretrained LLMs (2018-Present)

LLMs have several transformer units stacked together and are trained on a huge **corpus**



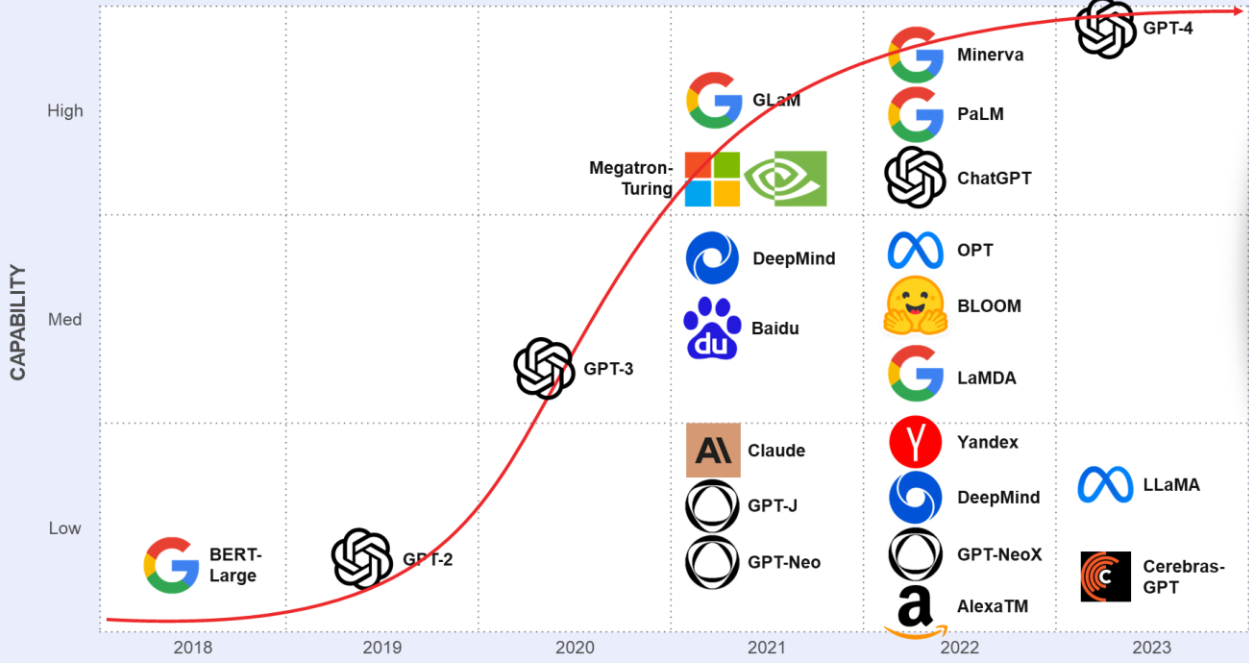
Limitation:

- Investment in infrastructure for training and inference

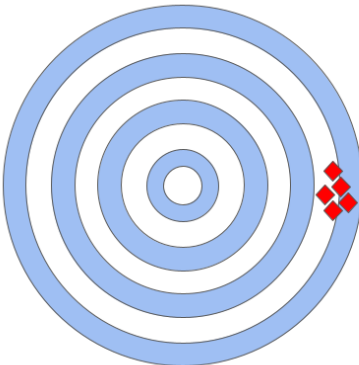


ChatGPT stands out with reinforcement learning via human feedback

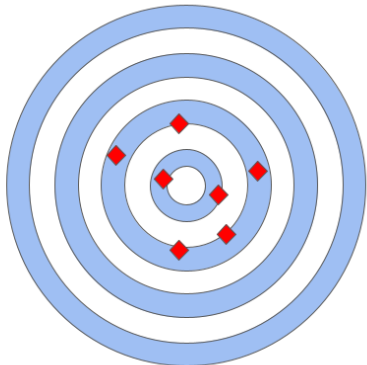
The capability of an LLM is not the only factor in determining its popularity (e.g., Google’s GLaM and Microsoft’s Megatron have been around longer than ChatGPT)



- Models must align their outputs to the intention and instruction of the user
- This requires Reinforcement Learning from Human Feedback (RLHF) – ChatGPT’s secret sauce



High capability
Low alignment



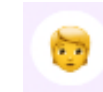
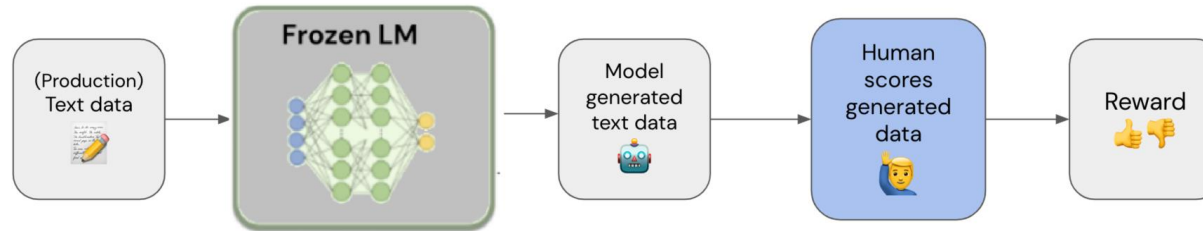
Low capability
High alignment



RLHF enables LLMs to communicate more like humans

1

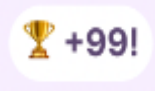
A small dataset of prompts and **human-ranked responses** is used to train a reward model



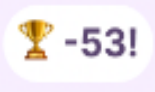
I want a story of Yoda fighting a robot



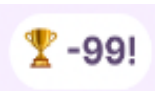
Long long ago, there was a robot...



Yoda waved hello. The robot died. The end.

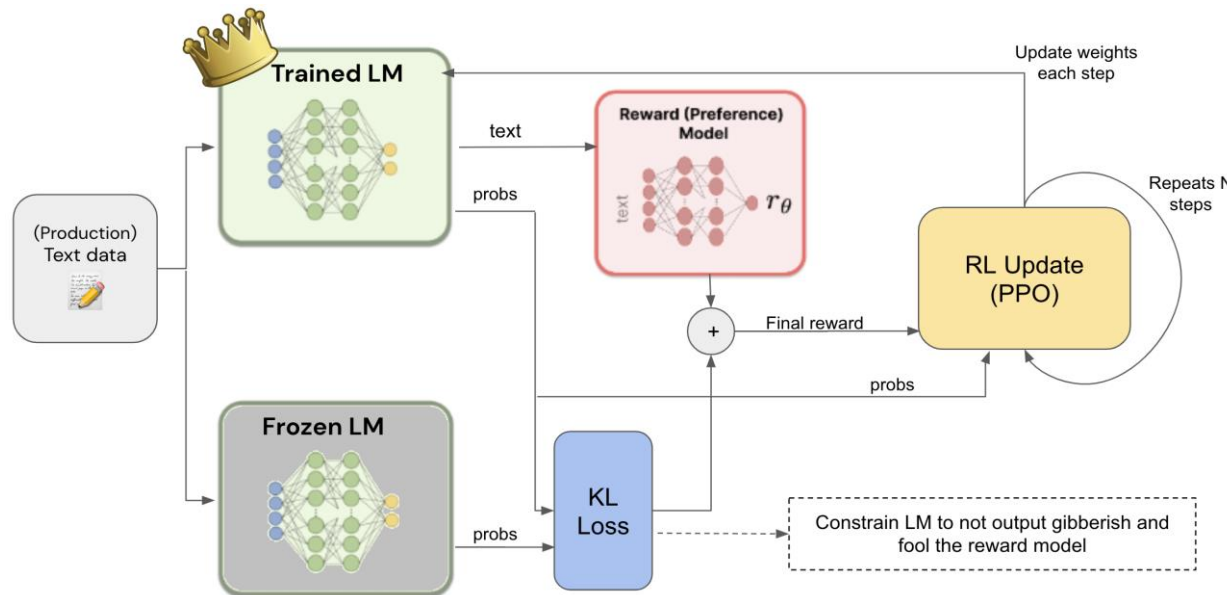


Yoda is a fictional character in Star Wars, first appearing in the 1980...



2

The Reward Model teaches the LLM to fine-tune weights to follow instructions with guardrails using **Reinforcement Learning**




1. Frozen model's response to input prompt



Write a poem about Iron Man



I'll be a wall and shield you from danger;
I'll take all the pains,
For to me they are no stranger.

2. Score the output using reward model 

3. Improve the response using reward and re-generate response & reward iteratively



Billionaire Tony Stark made a metal suit;
With a flamethrower and rockets in his boots.

Tailored solutions for every client

Typical questions from various customers

Training an LLM

Should I train an LLM with all my old records: claims, customer correspondence, transcriptions, etc.?



Fine-tuning an LLM

Can ChatGPT truly understand my unique business when it responds to my questions?



Prompt engineering

How can I use ChatGPT to help me write a SQL query given a question in English?



Questions to guide our clients on their LLM journey

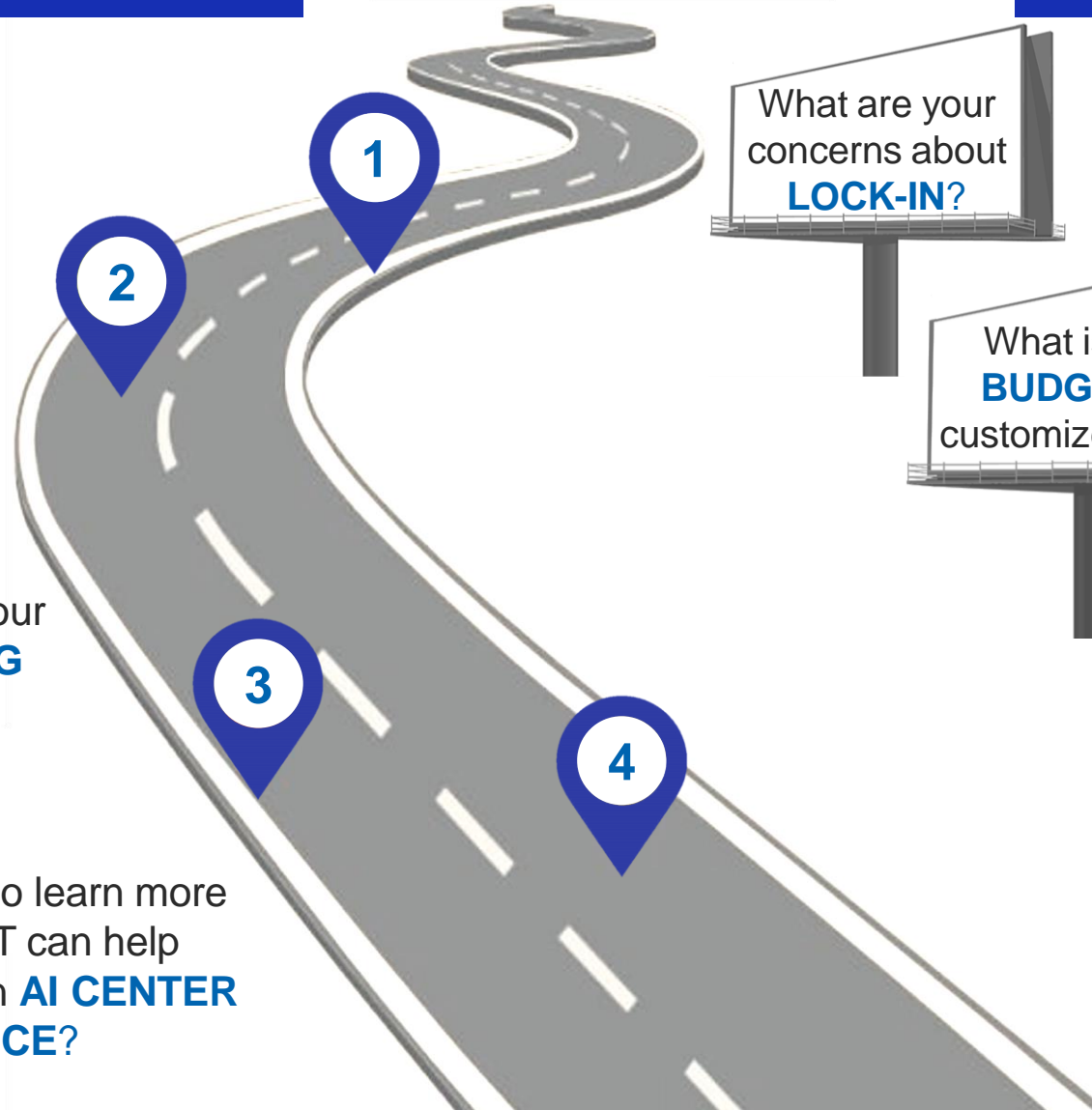
LLM and GPT acceleration

Who is responsible for **AI SOLUTIONS STRATEGY** at your organization?

Which **AI APPLICATIONS** or **GPT SOLUTIONS** are you most interested in deploying?

Have you assessed your readiness for **TESTING** and **ADOPTING** a customized LLM ?

Would you like to learn more about how WWT can help you establish an **AI CENTER OF EXCELLENCE**?



Reduce risks

What are your concerns about **LOCK-IN**?

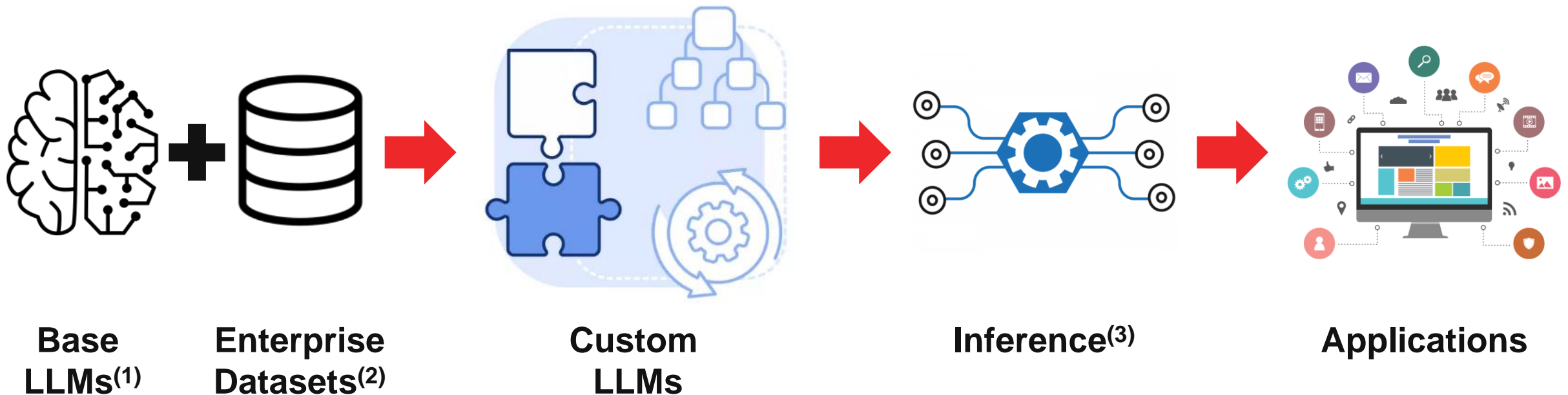
What is your **BUDGET** for customized LLM?

What are your concerns around **SHADOW AI**?

Do you have an existing **DATA SECURITY** plan?

Fine-tuning unlocks the full potential of your language model

Fine-tuning is used to modify the weights of an LLM based on information and patterns contained in large amounts of enterprise data

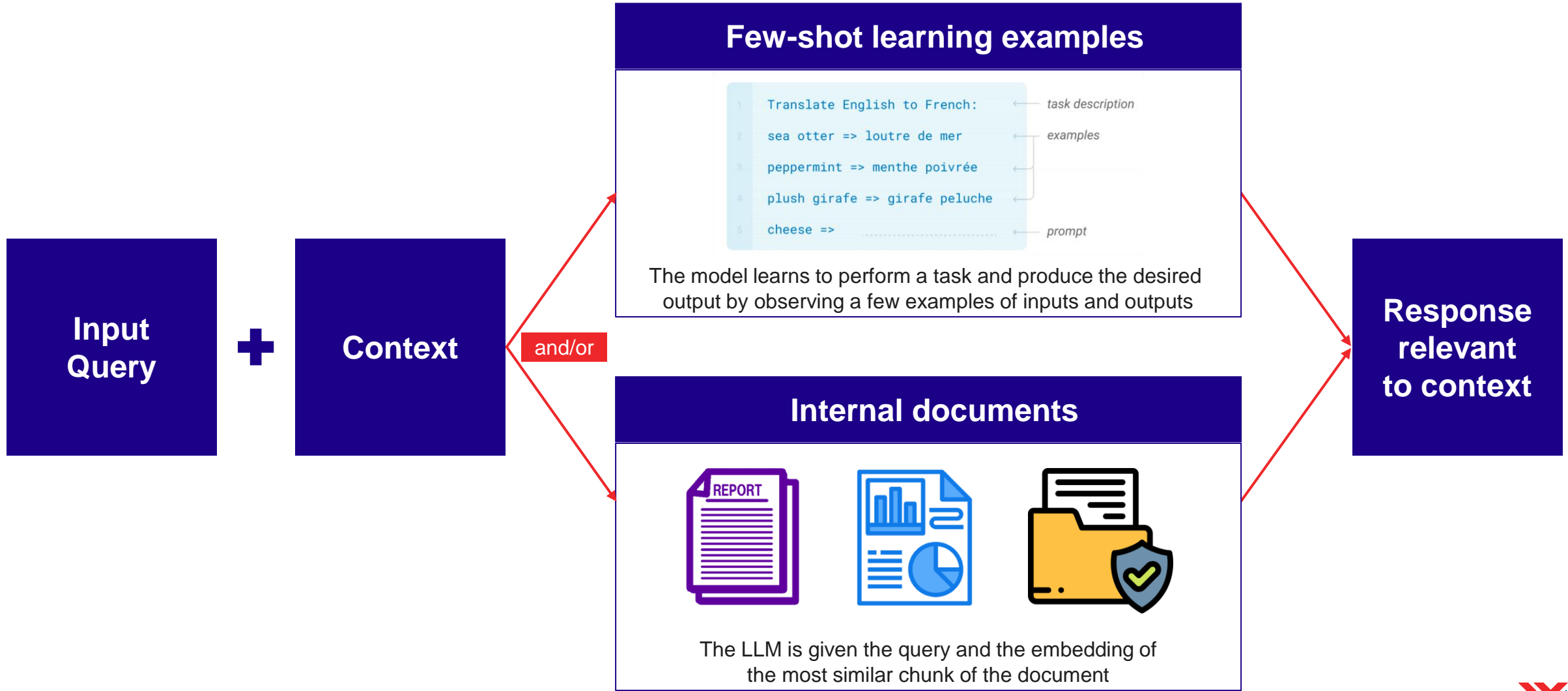


(1) Base LLMs may be state of the art proprietary LLMs (paid access on Cloud) or open-source LLMs

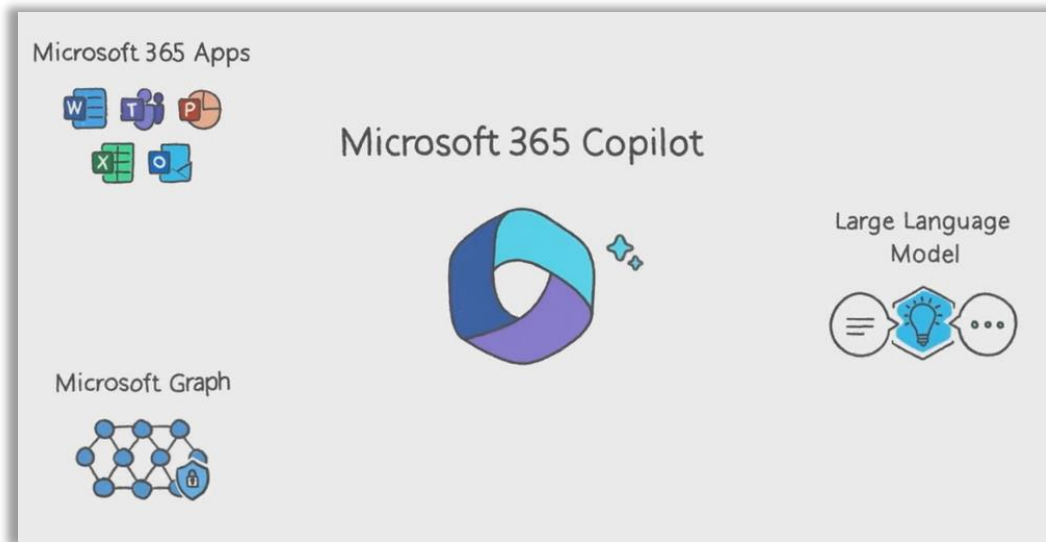
(2) Datasets need to be prepared and loaded appropriately to achieve fine-tuning

(3) Inferences may be free or billed per hit depending on whether the solution is deployed on a Datacenter (resource intensive capital investment) or Cloud

Two powerful strategies to engineer prompts can be used to produce customized responses



Microsoft Copilot offers a reference for customization



1

You provide Copilot with directions

User types a command – aka “prompt” - into the copilot assistant and a LLM generates a response

Microsoft 365 Copilot



2

Copilot retrieves 365 information with Graph

Copilot uses the Microsoft Graph to retrieve information from your enterprise Microsoft 365 app.

Microsoft Graph

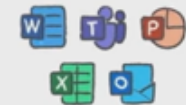


3

Copilot Processes 365 & Graph information

Copilot combines the response from the LLM and the information from the Microsoft Graph to generate a response

Microsoft 365 Apps



Microsoft Graph



4

You Receive a Final Response

Copilot then sends the final response to you

Large Language Model



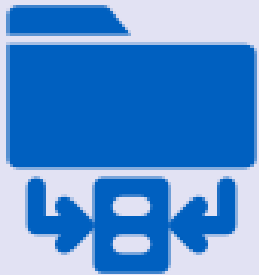
WWT solution: leveraging LLMs to maximize benefits



Proprietary Applications
Software applications that are owned and controlled by a single company or individual



Custom LLM Solution
WWT design and architecture services for LLM solutions



Domain-Centric Data Formats
A way of organizing data based on the concepts and relationships that are important to the domain.



Large Language Model
Trained baseline model to be fine-tuned on proprietary data



Improved data querying



Efficient use of LLM



Time saving



Increased flexibility

WWT is exploring ethical applications of AI that impact development and LLM operations



Bias & Discrimination

Training Implicit bias into AI
(discrimination, stereotypes)



Privacy

Data gathering, storage,
use and IAM



Transparency

Understanding AI functions
To establish trust



Misinformation

Produce convincing but
misleading information



Safety Concerns

Create damaging or violent content,
or offer directions for
unlawful or dangerous conduct



Accountability

Identify responsible
and accountable parties
for AI product and use

