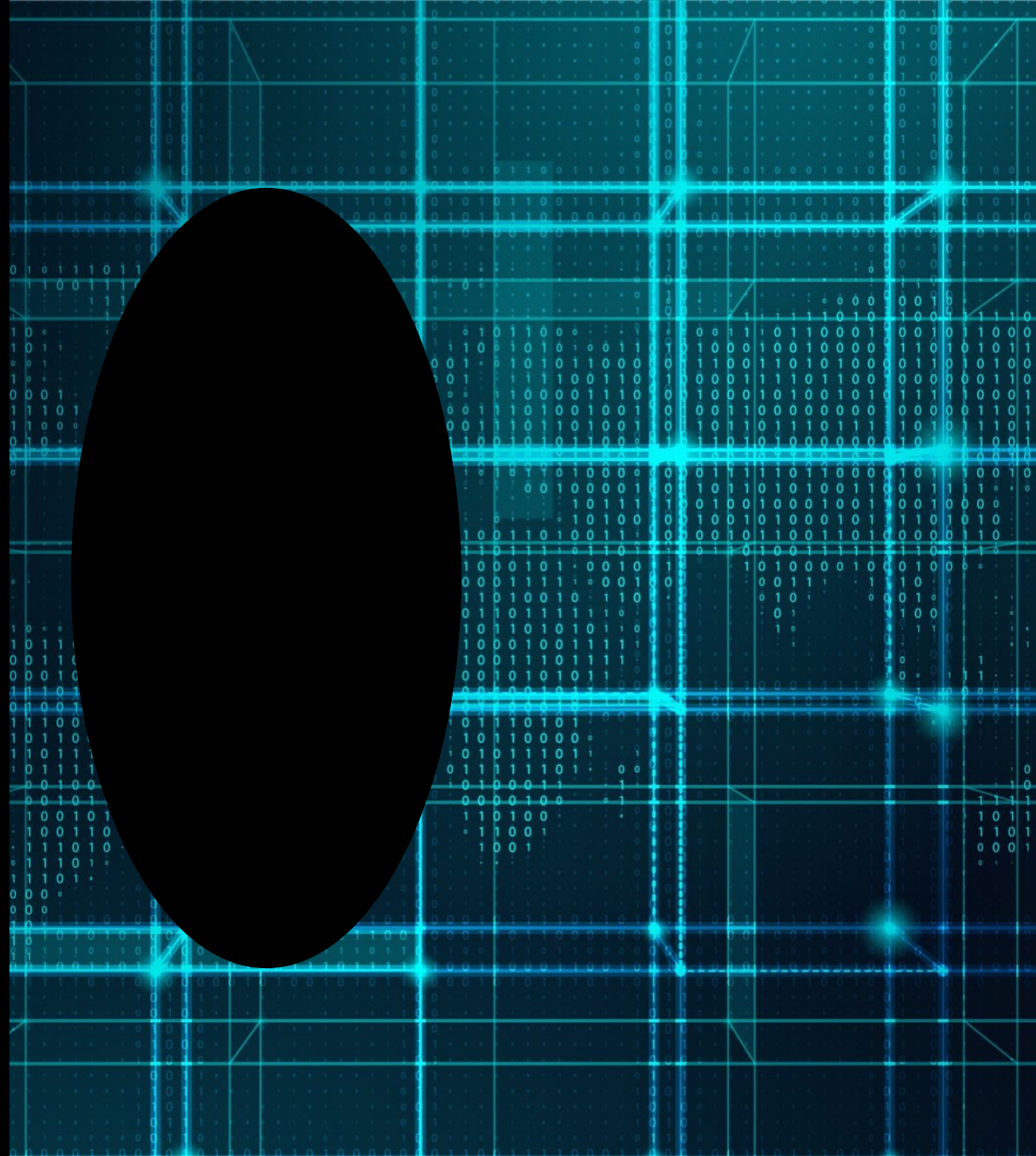
GENERATIVE AI

Surendra Panpaliya

GENERATIVE AI & AGENTIC AI WITH LANGCHAIN, LANGGRAPH & SEMANTIC KERNEL – 3-DAY PROGRAM

Intensive training on advanced AI frameworks and applications



PROGRAM INTRODUCTION

GENERATIVE AI FUNDAMENTALS

Training Program Overview

The program focuses on Generative AI and Agentic AI using LangChain, LangGraph and Semantic Kernel technologies.

Target Audience

Senior IT professionals from iLink Digital Pune are the primary participants aimed for this 3-day immersive training.

Trainer Credentials

Surendra Panpaliya is a Global AI Trainer and Consultant with distinguished communication expertise.

Learning Goals

Empowering participants to build enterprise-ready AI agents through interactive and practical sessions.

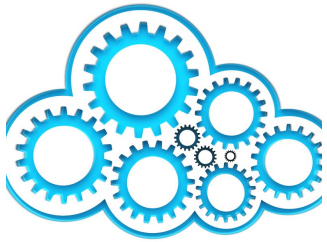
SURENDRA PANPALIYA

- **25 years of experience** driving AI-powered digital transformation across industries.
- **Founder & CEO, GKTCS Innovations** – building future-ready enterprises.
- **Empowered 35,000+ IT professionals** through training, mentoring, and consulting.
- **Partnered with 300+ multinational corporations** to accelerate business growth.
- Specialist in **AI-driven strategies, Generative AI, and advanced technology adoption**.

ABOUT ILINK DIGITAL

Company Overview

iLink Digital was founded in 2002 and has a global presence with an innovation hub in Pune.



Core Focus Areas

Key areas include Data Engineering, Generative AI, CloudOps, and Business Applications driving digital innovation.

Digital Transformation Strength

iLink Digital delivers pre-built frameworks that accelerate the pace of digital transformation for clients.

ICEBREAKER ACTIVITY

Share one

Interesting experience or

Challenge with AI or LLMs

PROGRAM OVERVIEW

PROGRAM OBJECTIVES

Understanding Generative AI

Explore the evolution and architecture of Generative AI for a strong foundational knowledge.

Building AI Agents

Focuses on creating and orchestrating AI agents

using LangChain, LangGraph, and

Semantic Kernel technologies.

PROGRAM OBJECTIVES

Enterprise-Grade RAG Pipelines

Learn to develop advanced Retrieval-Augmented Generation pipelines suitable for enterprise applications.

Capstone AI Agent Delivery

Demonstrating real-world application skills.

PROGRAM FLOW (3-DAY OVERVIEW)

Day 1: Foundational Concepts

The first day covers foundational AI concepts and frameworks that set the base for advanced topics.

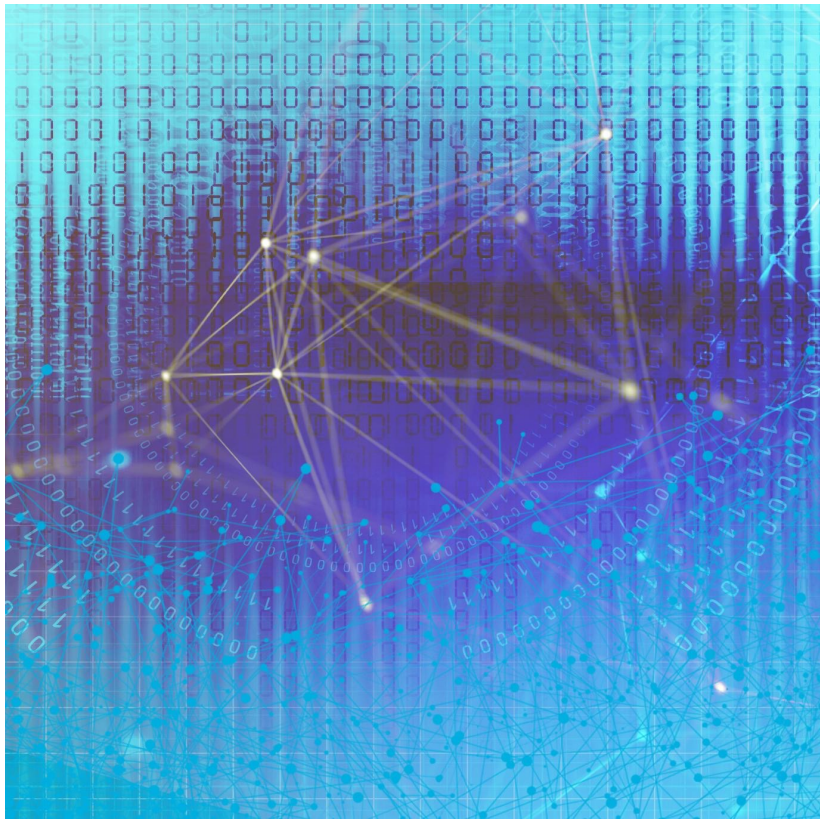
Day 2: Agent Design and Memory

Day two focuses on agent design, memory systems, and retrieval-augmented generation pipelines.

Day 3: Multi-Agent Orchestration

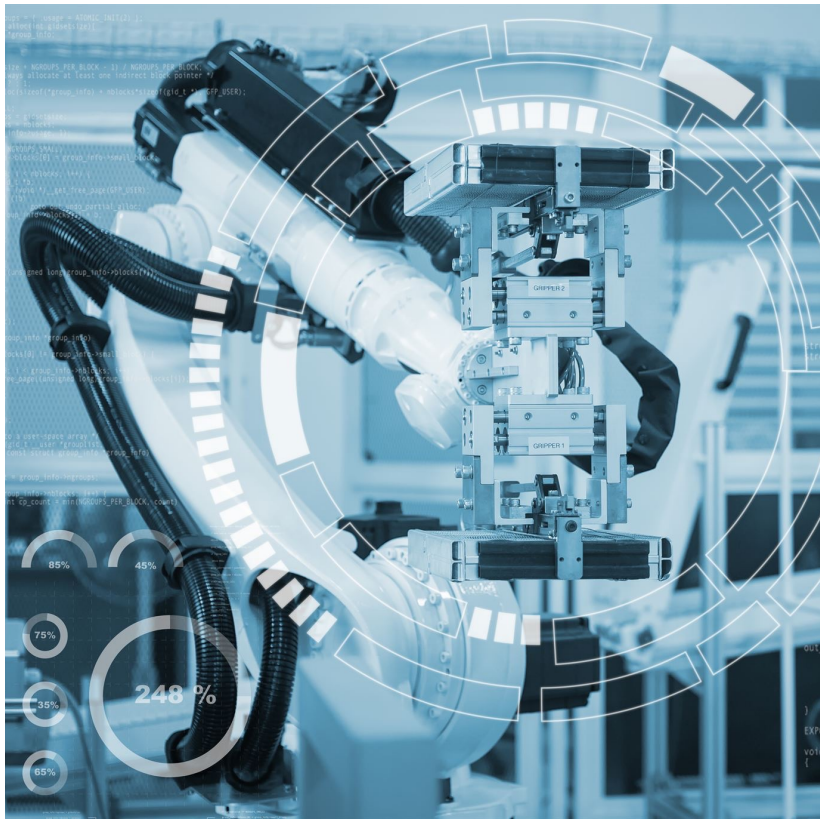
The final day emphasizes multi-agent orchestration and deployment of production-ready AI systems.

DAY 1 – FOUNDATIONS & FRAMEWORKS



- Kick-off & Program Overview
- Generative AI Evolution
- Transformer Architecture Demystified
- Generative AI with APIs
- LangChain Basics
- Semantic Kernel Basics
- LangChain + SK Together
- Intro to LangGraph

DAY 2 – AGENT DESIGN, MEMORY & RAG



- LangChain Agents Deep Dive
- LangGraph Advanced
- Memory & Personalization
- Semantic Kernel Advanced Features
- RAG Pipelines
- Real-World Integration

DAY 3 – MULTI-AGENT ORCHESTRATION & DEPLOYMENT

- Multi-Agent Orchestration with LangGraph
- Complex Tooling & Reasoning
- Advanced Features (Latest Updates)
- Observability & Governance
- Capstone Build
- Deployment Workshop

CAPSTONE & TOOLS

CAPSTONE PROJECT OVERVIEW

Project Objective

The Capstone Project challenges participants to build a production-ready enterprise AI agent.

Example Projects

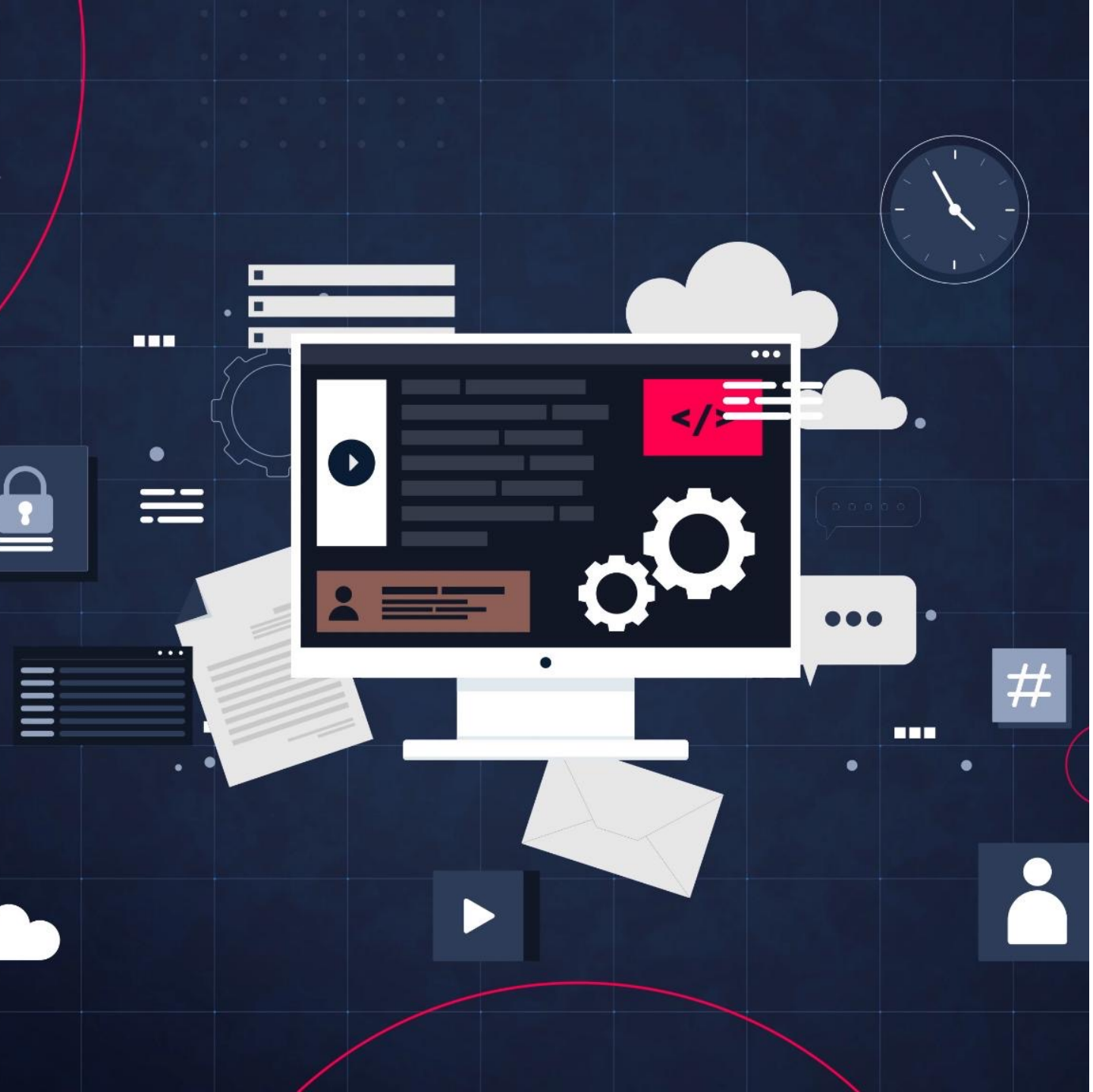
Sample projects include Policy Compliance Agent, Data Retrieval Assistant, and Risk Analysis Multi-Agent System.

Evaluation Criteria

Projects are evaluated based on innovation, architectural design, and clarity of deployment strategy.

Learning Application

The project allows participants to apply tools and concepts learned throughout the training effectively.



TOOLS & ENVIRONMENT SETUP

Core AI Tools

LangChain, LangGraph, Semantic Kernel, and vector databases like Milvus, FAISS and Chroma enable advanced AI agent development.

Development Environment

Python 3.11+, VS Code, and Jupyter Notebook provide a versatile and powerful coding environment for participants.

Deployment Options

Deployment can be done via cloud platforms like Azure and AWS, or local setups, offering flexible infrastructure choices.

OUTCOMES & CLOSURE

WRAP-UP & ACKNOWLEDGMENT

SESSION 2: GENERATIVE AI EVOLUTION

SESSION OBJECTIVES

Understanding Generative AI

Explore the historical evolution and fundamentals of Generative AI.

GPT Models Architecture

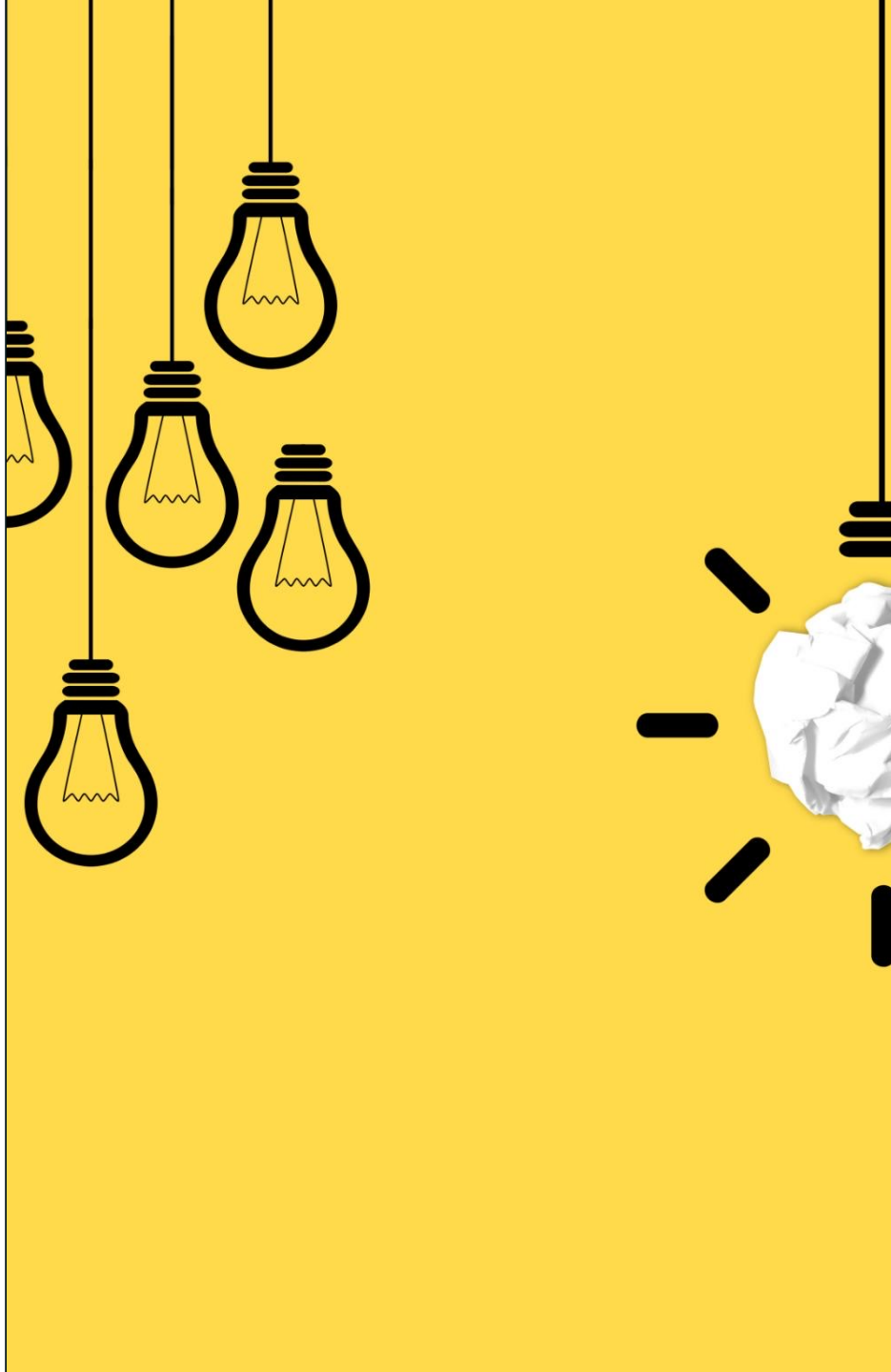
The session covers architecture and capabilities of GPT models including GPT-4 and GPT-5 comparisons.

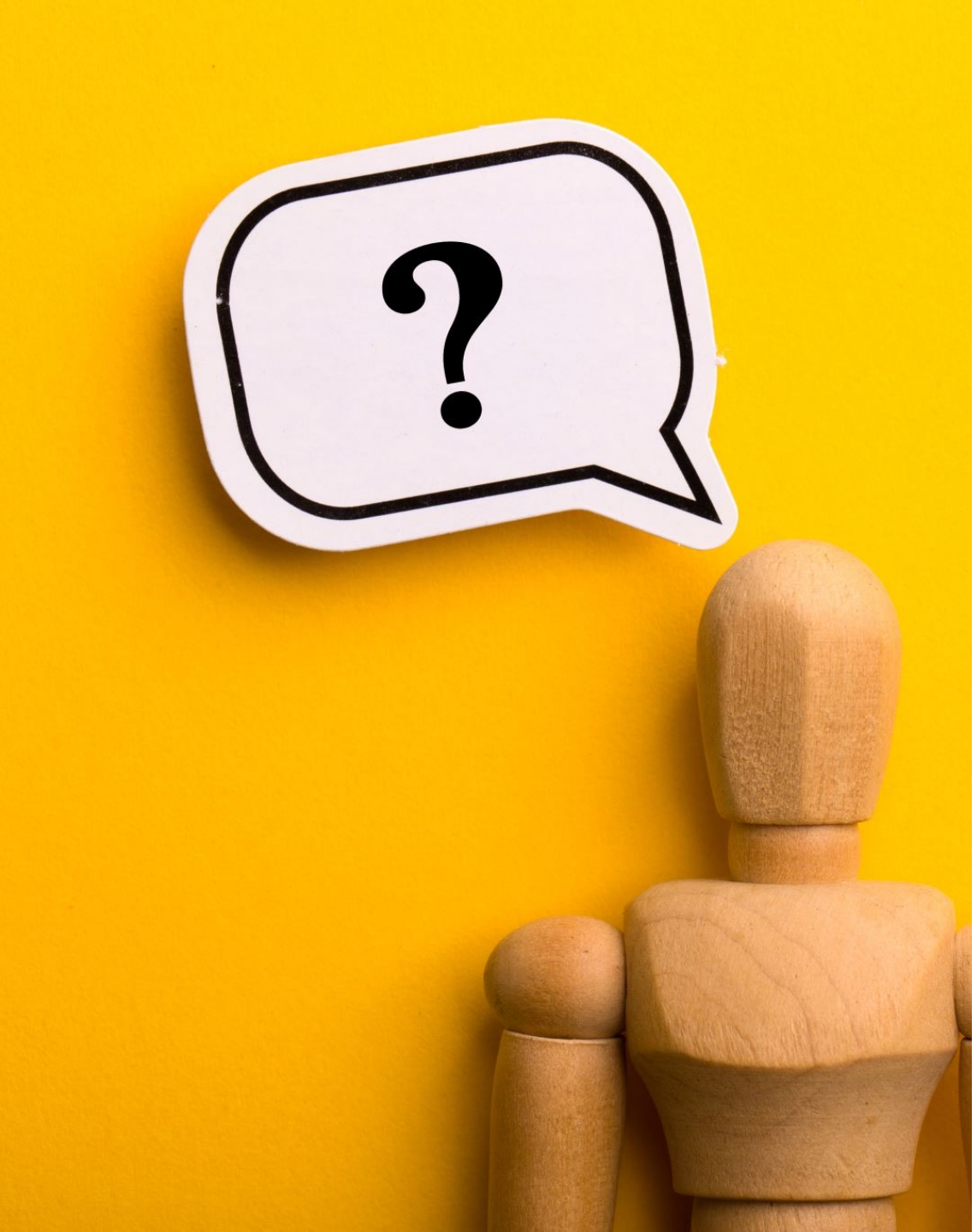
Hands-On GPT-5 Practice

Participants will engage in practical exercises using GPT-5 for summarization, Q&A, and creative generation.

Learning Path Visualization

A visual graphic depicts the learning journey from theory to practical application.





What’s your current exposure to Generative AI?

WHAT IS GENERATIVE AI?

Branch of artificial intelligence focuses on

Creating new content

Text, Images, Code, Audio, Video

by learning from **large datasets.**

WHAT IS GENERATIVE AI?

https://www.youtube.com/watch?v=rwF-X5STYks

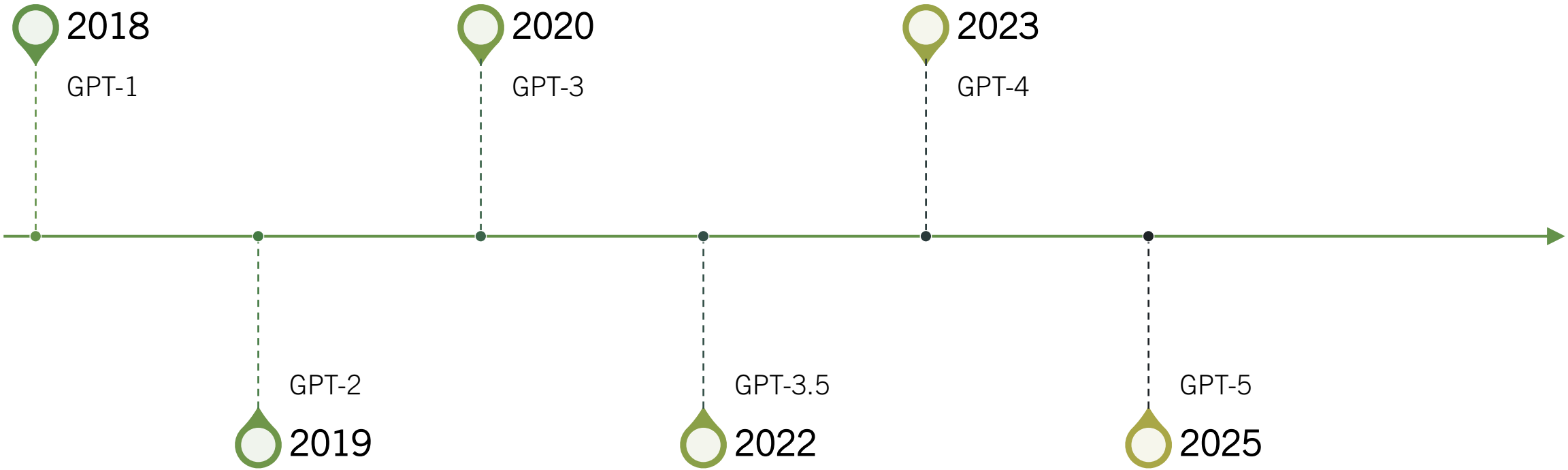
EVOLUTION OF GPT MODELS → GPT-5

Generative Pretrained Transformers (GPT)

are a family of large language models

developed by **OpenAI**.

EVOLUTION OF GPT MODELS



GPT-1: THE CURIOUS CHILD (2018)

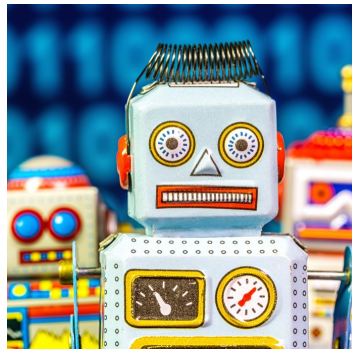
The proof-of-concept.

Introduced the transformer architecture

for natural language understanding.

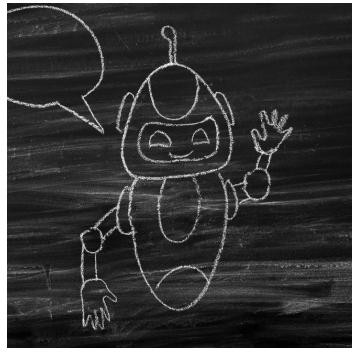
Trained on ~117M parameters.

GPT-1: THE CURIOUS CHILD



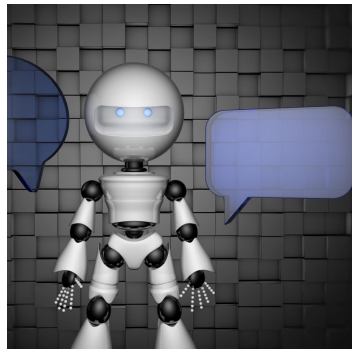
Early Language Model

GPT-1 was an initial step in large language models, launched in 2018 with 117 million parameters.



Limited Understanding

Much like a child learning to speak, GPT-1 often misunderstood context and created simple sentences, sometimes babbling unexpectedly.



Not Conversation Ready

GPT-1 could surprise but was not able to handle deep or serious conversations, similar to a toddler’s chatter.

🚶 GPT-2: THE TALKATIVE TEENAGER (2019)

Showed strong text generation capabilities.

1.5B parameters.

Famous for being partially withheld

due to concerns about misuse.

GPT-2: THE TALKATIVE TEENAGER

Enhanced Conversational Abilities

GPT-2 introduced 1.5 billion parameters, enabling longer, more coherent and context-aware conversations.

Mimicking Writing Styles

It could mimic various writing styles, making its generated text appear more human-like and adaptable.

Mixing Fact with Fiction

GPT-2 sometimes blended facts with fiction, resulting in creative but occasionally inaccurate information.

🏃 GPT-3: THE TALENTED YOUNG ADULT (2020)

Breakthrough in size (175B parameters).

Demonstrated zero-shot, few-shot learning.

Powered the first wave of real-world LLM applications.

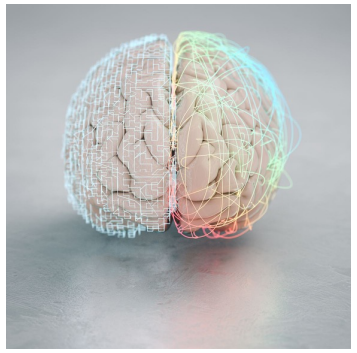
INTRODUCING GPT-3 BREAKTHROUGH

Massive AI Model Scale

With 175 billion parameters, GPT-3 set new standards in AI, enabling unprecedented language understanding and generation.

Versatile Content Creation

GPT-3 could generate essays, poems, code, and articles with fluent and creative language, appealing to enterprises.



Intelligence and Limitations

Despite its creativity and knowledge, GPT-3 still lacked real-world understanding and consistent reliability in complex tasks.

GPT-3.5 (2022)

Optimized variant of GPT-3

with reinforcement learning from human feedback (RLHF).

Provided better dialogue handling

the engine behind **ChatGPT's early versions.**

🚀 GPT-4: THE MATURE PROFESSIONAL (2023)

Huge improvement in reasoning,

multi-modal inputs (text + images),

reduced hallucinations,

stronger coding and enterprise reliability.

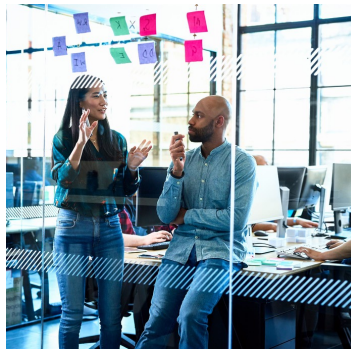
GPT-4: MULTIMODAL EVOLUTION

Multimodal Capabilities

GPT-4 processes both text and images, enabling richer and more accurate responses to diverse queries.

Superior Reasoning and Coding

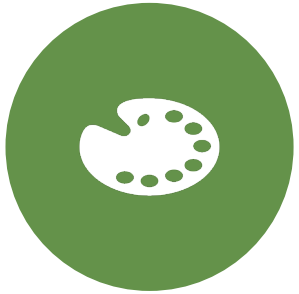
With enhanced reasoning, GPT-4 can generate production-level code and solve complex problems efficiently.



Enterprise-Ready and Reliable

GPT-4 offers robust, trustworthy support for enterprise needs and continuously improves its ability to explain decisions.

GPT-5: THE ENTERPRISE LEADER (2025)



Current state-of-the-art.

Expanded reasoning depth, memory,

multi-modal

text, images, audio, video

GPT-5: TRANSFORMING COLLABORATION

Deeper Contextual Understanding

GPT-5 features a larger context window, enabling richer comprehension and more insightful conversations for business needs.

Tailored Response Control

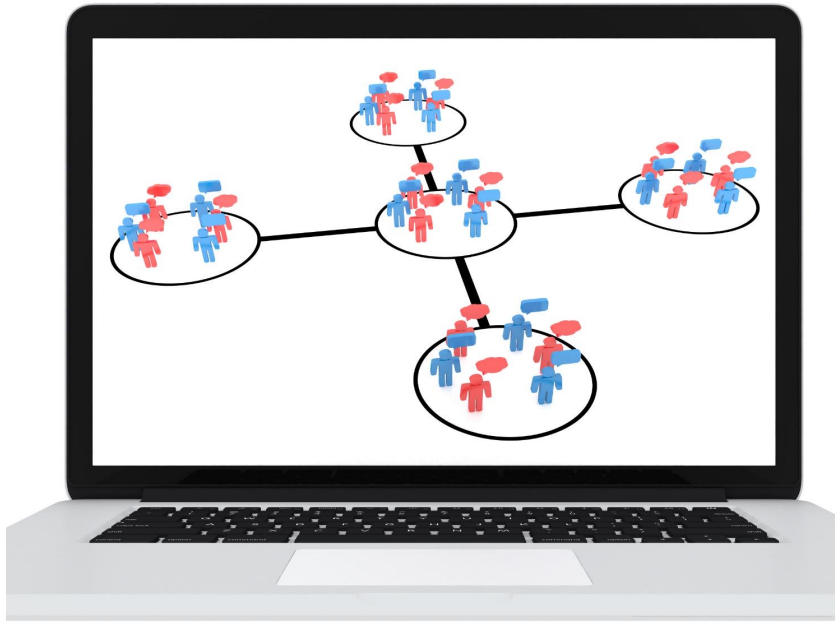
Verbosity control allows users to customise responses, making communication efficient and well-suited to different situations.

Advanced Collaboration and Governance

GPT-5 orchestrates multi-agent tasks, ensuring governance, compliance, and seamless teamwork across organisational roles.

Driving Digital Transformation

Acting as an organisational guide, GPT-5 leads teams in adopting new technologies and transforming business processes.

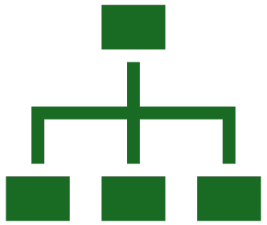


GPT-5: THE ENTERPRISE LEADER (2025)

- IMPROVED **FACTUAL ACCURACY**,
- **TOOL ORCHESTRATION**,
- **AGENTIC AI** SUPPORT.

GPT-5: THE ENTERPRISE LEADER (2025)

Designed for **enterprise integration**



with **compliance, governance, and**

security in mind.

COMPARATIVE CHART: GPT EVOLUTION

Feature / Model	GPT-3 (2020)	GPT-3.5 (2022)	GPT-4 (2023)	GPT-5 (2025)
Parameters	175B	~175B (optimized)	~1T est.	Multi-trillion scale (sparse/expert models)
Core Capability	Few-shot learning	Conversational fine-tuning (RLHF)	Advanced reasoning, multimodal	Deep reasoning, planning, multi-modal (text, images, audio, video)

COMPARATIVE CHART: GPT EVOLUTION

Feature / Model	GPT-3 (2020)	GPT-3.5 (2022)	GPT-4 (2023)	GPT-5 (2025)
Context Window	2K–4K tokens	8K tokens	32K–128K	**Up to millions (long-context memory)**
Use Cases	Chatbots, Q&A, apps	Conversational AI	Enterprise copilots, coding, multimodal	**Agentic AI, enterprise copilots, compliance-ready AI assistants**

COMPARATIVE CHART: GPT EVOLUTION

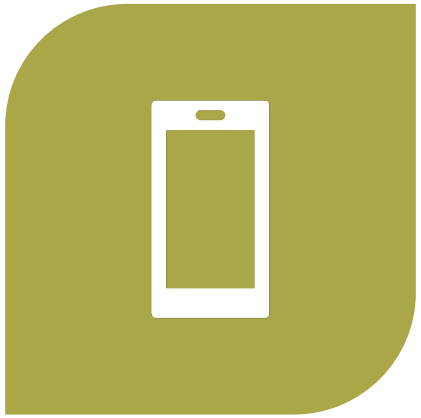
Feature / Model	GPT-3 (2020)	GPT-3.5 (2022)	GPT-4 (2023)	GPT-5 (2025)
Hallucinations	Moderate	Lower than GPT-3	Reduced further	Minimal with retrieval + reasoning
Fine-Tuning	Yes	Optimized	Yes, domain-specific	Advanced domain adaptation with guardrails

COMPARATIVE CHART: GPT EVOLUTION

Feature / Model	GPT-3 (2020)	GPT-3.5 (2022)	GPT-4 (2023)	GPT-5 (2025)
Enterprise Focus	Startup adoption	Developer adoption	Enterprise pilots	Enterprise-first (governance, observability, integration)

KEY TAKEAWAY

GPT-5 REPRESENTS A SHIFT FROM



SMART ASSISTANT TO

ENTERPRISE-READY AGENT.

CORE STRENGTHS OF GPT-5

Advanced Reasoning & Problem Solving

Verbosity & Tone Control

Expanded Context Window

Handles Large documents, Entire codebases

CORE STRENGTHS OF GPT-5

- Multi-Agent Orchestration (Agentic AI)
- Seamless Integration with Tools & APIs
- Enterprise-Grade Security & Governance
- Scalability for Enterprises

https://platform.openai.com/docs/overview

KEY TAKEAWAY

It not only generates but

reasons, plans, and

integrates securely

into enterprise ecosystems.

GPT-4 (CLASSIC) VS GPT-5 (NEXT-GEN)

Aspect	GPT-4 (Classic)	GPT-5 (Next-Gen)
Creativity Control	temperature, top_p	creativity (abstracted, tuned with verbosity + reasoning)
Length Control	max_tokens	verbosity (style-based length, not just a cap)
Repetition Control	frequency_penalty, presence_penalty	Integrated into discourse management (automatic coherence)
Tone/Style	Manual prompting	Parameters like formality, directness
Reasoning	Emergent, prompt-engineered	Explicit: reasoning=deep vs reasoning=fast
Context Handling	Fixed token window	Weighted prioritization + extended memory alignment

WHY GPT-5 FOR ENTERPRISES?

Key Capabilities:

- Larger context window for project docs/codebases
- Verbosity control for stakeholders
- Agentic AI for end-to-end workflows
- Secure deployment with compliance

WHAT IS GENERATIVE AI?

Generative AI Overview

Generative AI creates new content in text, images, code, and audio across multiple modalities.

Foundation Models and Training

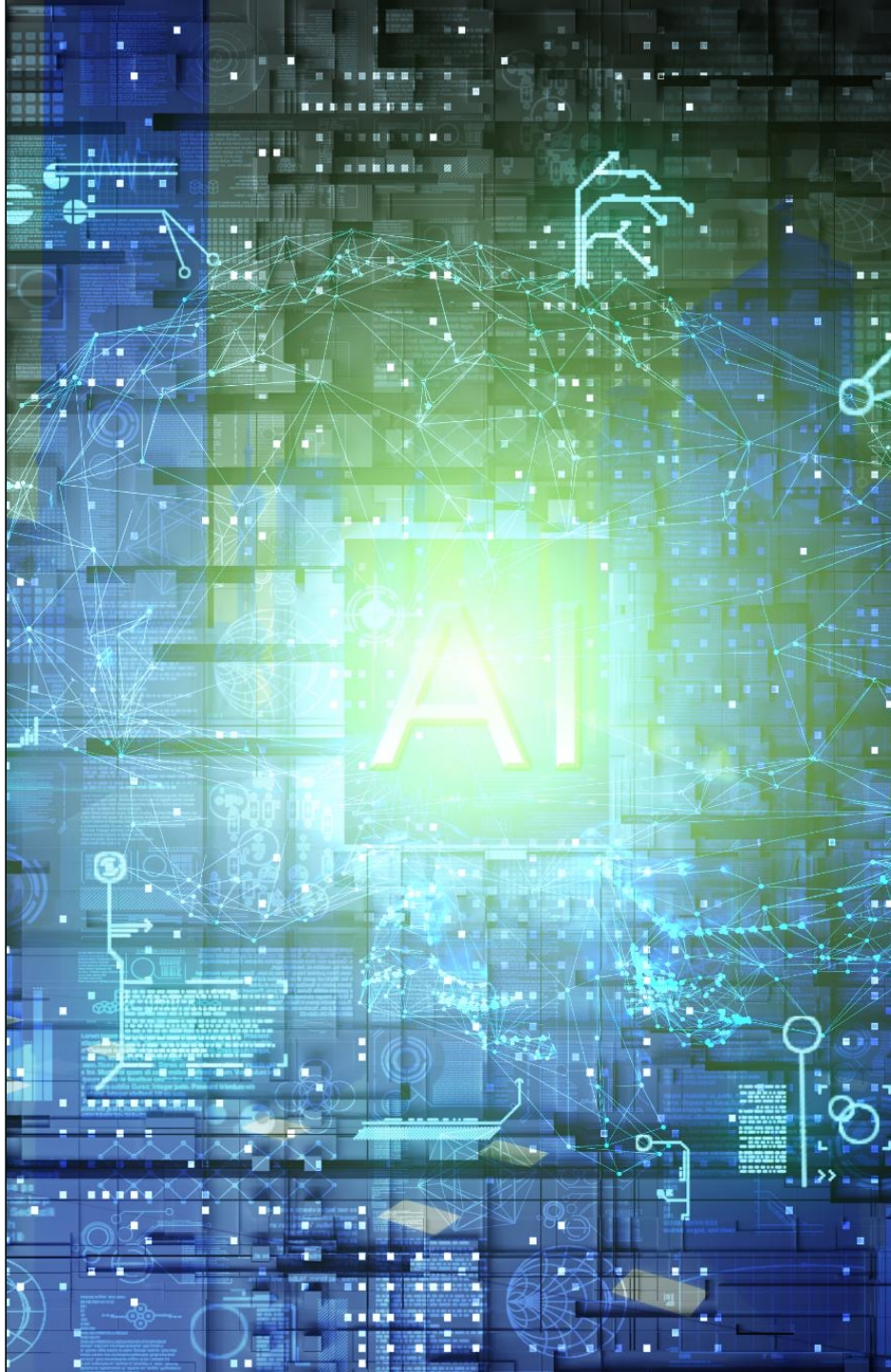
Foundation Models use transformers trained on large datasets with pre-training and fine-tuning stages.

Key Technical Concepts

Tokens, embeddings, and context windows form the basis for how generative AI understands and processes data.

Prompt Engineering

Prompt engineering guides model outputs by crafting effective inputs to shape AI-generated responses.



ENTERPRISE USE CASES

AI in Software Development

Generative AI automates code generation, creates test cases, and produces API documentation to boost development efficiency.

AI for Data Engineering

AI supports generating ETL queries and summarizing complex database schemas for streamlined data workflows.

CloudOps and Alert Analysis

Generative AI analyzes cloud operation alerts and summarizes logs to enhance monitoring and troubleshooting.

Business Applications

AI powers customer-facing chatbots and policy compliance assistants to improve service and governance.

EVOLUTION OF GPT MODELS

GPT-1: Transformer Pretraining

GPT-1 introduced pre-training transformers, laying the foundation for modern language models.

GPT-2: Enhanced Coherence

GPT-2 improved text coherence with a larger dataset and more parameters.

GPT-3: Few-Shot Learning

GPT-3 introduced few-shot learning enabling versatile reasoning with minimal examples.

GPT-4 and GPT-5 Advancements

GPT-4 added multimodal inputs and alignment; GPT-5 supports multi-agent systems and persistent memory.

GPT-4 VS GPT-5 COMPARISON TABLE

Architecture Evolution

GPT-4 uses a static transformer, while GPT-5 utilizes a dynamic memory transformer for enhanced flexibility.

Expanded Modalities

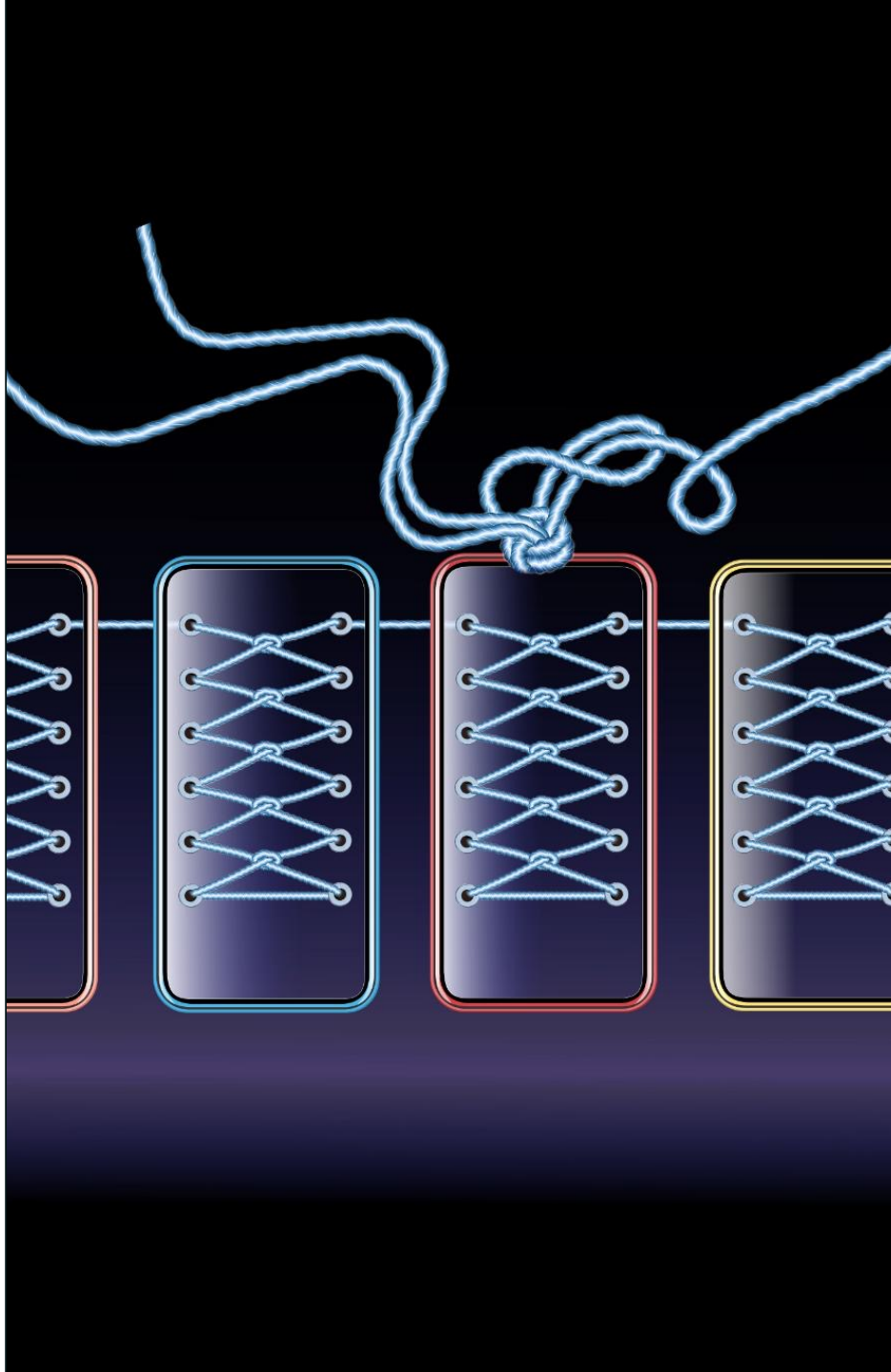
GPT-5 supports text, image, audio, video, and code, expanding beyond GPT-4's text and image capabilities.

Enhanced Context Window

Context window increased from 128K tokens in GPT-4 to over 1 million tokens in GPT-5 for full-document reasoning.

Advanced Features

GPT-5 introduces adaptive verbosity, multi-agent orchestration, and API integration via Model Control Plane.



GENERATIVE AI EVOLUTION

Advancements shaping intelligent content creation

UNDERSTANDING GENERATIVE AI

WHAT IS GENERATIVE AI?

Definition and Capabilities

Generative AI creates new content like text, images, audio, or code by learning from existing data patterns.

Deep Learning and Architectures

It uses deep learning models, especially transformers, to understand and replicate complex data patterns effectively.

Applications and Impact

Generative AI automates content creation, boosts productivity, and enables new human-computer interaction forms.

REAL-WORLD USE CASES

Customer Support Automation

AI-powered chatbots efficiently manage customer queries, reducing response times and improving satisfaction levels.

Content Creation Assistance

Generative AI tools help writers produce articles, marketing content, and social media posts with ease.

Software Development Aid

Generative AI suggests code snippets, debugs, and writes functions, streamlining developer workflows.

Healthcare Innovations

AI supports medical imaging analysis and personalized treatment plans to enhance healthcare outcomes.

EVOLUTION OF GPT MODELS

GPT-1 TO GPT-5 OVERVIEW

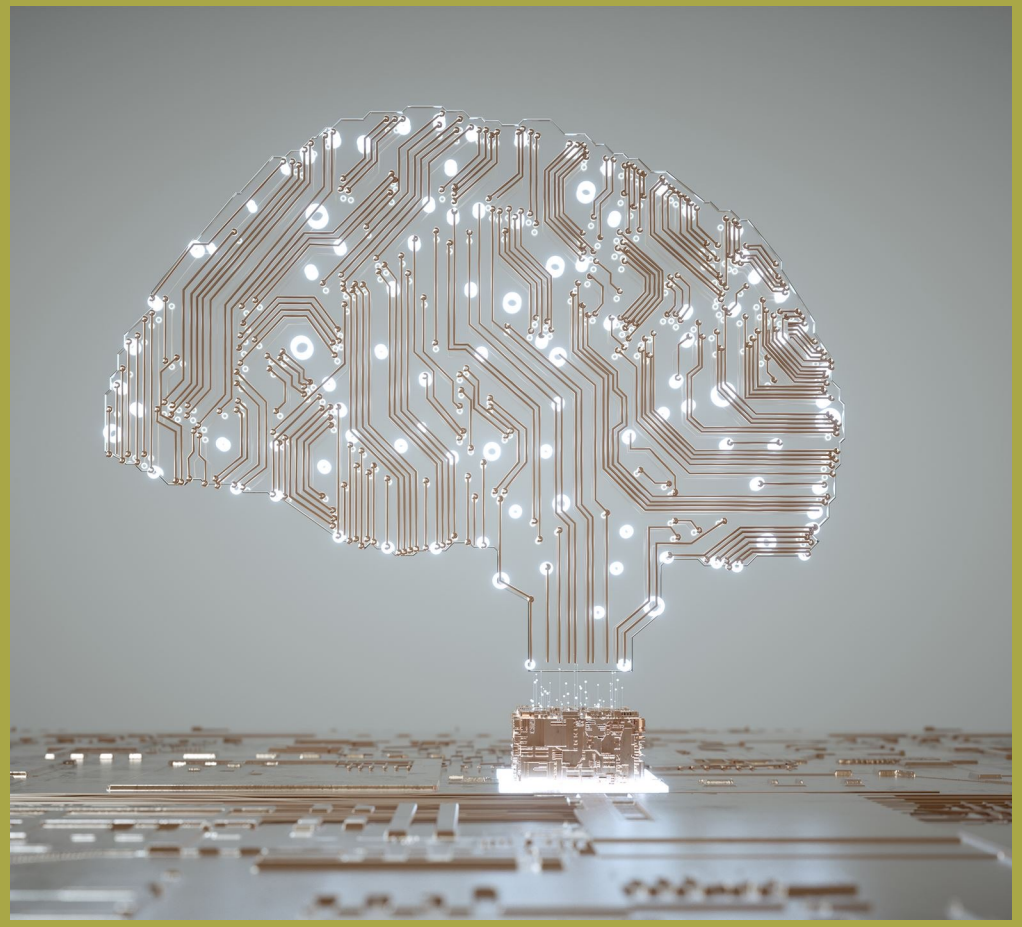
MODEL	YEAR	PARAMETERS	KEY FEATURE
GPT-1	2018	117M	Basic LM
GPT-2	2019	1.5B	Text generation
GPT-3	2020	175B	Few-shot learning
GPT-4	2023	Trillions	Multimodal
GPT-5	2025	Advanced	Planning + Tools

PARAMETER COMPARISON AND YEARLY EVOLUTION

YEAR	MODEL	PARAMETERS (MILLIONS)
2018	GPT-1	117
2019	GPT-2	1500
2020	GPT-3	175000
2023	GPT-4	1000000
2025	GPT-5	1500000

COMPARING GPT-4 AND GPT-5

CAPABILITIES AND PARAMETERS



Multimodal Capabilities

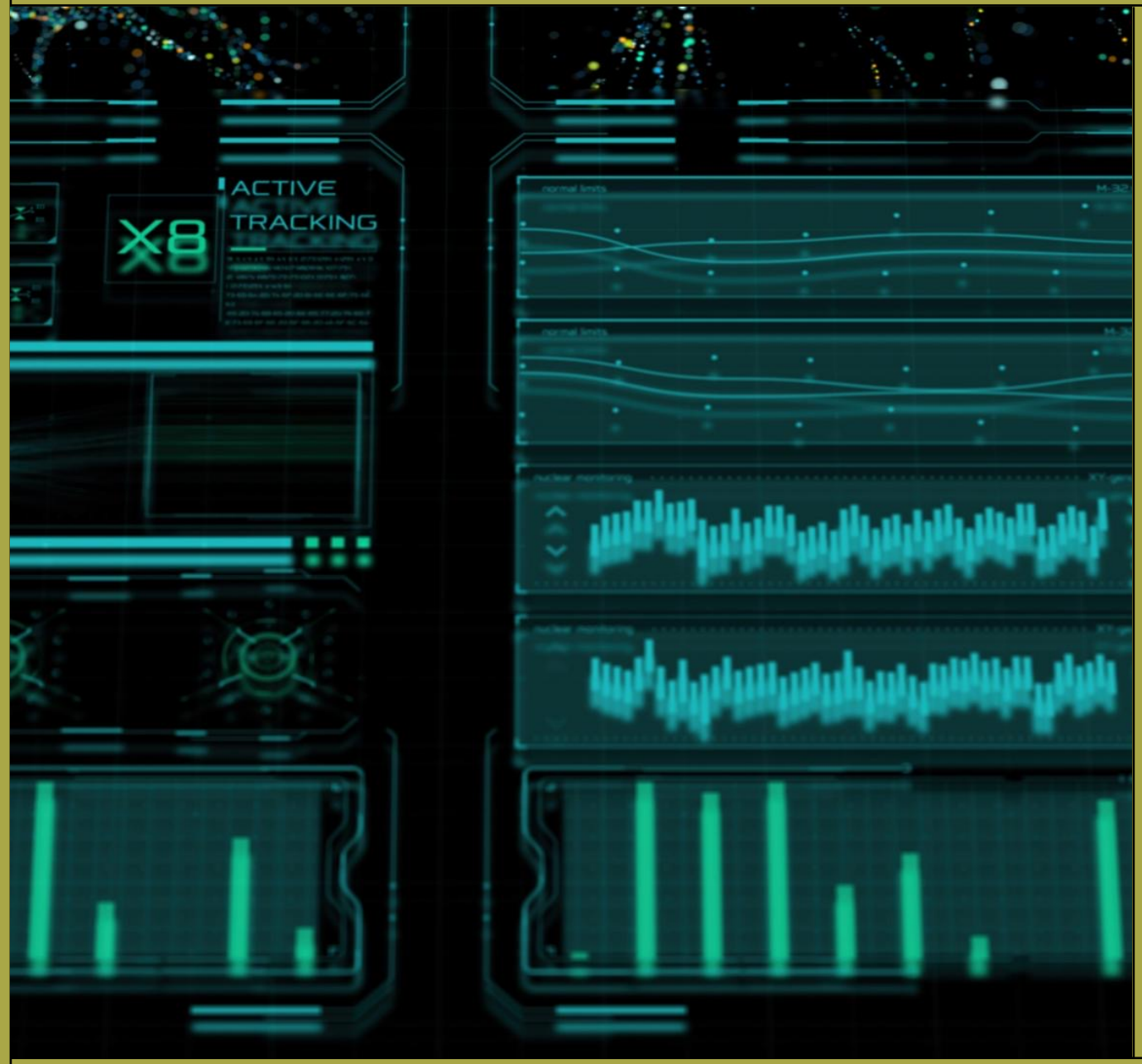
GPT-4 introduced multimodal capabilities, processing both text and images for enhanced understanding.

Advanced Planning and Integration

GPT-5 enhances planning features and integrates tools more effectively for complex task execution.

Expanded Training Data

GPT-5 uses more diverse datasets than GPT-4, improving accuracy and generalization across tasks.



VERBOSITY VS TEMPERATURE

Role of Verbosity

Verbosity controls the length and detail of AI-generated responses, balancing clarity and information depth.

Temperature Settings

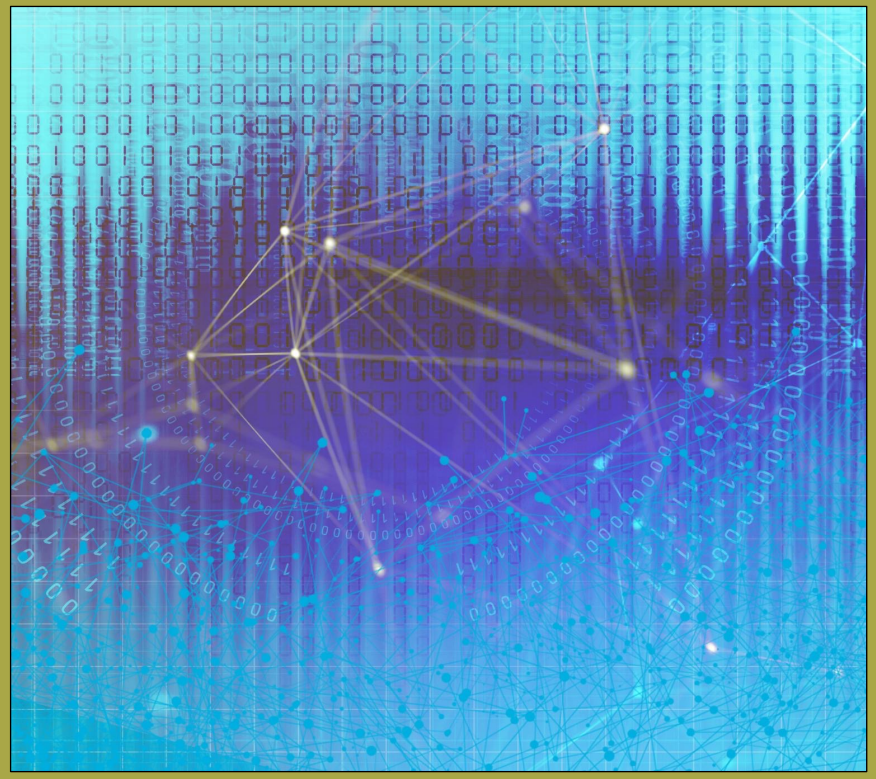
Temperature adjusts response randomness; low values yield deterministic outputs, high values increase creativity.

Advancements in GPT-5

GPT-5 introduces tunable verbosity and adaptive temperature for greater user control and context-aware output.

HANDS-ON PRACTICE

GPT-5 PROMPT EXAMPLES



Exploring GPT-5 Capabilities

Participants engage in exercises showcasing GPT-5's ability to summarize, generate itineraries, and create creative content.

Prompt Engineering Techniques

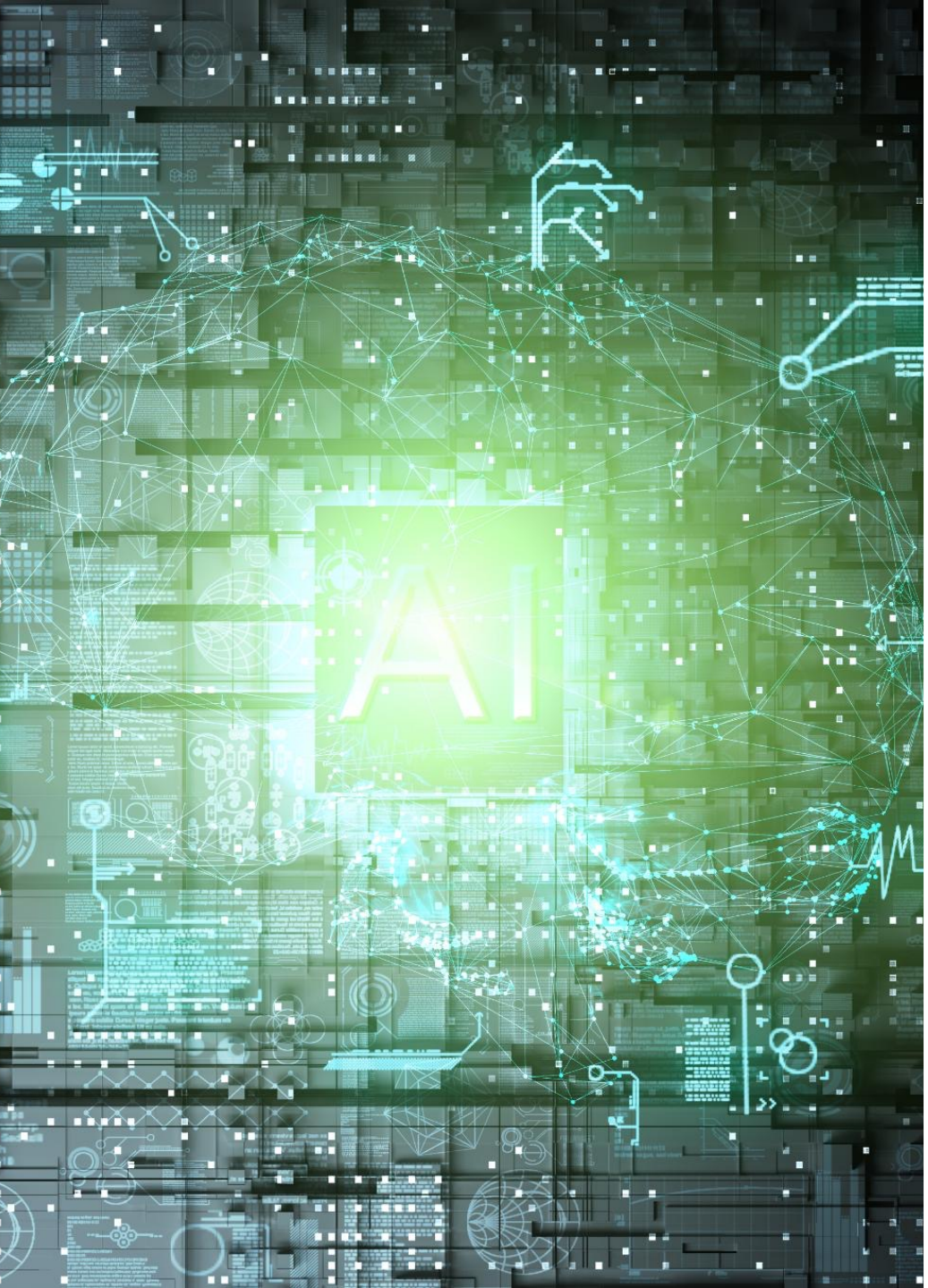
Learn to craft effective prompts using system messages and control parameters like temperature and verbosity.

Leveraging Few-Shot Examples

Incorporate few-shot examples to guide model responses for enhanced context and accuracy.

Building User Confidence

Session aims to empower users to utilize GPT-5 effectively for diverse real-world tasks.



PROMPT ENGINEERING PRIMER

Core Elements of a Prompt

Prompts consist of role instructions, contextual examples, and constraints to guide model outputs effectively.

Key Prompt Variables

Variables like temperature, top-p, maximum tokens, and frequency penalty control output creativity and specificity.

Prompt-Response Loop

The prompt-response loop highlights how input prompts directly influence generated outputs in AI models.

ACTIVITY 1: REAL-WORLD APPLICATIONS

Group Brainstorming

Participants work in teams to brainstorm real-world Generative AI use cases relevant to their projects.

Mapping Ideas to Patterns

Trainer categorizes ideas into broader AI application patterns like data summarization and automation.

Collaborative Learning

The activity fosters collaboration and critical thinking on integrating Generative AI into workflows.

ACTIVITY 2: HANDS-ON WITH GPT-5

Prompt Writing and Testing

Write and test prompts for tasks like summarization, question answering, and creative writing.

Temperature Parameter Variation

Varying the temperature between 0.2 and 0.8 helps explore different creativity levels and output styles.

Learning by Doing

Hands-on activity bridges theory and practice, enhancing participants' prompt engineering skills.

KEY OUTCOMES

Understanding Generative AI

Participants will be able to clearly explain the concept and applications of Generative AI.

GPT Model Evolution

Comprehend the progression from GPT-1 through GPT-5, highlighting key improvements and differences.

Prompt Crafting Skills

Gain the ability to craft and test effective prompts for AI applications.

Enterprise Application Identification

Identify relevant enterprise uses of Generative AI within iLink Digital.

TRANSFORMER ARCHITECTURE & ATTENTION MECHANISM

Self-Attention and its role in LLMs

Encoder-only, decoder-only, and encoder-decoder variations

Use cases and model examples for each type

WHAT IS A TRANSFORMER IN LLM?

A Deep learning model architecture

introduced by Vaswani et al.

in the paper

“Attention Is All You Need” (2017)

WHAT IS A TRANSFORMER IN LLM?

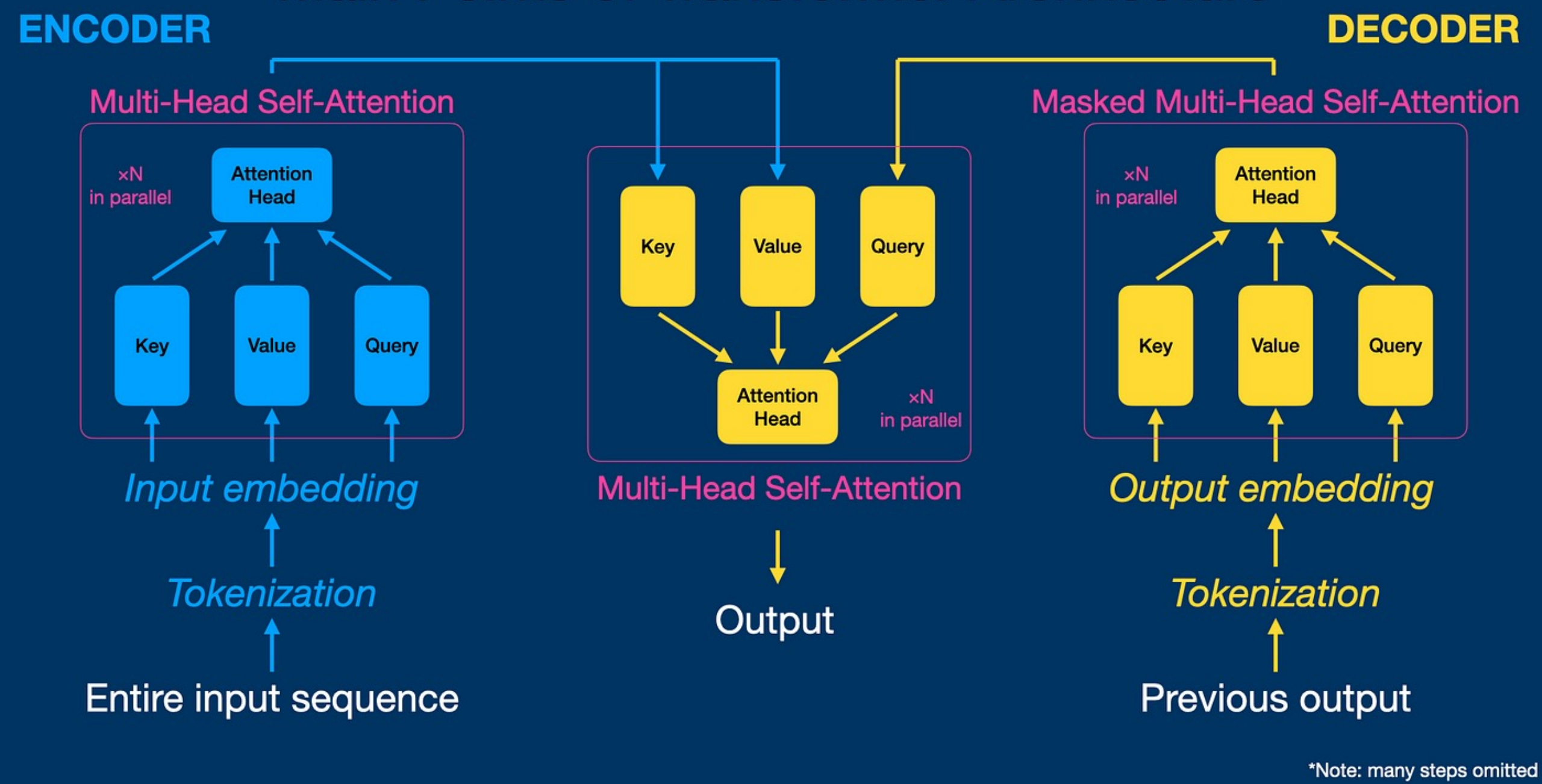
Foundation behind LLMs

like GPT (OpenAI),

BERT (Google), and

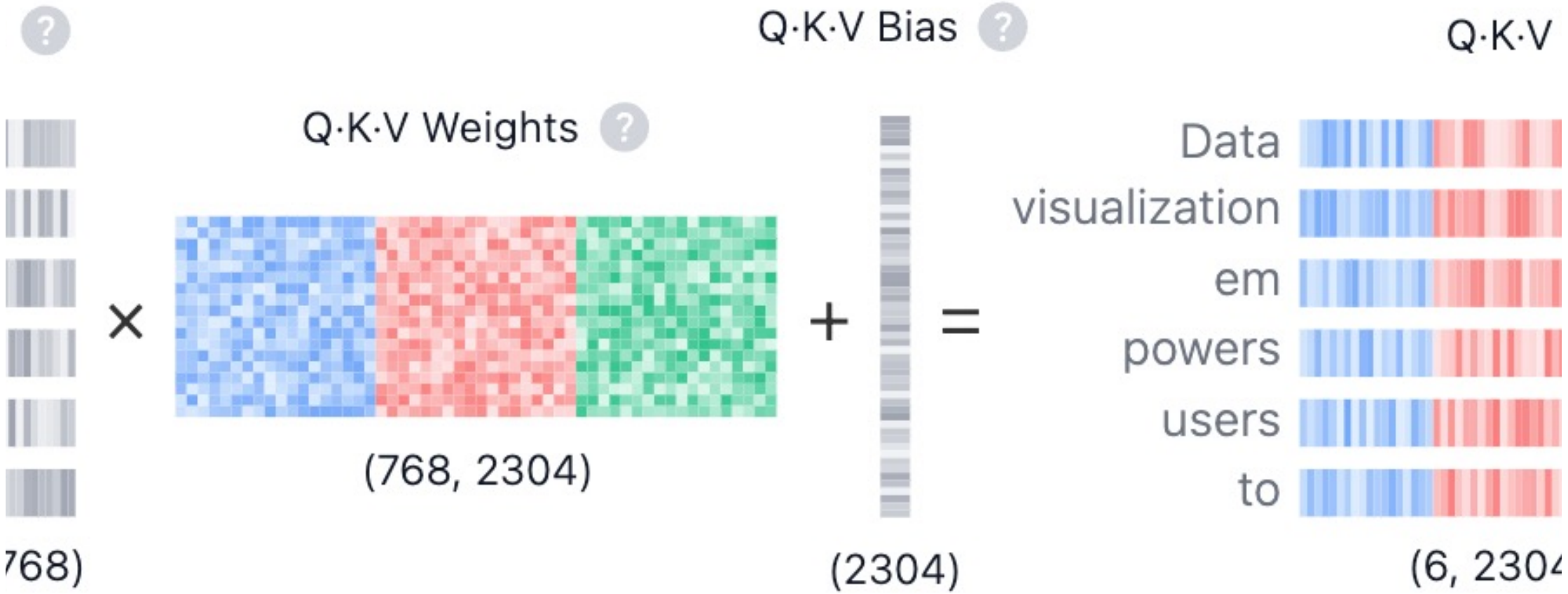
Gemini (Google DeepMind).

Main Points of Transformer Architecture



*Note: many steps omitted

QUERY, KEY, AND VALUE MATRICES



SELF-ATTENTION: THE CORE MECHANISM

WHAT IT DOES?

EACH WORD EXAMINES ALL OTHER WORDS

IN A SENTENCE TO DECIDE WHAT MATTERS MOST.

WORDS BECOME **QUERIES**, **KEYS**, AND **VALUES**.

TRANSFORMER VARIANTS

a) Encoder-Only

Structure: Stack of self-attention layers (like BERT)

Best for:

Understanding tasks—classification,

search embedding, document summarization.

TRANSFORMER VARIANTS

b) Decoder-Only

Structure:

Masked self-attention + feed-forward layers (like GPT)

Best for:

Generating text, chatbots, email copy.

TRANSFORMER VARIANTS

c) Encoder-Decoder

Structure:

Encoder processes input →

Decoder uses both self- and cross-attention

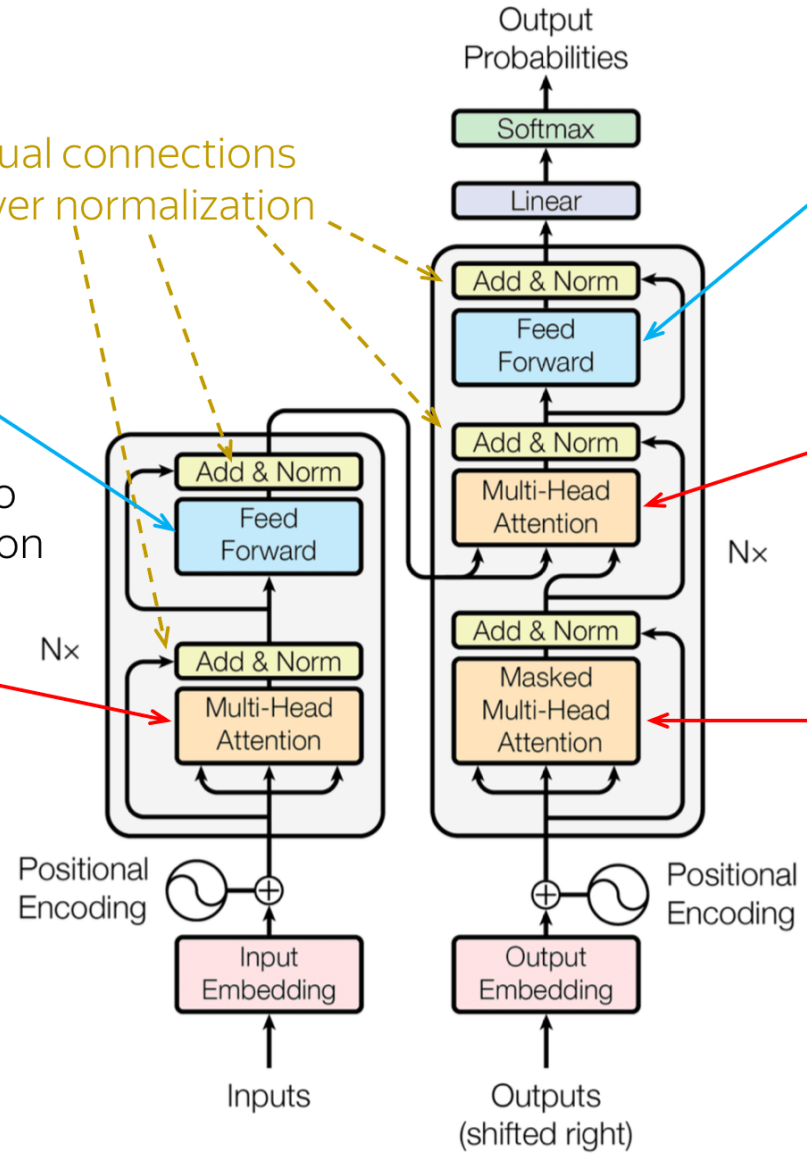
(original “Attention is All You Need”)

Residual connections and layer normalization

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states



Feed-forward network: after taking information from other tokens, take a moment to think and process this information

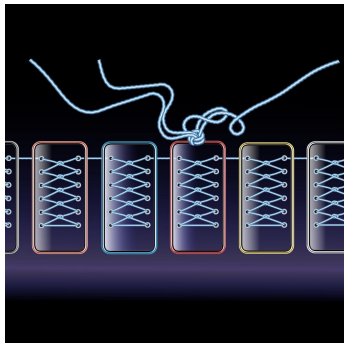
Decoder-encoder attention: target token looks at the source

queries – from decoder states; keys and values from encoder states

Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states

ENCODER VS DECODER BASICS

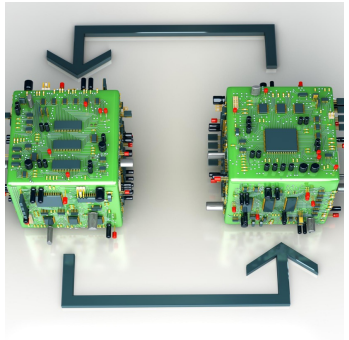


Encoder Functionality

The Encoder converts input text into contextual embeddings using token embedding, multi-head attention, and feed-forward layers.

Decoder Functionality

The Decoder generates output text token by token using masked self-attention, cross-attention, and feed-forward layers.



Encoder-Decoder Interaction

Combined encoder-decoder architectures excel at translation and summarization by capturing context and predicting next tokens.

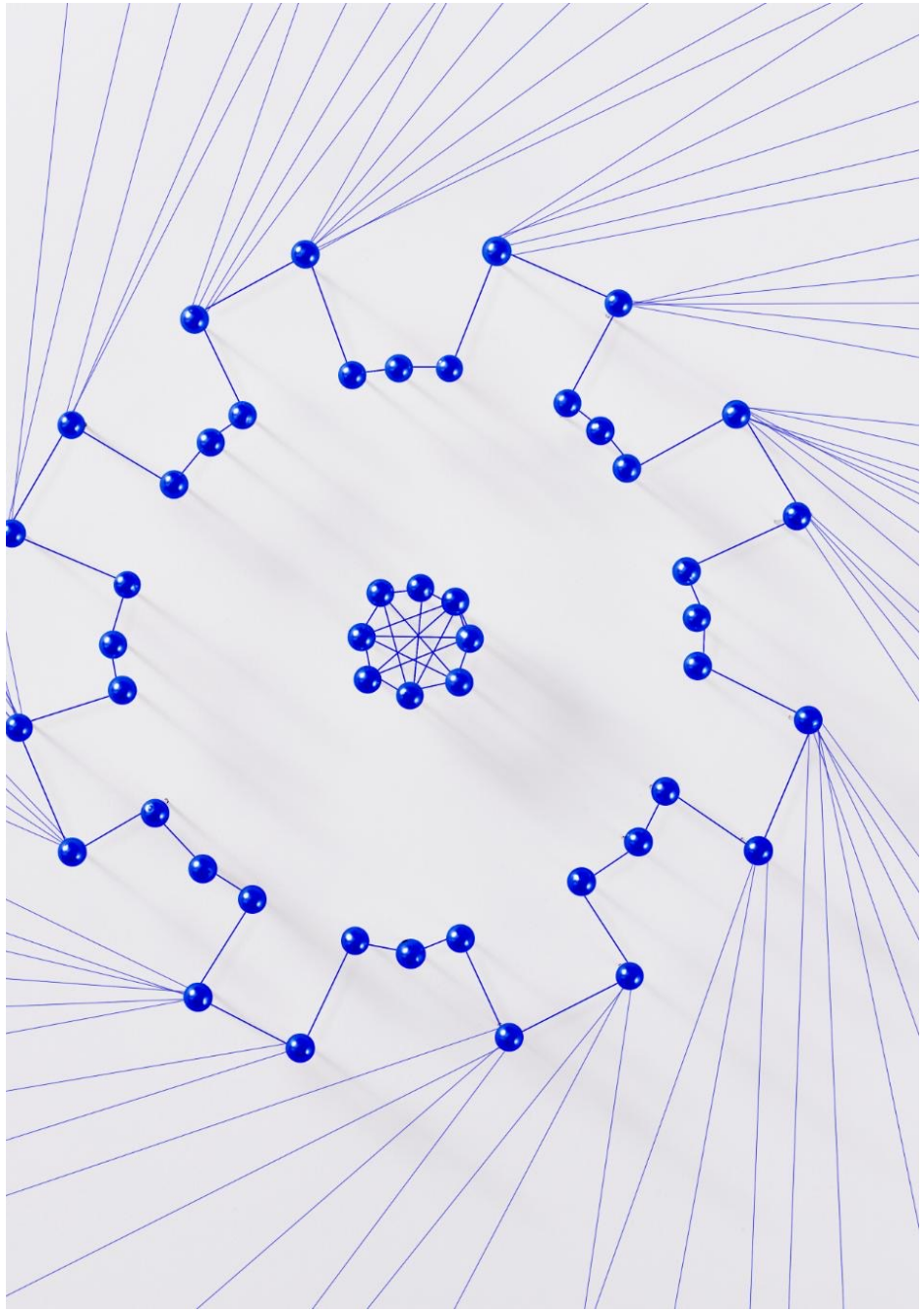
SELF-ATTENTION: THE CORE MECHANISM

WHAT IT DOES?

EACH WORD EXAMINES ALL OTHER WORDS

IN A SENTENCE TO DECIDE WHAT MATTERS MOST.

WORDS BECOME **QUERIES**, **KEYS**, AND **VALUES**.



INSIDE THE ATTENTION MECHANISM

Core Concept of Attention

Attention mechanisms enable words in a sentence to focus on relevant context, capturing relationships effectively.

Mathematical Formulation

Attention is computed using query, key, and value matrices with a softmax function to weigh importance.

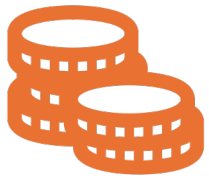
Types of Attention

Self-attention relates tokens within the same sequence, cross-attention connects encoder to decoder, while multi-head attention captures diverse relationships.

Visualizing Token Interactions

Animated visuals illustrate token-to-token interactions, making attention mechanisms easier to understand.

ENCODER SELF-ATTENTION (MULTI-HEAD ATTENTION)



Each word (token)

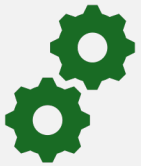
attends to all other words

in the sentence.

Helps understand context.

ENCODER (LEFT SIDE)

Takes the input sentence and



processes it into

an abstract representation

INPUT EMBEDDING + POSITIONAL ENCODING

Words are converted into

vectors (Input Embedding).

Positional information is added

so the model knows the order of words.

ENCODER SELF-ATTENTION EXAMPLE

in “he saw a bat,”

“bat” can mean an animal or

a sports item depending

on the context.

WHY IS IT CALLED MULTI-HEAD ATTENTION?

Multi-head means

We split the attention into

Multiple parallel “heads”

Each head learns to focus on

different types of relationships or patterns.

ADD & NORM

Output of attention is added back

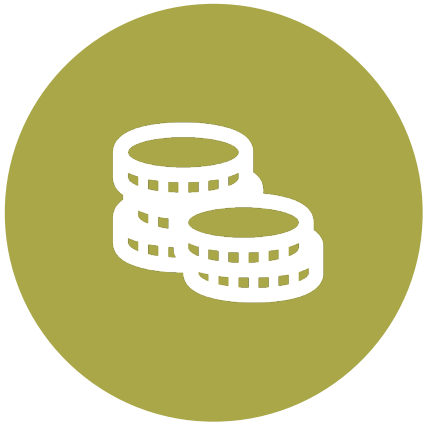
to the input (residual connection).

Layer normalization is applied

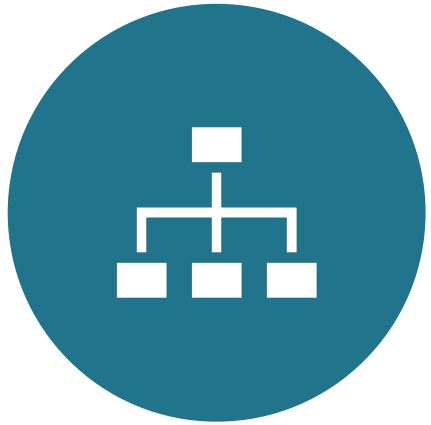
to stabilize training.

FEED-FORWARD NETWORK (FFN)

A SMALL NEURAL NETWORK APPLIED



INDEPENDENTLY TO EACH TOKEN.

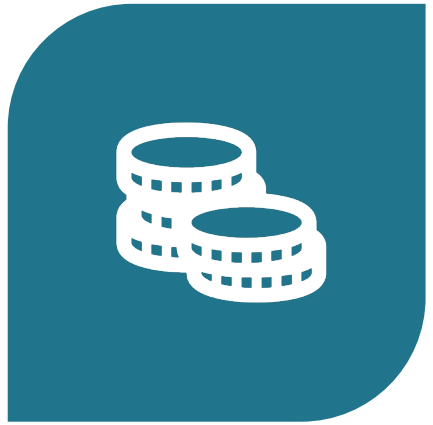


PROCESSES THE INFORMATION FURTHER.

DECODER (RIGHT SIDE)

GENERATES THE OUTPUT

LIKE A TRANSLATED SENTENCE OR



NEXT TOKEN IN LLM.

OUTPUT EMBEDDING + POSITIONAL ENCODING

Previous output tokens are embedded

shifted right for training.

Positional encoding is added.

MASKED MULTI-HEAD SELF-ATTENTION

Like encoder attention,

but **masked** to prevent future tokens

from being seen

Add & Norm

DECODER-ENCODER ATTENTION

The decoder token now **attends to encoder outputs.**

This is where the decoder “looks”

at the encoded input sentence

to decide what to generate.

Add & Norm

FEED-FORWARD NETWORK

Further processes the information

(same as encoder’s FFN).

Add & Norm

Also repeated N times.

FINAL STEP: OUTPUT PROBABILITIES

A Linear layer + Softmax converts

the final decoder output

to probabilities over vocabulary.

FINAL STEP: OUTPUT PROBABILITIES

The token with

the highest probability

is selected as

output (e.g., next word).

ATTENTION FLAVORS

Self-Attention:

Words attend to words

within the same sentence

(question or policy).

ATTENTION FLAVORS

Masked Self-Attention:

Decoder-only models only

look at **prior words**

keeps output coherent in auto-generation

ATTENTION FLAVORS

Cross-Attention:

Decoder aligns each output token

with the encoded input

from the encoder

ENCODER AND DECODER BASICS

FEATURE	ENCODER	DECODER
Primary Function	Processes input text	Generates output text
Usage	BERT, T5 (input side)	GPT, T5 (output side)
Directionality	Bidirectional	Autoregressive
Context	Understands input context	Uses encoder context

ATTENTION MECHANISM AND VISUALIZATION

ATTENTION TYPE	DESCRIPTION	APPLICATION
Self-Attention	Each token attends to all others in the same sequence	Contextual embedding generation
Cross-Attention	Decoder attends to encoder outputs	Sequence-to-sequence tasks
Multi-Head Attention	Multiple attention layers run in parallel	Enhanced feature extraction

TRANSFORMER ARCHITECTURE VISUAL AND EVOLUTION

MODEL	YEAR	PARAMETERS	KEY FEATURES
GPT-1	2018	117M	Basic transformer decoder
GPT-2	2019	1.5B	Improved text generation
GPT-3	2020	175B	Few-shot learning
GPT-4	2023	~1T (estimated)	Multimodal capabilities
GPT-5	2025	~10T (estimated)	Advanced reasoning and memory

WHY TRANSFORMERS POWER LLMS

Scalability and Parallel Processing

Transformers enable scalable training by processing entire sequences simultaneously, enhancing efficiency over traditional models.

Capturing Long-Range Dependencies

Transformers effectively capture contextual relationships between distant words for coherent and contextually accurate outputs.

Versatility in NLP Tasks

Transformers power diverse NLP applications including translation, summarization, question answering, and conversational AI.

HANDS-ON ACTIVITIES AND Q&A

Visualizing Attention Flow

Participants use diagrams and interactive tools to understand the flow of attention in transformer models.

Encoder and Decoder Comparison

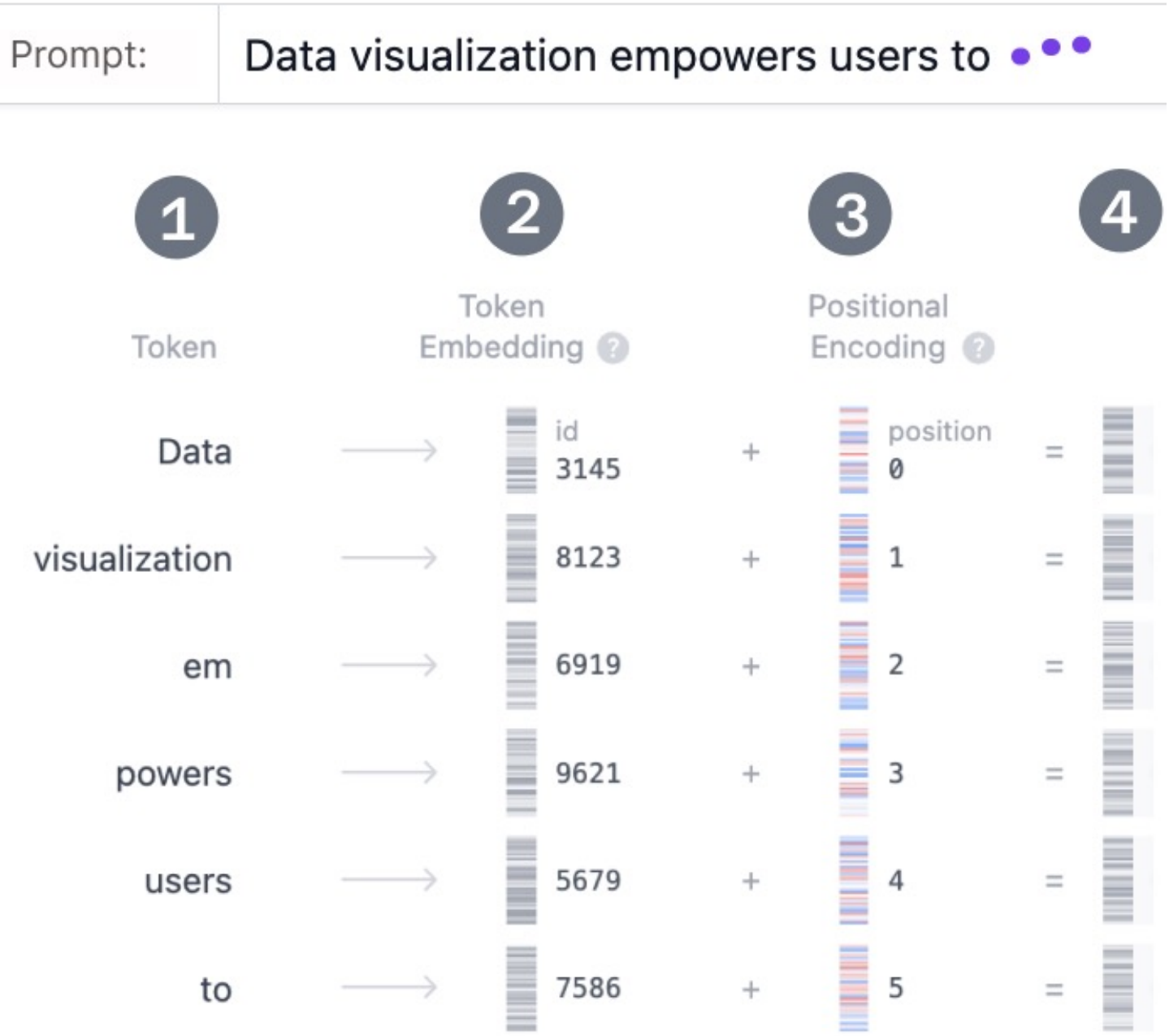
Activities include comparing the functions and roles of encoder and decoder components in transformers.

Exploring GPT Architecture

Participants explore the structure and workings of GPT models to deepen understanding of transformer applications.

Interactive Q&A Session

The session ends with a Q&A where participants clarify concepts and discuss real-world transformer applications.



VISUAL ACTIVITY – ATTENTION FLOW

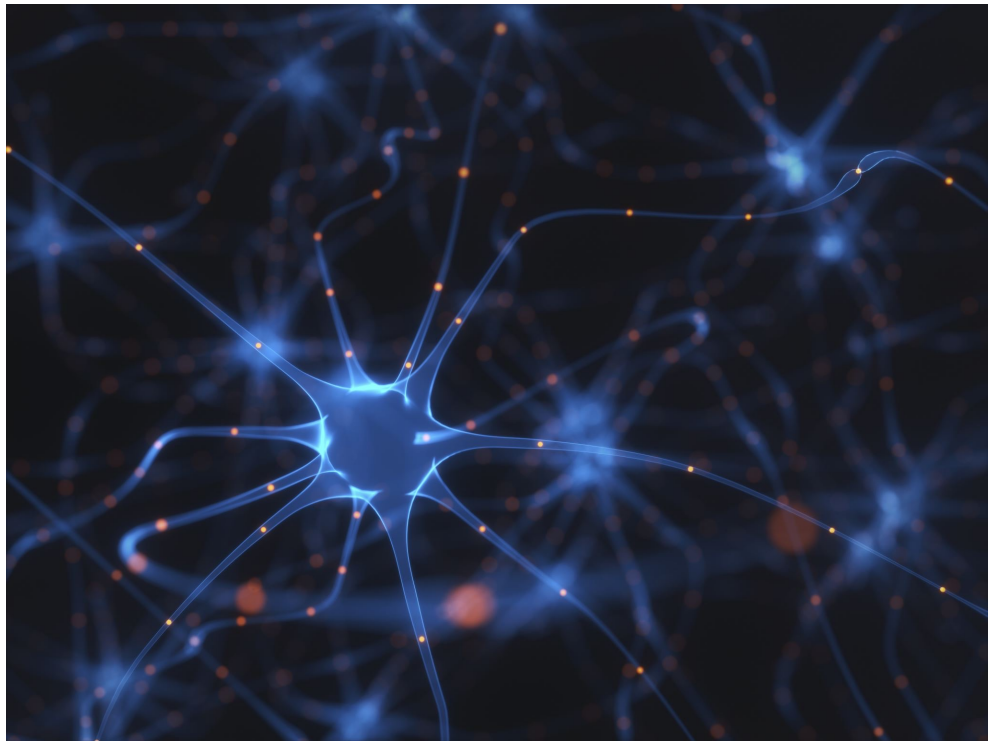
Objective: Visualize how tokens interact via attention heads.

https://youtu.be/ECR4oAwocjs?si=OMym_5OrPue2vHrk

Tools: Use interactive tool

https://poloclub.github.io/transformer-explainer/)

SESSION OUTCOMES



Understanding Transformer Architecture

Participants will grasp the structure of transformers including encoders and decoders roles.

Mechanics of Attention

Session covers how attention mechanisms work and their importance in transformers.

Evolution of Transformer Models

Overview of progression from early models like GPT-1 to advanced GPT-5.

Preparation for Advanced Topics

Foundational knowledge supports learning advanced generative AI and agentic systems.

WRAP-UP & ACKNOWLEDGMENT