**1. Core AI and ML Concepts**

* **Artificial Intelligence (AI):** Refers to building systems that mimic human intelligence—used across Google services like Google Assistant, Translate, and Search.
* **Machine Learning (ML):** A subset of AI where systems learn from data. Google Cloud AI Platform and TensorFlow are key tools here.
* **Generative AI:** Focuses on creating new content (text, images, etc.). Google’s Bard and Imagen are examples.
* **Deep Learning:** Uses neural networks with many layers—central to Google’s speech recognition, image classification, and LLMs.
* **Foundation Models:** Large-scale models trained on vast data, such as Google’s PaLM (Pathways Language Model).
* **Large Language Models (LLMs):** Specialized in understanding and generating human language—used in Google Search, Docs (smart compose), and Bard.

**2. Data and Its Role in AI**

* **Structured Data:** Organized data (e.g., spreadsheets, SQL databases)—used in Google BigQuery and Cloud SQL.
* **Unstructured Data:** Includes text, images, audio—processed using Google Cloud Vision, Natural Language API, and Speech-to-Text.
* **Data Quality:** Emphasized for accurate model training.
* **Data Accessibility:** Ensures data is usable and in the right format—Google Cloud Storage and DataPrep help with this.

**3. ML Lifecycle (Aligned with Google Cloud AI Platform)**

* **Data Ingestion & Preparation:** Google Cloud Dataflow, Dataprep, and BigQuery support this phase.
* **Model Training:** Google Vertex AI and TensorFlow are used to train models.
* **Model Deployment:** Vertex AI and AI Platform Prediction allow models to be served at scale.
* **Model Management:** Tools like Vertex AI Model Monitoring help track model performance over time.

**4. Responsible AI (Aligned with Google’s AI Principles)**

* **Secure AI:** Protects models from misuse—Google’s Secure AI Framework (SAIF) is referenced.
* **Ethical AI:** Ensures fairness, transparency, and accountability—Google’s AI Principles guide this.

**5. Learning Approaches in ML**

* **Supervised Learning:** Used in Google Photos (face recognition), Gmail (spam detection).
* **Unsupervised Learning:** Powers Google News clustering and YouTube recommendations.
* **Reinforcement Learning:** Applied in robotics and game-playing AI (e.g., DeepMind’s AlphaGo).

**Google Cloud Platform: Vertex AI**

**Vertex AI** is Google Cloud’s unified ML platform that supports the full ML lifecycle:

* **Model Garden**: A repository of pre-trained models from Google, third parties, and open-source communities.
* **Model Builder**:
  + **Custom Models**: Train models from scratch using ML frameworks.
  + **AutoML**: Build models with minimal technical expertise.

Vertex AI simplifies:

* Model training
* Deployment
* Management of ML and Gen AI solutions

**Infrastructure**

The foundational layer includes:

* **Hardware**: Servers, GPUs, TPUs
* **Software**: Tools to train, store, and run AI models

**AI on the Edge**

Running AI closer to the data source or user:

* **Lite Runtime (LiteRT)**: Google’s tool for deploying AI models on edge devices.
* **Gemini Nano**: A compact, efficient AI model optimized for on-device performance.

**Gemini Tooling for Personal Productivity**

**a. Gemini App**

* A generative AI chatbot designed to assist with:
  + Writing
  + Planning
  + Learning
  + General productivity tasks

**b. Gemini for Google Workspace**

* Integrates generative AI into familiar Google Workspace apps:
  + **Gmail**: Compose emails using AI assistance.
  + **Slides**: Generate images directly within presentations.
  + **Meet**: Summarize meeting notes automatically.

**c. Gemini for Google Cloud**

* Acts as an AI assistant for cloud developers and administrators:
  + Write and debug code
  + Manage and optimize cloud applications
  + Analyze data in **BigQuery**
  + Strengthen security posture

**d. NotebookLM**

* A research assistant tool that:
  + Allows file uploads
  + Summarizes key points
  + Answers questions
  + Generates ideas
  + Maintains grounding in the uploaded source material

**Streamlining Prompting Workflows**

Techniques to enhance productivity and consistency when interacting with AI:

* **Reusing Prompts**: Save prompts as templates for repeated use.
* **Prompt Chaining**: Maintain context by continuing conversations within the same chatbot session.
* **Saved Info in Gemini**: Store specific information for consistent reuse by the model.
* **Gems**: Personalized AI assistants within Gemini that:
  + Follow specific instructions
  + Streamline workflows
  + Use templates and guided interactions

**Prompting Techniques**

Different strategies to guide the AI’s behavior and output:

* **Zero-shot**: No examples provided; the model infers the task from the prompt alone.
* **One-shot**: One example is given to guide the model.
* **Few-shot**: Multiple examples are provided to improve accuracy.
* **Role Assignment**: Assign a persona to the model to influence tone, style, and focus.
* **Prompt Chaining**: Engage in iterative dialogue to refine responses.

**Model Guidance and Refinement**

Techniques to ensure the AI’s output is accurate and trustworthy:

* **Grounding**: Ensures AI responses are connected to verifiable sources.
* **RAG (Retrieval-Augmented Generation)**:
  1. **Retrieving**: The model pulls relevant information from a large knowledge base.
  2. **Generating**: It then uses this information to produce a grounded and accurate response.

| **Technique** | **When to Use It** |
| --- | --- |
| **Zero-shot** | When no examples are needed. E.g., “Summarize this article.” |
| **One-shot** | When one example helps clarify the task. |
| **Few-shot** | When multiple examples improve accuracy. |
| **Role Prompting** | When tone or perspective matters. E.g., “As a project manager, write a status update.” |
| **Prompt Chaining** | When refining or continuing a conversation with the AI. |

**Gemini Ecosystem**

| **Concept** | **Practical Example** |
| --- | --- |
| **Gemini App** | You’re writing a project proposal. You ask Gemini: “Draft a proposal for a new quality assurance tool for our engineering team.” |
| **Gemini for Google Workspace** | In Gmail, you type “Summarize this customer complaint and draft a polite response.” Gemini generates a reply. In Slides, you say “Create an image of a QA workflow” and it inserts a visual. |
| **Gemini for Google Cloud** | You’re deploying a new microservice. You ask Gemini: “Generate a Cloud Run deployment YAML for a Node.js app with 512MB memory.” |
| **NotebookLM** | You upload a test plan document and ask: “Summarize the key test cases and identify any missing coverage areas.” |

| **Technique** | **Description** | **Example** |
| --- | --- | --- |
| **Zero-shot** | No examples | “Summarize this article.” |
| **One-shot** | One example | “Here’s one test case. Write another.” |
| **Few-shot** | Multiple examples | “Here are 3 bug reports. Write a 4th.” |
| **Role Prompting** | Assign a persona | “As a senior engineer, write a review.” |
| **Prompt Chaining** | Iterative refinement | “Generate a report → Add summary → Add recommendations.” |

**Customer Engagement Suite (Built on Google CCaaS)**

This suite is designed to enhance customer interaction and support, leveraging cloud-native tools:

* **Conversational Agents**: AI-powered chatbots that interact with customers using natural language.
* **Agent Assist**: Provides real-time support to human agents by suggesting responses and surfacing relevant information.
* **Conversational Insights**: Analyzes customer interactions to extract actionable insights and trends.

These tools are part of Google’s **Contact Center as a Service (CCaaS)** platform, which is enterprise-grade and cloud-native.

**Vertex AI Search**

A powerful tool for implementing search and recommendation systems within your business. It enables:

* Semantic search capabilities.
* Personalized recommendations.
* Integration with internal data sources.

**Agentspace**

A framework to embed intelligent agents into internal dashboards or websites. These agents can:

* Access and understand data from various internal sources.
* Provide conversational interfaces for internal users.

**RAG: Retrieval-Augmented Generation**

A key architecture for enhancing LLM responses with external data:

1. **Retrieval**: Pulls relevant information from external sources.
2. **Augmentation**: Injects retrieved data into the prompt.
3. **Generation**: LLM generates a response based on the enriched prompt.
4. **Iteration (optional)**: The process can repeat to refine the response.

**Types of Agents**

* **Deterministic Agents**: Follow predefined rules and paths (traditional approach).
* **Generative Agents**: Use LLMs and natural language to create dynamic, conversational experiences.

**Reasoning Loop & Prompt Engineering Techniques**

These techniques enhance the reasoning capabilities of LLMs:

* **ReAct (Reason and Act)**: Combines reasoning with actions to solve user queries.
* **CoT (Chain-of-Thought)**: Breaks down complex problems into intermediate reasoning steps.
* **Metaprompting**: Uses prompts to generate or modify other prompts, guiding the model’s behavior.

**Tooling for Gen AI Agents**

* **Extensions**: Connect agents to external APIs.
* **Functions**: Define specific tasks the agent can perform.
* **Data Stores**: Provide structured access to information.
* **Plugins**: Add new capabilities or integrations to the agent.

**Types of Agents**

| **Type** | **Description** | **Use Case** |
| --- | --- | --- |
| **Deterministic** | Rule-based, predefined logic | IVR systems, form-based bots |
| **Generative** | LLM-powered, natural language | AI chatbots, virtual assistants |

**Reasoning Loop & Prompt Engineering Techniques**

| **Technique** | **Description** |
| --- | --- |
| **ReAct (Reason and Act)** | Combines reasoning with actions to solve queries |
| **CoT (Chain-of-Thought)** | Breaks down problems into intermediate reasoning steps |
| **Metaprompting** | Prompts that generate or modify other prompts |

Gemini

Trained on a massive dataset of text, images, code, audio, video, and more. This multimodal training allows Gemini to perform tasks across various domains, including language understanding, image generation, code generation, and more.

Chirp

Trained on a large dataset of audio in various languages. Chirp is designed for speech recognition and can be used for tasks like voice assistants, transcription, and translation. Find information at the right time.

Imagen

Trained primarily on a massive dataset of images and text descriptions. This enables Imagen to generate high-quality images from text descriptions, perform image editing tasks, and understand the content of images.

Google introduces Gemini, a multimodal AI model capable of understanding and integrating various information types. The following year, Google expands the Gemini ecosystem with Gemini 1.5, bringing it to more products and launching Gemini Advanced, which provides access to Google's most capable AI models.

**2024: Gemma**

Google announces Gemma, a family of lightweight state-of-the-art open models built from the same research and technology used to create the Gemini models.

Gen AI has been built into tooling across Google’s comprehensive ecosystem to support your organization in their day-to-day work. Here are some examples:

* **Google Search:** Leverage the power of Gemini in Google Search to find information faster and with more accuracy.
* **Gemini for Google Workspace:** Draft emails in Gmail, generate presentations in Slides, summarize meeting notes in Docs, and automate tasks in Sheets with integrated gen AI features.
* **Gemini App:** Access the power of Gemini directly through a dedicated app for personalized assistance and creative exploration.
* **Gemini for Google Cloud:** Build applications and services on Google Cloud using powerful Gemini models.