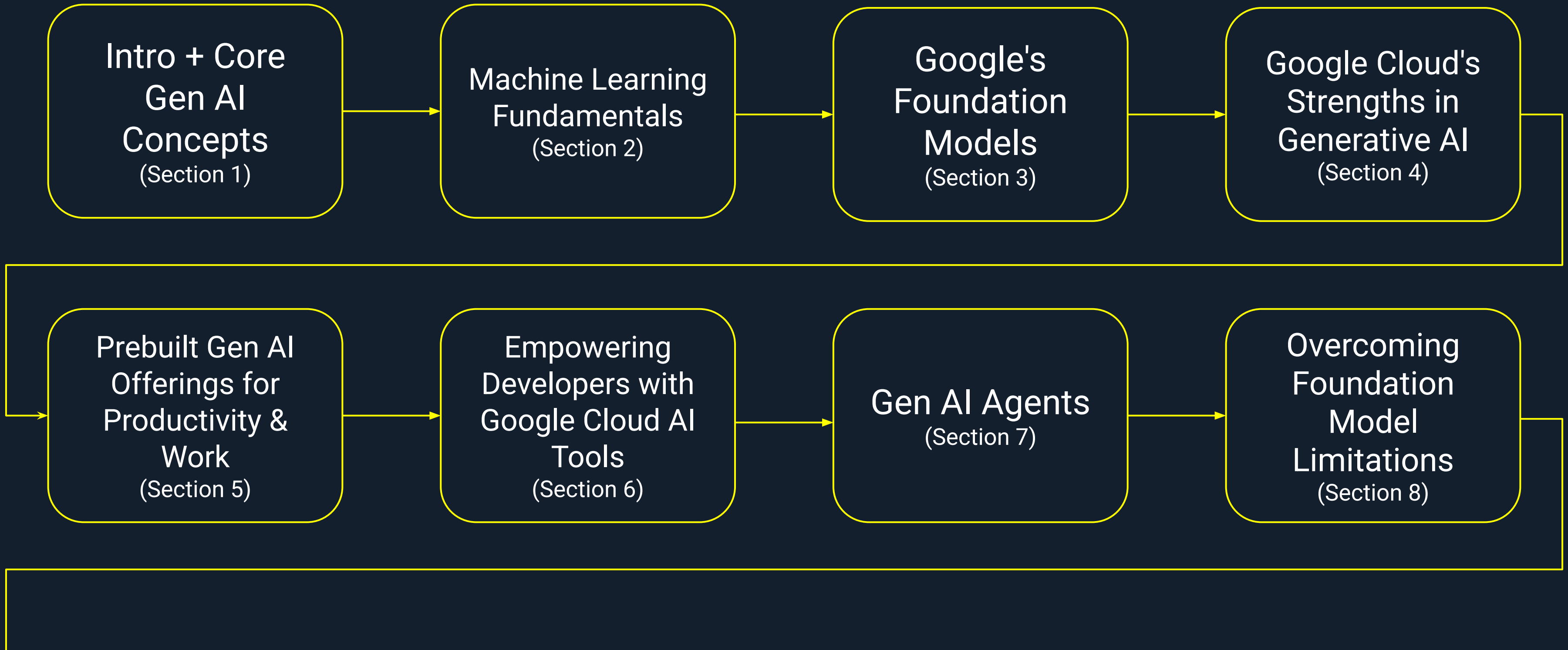
Become Google Cloud Generative AI Leader Certified

The Roadmap To Success by Vladimir Raykov



Intro + Core Gen AI Concepts (Section 1)

Machine Learning Fundamentals (Section 2)

Google's Foundation Models (Section 3)

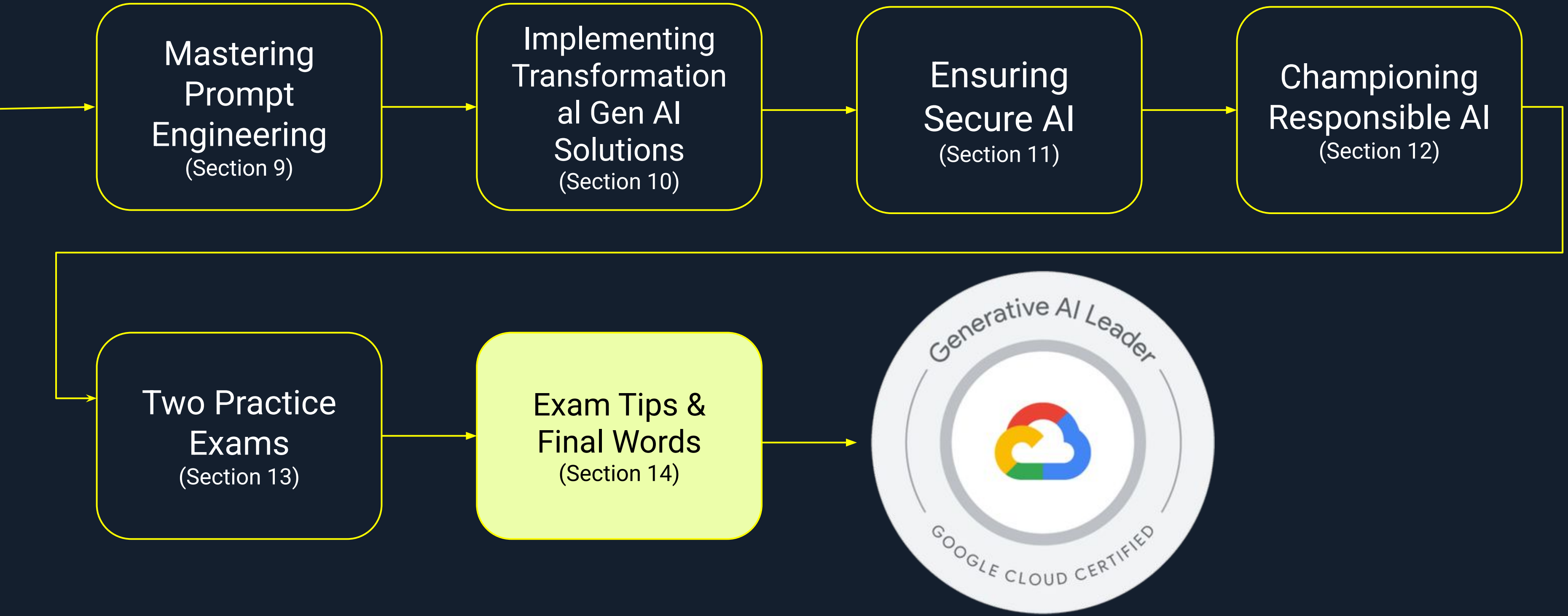
Google Cloud's Strengths in Generative AI (Section 4)

Prebuilt Gen AI Offerings for Productivity & Work (Section 5)

Empowering Developers with Google Cloud AI Tools (Section 6)

Gen AI Agents (Section 7)

Overcoming Foundation Model Limitations (Section 8)



Mastering Prompt Engineering (Section 9)

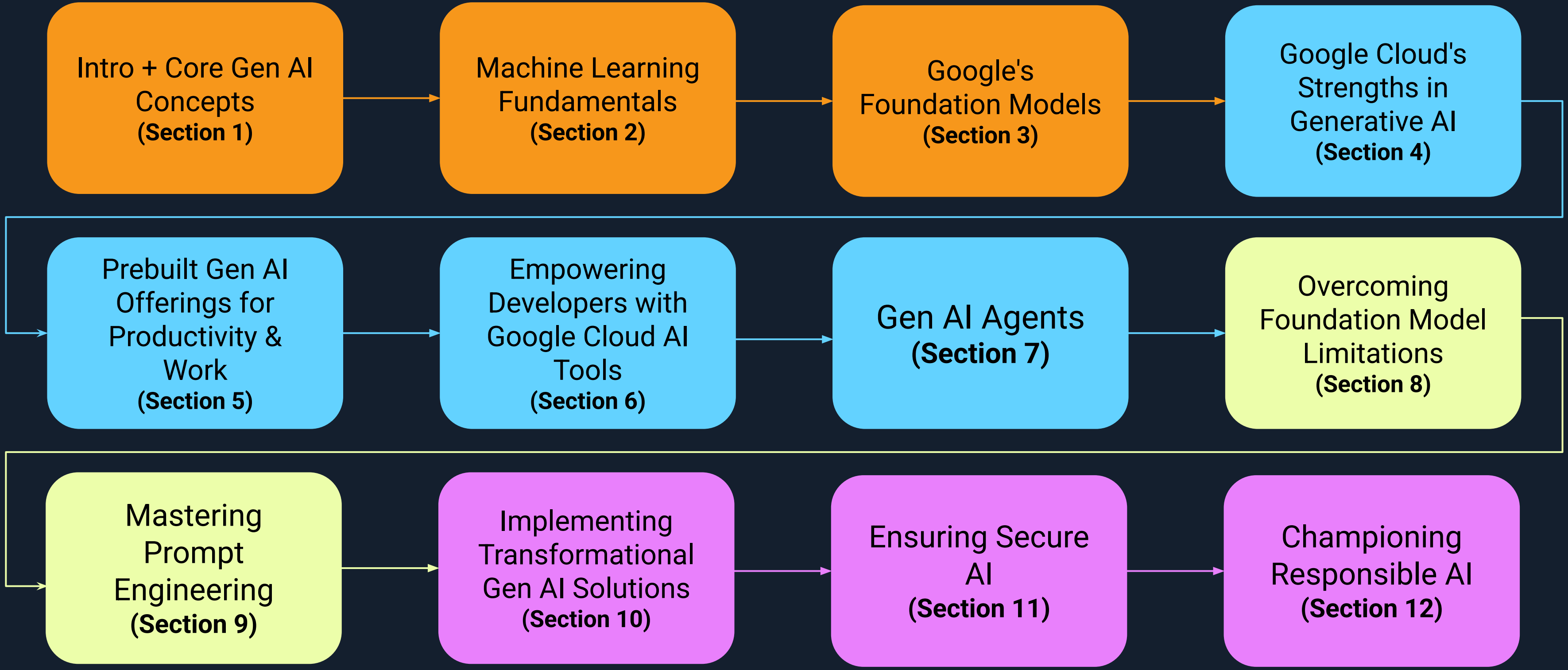
Implementing Transformational Gen AI Solutions (Section 10)

Ensuring Secure AI (Section 11)

Championing Responsible AI (Section 12)

Two Practice Exams (Section 13)

Exam Tips & Final Words (Section 14)



Intro + Core Gen AI Concepts **(Section 1)**

Machine Learning Fundamentals **(Section 2)**

Google's Foundation Models **(Section 3)**

Google Cloud's Strengths in Generative AI **(Section 4)**

Prebuilt Gen AI Offerings for Productivity & Work **(Section 5)**

Empowering Developers with Google Cloud AI Tools **(Section 6)**

Gen AI Agents **(Section 7)**

Overcoming Foundation Model Limitations **(Section 8)**

Mastering Prompt Engineering **(Section 9)**

Implementing Transformational Gen AI Solutions **(Section 10)**

Ensuring Secure AI **(Section 11)**

Championing Responsible AI **(Section 12)**

Domain 1: Fundamentals of gen AI
Domain 2: Google Cloud’s gen AI offerings
Domain 3: Techniques to improve gen AI model output
Domain 4: Business strategies for a successful gen AI solution

SECTION 1

AI vs ML vs Deep Learning Overview - Part 1

1. **AI is technology that mimics human intelligence.**
2. AI is the umbrella term that includes **Machine Learning** and **Deep Learning**.
3. AI learns and improves through massive amounts of training data (particularly true for Deep Learning Applications) - far more examples than humans need to learn similar tasks.
4. AI solves complex problems, with varying degrees of success, across many industries (e.g., healthcare, finance, retail, manufacturing, education, entertainment, transportation, energy, and more)
5. AI excels at pattern recognition and processing large amounts of data.
6. AI operates on **probability-based decision making**, providing confidence levels rather than absolute answers
7. Responsible AI development requires **transparency**, **fairness**, and **human oversight** to prevent misuse.

AI vs ML vs Deep Learning Overview - Part 2

1. **Machine Learning (ML)** is the subset of AI focused specifically on learning from data.
    a. It transforms traditional programming by **learning patterns** rather than following explicit rules.
2. The **three main types** are:
    a. Supervised Learning
    b. Unsupervised Learning
    c. Reinforcement Learning
3. **Data quality and quantity** are even more crucial in ML than in general AI applications.
4. ML models improve continuously through exposure to more data.
5. The field faces unique challenges in **computation**, **algorithm selection**, and **explainability**.

AI vs ML vs Deep Learning Overview - Part 3

1. **Deep Learning (DL)** is a specialized subset of ML that uses **Neural Networks** with multiple layers.
    a. It excels at **processing unstructured data** like images, text, and sound.
2. Deep Learning can automatically discover important features in data **without** human guidance.
    a. It requires massive amounts of data and computational power to train effectively.
3. The technology powers many modern AI applications, including content generation and natural language processing.
4. Deep Learning systems can be more complex and harder to interpret than traditional Machine Learning models.
    a. Its architecture is inspired by the human brain's neural networks.

Computer Vision & Natural Language Processing

1. **Computer Vision** enables machines to interpret visual data, using deep learning models (that use CNNs architecture) for tasks such as **image classification** and **object detection**.
2. **Natural Language Processing (NLP)** allows computers to understand and generate human language, leveraging architectures like RNNs and Transformers for tasks like **translation** and **sentiment analysis**.
3. **Deep Learning** *is the driving force behind the advancements in both Computer Vision and NLP, enabling machines to learn from raw data and perform complex tasks.*

Foundation Models (FMs) - Overview

1. **Definition:** *FMs are large-scale, **pre-trained models** adaptable to a wide range of tasks through fine-tuning. They learn general patterns from massive datasets.*
2. **Key Architectures:** *Common architectures include Transformers (especially for language), CNNs (for images), RNNs (for sequential data), and GNNs (for graph data).*
   a. ***Transformers (designed by Google)** are crucial for many modern FMs.*
3. **Training Process:** FMs are pre-trained using **self-supervised learning** on **large amounts of unlabeled data**. This allows them to learn general representations, which are then adapted through **fine-tuning with smaller labeled datasets for specific tasks.**
4. **Multimodality:** *Modern FMs are increasingly multimodal, processing and generating information across modalities like text, images, and code (e.g., Google's Gemini family, Amazon Nova).*
5. **Key Capabilities:** *FMs excel at language processing, visual comprehension, code generation, human-centered engagement (chatbots), and speech-to-text.*
6. **Prompt Engineering:** *Effective prompt engineering is crucial for eliciting desired outputs from FMs.*

Large Language Models (LLMs) - Overview

1. **What are LLMs?** Large Language Models (LLMs) are specialized Foundation Models designed to **understand and generate human language**.
2. **How do LLMs work?**
   a. **Tokenization:** *LLMs break text into smaller units called **tokens**, like words or parts of words, to process language more effectively.*
   b. **Transformer Architecture:** *Powered by **self-attention**, transformers excel at understanding word relationships, keeping context over long text, and processing information quickly.*
   c. **Pre-training:** *LLMs learn language patterns and context by processing massive datasets using **self-supervised learning**.*
3. **Why are LLMs important?** *They have transformed technology by enabling advanced applications like:*
   a. **Content Creation:** *Writing, summarizing, and analyzing text.*
   b. **Code Generation:** *Tools like GitHub Copilot assist developers with coding tasks.*
   c. **Language Translation:** *Accurate and context-aware translations.*
   d. **Customer Service:** *Chatbots and virtual assistants for natural, human-like interactions.*

Multimodal Models

1. **Multimodal Models** can **process** and **generate** **multiple types of data simultaneously**, such as text, images, audio, and video.
2. They learn the relationships between **different modalities**, allowing them to combine and understand data in a more holistic way.
3. They differ from **LLMs (unimodal models)**, which are limited to **text-based input and output**.
4. They represent a major step forward in AI, enabling more human-like understanding and interaction with the world.

Prompt Engineering vs Prompt tuning

1. **Prompt Engineering:** *The art of crafting effective text prompts to guide model outputs – accessible to anyone and requires no technical training or additional costs.*
2. **Prompt Tuning:** *A technical method that trains additional "soft prompt" parameters attached to the input, while keeping the base model's core weights unchanged.*
3. **Key Difference:** *Engineering optimizes your text inputs; tuning trains specialized parameters for consistent, domain-specific performance.*
4. **Google Cloud Support:** *Vertex AI provides tools for both approaches – prompt engineering experimentation and prompt tuning capabilities for deeper customization needs.*

Diffusion Models (E.g. Imagen by Google DeepMind)

1. **Diffusion Models** start with random noise and gradually refine it into meaningful outputs, like images, text, or audio.
2. They operate in two main steps:
   a. **Forward Diffusion:** Gradually adds noise to structured data until it becomes pure noise.
   b. **Reverse Diffusion:** Gradually removes noise from random data to generate a coherent output.
3. Their ability to learn patterns through noise makes them **incredibly powerful and versatile**.
4. **Important:** Diffusion models **cannot** interpret image content.

SECTION 2

The Machine Learning Process

1. The **ML Process** consists of three main steps:
    a. **Training Data** → **ML Algorithm** → **Model**

Data Types in Machine Learning

1. The ML process starts with **collecting** and **processing** **training data**.
2. Data **quality** and **preparation** are crucial for model success.
3. **Data Categories by Labels:**
    a. **Labeled data**: Comes with predefined tags/labels (used in Supervised Learning, where the labels guide the learning process)
    b. **Unlabeled data**: No predefined labels (used in Unsupervised Learning to discover hidden structures or relationships within the data). The absence of labels requires different learning algorithms to find patterns.
4. **Data Categories by Structure:**
    a. **Structured data**: Organized in tabular formats (like SQL databases, spreadsheets). This organization makes it easy to query and analyze the data.
        i. **Time-series data**: Data points collected at successive points in time, used for analyzing trends and patterns over time.
5. **Unstructured data**: No predefined format (like social media posts, images, text, videos). This type of data often requires specialized techniques for processing and analysis.

Learning Types - Supervised Learning

1. **Supervised Learning** involves training algorithms on **labeled data** to predict outcomes for new data.
   a. **Classification** assigns input data to predefined categories (labels).
   b. **Regression** predicts continuous values (numbers).
2. Labeled data is crucial for Supervised Learning, providing the necessary information for models to learn.
3. Supervised Learning is one of the **three main categories** of ML, alongside **Unsupervised Learning** and **Reinforcement Learning**.

Learning Types - Unsupervised Learning

1. **Unsupervised learning** works with **unlabeled data** to discover **hidden patterns** and **structures**.
    a. **Clustering** groups similar data points together based on their characteristics (e.g., customer segmentation).
    b. **Dimensionality Reduction** simplifies complex data by reducing the number of features while retaining essential information (e.g., simplifying user preferences).
    c. **Anomaly Detection** identifies unusual data points or patterns that deviate from the norm (e.g., fraud detection).
    d. **Density Estimation** analyzes the distribution of data points to identify areas of high and low concentration (e.g., location planning).
2. A key difference from Supervised Learning is that Unsupervised Learning works **without** predefined labels or "correct answers."

Learning Types - Reinforcement Learning & RLHF

1. **Reinforcement Learning [RL] is about learning through interaction and feedback**.
2. Key components include **agent**, **environment**, **state**, **action**, and **reward**.
3. RLHF incorporates **human preferences** into the learning process, specifically:
    a. Collects human feedback on model outputs.
    b. Trains a reward model based on human preferences.
    c. Fine-tunes the model to align with these preferences.
    d. Helps reduce undesirable behaviors.
    e. Particularly valuable for tasks where success is hard to define mathematically.
4. [RL] Common applications include:
    a. Self-driving cars
    b. Game AI
    c. Robotics
5. [RLHF] Important considerations:
    a. Human feedback can be valuable but also **expensive** and **subjective**.
    b. RLHF can be used for fine-tuning after self-supervised learning.
    c. The field combines elements of behavioral psychology and machine learning.

Machine Learning Lifecycle

1. The **Machine Learning Lifecycle** guides projects from data to a functioning model.
2. **Key Stages:**
   1. **Data Ingestion & Preparation**: *Collecting, cleaning, and transforming data. Crucial for model success.*
   2. **Model Training**: *Selecting an algorithm, training it on data, tuning, and evaluating.*
   3. **Model Deployment**: *Making the trained model available for use in production.*
   4. **Model Management**: *Continuously tracking performance, versioning, and retraining.*
3. **Iterative Nature:** *The process often involves looping back between stages.*
4. **Google Cloud Support:** *Google Cloud, particularly through **Vertex AI**, offers tools and services to support each stage of this lifecycle, enabling efficient and scalable ML operations.*

Data Quality and Accessibility

1. **Data Quality is Paramount:** *Poor data leads to poor model performance*
2. **Key Characteristics of Quality Data:**
    a. **Completeness**: No critical missing values.
    b. **Consistency/Accuracy**: Data is correct and free of contradictions.
    c. **Relevance**: Data is appropriate for the task.
    d. **Timeliness**: Data is sufficiently up-to-date.
3. **Data Accessibility is Essential:** *Data must be easily obtainable and usable by those who need it.*
4. **Key Aspects of Accessibility:**
    a. **Availability**: Data can be accessed when needed.
    b. **Usability/Format**: Data is in a usable form.
    c. **Cost**: Access costs are manageable.
    d. **Security** & **Governance**: Access is secure and compliant.
5. **Leader's Role:** *Champion data quality, advocate for accessible data infrastructure, and understand that both require continuous effort.*

SECTION 3

The Generative AI Landscape

1. **There are five core layers of the Generative AI Landscape**
2. **Infrastructure**: The foundational computing resources (e.g. GPUs, TPUs, storage).
    a. *Business Implication: Cost, scalability, access.*
3. **Models**: The AI algorithms that generate content.
    a. *Business Implication: Core capabilities, development effort, customization.*
4. **Platforms**: Tools for building, training, and deploying models.
    a. *Business Implication: Efficiency, democratization, governance.*
5. **Agents**: AI systems that perceive, reason, and act.
    a. *Business Implication: Automation, enhanced user experience, personalization.*
6. **Applications**: End-user software delivering Gen AI capabilities.
    a. *Business Implication: Value realization, user experience, market differentiation.*

Gemini: Capabilities and Use Cases

1. **Gemini is a family of natively multimodal AI models** that work across text, code, images, audio, and video seamlessly.
2. You have three main options:
   a. **Pro** for versatile high-end performance
   b. **Flash** for speed and efficiency
   c. **Nano** for edge deployment.
3. The core strengths are *sophisticated reasoning, long-context understanding, high-quality generation, and flexible performance options.*
4. Your job as a leader is to *match these capabilities to your specific business requirements*.
5. Google continues to evolve these models, s**o stay tuned for new developments** and enhanced capabilities.

Gemma: Capabilities and Use Cases

1. **Gemma offers lightweight, open AI models** based on Google's advanced research, extending beyond text to specialized areas like coding and vision.
2. **The key advantages are** *open weights for flexibility, responsible AI principles, excellent performance-to-size ratios, and strong developer support.*
3. **You have specialized variants emerging:** ***CodeGemma*** *for programming tasks and* ***PaliGemma*** *for vision-language applications,* ***ShieldGemma*** *for content moderation with more likely coming.*
4. **The core strengths include** *solid language fundamentals, domain-specific expertise in specialized variants, high adaptability through fine-tuning, and deployment flexibility.*
5. **Strategically, Gemma excels in** *rapid prototyping, controlled deployments, cost-effective custom solutions, and building internal AI capabilities.*

Imagen: Capabilities and Use Cases

1. **Advanced Text-to-Image Models:** ***Imagen*** *specializes in generating high-fidelity images from text, utilizing the diffusion techniques we discussed earlier.*
2. **Key Strengths:** *Produces high-quality, realistic, and artistic images based on a deep understanding of complex text prompts. It's integrated with Google Cloud for scalable use.*
3. **Core Capabilities:** *Includes text-to-image generation, image editing, style application, and creating image variations.*
4. **Significant Business Use Cases:** *Transforms marketing, product design, media content creation, e-commerce visuals, and educational materials.*
5. **Imagen** *offers strategic advantages by accelerating visual content creation, reducing costs, and enabling new forms of creative expression.*

Veo: Capabilities and Use Cases

1. **Advanced Text-to-Video Model:** ***Veo*** *specializes in generating high-definition video clips from text prompts, and can also incorporate image or video inputs.*
2. **Key Strengths:** *Aims for high-quality, coherent video with nuanced understanding of prompts and visual consistency.*
3. **Core Capabilities:** *Text-to-video, image-to-video, video-to-video generation/editing, stylistic control, and audio generation capabilities.*
4. **Exciting Business Use Cases:** *Revolutionizing marketing content, media production, educational materials, product visualization, and rapid video prototyping.*
5. **Veo** *offers the potential to dramatically accelerate and customize video production, enabling richer storytelling and engagement.*

SECTION 4

Google's AI-First Vision and Commitment to Innovation

1. **AI-First Philosophy:** Google rethinks its entire business with AI at the core.
2. **Decades of Pioneering Research:** Their deep history, including breakthroughs like the Transformer, underpins their approach.
3. **Commitment to Innovation:** This includes translating research into products, building foundational infrastructure (like TPUs), and fostering an open ecosystem.
4. **Focus on Responsible AI:** A core tenet for trust and broad benefit.
5. **Strategic Advantage for You:** Provides access to cutting-edge tech and a proactive partner in AI's future.

The Google Cloud Enterprise-Ready AI Platform

1. **Google Cloud's Enterprise-Ready AI Platform is...**
   a. **Responsible**: *Built with principles and tools for fairness, ethics, explainability, and human oversight.*
   b. **Secure**: *Leverages Google's secure-by-design infrastructure to protect your data and AI models.*
   c. **Private**: *Provides you with control over your data, supporting privacy and data governance.*
   d. **Reliable**: *Engineered for high availability and consistent performance for business continuity.*
   e. **Scalable**: *Designed to seamlessly scale your AI initiatives as your business needs grow.*

Powering Gen AI: Google Cloud's AI-Optimized Infrastructure

1. G**oogle Cloud's AI-Optimized Infrastructure** *provides essential hardware for demanding Gen AI workloads, including…*
    a. **GPUs (Graphics Processing Units):** *Powerful parallel processors crucial for deep learning.*
    b. **TPUs (Tensor Processing Units):** *Google's custom-designed AI accelerators for high performance and efficiency.*
    c. **AI Hypercomputer:** *An integrated supercomputing architecture combining compute, networking, and software for extreme-scale AI.*
    d. **Delivered via Cloud:** *Accessible through Google's global data centers, offering on-demand power and scalability.*

Data Control & Democratizing AI on Google Cloud

1. **Google Cloud's AI Platform enables...**
2. **Data Control by:**
   a. *Ensuring robust security for your data and models.*
   b. *Providing strong privacy commitments, keeping your data yours.*
   c. *Offering governance features to manage data per your policies.*
   d. *Aiming for transparency in data usage.*
3. **Democratizing AI through:**
   a. *Pre-trained Models: Offering powerful, ready-to-use AI capabilities.*
   b. *APIs: Allowing easy integration of AI into applications.*
   c. *Low-code/No-code Tools: Empowering a broader range of users to build AI solutions.*

SECTION 5

Gemini App and Google AI Subscription Plans

1. **Gemini App:** *Provides accessible, chat-based interaction with advanced AI models, now featuring real-time camera integration and superior image generation capabilities.*
2. **Premium Subscription Plans:**
    a. **Google AI Pro** *offers enhanced features for everyday users.*
    b. **Google AI Ultra** *provides maximum capabilities for power users and early adopters.*
3. **Advanced Capabilities:** *Include customizable Gems, native audio output, video generation, enhanced reasoning modes, and seamless integration across Google's ecosystem.*
4. **Professional Value:** *While designed for individual use, these tools can significantly enhance team productivity, creativity, and AI literacy when adopted by organization members.*

Google Agentspace: Features and Applications

1. **Agentspace as Enterprise AI Hub:** *A comprehensive platform that serves as the central source of enterprise truth, integrating all organizational data sources with advanced AI capabilities.*
2. **Multi-Tier Architecture:** *Available in Enterprise and Enterprise Plus configurations to meet varying organizational needs and security requirements.*
3. **Integrated Workflow Experience:** *Seamless Chrome integration brings AI assistance directly into employees' natural work patterns.*
4. **Comprehensive Agent Ecosystem:** *Supports Google-built expert agents, partner-developed solutions, and custom internal agents with sophisticated communication capabilities.*
5. **Advanced Automation:** *Enables complex, multi-step workflows through tool integration and Agent2Agent communication protocols.*
6. **Flexible Deployment:** *Available as cloud SaaS or on-premises solutions to meet diverse organizational requirements.*
7. **Strategic Business Impact:** *Transforms information access, enhances decision-making, and enables intelligent task automation across the enterprise.*

Gemini for Google Workspace

1. **Gemini for Google Workspace** embeds **Gemini AI capabilities** directly into familiar productivity apps, offering…

    a. ***AI assistance*** *across Gmail, Docs, Sheets, Slides, Meet, and Drive for tasks like drafting, summarizing, brainstorming, data organization, and presentation creation.*

    b. ***Significant business value*** *through* ***increased productivity****,* ***enhanced creativity****,* ***improved collaboration****, and easy adoption by employees.*

Google Cloud's External Search Offerings (Vertex AI Search, Google Search)

1. **Google Cloud's External Search Offerings** *enhance how external users find information.*
2. **Google Search (in Enterprise Context):** *Its vast knowledge base can be used for **grounding** Generative AI models to provide factual, up-to-date answers based on public information.*
3. **Vertex AI Search:** *Enables businesses to create custom, AI-powered search engines **for their external audiences** (e.g., on websites, in apps) using their own company data.*
4. **Key Features of Vertex AI Search:** *Indexing company content, AI-driven relevance, generative AI summaries/answers, multimodal capabilities, and customization.*
5. **Business Benefits of Vertex AI Search:** *Improved customer experience, higher conversion rates, reduced support costs, and enhanced brand perception.*

Google's Customer Engagement Suite

1. **Google's Customer Engagement Suite** *is a collection of AI-powered solutions for enhanced customer interactions, featuring…*
   a. **Conversational Agents:** *AI virtual agents providing 24/7 support and handling routine inquiries.*
   b. **Agent Assist:** *Real-time AI support for human agents, boosting their productivity.*
   c. **Conversational Insights:** *AI-driven analytics for valuable understanding from customer interactions.*
2. **Google Cloud CCaaS:** *is the foundational platform for these solutions.*
3. **Business Value:** *Improved customer satisfaction, increased operational efficiency, and data-driven strategic insights.*

SECTION 6

Vertex AI Platform: Unified ML Platform

1. **The Google Cloud's unified ML platform is called Vertex AI.**
2. **An Integrated Environment:** *Bringing together tools for the entire machine learning lifecycle.*
3. **End-to-End MLOps:** *Supporting data management, training, deployment, monitoring, and pipeline automation.*
4. **Support for All Skill Levels:** *With tools for custom coding and no-code/low-code options like AutoML.*
5. **Access to Foundation Models:** *Through features like the Model Garden and tools for fine-tuning and deployment.*
6. **Key Business Value:** *Increased developer productivity, faster time-to-market, improved model quality, democratization of AI, and cost efficiency.*

What Are RAG And Grounding

1. **RAG** and **Grounding** are vital techniques for enhancing LLM performance…
2. **Grounding:** *<u>The principle of connecting an AI model's responses to verifiable sources of information</u> (like Google Search, as seen in tools like Google AI Studio, or private data) to improve accuracy.*
3. **Retrieval-Augmented Generation (RAG):** *An architectural pattern implementing grounding with custom knowledge bases. It involves:*
   a. ***Retrieving** relevant information.*
   b. ***Augmenting** the query with this information.*
   c. *Having the LLM **generate a response** based on this context.*
4. **Benefits of RAG:** *Leads to more accurate, up-to-date, contextually relevant, and often verifiable answers from LLMs.*
5. *RAG allows leveraging existing knowledge bases, offering an efficient way to build specialized AI solutions.*
6.

Vertex AI Search (as a Developer Tool)

1. Within the Vertex AI platform, **Vertex AI Search** serves as…
   1. **A Key Enabler for RAG:** *Providing the powerful "Retrieval" mechanism needed for Retrieval-Augmented Generation applications.*
   2. **A Tool for Unstructured Data:** *Capable of indexing and making searchable diverse enterprise content.*
   3. **An API-Driven Service:** *Allowing deep integration into custom AI applications and workflows.*
   4. **A Managed Service:** *Simplifying development by handling underlying search infrastructure.*

AutoML on Vertex AI

1. **AutoML** on Vertex A a key feature for democratizing AI development, offering…
   a. **Automated Machine Learning:** *Automates many complex steps in building custom ML models.*
   b. **Low-Code/No-Code Approach:** *Enables users with not much development experience, such as subject matter experts, to create models.*
   c. **How it Works:** *Users provide labeled data, and AutoML handles much of the preprocessing, model selection, and tuning.*
   d. **Key Benefits:** *Democratizes AI, increases development speed and efficiency, can produce high-quality models, and allows teams to focus on business problems.*

SECTION 7

Vertex AI Agent Builder: Creating Custom AI Agents

1. **Vertex AI Agent Builder** is Google Cloud's platform for creating custom Generative AI agents, offering…
   a. **A Unified, Low-Code Environment:** *Simplifying the development of sophisticated AI agents.*
   b. **Key Functionalities:** *Easy data connectivity, foundation model integration, defining and using tools for actions, conversation design, and straightforward testing/deployment.*
   c. **Enables Custom AI Agents:** *For applications like advanced customer service, internal helpdesks, specialized knowledge experts, and task automation.*
   d. **Business Value:** *Faster agent development, broader team empowerment, improved operational efficiency, and enhanced user experiences through tailored AI solutions.*

Key Google Cloud Services & APIs for Agent Tooling

1. **The Core Idea:** *Agents use these services and APIs as 'tools' to access data, trigger actions, and leverage specialized AI.*
2. **Key Service Categories for Tooling:**
    a. **Data Services:** *Cloud Storage, Databases (Cloud SQL, etc.) for information access.*
    b. **Compute Services:** *Cloud Functions, Cloud Run for custom logic and actions.*
    c. **Vertex AI Platform:** *Access to foundation models and other ML services.*
    d. **Pre-built AI/ML APIs:** *Speech-to-Text, Text-to-Speech, Translation, Document AI, Vision AI, Video AI, Natural Language API for specialized understanding.*
3. **Business Value:** *Enables the creation of highly capable, integrated agents that can automate complex tasks and interact intelligently with diverse information and systems.*

How Gen AI Agents Interact with the External Environment

1. **Gen AI Agents** use **various interaction mechanisms** to connect and act, including…
   a. **Functions (Function Calling):** *Allowing agents to execute pre-defined code or call external APIs to perform actions or get dynamic data.*
   b. **Extensions/Plugins:** *Pre-built or custom add-ons that provide standardized connections to external services or data sources.*
   c. **Data Stores (Knowledge Bases):** *Enabling agents to query specific databases or document repositories for grounding and accessing up-to-date or proprietary information (key for RAG).*
2. **Purpose:** *These mechanisms allow agents to access real-time data, perform actions in other systems, and personalize interactions, making them effective digital workers.*

Choosing Your Development Environment: Vertex AI Studio vs. Google AI Studio

1. **Google AI Studio:** *Best for quick prototyping, prompt experimentation, and learning with Google's latest generative models; often free and highly accessible.*
2. **Vertex AI Studio (on Vertex AI Platform):** *The choice for building production-ready, enterprise-grade AI solutions, offering advanced security, data governance, model fine-tuning, MLOps, and integration with the broader Google Cloud ecosystem.*

SECTION 8

Common Limitations of Foundation Models

1. **Foundation Model Limitations include:**
   a. **Data Dependency:** *Model performance is tied to the quality and nature of its training data.*
   b. **Knowledge Cutoff:** *Models have no information about events post their training date.*
   c. **Bias and Fairness:** *Models can learn and amplify societal biases from training data.*
   d. **Hallucinations:** *Generating plausible but factually incorrect or nonsensical outputs.*
   e. **Edge Cases:** *Difficulty handling rare, unusual, or novel input scenarios.*

Google Cloud Recommended Practices to Address Limitations

1. Google Cloud recommended practices to address **foundation model limitations** include:
2. **Grounding & RAG (Retrieval-Augmented Generation):** *Connect models to external knowledge to improve accuracy, currency, and reduce hallucinations.*
3. **Prompt Engineering:** *Craft effective inputs to guide models towards desired, accurate, and unbiased outputs.*
4. **Fine-tuning:** *Adapt pre-trained models to specific domains or tasks using custom datasets.*
5. **Human in the Loop (HITL):** *Incorporate human oversight for review, validation, and correction, especially for critical tasks and to catch errors.*
6. **Combined Approach:** *Depending on the context, these practices are often used together for the best results.*

Data Sources: Understanding Grounding in LLMs

1. **Grounding** information can come from…
    a. **First-Party Enterprise Data:** *Your organization's internal, proprietary information (e.g., wikis, databases, support logs) for context-specific AI.*
    b. **Third-Party Data:** *Licensed or acquired external datasets (e.g., industry reports, market data) for specialized knowledge.*
    c. **World Data (Public Information):** *Information from the public internet, often accessed via tools like Google Search, for current events and general knowledge.*
    d. **Context is Key:** *The choice of data source depends on the specific AI application and its information needs.*

How Retrieval-Augmented Generation (RAG) Improves Output

1. **RAG enhances LLM responses by making them…**
   a. **More Accurate and Factual:** *Basing answers on retrieved, verifiable data, reducing hallucinations.*
   b. **Up-to-Date:** *Overcoming knowledge cutoffs by accessing current information from dynamic knowledge bases.*
   c. **Contextually Relevant:** *Providing precise answers on niche or proprietary topics using specific data.*
   d. **More Transparent:** *Often enabling citations to source documents for verifiability.*
   e. **Better at Handling Complexity:** *By providing focused context for complex queries or edge cases.*

Google Cloud Grounding Offerings

1. **Google Cloud helps implement grounding and RAG techniques through…**
    a. **Pre-built RAG with Vertex AI Search:** *A streamlined way to build RAG applications by connecting Vertex AI Search (indexing your data) with foundation models.*
    b. **RAG APIs and Building Blocks:** *For more custom and flexible RAG implementations, allowing developers to control the retrieval and generation pipeline.*
    c. **Grounding with Google Search:** *An option in tools like Google AI Studio and model APIs to ground responses in real-time public web information.*
    d. **Supporting Services:** *Cloud Storage, Databases, and the broader Vertex AI platform play crucial roles in supporting RAG architectures.*

The Importance of Continuous Monitoring & Evaluation of Gen AI Models

1. *Continuous **monitoring and evaluation** are essential for the long-term success of Gen AI Models because…*
2. **Ensures Ongoing Performance:** *Tracks KPIs and detects degradation or drift.*
3. **Maintains Fairness and Manages Bias:** *Helps identify and address fairness issues that may emerge over time.*
4. **Upholds Security and Compliance:** *Includes applying updates and monitoring for misuse.*
5. **Google Cloud Support:** *Vertex AI provides tools for versioning (Model Registry), performance/drift monitoring, and feature management (like Vertex AI Feature Store).*
6. **Iterative Improvement:** *Creates a feedback loop for refining models.*

SECTION 9

The Art and Science of Prompt Engineering for LLMs

1. **Prompt engineering is a crucial skill for interacting with LLMs…**
2. **Definition:** *The practice of designing and refining input prompts to guide LLMs toward desired outputs.*
3. **Art & Science:** *Combines structured approaches with creative intuition.*
4. **Significance:** *Essential for obtaining accurate, relevant, and useful responses, controlling output style, and unlocking the LLM's full potential for specific tasks.*
5. **Impact of Ineffective Prompts:** *Can lead to vague, incorrect, or biased outputs.*

Essential Prompting Techniques

1. **Essential Prompting Techniques**
   a. **Zero-shot Prompting:** *Ask directly without examples (e.g., Summarize this article:).*
   b. **One-shot Prompting:** *Provide one example to guide the format/style (e.g., Input: happy / Output: joyful. Input: sad / Output:).*
   c. **Few-shot Prompting:** *Provide multiple examples for better context and generalization (e.g., several Q&A pairs before the actual question).*
   d. **Role Prompting:** *Instruct the LLM to adopt a specific persona for tailored tone and style (e.g., Act as a historian and explain...).*

Advanced Prompting: Prompt Chaining, Chain-of-Thought, and ReAct Prompting

1. **Advanced Prompting Techniques…**
   a. **Prompt Chaining:** *Breaking down complex tasks into a sequence of simpler, interconnected prompts.*
   b. **Chain-of-Thought (CoT) Prompting:** *Encouraging the LLM to articulate its step-by-step reasoning process before giving a final answer, improving accuracy for reasoning tasks.*
   c. **ReAct (Reason and Act) Prompting:** *Enabling a synergy between reasoning and taking actions (using tools or APIs) to gather information and solve multi-step problems.*

Temperature, Top K, Top P, Input/Output Length - 1

1. AI models generate text by assigning probabilities to possible word choices. **Inference parameters** control how the model makes these selections:
2. **Temperature (Creativity Control):**
   a. *Low (0.2): Predictable outputs (e.g., reports, summaries).*
   b. *Medium (0.7): Balanced creativity and reliability (e.g., emails, ideas).*
   c. *High (1.0): Creative, diverse outputs (e.g., creative writing).*
3. **Top P (Nucleus Sampling): Controls cumulative probability.**
   a. *Low (0.3): Uses only top options until 30% cumulative probability.*
   b. *High (0.9): Includes more diverse options.*
4. **Top K: Limits the number of options considered.**
   a. *Top K = 5: Uses only top 5 options.*
   b. *Top K = 50: More variety, controlled choices.*

Temperature, Top K, Top P, Input/Output Length - 2

5. **Context Window & Length:**
   a. *Total input and output tokens must fit within the model's context window.*
   b. *Example: With a 4,000-token window, 3,000 tokens input leaves room for 1,000 tokens output.*
6. **Common Combinations:**
   a. *Creative Writing: High Temperature (0.7–1.0), High Top P (0.9)*
   b. *Factual/Technical: Low Temperature (0.2–0.4), Low Top P (0.3)*
   c. *Balanced: Medium Temperature (0.5–0.7), Medium Top P (0.7)*

SECTION 10

Identifying Types of Gen AI Solutions for Business

1. **Generative AI can power diverse solutions, including…**
2. **Content Creation & Augmentation:** *For text, reports, and summaries.*
3. **Media Generation:** *Creating images, designs, and videos.*
4. **Code Generation & Assistance:** *Assisting software development.*
5. **Data Synthesis & Augmentation:** *For training ML models and testing.*
6. **Personalized User Experiences:** *Tailoring content and interactions.*
7. **Automation of Complex Tasks:** *Powering intelligent agents and workflows.*

Key Factors Influencing Gen AI Needs

1. Before choosing a solution, carefully consider your organization's…
2. **Business Requirements:** *The specific problems to solve or opportunities to seize.*
3. **Technical Constraints & Capabilities:** *Existing infrastructure and team skills.*
4. **Scale of Deployment:** *From small teams to enterprise-wide or customer-facing.*
5. **Customization & Specificity:** *The need for tailored vs. general-purpose AI.*
6. **Data Privacy & Security Requirements:** *Based on data sensitivity and regulations.*
7. **Latency & Performance Needs:** *Required speed and responsiveness of the solution.*

Choosing the Right Gen AI Solution: A Strategic Approach

1. Strategic Gen AI selection involves **six key elements**:
   1. **Aligning with business objectives** - *ensuring direct goal alignment*
   2. **Assessing technical feasibility -** *considering capabilities and integration needs*
   3. **Evaluating build-versus-buy options** - *choosing between off-the-shelf, platform, or custom solutions*
   4. **Ensuring data strategy alignment -** *prioritizing privacy, security, and governance*
   5. **Starting with pilot projects** - *testing and learning before full deployment*
   6. **Analyzing total cost and ROI** - *weighing all expenses against anticipated benefits*

Identifying the Steps to Integrate Gen AI into an Organization

1. **Seven key steps for Gen AI organizational integration:**
    a. **Establishing ownership and governance** - *defining roles and ethical frameworks*
    b. **Preparing data and infrastructure** - *ensuring technical readiness*
    c. **Developing the AI solution** - *building or configuring the system*
    d. **Pilot testing and iteration** - *controlled testing with refinement*
    e. **Change management and training** - *preparing employees for transformation*
    f. **Scaled deployment** - *broader rollout with process integration*
    g. **Continuous monitoring** - *ensuring long-term effectiveness*

Measuring the Impact and ROI of Gen AI Initiatives

1. <u>**Six key aspects of measuring Gen AI impact and ROI:**</u>
   a. **Defining clear, aligned metrics** - *SMART KPIs tied to business objectives*
   b. **Establishing baselines** - *measuring pre-AI performance for comparison*
   c. **Implementing tracking mechanisms** - *reliable post-deployment monitoring systems*
   d. **Analyzing quantitative and qualitative impact** - *numbers plus softer benefits*
   e. **Calculating ROI** - *comparing total benefits to total costs*
   f. **Iterating and communicating** - *continuous improvement and stakeholder updates*

SECTION 11

The Importance of Security Throughout the ML Lifecycle

1. **Security is vital at every ML lifecycle stage:**
    a. **Protecting valuable assets** - *safeguarding sensitive data and AI models*
    b. **Mitigating risks** - *preventing breaches, theft, and malicious manipulation*
    c. **Critical at each phase** - *from secure data ingestion through ongoing monitoring*
    d. **Requiring holistic approach** - *implementing defense-in-depth strategies*

Understanding Google's Secure AI Framework (SAIF)

1. SAIF is Google’s **conceptual framework** for **secure AI**, based on six core elements…
    a. **Expanding Security Foundations:** *Leveraging robust infrastructure and adapting to new AI threats.*
    b. **Extending Detection & Response:** *Monitoring AI systems and using threat intelligence.*
    c. **Automating Defenses:** *Using AI to improve security response scalability and speed.*
    d. **Harmonizing Platform Controls:** *Ensuring consistent security across AI tools and platforms.*
    e. **Adapting Controls & Feedback Loops**: *Continuously learning and evolving defenses.*
    f. **Contextualizing AI Risks in Business Processes:** *Conducting end-to-end risk assessments.*

Leveraging Google Cloud Security Tools for AI

1. **Key Google Cloud security tools for AI include…**
   a. **Identity and Access Management (IAM):** *Controls who has access to your AI resources and data, enforcing the principle of least privilege.*
   b. **Security Command Center:** *Provides centralized visibility into security posture and helps detect threats and misconfigurations.*
   c. **Workload Monitoring Tools (e.g., Cloud Monitoring, Cloud Logging):** *Enable detection of anomalies and operational oversight of AI systems.*

SECTION 12

The Imperative of Responsible AI and Transparency in Business

1. **Responsible AI** and **transparency** are **fundamental for trustworthy AI:**
   a. **Responsible AI definition** - *developing AI ethically to benefit society while minimizing harm*
   b. **Transparency importance** - *providing clarity on AI systems, data, and limitations*
2. **Business benefits** - *building trust, mitigating risks, enhancing reputation, driving innovation*
3. **Principled approach** - *essential framework for guiding AI development and deployment*

Navigating Privacy Considerations in Gen AI

1. **Protecting privacy** in Gen AI involves four key areas:
   a. **Awareness of privacy risks** - *data leakage, sensitive attribute inference, and content misuse*
   b. **Privacy-enhancing techniques** - *data minimization, anonymization, and pseudonymization*
   c. **Secure governance** - *robust data handling, access controls, and user consent*
2. **Proactive approach** - embedding privacy by design into AI development

Describing the Implications of Data Quality, Bias, and Fairness

1. **Three key data-related aspects** crucial for **responsible AI**:
    a. **Data quality** - *poor quality data leads to flawed outputs and ineffective systems*
    b. **Bias** - *training data biases get learned and amplified, causing discriminatory outcomes*
    c. **Fairness** - *ensuring AI systems don't produce unjustly discriminatory results for different groups*
2. **Business impact** - *neglecting these leads to trust loss, reputational damage, and legal issues*
3. **Leadership role** - *champion practices ensuring data integrity, bias mitigation, and fairness promotion*

Describing the Importance of Accountability and Explainability in AI Systems

1. **The Two pillars of responsible AI:**
   a. **Accountability** - establishing clear responsibility for AI development, deployment, and outcomes to enable risk management
   b. **Explainability** - the ability to understand and describe how AI models make decisions in human terms
2. **Combined importance** - *accountability ensures oversight while explainability builds trust, aids debugging, supports compliance, and enables better design*
3. **Leadership role** - *foster culture valuing accountability in governance and appropriate explainability in solutions*