LIVE AI Workshops.

Unlocking the Power of Modern AI:
Mastering LLMOps

...For Business Success

Steve Taplin
CEO of Sonatafy Technology

Antonio Tamayo
Lead AI Engineer, PHD

Brought To You By:

Sonatafy
Nearshore Software Done Right.

Event Agenda & Introductions

Event Agenda

Introduction
- What is a Generative Model?
- Why now?
- Why is hardware and infrastructure relevant?

LLM General Applications

Understanding LLMs
- How do LLMs work?

LLMs in Production
- What is LLMOps
- LLMs limitations
- How can we mitigate the limitations of LLMs?
- Challenges
- Integration process
- Approaches (Prompt Engineering, Fine-tuning)
- Data quality
- Leveraging LLMs with LLMOps

Steve Taplin
CEO of Sonatafy Technology

Lifetime Entrepreneur
Former IBM Executive
Forbes & Entrepreneur Featured Author
Top 100 Influential Tech Executives by CIO Magazine
Forbes Technology Council Member
Software Podcast Host
Entrepreneur Leadership Network
Frequent Industry Speaker

Antonio Tamayo
Lead AI Engineer, PHD

PHD in Computer Science & Machine Learning

AWS Academy Graduate
AWS Fundamentals: Addressing Security Risk
AWS Academy Cloud Foundations
AWS Fundamentals: Going Cloud-Native

Google Cloud Platform Certifications
Google Sentiment Analysis with Deep Learning using BERT
Google Cloud Platform Fundamentals: Core Infrastructure
Google Cloud Platform Big Data and Machine Learning

IBM Cognitive Class Certifications
Machine Learning with Python
Machine Learning Dimensionality Reduction
Data Analysis with Python
Deep Learning with TensorFlow
Deep Learning Fundamentals

Who Is Sonatafy Technology?

Award-Winning Services

Inc. 500 Regionals Rocky Mountain

Inc. 500 Power Partner 2023

Clutch Global Fall 2023

Clutch Champion Fall 2023

Startup of the Year
Globee Awards

< We've Bridged The Gap Between Onshore & Nearshore >

Custom Software Development

Currently Have Over 100 Active Full Time, Salaried Software Engineers and Are **Growing Rapidly**

Forbes & Entrepreneur | Inc. Regionals ROCKY MOUNTAIN

Inc. Power Partner 2023 | Top 100 Influential Tech Executive of 2023 by CIO Today

A Perfect Match

Affordable, English-Proficient and Time Zone Aligned To **Meet Your Needs**

Time Zone Alignment
Limiting Burnout Of Client Teams By Having Same Time Zone Support

Access To Talent
Access To Affordable And Skilled Resources To Meet Current Budgets & Timelines

Hiring Efficiency
Companies Not Able To Hire Fast Enough Due To Demand And Internal Processes

Company Expertise
Executive-Level Expertise And Commitment To Ensure Software Success

Staying Compliant
We Meet Strict Compliance In Industries Like Healthcare, Life Sciences & FinTech

LIVE AI Workshops.

Unlocking the Power of Modern AI: Mastering LLMOps

Accelerate Your AI Journey | FREE Assessment & MVP Models

Advanced Solutions ✨

To Common AI Challenges

Our team specializes in guiding organizations through their AI journey, addressing these challenges head-on.

We offer expertise in simplifying complex AI concepts, sourcing top-tier talent, ensuring data readiness, crafting strategic roadmaps, and defining clear success metrics.

Partnering with us means transforming your AI ambitions into achievable, impactful realities. Let us help you navigate the intricacies of AI implementation with confidence and precision, ensuring your investment translates into real-world success.

Next Steps?

Scheduling An AI Assessment

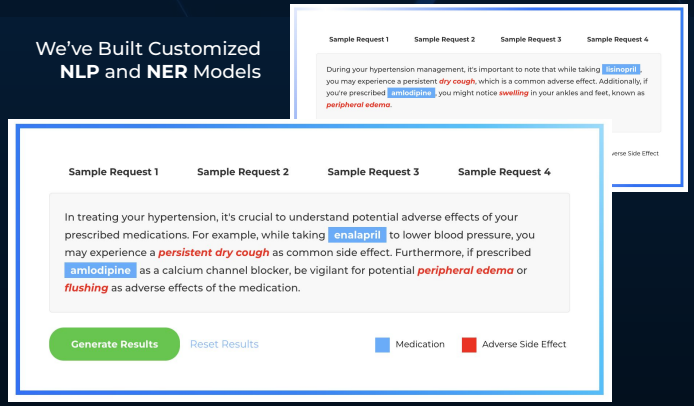
Led By Leaders of AI Engineering

Our AI service offering is specifically **designed to help clients rapidly accelerate their AI initiatives**, enabling them to stay ahead in the highly competitive tech landscape.

Whether you're starting from scratch or looking to enhance your existing AI capabilities, our expert team is here to guide you through every step of the process.

Our AI development team is led by **Dr. Antonio Tamayo**, who has a Ph.D. in Computer Science and is an AI and Data Scientist leading expert.

We've Built Customized **NLP** and **NER** Models



Our Team's Verified AI Certifications

AWS Fundamentals: Addressing Security Risk

AWS Academy Graduate - AWS Academy Cloud Foundations

AWS Fundamentals: Going Cloud-Native, Coursera

Google Sentiment Analysis with Deep Learning using BERT, Coursera

IBM Machine Learning with Python, Cognitive Class

IBM Statistics 101, Cognitive Class

IBM Machine Learning Dimensionality Reduction, Cognitive Class

IBM Data Analysis with Python, Cognitive Class

Google Cloud Platform Fundamentals: Core Infrastructure, Coursera

Google Cloud Platform Big Data and Machine Learning Fundamentals, Coursera

IBM Deep Learning with TensorFlow, Cognitive Class

IBM Deep Learning Fundamentals, Cognitive Class

Check out our AI Demo's

Sonatafy Nearshore Software Done Right.

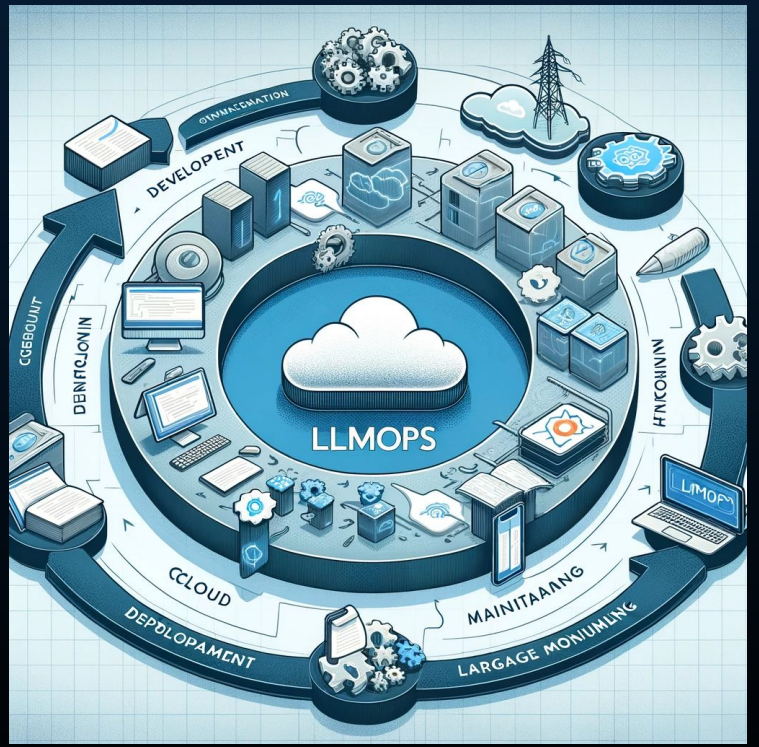
What are LLMs?

Large Language Models (LLMs) are advanced AI systems that understand and generate human language. Trained on massive text datasets, they can write, translate, summarize, and answer questions.
Examples: ChatGPT, BERT, LLaMA

What are LLMOps?

LLMOps refers to the lifecycle management of Large Language Models

- **Development /Training of the Model**
  - Identify specific goals and preparing dataset
- **Deploying the Model**
  - In the appropriate infrastructure
- **Monitoring the Model**
  - Continuous monitoring is needed
- **Maintenance of the Model**
  - Adjusting model to maintain accuracy and relevance



New Client Offers in 2024

Free One-On-One
AI Workshops With Clients

- **Free** AI Proof Of Concepts
- **Free** AI Roadmap
- **Free** In-Depth Code Reviews

Monthly
AI Webinars

300+
Tech Leaders Attended Our Last Event!

360+
Registered for our 7/25 event

What Is A Generative Model?

What is a Generative Model?

It is important to delineate clearly between the terms generative model, artificial intelligence, machine learning, deep learning, and language model:

Artificial Intelligence (AI)

This is a broad discipline within computer science dedicated to the development of intelligent agents capable of reasoning, learning, and autonomous action.

Machine Learning (ML)

A subfield of AI, ML is concerned with creating algorithms that enable systems to learn from and make predictions based on data.

Deep Learning (DL)

A specialized area of ML that employs deep neural networks, characterized by multiple layers, to learn complex data patterns.

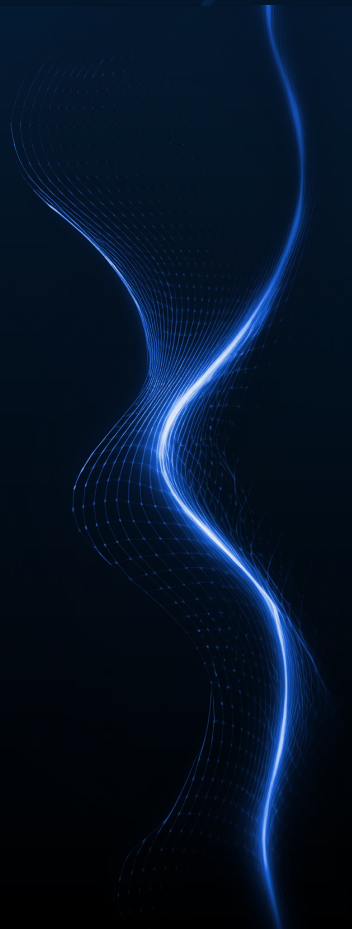
Generative Models

A category of ML models designed to generate new data by learning patterns from existing data.

Language Models (LMs)

Statistical models that predict the sequence of words in natural language. Some LMs employ deep learning techniques such as the transformer, and are trained on extensive datasets, evolving into large language models (LLMs).

Why Now?



The Success of Generative AI

- AI in 2022 is due to improved algorithms, enhanced computational power, hardware design, large labeled datasets, and collaborative research efforts.

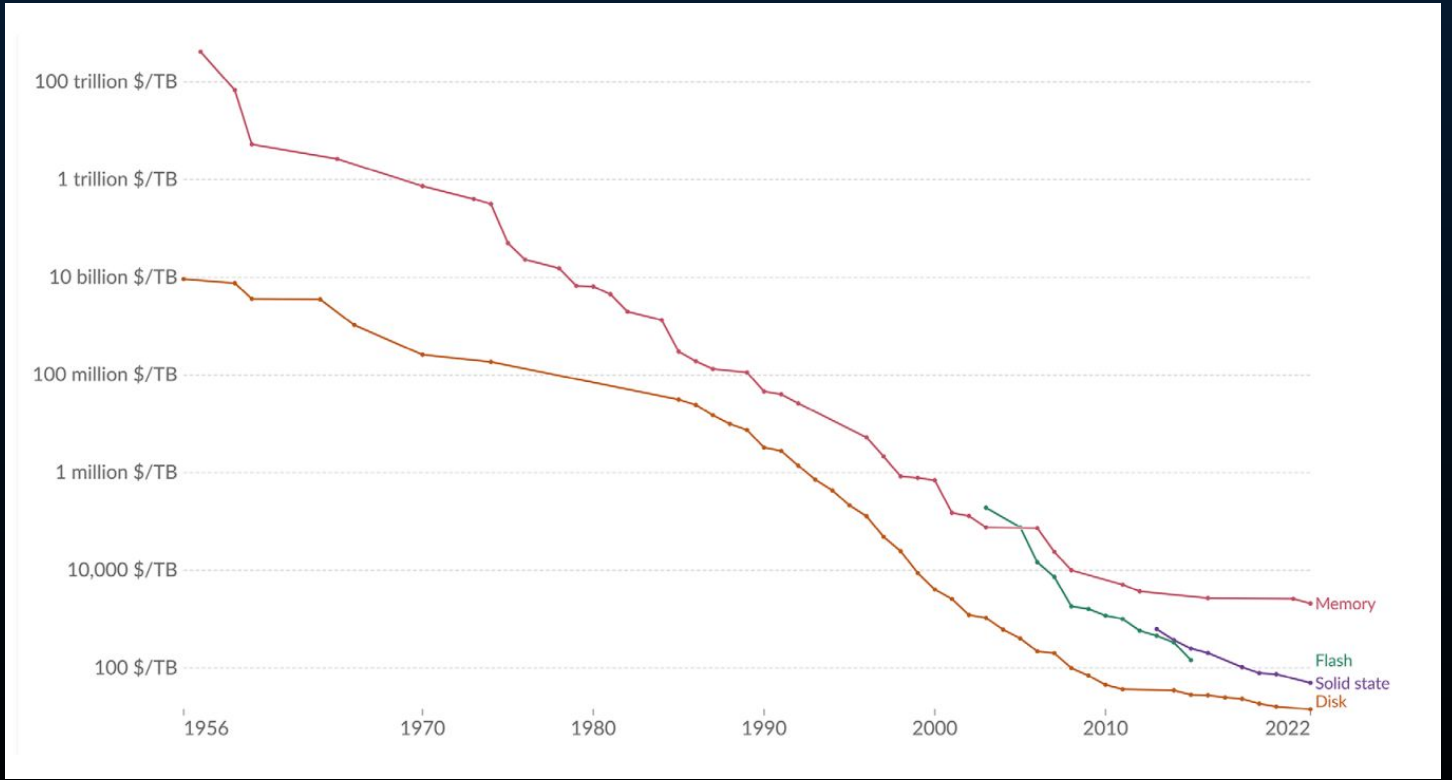
Sophisticated mathematical and computational methods, such as the backpropagation algorithm introduced in the 1980s, have been crucial.

The Rise of Deep Learning

- Deep Learning Breakthroughs: Multi-layered neural networks improved generative models.
- Hardware Advances: GPUs provided essential computational power.
- Cost Reduction: Lower hardware costs enabled deeper model development.

Why Now?

Figure 1. Cost of computer storage since the 1950s in dollars per terabyte



Why Is Hardware And Infrastructure Relevant?

Why is Hardware and Infrastructure Relevant?

Parameter Significance:
More parameters capture complex word relationships.

Pattern Recognition:
Predicts words like "dog" after "chase" and "cat."

Perplexity:
Lower perplexity indicates better model performance.

Emerging Abilities:
Models with 2-7 billion parameters generate diverse texts (poems, code, scripts) and answer complex questions.

LLM General Applications

Generative Models

Encompass a variety of types, each tailored to handle different data modalities within distinct domains.

These types include:

Text-to-Text Models:
- Notable examples include LLaMa 2, GPT-4, Claude, and PaLM 2.

Text-to-Image Models:
- Prominent examples are DALL-E 2, Stable Diffusion, and Imagen.

Text-to-Audio Models:
- Examples include Jukebox, AudioLM, and MusicGen.

Text-to-Video Models:
- Examples are Phenaki and Emu Video.

Text-to-Speech Models:
- Examples include WaveNet and Tacotron.

Speech-to-Text Models:
- Examples are Whisper and SpeechGPT.

Image-to-Text Models:
- Examples include CLIP and DALL-E 3.

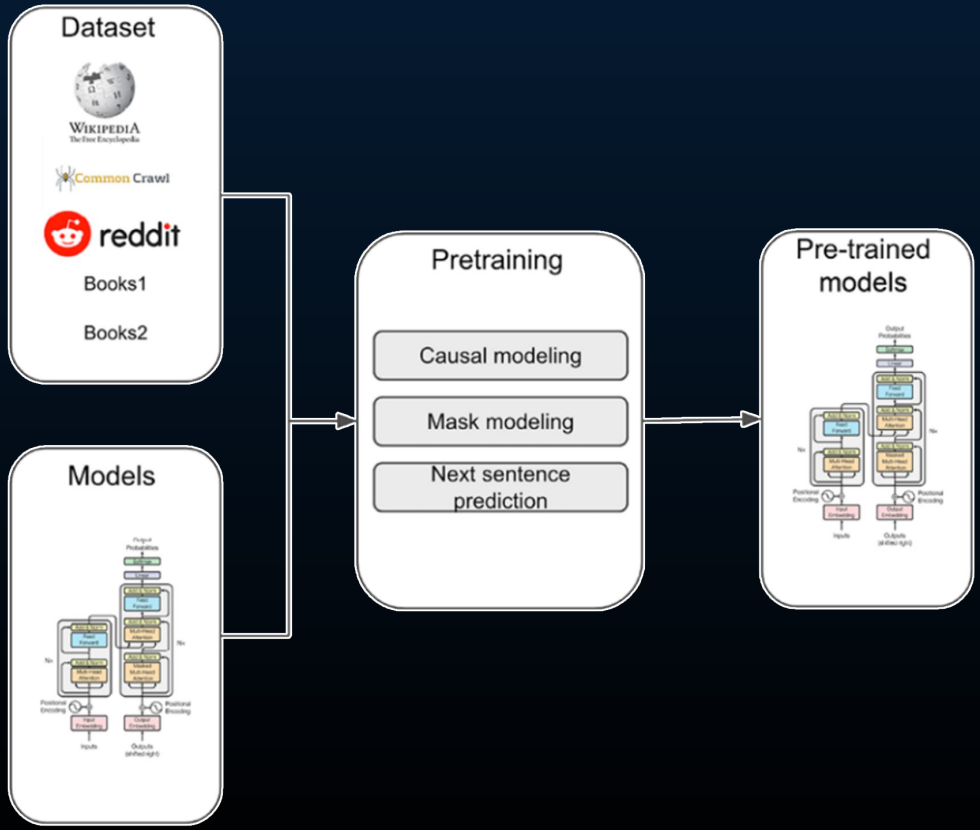
Image-to-Image Models:

Text-to-Code Models:
- Examples include Stable Diffusion and DALL-E 3.

Video-to-Audio Models:
- An example is Soundify.

Understanding LLMs

How Do LLMs Work?



How Do LLMs Work?

Extensive Neural Networks:
LLMs use many layers to process and learn from vast text data.

Sophisticated Algorithms:
Trained with algorithms like backpropagation to optimize performance.

Attention Mechanisms:
Focus on relevant parts of input text to capture complex patterns.

Massive Datasets:
Learn linguistic nuances by predicting the next word in sequences.

Fine-Tuning:
Refined on specific tasks to enhance contextual accuracy.

What Are The Limitations Of LLMs?

What are the Limitations of LLMs?

While LLMs showcase impressive capabilities, they also exhibit certain limitations that can impede their effectiveness in various scenarios.

Recognizing these limitations is essential for developing robust applications.

Some key challenges associated with LLMs include:

Outdated Knowledge

LLMs depend entirely on their training data and, without external integration, cannot provide up-to-date real-world information.

Inability to Take Action

LLMs cannot perform interactive actions such as searches, calculations, or lookups, which significantly limits their functionality.

Lack of Context

LLMs often struggle to maintain and incorporate relevant context from previous conversations and supplementary details, leading to less coherent and useful responses.

Hallucination Risks

Without adequate grounding, LLMs can generate incorrect or nonsensical content due to insufficient knowledge on certain topics.

Biases & Discrimination

LLMs can reflect biases present in their training data, potentially exhibiting religious, ideological, or political biases.

Lack of Transparency

The behavior of large, complex models can be opaque and challenging to interpret, making alignment with human values difficult.

Context Limitations

LLMs may have difficulty remembering previously mentioned details or providing additional relevant information beyond the given prompt.

Token Context Length

LLMs typically process up to 4096 tokens.

How Can We Mitigate The Limitations Of LLMs?

Addressing these limitations involves several techniques, ***including:***

Retrieval Augmentation

Accessing external knowledge bases to supplement the outdated training data of LLMs, thus providing additional context and reducing the risk of hallucinations.

Chaining

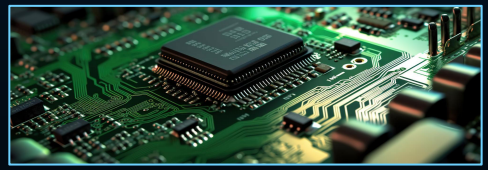
Integrating actions such as searches and calculations to enhance the LLM's capabilities.

Prompt Engineering

Carefully designing prompts to include essential context, guiding the model to generate appropriate responses.

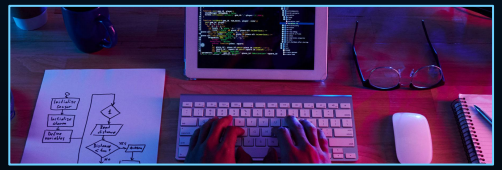
Monitoring, Filtering, and Reviews

Implementing continuous oversight of application inputs and outputs to identify issues, using both manual reviews and automated filters.



Memory

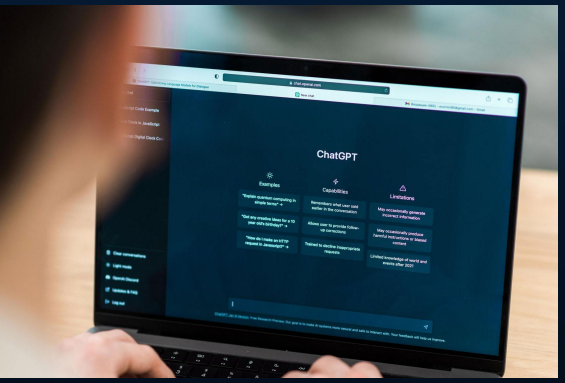
Preserving the context of conversations by storing interaction data across sessions.



Fine-Tuning

Adapting the LLM by training it on domain-specific data to align its behavior with the application's requirements.

How Can We Mitigate The Limitations Of LLMs?



Challenges

- **Compositional Reasoning**: Use elicit prompting and chain-of-thought techniques.
- **Problem Breakdown**: Employ self-ask prompting to methodically address problems.
- **Training Enhancements**: Integrate these tools into training pipelines for better reasoning.

Integration Process

- **Context and Inference**: Use prompting for context, chaining for inference, and retrieval for facts.
- **Ethical Safeguards**: Apply filters and constitutional AI principles.
- **External Data**: Connect to external data sources to reduce hallucination risks.
- **LLMOps**: Provide structure and oversight for responsible LLM use.

The Approach & Data Quality

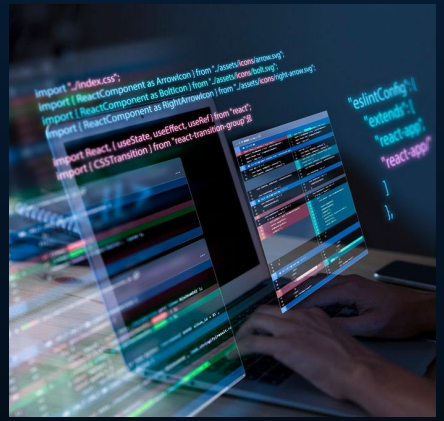
The Approach

A comprehensive approach combining preparation, vigilance, and ethical design is necessary for effective LLM integration. By augmenting LLMs with external data and thoughtful engineering, AI systems can overcome their innate limitations and achieve previously unattainable capabilities.

Data Quality

Data quality is critically relevant for LLMOps because it directly impacts the performance, reliability, and ethical considerations of large language models (LLMs). High-quality data is essential for training, fine-tuning, and deploying LLMs effectively.

Leveraging LLMs With LLMOps



Integration with Other

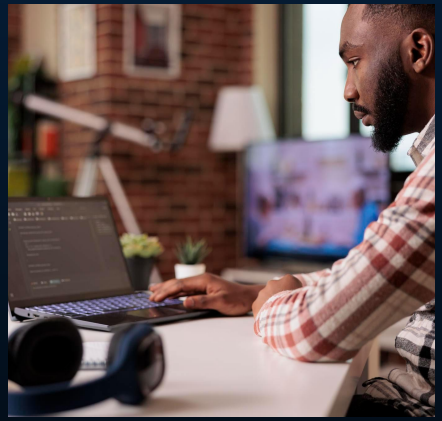
Data Sources and Tools

LLMOps integrates LLMs with various data sources and tools, enhancing their capabilities beyond text generation. This enables the creation of more powerful and versatile applications by combining LLMs with external knowledge bases and functionalities.

Solving

Operational Challenges

LLMOps addresses common challenges associated with using LLMs alone, such as lack of external knowledge, incorrect reasoning, and inability to take action. By providing solutions through integrations and off-the-shelf components, LLMOps practices streamline the operational management of LLMs.



Customization

& Flexibility

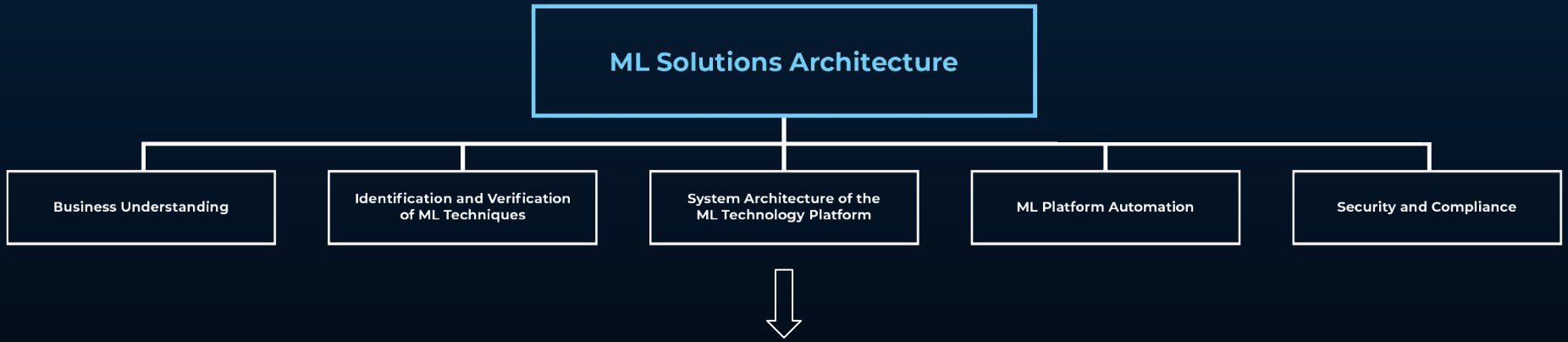
LLMOps allows developers to build customized natural language processing solutions, providing the flexibility needed to tailor LLM applications to specific tasks. This customization is a key aspect, which aims to optimize the deployment and use of LLMs in various contexts.

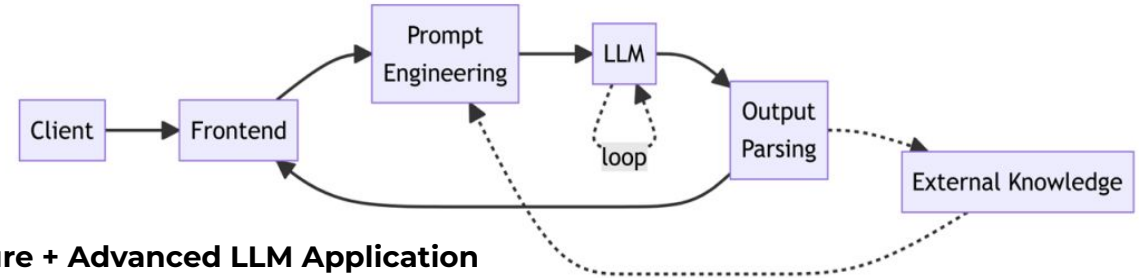
Dynamic &

Data-Aware Applications

By facilitating the creation of dynamic, data-aware applications, LLMOps helps leverage the full potential of LLMs. This dynamic capability is crucial for modern AI/NLP, which seeks to maximize the effectiveness and efficiency of LLM deployments.

Leveraging LLMs With LLMOps





ML Architecture + Advanced LLM Application

Leveraging LLMs with LLMOps

LLMOps Pipeline	Model Operations Hub	Continuous Improvement & Monitoring
Model Training	Model Management	Monitoring
Data Processing	Model Serving	Feature Engineering
Testing & Debugging	Resource Management	Performance Monitoring
Visualization	Optimization	Automation

Leveraging LLMs With LLMOps

LangChain

- It is an open-source Python framework for building LLM-powered applications.
- It simplifies the development of sophisticated LLM applications by providing reusable components and pre-assembled chains.
- Beyond basic LLM API usage, LangChain facilitates advanced interactions like conversational context and persistence through agents and memory.
  - This allows for chatbots, gathering external data, and more.

How Does It Work?

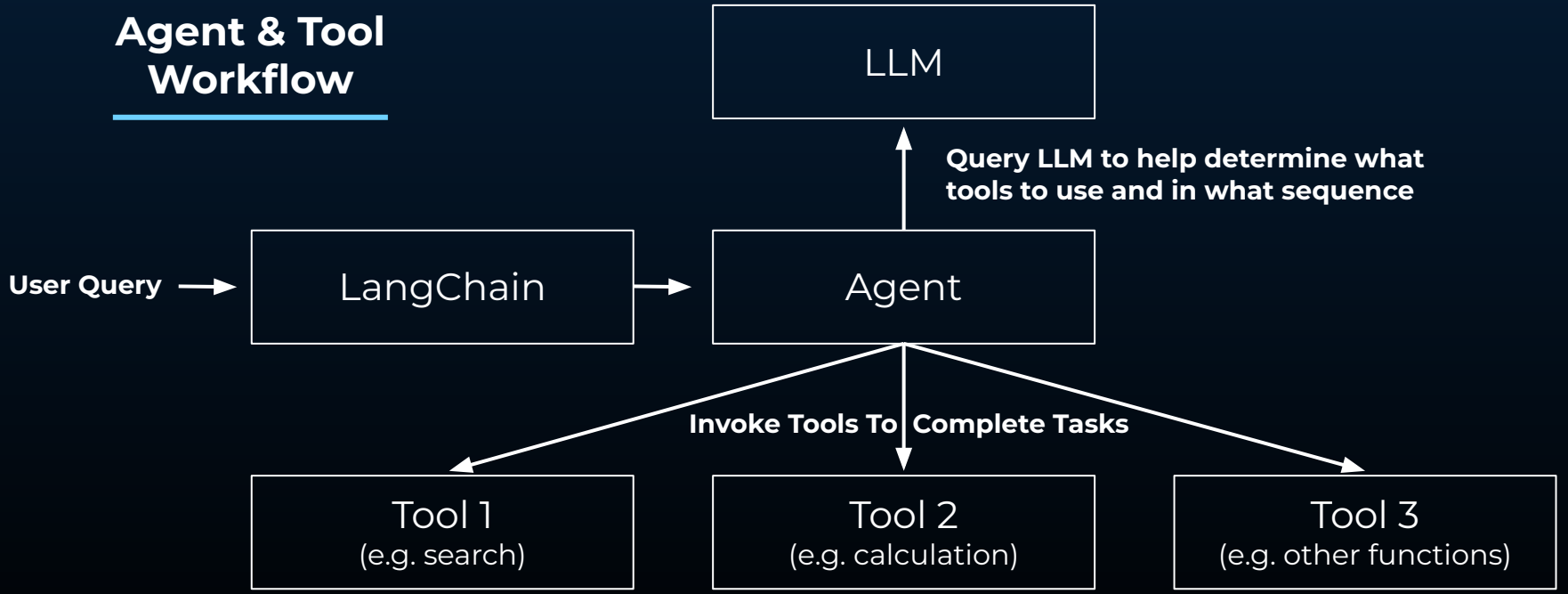
Chains: a chain is a sequence of calls to components, which can include other chains. It allows to make several combined queries to LLMs and specific tools.

Agents: They enable decision making by orchestrating the integration of chains and calling on tools.

Memory: It prevents chains from running in isolation by storing interaction information, making the systems much more powerful.

Tools: modular interfaces for agents to integrate external services like databases and APIs.

Agent & Tool Workflow



Question From Our Last AI Workshop

Q. Since the LLMs are constantly evolving and changing, given that you build an agent using, say, Llama, how easy is it to change or update your data and switch LLM?

A. That is a common use case in modern AI-based systems. It can be faced with a well-designed data pipeline following the best practices of LLMOps. That is, having an agent in production, you should have a data pipeline running behind the scenes waiting for an updated dataset to retrain an LLM or to adjust the prompts. If the metrics of this new model are better than those of the model in production, then the pipeline will change the model or prompts and the system will be updated with the new version of the model. The system must have a method to evaluate its responses and compare whether they are better or not compared to the current model in production. Finally, it can be considered easy if you have a well-designed architecture, otherwise, it could be a complex and highly demanding task.

The whole process might include the following steps:

- Model Registry
- Data Pipeline
- Data Versioning
- Continuous Automated
- Training Pipelines
- Testing and Validation
- Benchmarking
- Model Deployment

Question From Our Last AI Workshop

Q. How are you working to measure/ensure data quality?

A. It depends on the type of project you are working on. In general, there are two types of data validation to ensure data quality:

1. Manual validation performed by experts.

- For example, clinical data are usually validated by several experts and metrics such as inter-annotator agreement are used to measure the degree of agreement between experts and validate the data quality before training an AI model.

2. Automatic validation.

The validation process might include some of the following steps:

- Data Collection and Ingestion
- Data Cleaning and Preprocessing
- Data Quality Metrics
- Data Quality Tools and Techniques
- Bias & Fairness Checks
- Continuous Monitoring & Maintenance
- Feedback Loops

Frequently Asked Questions

Most Commonly Asked...

Where would a demo environment be hosted?

The demo environment can be hosted in the Sonatafy Demo environment on AWS or a client-specified environment.

How long will it take to get a demo environment up and running?

Once we have sample data, we can typically have a demo up and running within 2-3 weeks.

What costs are involved with Sonatafy building a demo?

If hosted in the Sonatafy Demo environment, there are no costs.
If hosted in a client environment, there will be cloud hosting costs (e.g., AWS, Azure) and minimal software costs (e.g., Google Collab Pro Plus, Hugging Face Pro). These costs will vary based on the specifics of the solution.

Who owns the IP?

For the Demo environment, Sonatafy retains ownership of the IP. If we are hired to implement a complete solution, clients typically own 100% of the IP.

How do we ensure compliance with Data (i.e., HIPAA)?

Many clients have us sign an NDA and a Business Associate Agreement (BAA). Clients often scrub the sample data before sending it to us to exclude sensitive information.

Will your confidential data be Secure or Exposed?

Unlike solutions like ChatGPT, this is a private environment, and information is not shared externally.

How will this solution integrate with our existing environment?

This is a demo environment to demonstrate capabilities. If you choose to proceed, a custom solution will be architected.

Will my internal team need training or education on how to use the AI environment?

This is a demo environment to demonstrate capabilities. If you choose to proceed, a custom solution will be architected.

Thank You ✨

Closing Message For Our Guests

Thank you for attending our AI Workshop.

Our goal was to showcase our capabilities and learn about your business challenges. We welcome any feedback on how well the workshop met your expectations and if there are additional topics you would have liked us to cover.

BONUS For attending the workshop, we'd like to:

1. Provide you with a 60-minute follow-up session to dive into your tech stack, product, and data, culminating in a **Free AI Analysis/Roadmap.**
2. Build a **Free Customized Demo** of our capabilities using a dataset you provide, where we will train a model to extract data.
3. Send you this **Free LLMOps EBook**

We look forward to your feedback and to exploring these opportunities further.

LIVE AI Workshops.

Thank You For Joining!

Steve Taplin
CEO of Sonatafy Technology

Antonio Tamayo
Lead AI Engineer, PHD

Brought To You By:

Sonatafy
Nearshore Software Done Right.