

LLMOps For ECommerce

20th June 2024

Who am I?



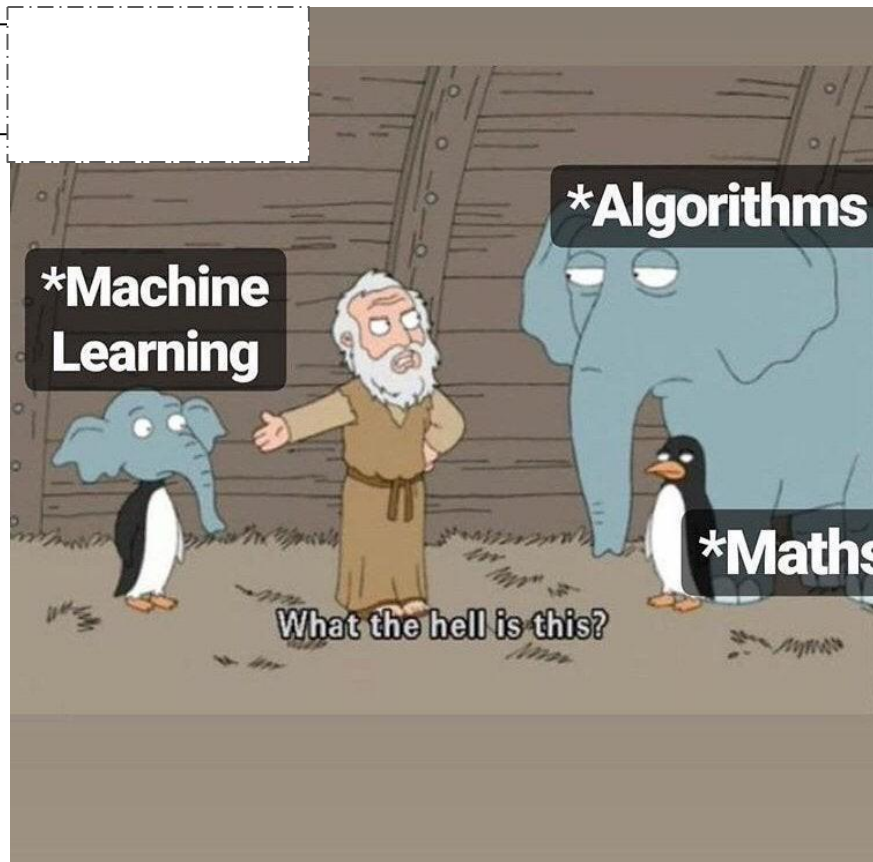
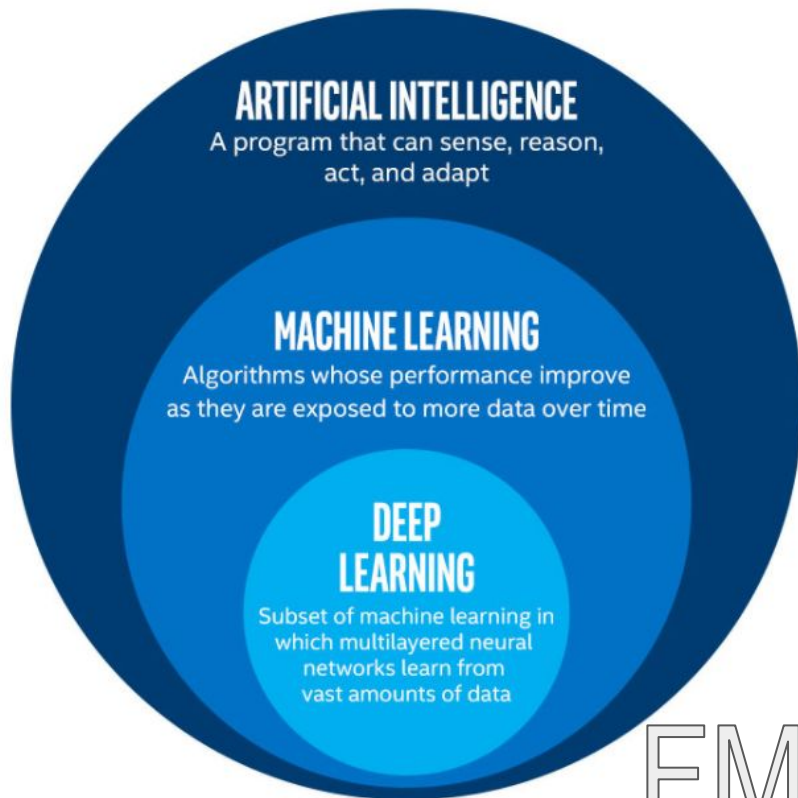
Officially

18+ years of industry experience in tech and product roles across The US, Africa, South-East Asia, and Europe with organizations such as Booking.com, AppsFlyer, GoJek, Rocket Internet, and National Instruments. Currently Director-Big Data & ML/AI with Booking.com

Un-Officially

Amateur Chess (ELO 2241) and Go (4kyu) player, reach out if you want to discuss either :)

Terminology



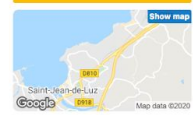
FMOPs?? LLMOPs??

Autocomplete

Search
 Destination/property name: Saint-Jean-de-Luz
 Check-in date: Saturday, September 12, ...
 Check-out date: Monday, September 14, 2...
 2-night stay
 2 adults
 2 children 1 room
 0 years old
 2 years old
 I'm traveling for work
 Search

Intent detection

Content selection



Filter by:

Your Budget

- € 50 - € 100 per night 4
- € 100 - € 150 per night 7
- € 150 - € 200 per night 13
- € 200 + per night 48

Popular Filters

- Parking 42
- Swimming pool 13
- Family rooms 39
- Free cancellation 17
- Apartments + homes 52
- Breakfast included 3
- Very Good: 8+ 34
- Hotels 17

Star Rating

- 2 stars 4
- 3 stars 20
- 4 stars 5
- 5 stars 2
- Unrated 39

Distance from center of Saint-Jean-de-Luz

75% of places to stay are unavailable for your dates on our site.

If you're flexible, check out these alternative dates:

- Sept 10 - Sept 12 From € 74 per night
- Sept 11 - Sept 13 From € 74 per night
- Sept 12 - Sept 14 From € 69 per night
- Sept 13 - Sept 15 From € 69 per night
- Sept 14 - Sept 16 From € 69 per night

Traveler flexibility

Topic extraction

Personalised UI

Saint-Jean-de-Luz: 70 properties found
3 reasons to visit: beaches, oceanside & scenery

Nearby Beaches: Lafitena Beach Mayarco Beach Grande Beach

Top Picks for Families Show homes first Lowest Price First Cleanest properties first Genius

Commission paid and other benefits may affect an accommodation's ranking. Learn more.

Properties in Saint-Jean-de-Luz with Genius Level 2 benefits

- Apartamentos de C...**
8.5 Very Good
Starting from € 10,071
- Port nivelle IBAIAN**
8.7 Excellent
Starting from € 266
- Mosaikhotel (ex Le ...**
8.3 Very Good
Starting from € 405
- Hôtel Parc Victoria**
8.9 Excellent
Starting from € 1,161

Recommendations

Traveler preferences

Hôtel Madison Saint Jean de Luz ★★★★★
 Wonderful 9.1
 560 reviews
 Cleanliness 9.6
 Guest Favorite

Recommended for 2 adults, 2 children

Family Suite
 1 bedroom + 1 living room
 2 beds (1 king, 1 sofa bed)
 Only 1 left at this price on our site

2 nights, 2 adults, 2 children
 € 774
 Additional charges may apply
 Select your room

Hotel Arena ★★
 Very Good 8.6
 1,793 reviews
 Cleanliness 9.0

Recommended for 2 adults, 2 children

Family Room
 3 beds (2 bunk beds, 1 queen)
 FREE cancellation - no prepayment needed
 Risk Free: You can cancel later, so lock in this great price today!

2 nights, 2 adults, 2 children
 € 203
 Includes taxes and charges
 Select your room

Ranking

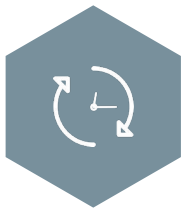
And translations, UI optimisation, content highlighting, content augmentation, fraud detection...

ML Platform in Numbers

<20 ms

Latency

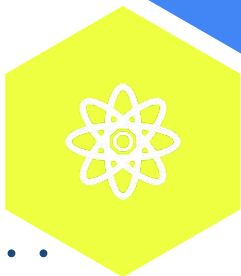
In Many Scenarios



~400

Models

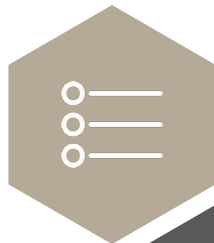
In Production



N

Machine Learning
Practitioners

Up to **~3B**
Training Examples



~10

Frameworks
Supported



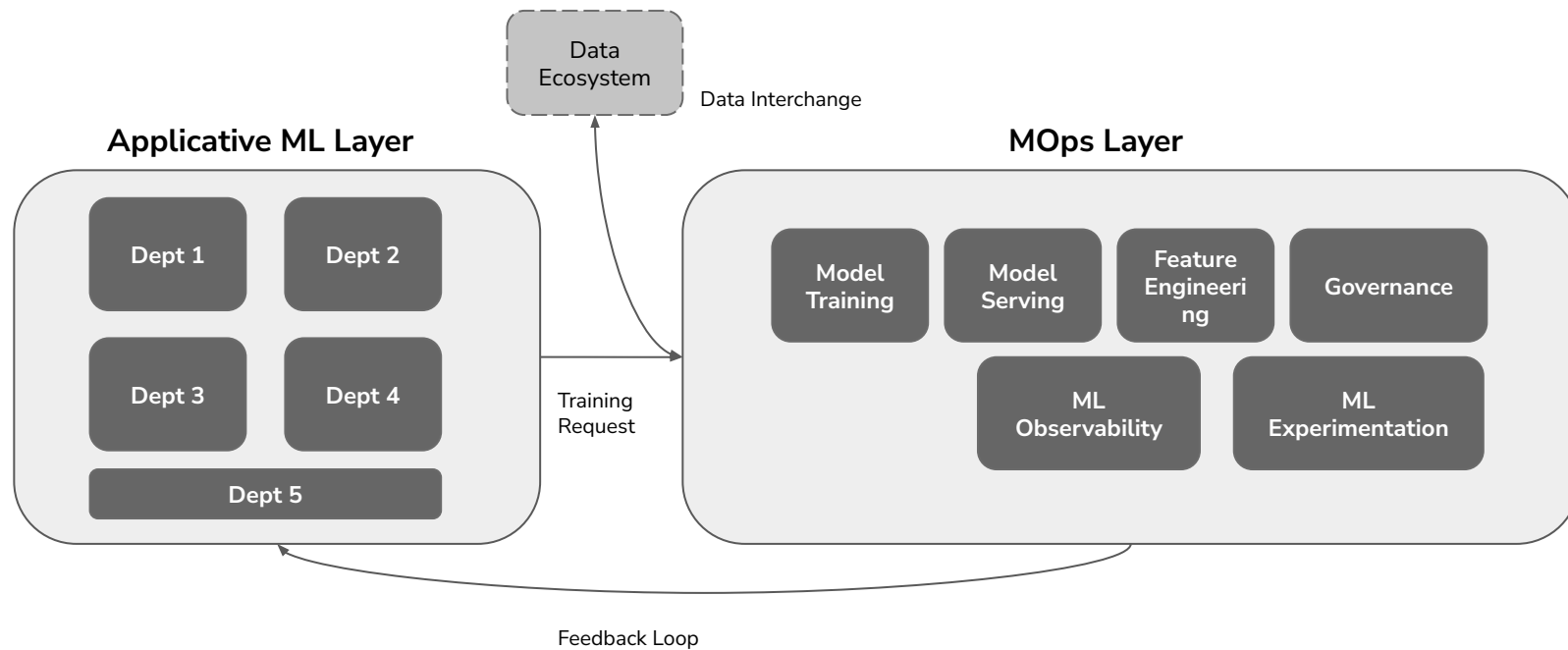
**220
Billion**

Predictions
Per Day



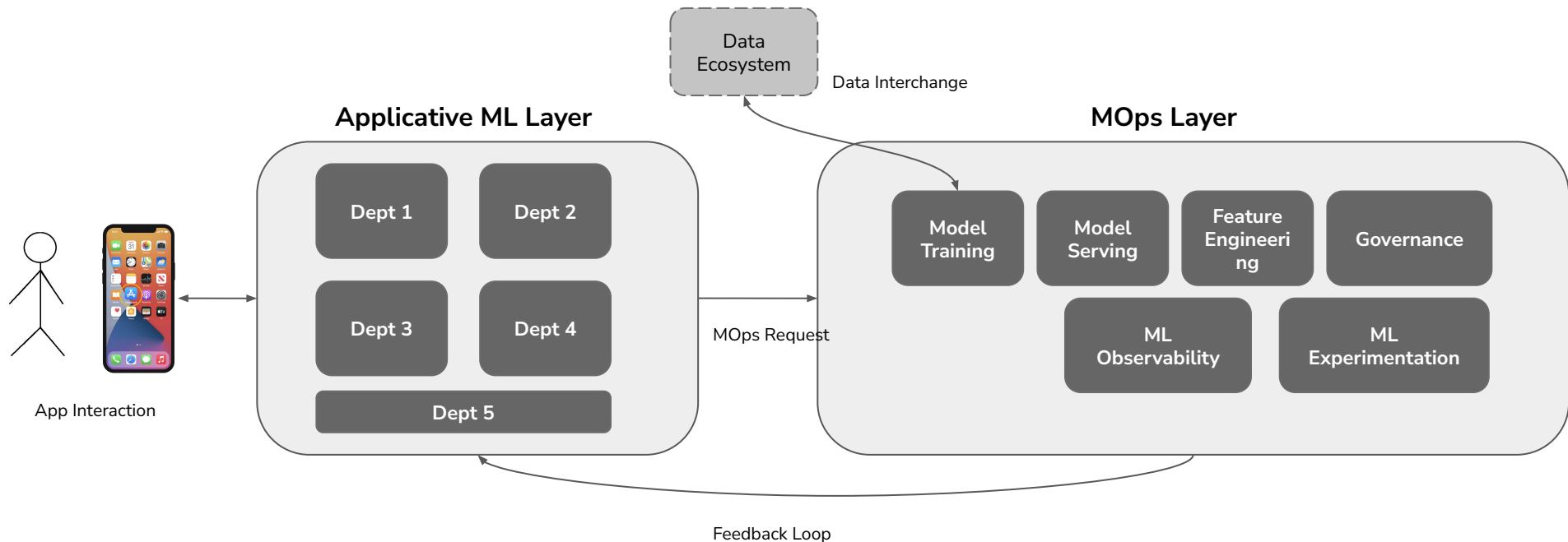
Typical MOps Flow

Model Training



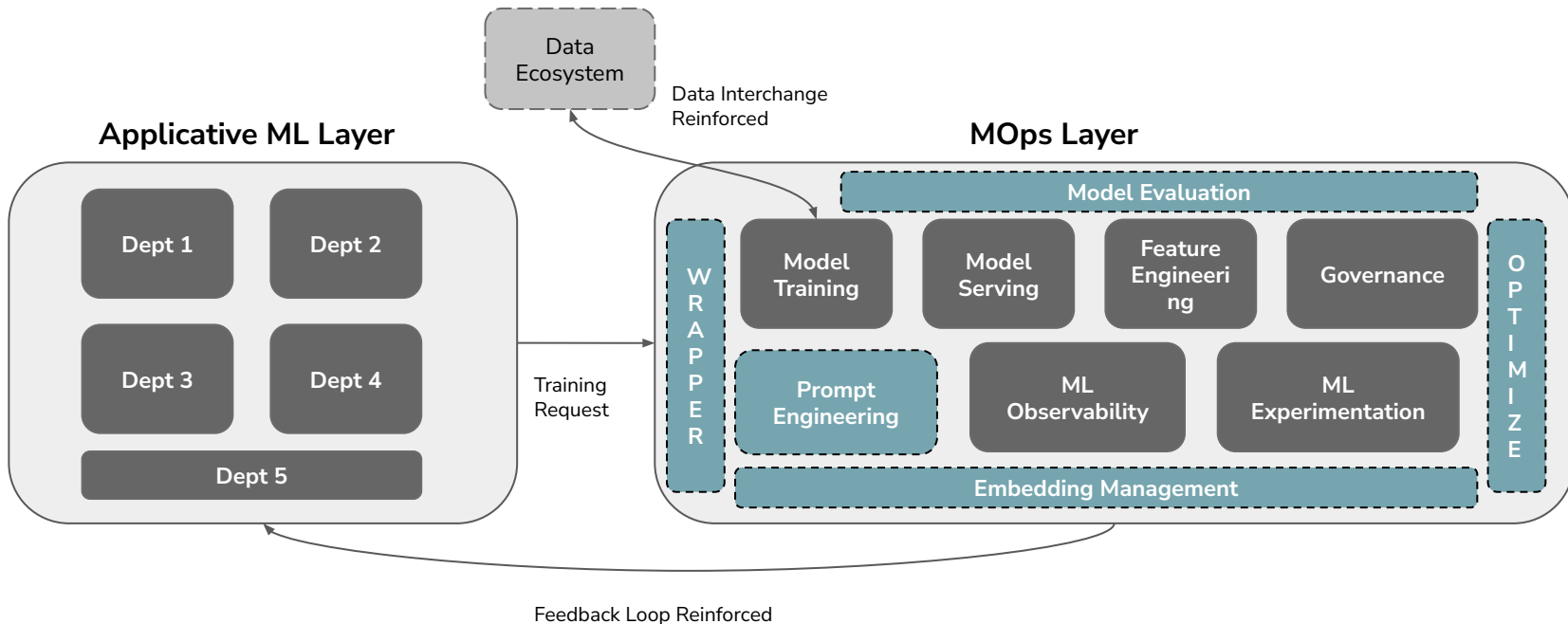
Typical MOps Flow

Model Serving



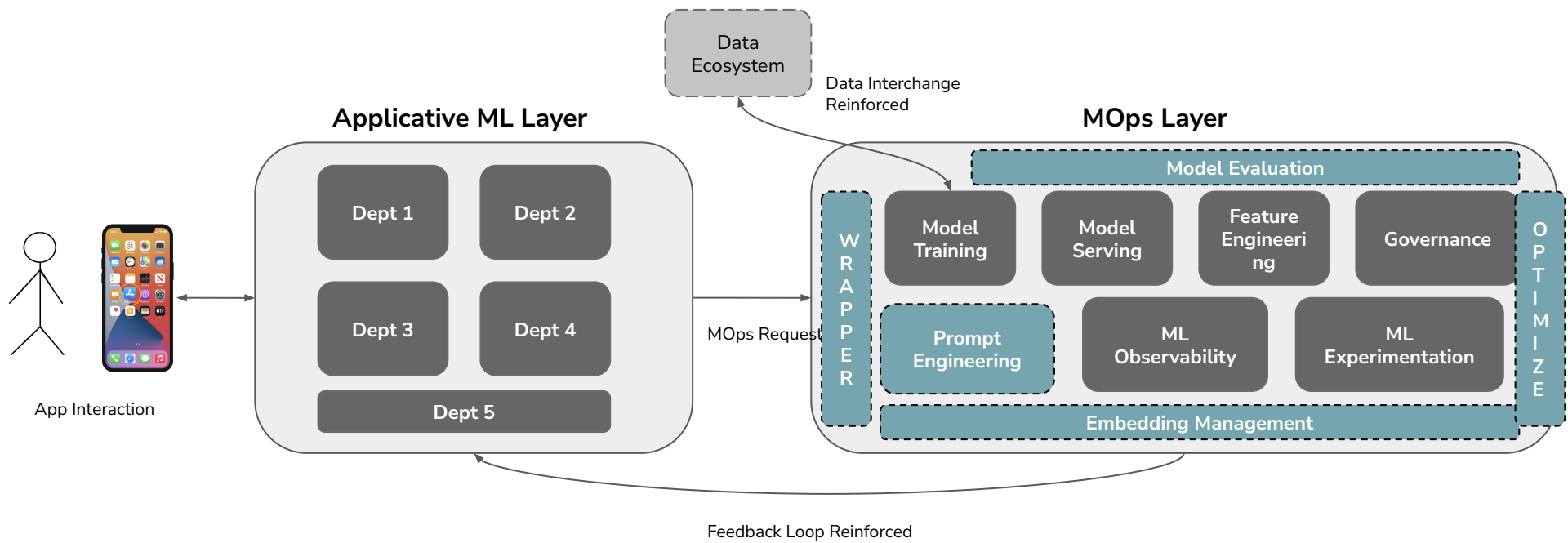
With LLMs, What has Changed?

Model Training



With LLMs, What has Changed?

Model Serving

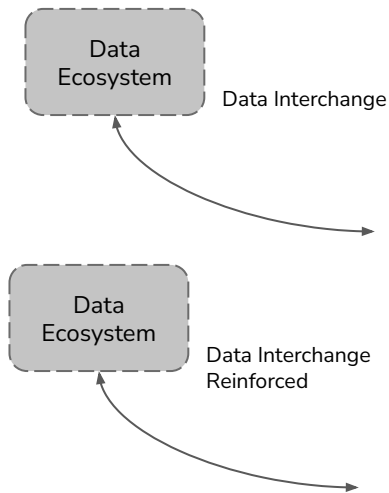


Layer Breakdown

Let's Take Each Layer one by one and Discuss

1. What each Layer does ?
2. Complexity of each layer before and after Large Models
3. Opportunities and Product Lens on each Layer before and after the advent of GenAI

LLMOps-Data Layer



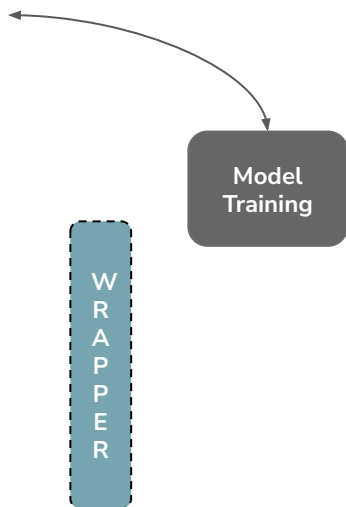
Bedrock of any ML ecosystem is its data ecosystem for training data. The data ecosystem should be robust, with well-labelled, well-lineaged data, allowing for proper tracking of results generated by our ML models (you cannot measure, what you cannot track !!)

1. The layer has many nuances, and depths with evolution to Data lakes, Data Fabrics vs Mesh, etc.
2. A lot of VC investment has always targeted “Big Data”
3. In a typical Non-LLM world, most of the data in the lake has been “structured”

What Changes with LLMs?

With an LLM ecosystem, the data layer interaction is no longer limited to only structured data, but extends to unstructured and semi-structured data (direct chat interactions, images uploaded, etc.) bringing additional complexity

LLMOps-Model Building/Training



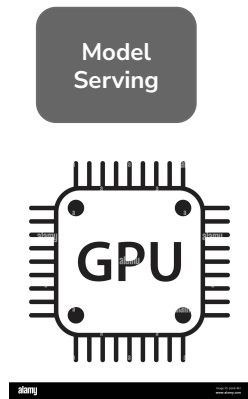
Model Training/Building is the assembly line for ML user to create models. The MOps ecosystem enables the tooling for Multi-modal deployments with various model-building formats supported (LightGBM, Tensorflow, etc.)

1. The layer interfaces with the data ecosystem, and has newer and more optimised formats come up regularly
2. A lot of VC investment has always targeted Model Building
3. In a typical Non-LLM world, model training is a “middleware” to building tools

What Changes with LLM?

Building a model is out of scope for now, but serving publicly available models via an enterprise wrapper is a possibility. A wrapper/middleware addition could be envisaged here

LLMOps-Model Serving Layer



Model-Serving is the production line of the platform. Wherein multi-modal serving is deployed on specialized hardware (CPU/GPU/XPUs, etc.)

1. The layer is highly extensible, and requires scaleup engineering for scaling number of models
2. A lot of VC investment and optimizations go into both the software as well as hardware side of this ecosystem
3. In a typical Non-LLM world, most of the usecases used to be enough for 1 GPU, not anymore

What Changes with LLM?

For LLMs, singular GPUs are not enough (mostly), reinforcement of this team for Distributed GPU Training needs to be thought off. Additionally a lot of investment has gone into the custom hardware space for this (Trainium, XPUs, etc.)

LLMOps-Feature Engineering Layer

Feature
Engineering

Embedding Management

Feature Engineering is an encompassing term for both the feature compute as well as feature discoverability part of the platform (Discuss What is a Feature or ML asset)

1. The standard evolution of this ecosystem has been a Feature Store—both offline and online serving
2. A lot of VC investment and optimizations go into both the software as well as hardware side of this ecosystem
3. In a typical Non-LLM world, most of the use-cases used to be scalar

What Changes with LLM?

Whilst upto now, Online and Offline feature serving was needed, Vector Store and Prompt-Store use-cases (with embeddings) will crop up as we go along as well

LLMOps-ML Governance Layer



Governance

ML governance is a regulatory part of the ML ecosystem, which has become increasingly more important in the active- regulation ecosystem (EU AI Act, DAC7, eprivacy, DSA/DMA, etc.)

1. MOps ecosystem needs to build tooling to adhere to these regulations
2. A lot of solution providers have come forth, building E2E solutions in their MLOps chains (acquired or self-built) to incorporate compliance flows

What Changes with LLM?

With LLMs, increased scrutiny on Data and ML systems has come into place, requiring renewed focus. Some example of investment areas are model versioning, model drift, data versioning, etc.

LLMOps-ML Observability Layer

ML
Observability

ML observability demarcates the 30k feet view of your pipelines in an increasingly complex ML ecosystem

1. MOps ecosystem, with increased complexity, has renewed interest in specific observability tools (sitting on top of the system observability tools)
2. A lot of VC funding has been attracted into these areas with vendors such as Arize, Aporia, Whylabs, etc.
3. With Advent of LLMs, this layer has gained particular significance

What Changes with LLM?

LLMs being non-deterministic, Output moderation as well guard-railing has come up as a renewed focus for MOps ecosystem

LLMOps-ML Experimentation Layer

ML
Experimentation

ML Experimentation, both with the models, as well as with different models is an important part of the ML ecosystem


1. Multi-model deployment as well as multi-hypothesis deployment has been the bread and butter for the industry
2. Multiple tools in the industry allow tracking of said models already (W&B, amongst others)

What Changes with LLM?

This layer has gained renewed focus as well, especially with prompt-experimentation, and multi-model experimentation added into the mix

LLMOps-Feedback Loop

Feedback loops is feedback collection mechanism from an in-situ model to improve the model's performance. Such a flow is used for multiple purposes in the ML ecosystem



Feedback
Loop

1. Label Collection, knowledge reinforcement are common usecases
2. Multiple tools in the industry allow this flow to be perpetuated
3. With Advent of LLMs, this layer has gained particular significance

What Changes with LLM?

This layer has gained pivotal focus with Reinforcement Learning gaining a lot of traction in the ecosystem

LLMOps-Model Evaluation and Optimization

Model Evaluation

O
P
T
I
M
I
Z
E

Model Evaluation had been a subcomponent of ML experimentation, as Model-as-a-Service was never a thing, also optimization was something only big organizations invested in !!

What Changes with LLM?

Advent of new product capabilities in Model Evaluation (Open-sourced as well as closed) for evaluation against toxicity, bias, any 3rd party parameter (Sagemaker Clarify, say). Optimization is another area that has popped up with cost and compute required for LLMs (Nemo Microservices and Toolkit by Nvidia, say)

Additional Paradigms: SLMs and Frugal AI

What More? SLMs now?

Discuss hyper-focused SLMs who can far outstrip frontier models→ How would MLOps ecosystem work for a model on a mobile, say? Gemini Nano, Phi Models, OpenAI in the future

Cost Matrices	Way of Optimization
Hardware	CPU deployments instead of GPUs, custom hardware such as Groq, Inferentia , Trillium, etc.
Software	Deployment and cost-optimization so the best out of hardware (NeMO microservices, etc.)
Use-case Market	In essence, not everyone needs to solve for the big market, Mobile market inherently relies on SLMs

Are we Done?

Another "Death by PowerPoint"?

