ML Introduction
Vectors and Embeddings
RNNs
CNNs
Transformer Architectures
Attention
Positional Encoding

GenAI Introduction
Responsible AI Practices
Use-Cases and Limitations
Agents
Agentic Workflows
Agentic Coding

Modalities
Multi-modal
Multi-model

Text Generation
Text-to-Text
Single-Turn
Multi-Turn

LLM Basics
LLM Use-Cases
Anatomy of a Prompt
Tokenization
LLM Hyper Parameters
Principles of Prompting
Ethics / Bias
Model Cards
Leaderboards
ML Commons and ML Perf
https://artificialanalysis.ai/
Prompt Roles
Context Window
Hallucinations
Topic Controls
Truthfulness Checks
Summarisation
Code Generation
Reasoning
Classification
QnA
Sentitent Analyst

Image Generation
Stable Diffison
Dall-e
Text-to-Image
Image-to-Image

Video Generation
Sora
Text-to-Video
Image-to-Video
Video-to-Video

3D Generation
Model Gen
Texture Gen
Animation Gen

Audio Generation
Descript
EleventLabs
Whisper
Text-to-Speech
Speech-to-Speech
Speech-to-Text

LEVEL 100

Workbenches / Playgrounds
OpenAI Playground
Anthropic Workbench
Google AI Studio
AWS Bedrock Playground
Cohere Playground

AI-Powered Assistants
ChatGPT
Anthropic Claude
Google Gemini
Meta AI
Mistral AI

Prompt Engineering
General Knowledge
Chaining
Tree of Thought
CO-STAR
Zero-Shot
Few-Shot
Chain of Thought
Self-Consistency
reAct
Reflexion

Types of LLMs Models
Foundational models
Fined-Tuned Models
Instruct Models
Embedd Models
OpenSource Models
OpenWeight Models
Edge-Optimized Models
Uncensored LLMs
ClosdeSource Models

Model-as-a-Service
Google VertexAI
Amazon Bedrock
Azure AI Studio
Alibaba Cloud Model Studio

Servering
Kserve
Ray
vLLM
TGI
picoLLM

Containers/K8s
OPEA AI
Nvidia Workbench

LLM DevTools and Workflow
GenAI Lifecycle
LLMOps
Hugging Face
Ollama
LMS
Llamafile
LangChain
Llama Index
Llama.cpp
GGML/GGUF
Azure AI Promptflow
Automatic Reasoning and Tool Use (ART)
Context Caching
ONNX

AI-Code Assistants
Amazon Q Developer
GitHub Copilot
Gemini Code Assist
Codeium/WindSurf
Cursor

Env Setup
Conda
Jupyter Server
Lighting.AI
Google Colab
SageMaker Studio Lab
SageMaker Notebooks
Azure ML Notebooks

GenAI Security
Prompt Injections
Guard Rails
OWASP Top Ten LLM
Monitoring (Lang Fuse / Lang Smith)
ISO/IEC 42001:2023

Model Optimization
WebNN
OpenVINO
DeepSpeed
Megatron

GenAI Hardware
CPU Architecutres
x86
ARM

AI Delivery Platform
Intel Tiber AI Cloud
Scale AI
Runpod.io
Groq
Replicate
Ori
Nebulis

App Prototyping
Streamlit
Gradio
FastHTML
v0
Lovable

Frameworks
WebNN
DirectML
CoreML
MLX
Tensorflow
Keras

heterogeneous computing platforms
oneAPI
CUDA
Apple Metal
AMD ROCm
SYCL
OpenCL

AI Accelerators
XPUs
TPUs
GPUs
iGPU
NPUs
VPUs
FPGAs
ASICs

Instruction Masking
Structure Generation https://blog.dottxt.co/coalescence.html

LEVEL 500

Knowledgebases
Sythentic Data Generation
Vectorstores
Knowledge graph
Session Management
Search Engine for LLMs SERP / Tavily

RAGs
RAG Strategies
Chunking
Document Cracking
Type of Search
Semantic Search
Keyword Search
Density Retreival
Reranking
HNSW

LLM Customization
Hyperparemter optimzation
Prompt Tuning
ICL
MoE
Weights and Biases

SFT
Instuct / Domain FT
PEFT
LoRA
qLoRA
Full-Fine Tuning
Last Layer
RLHF

Size Optimization
Qauntiization
Distillation
Pruning

Flash Attention

litGPT

Axolotl

Evals
Traditional ML Metrics
AI-Assisted Metrics
Generation quality metrics
Risk and safety metrics
Ragas

AlexNet ~2012 62.3M params

LLM facts:
BERT-L ~ 2018 340M params
GPT-2 ~ 2019 1.5B params
Gemini small - 1B?
Llama3.1 8B params
Llama3.1 70B params
GPT-3 ~ 2020 175B params
Llama3.1 408B params
PaLM ~ 2022 540B params
GPT-4 ~2023 1.8T params
GPT-4o ~2024 1.8T params
Claude Haiku 2T estimated

Defining the size of model:
- Cost
- Capabilities
- RAM

Model weights 4 bytes per param
Adam optimizer 8 bytes per param
Gradients 4 bytes per param
Activation & temp 8 bytes per param
==> 4 bytes per params + 20 extra bytes per param
~ **20X more RAM** than space needed for the params

Haiku (~**20B**), Sonnet (~70B), and Opus (~2T).