PUC Minas

LLM

THE TECHNOLOGY BEHIND CHAT GPT

BY: BEATRIZ FULGENCIO

PUC Minas

Agenda

- Main concepts
  - Definition of LLMs
  - Tokenization & Embeddings
- LLM pipeline
  - Pre-training
  - Post-training
  - RLHF
- Adapting LLMs for you
  - Retrieval-Augmented Generation
  - PEFT Fine tuning
  - RAG vs Fine-Tuning
- Future

PUC Minas

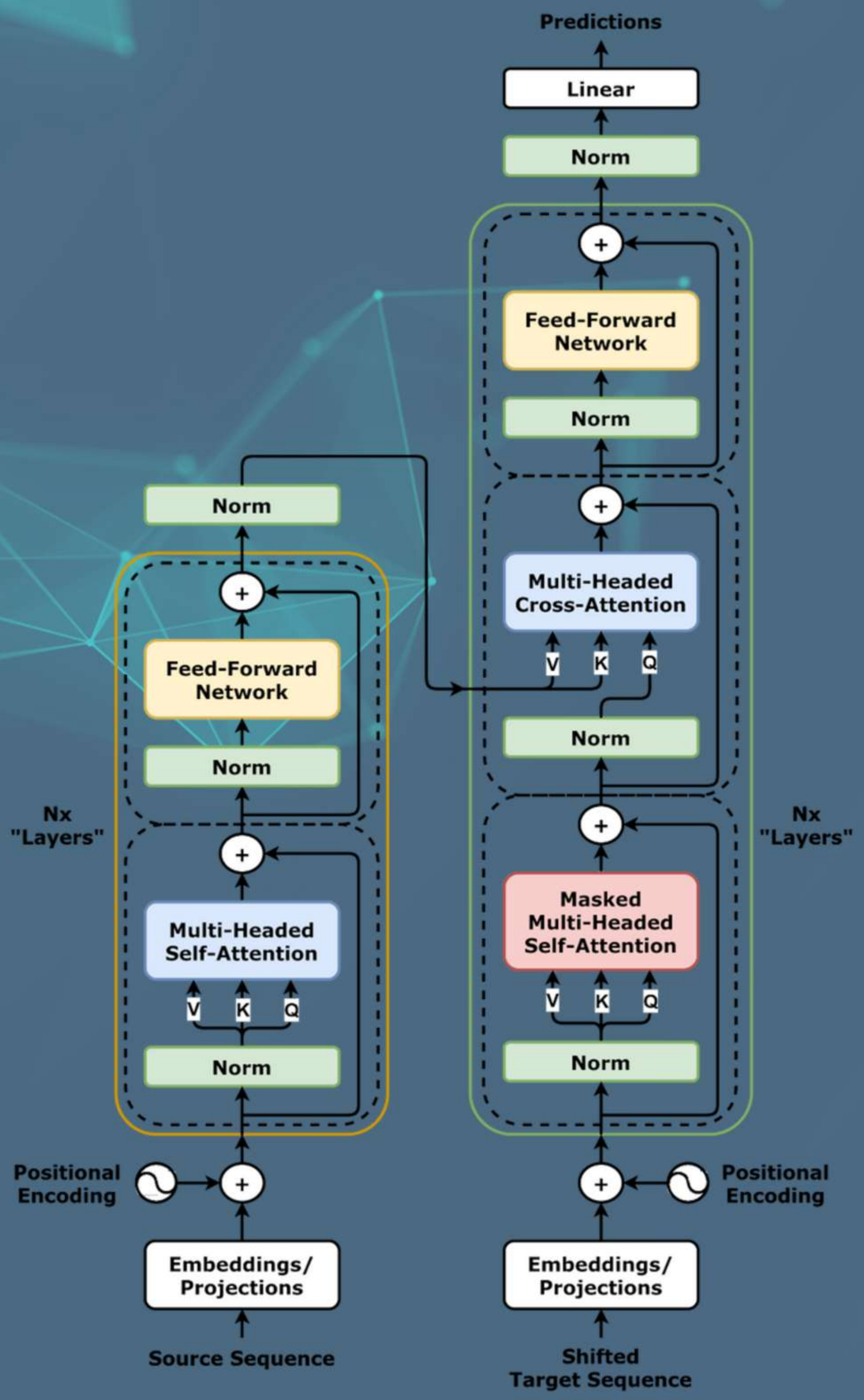
What is a Large Language Model?

- A Large Language Model (LLM) is a language model trained with self-supervised learning on large amouts of text data, designed for natural language processing tasks, especially text generation.

- The most capable LLMs are generative pre-trained transformers (GPTs) used in chatbots such as ChatGPT, Gemini and Claude.

- LLMs can be fine-tuned for specific tasks or guided via prompt engineering, but they may inherit inaccuracies and biases from their training data.

PUC Minas



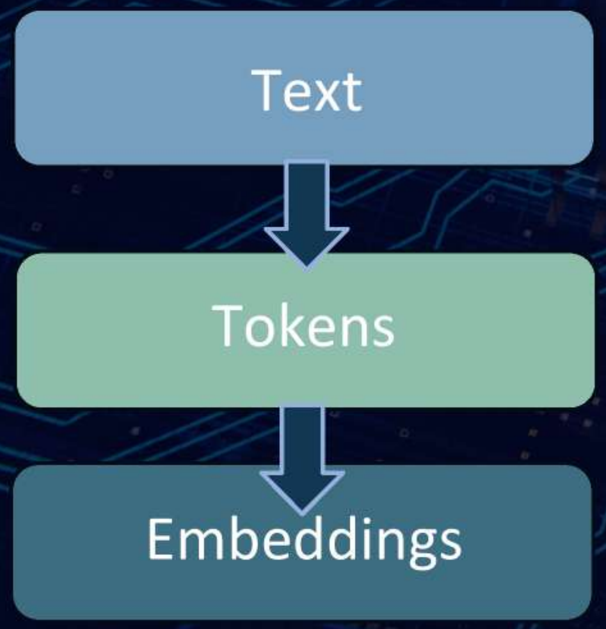
The Transformer Architecture

LLM architecture:https://bbycroft.net/llm

→

PUC Minas

Tokenization & Embeddings



Tokenization

- Converts text into smaller units called tokens (sentences, words, subwords or characters).
- Sentence & word tokenizers split at natural boundaries; subword tokenization (e.g. BPE) handles rare words.
- Tokens are assigned IDs and padded to uniform length for batch processing.

https://tiktokenizer.vercel.app

Embeddings

- Dense vectors that represent tokens numerically for neural networks.
- Traditional word embeddings assign a fixed vector to each word.
- Contextual embeddings vary with surrounding words, capturing semantics.
- Positional embeddings encode word order so the model knows positions.

https://cleverzone.medium.com/what-are-embeddings-a-simple-guide-with-visuals-4dc8689e89d1

PUC Minas

LLM pipeline

Pre-training

"internet document simulator"

Post-training

An assistant, trained by Supervised Finetuning

RLHF

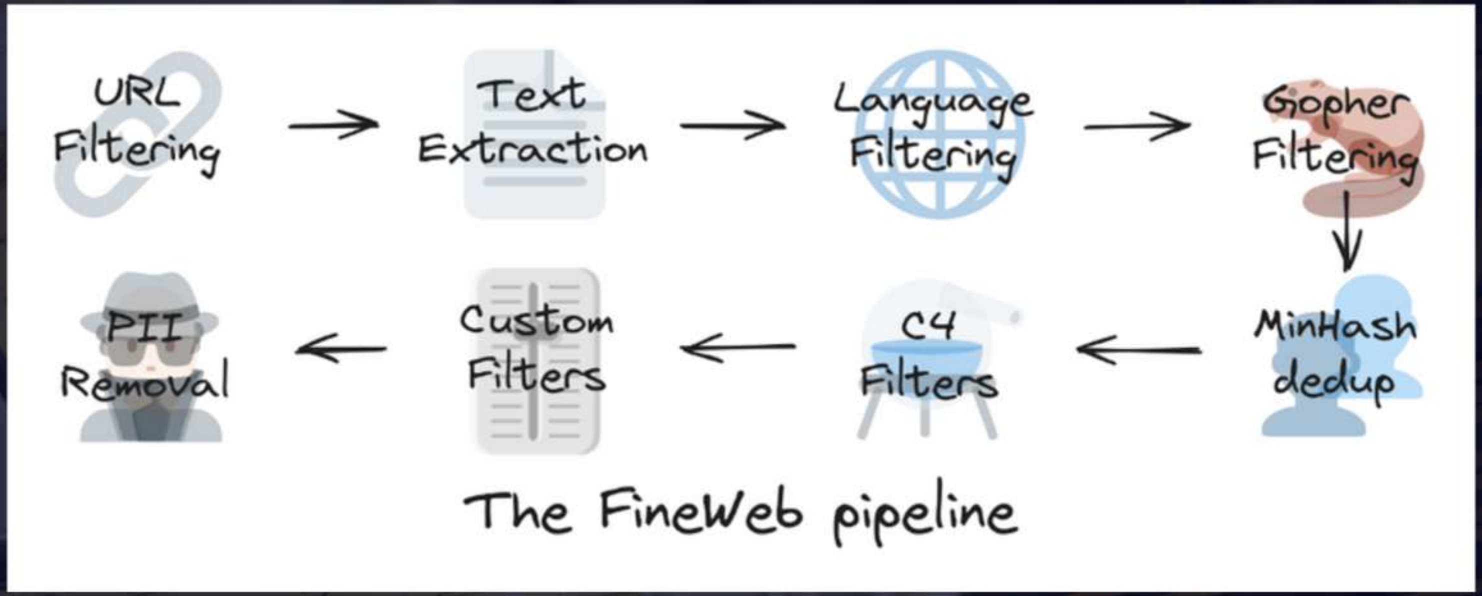
RL model

PUC Minas

Pre-training

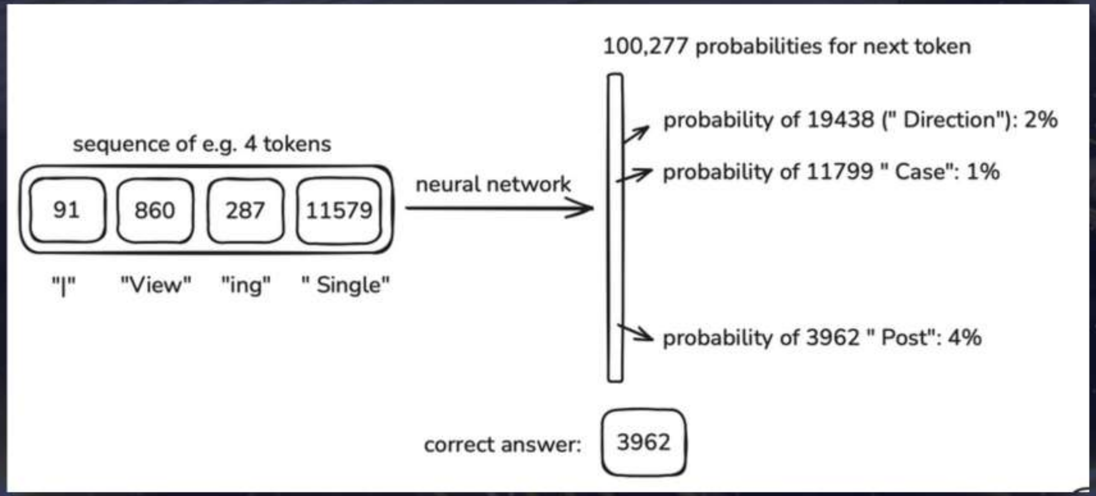
- Pre-training exposes the model to large, diverse and unlabeled text to learn general language patterns and semantics. (internet)
- Common objectives include Masked Language Modeling (e.g. BERT) and Causal Language Modeling (e.g. GPT), which teach the model to predict missing or next tokens.
- Pre-training is resource-intensive: massive datasets and weeks of GPU time, but it creates a versatile foundation that can be adapted for many tasks.

Step 1: download and preprocess the internet



Step 2: tokenization

Step 3: neural network training



Step 4: inference

LLM architecture:https://bbycroft.net/llm

PUC Minas

Post-Training

Supervised Fine-tuning

Training Data & Protocol

After pre-training, models are fine-tuned on curated conversation data (supervised fine-tuning) to become assistants. All dialogues are consistently tokenized with special role markers under clear labeling rules.

Limitations & Tooling

Compensate for limited working memory, counting/spelling brittleness, and token-by-token reasoning by integrating external tools like web search and a code interpreter

Hallucinations & Mitigations

Hallucinations are known wrong answers that models give when asked about something it doesn't know. To prevent this we have two mitigations:

1. **Refusal Tuning**: teaching the model to refuse unknown queries
2. **Search Tags**: adding tags such as <SEARCH_START>... <SEARCH_END> to trigger real-time lookups.

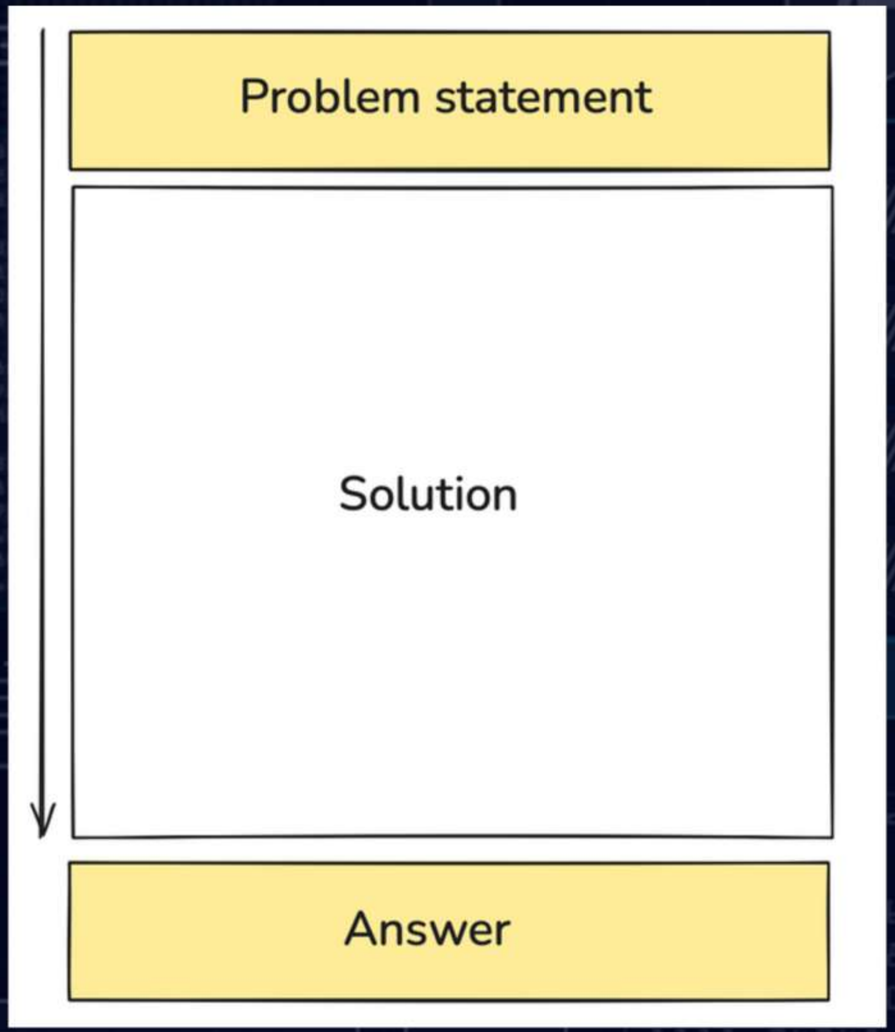
PUC Minas

Post-Training

Reinforcement Learning from Human Feedback (RLHF)

- Reinforcement Learning from Human Feedback (RLHF) refines the assistant using a reward model trained from human preference comparisons.
- RLHF involves three phases: choose a base model, collect human feedback on outputs, and optimize the model with reinforcement learning guided by the reward model.
- This are called thinking models, and they are available in ChatGPT o.. models and deep-seek-R1. It is the fronteer development in the area.

We are given problem statement (prompt) and the final answer.We want to practice solutions that take us from problem statement to the answer, and "internalize" them into the model.



PUC Minas

Pre-training X Post-training

Aspect	Pre-Training	Post-Training
Objective	Acquire general linguistic knowledge	Optimize for a specific task or domain
Data	Large, diverse and mostly unlabeled	Smaller, labeled and domain-specific
Techniques	Unsupervised / self-supervised (MLM, CLM, NSP)	Supervised learning, transfer learning, domain adaptation
Resources	Extensive compute and long training	Moderate compute and shorter training
Challenges	Cost, data availability, balancing generalization	Overfitting, data quality, task alignment

PUC Minas

Retrieval-Augmented Generation (RAG)

- RAG retrieves relevant documents from an external corpus and concatenates them with the input before generation, allowing the model to access up-to-date information.
- It improves accuracy, controllability and reduces hallucinations by grounding responses in retrieved evidence.
- RAG is adaptive: the retriever indexes documents (via embeddings) and the generator composes the answer.

PUC Minas

Parameter-Efficient Fine-Tuning (PEFT)

Ideal for cost-effective, agile domain adaptation of large LLMs

What Is PEFT?

- Fine-tune only a small set of added parameters
- Keep the majority of the pre-trained model frozen

Common Techniques

- **LoRA** – injects low-rank trainable weight updates
- **Adapters** – small bottleneck layers between transformer blocks
- **Prefix-Tuning** – learned prompt vectors prepended to each layer
- **BitFit** – tune only bias terms

PUC Minas

RAG VS PEFT

Aspect	Retrieval-Augmented Generation (RAG)	Parameter-Efficient Fine-Tuning (PEFT)
Data Source	External documents ingested into a vector store and retrieved at query time	Task-specific labeled datasets used to train small added modules
Objective	Inject up-to-date or proprietary knowledge on-the-fly and reduce hallucinations	Adapt a pre-trained model to domain/task with minimal parameter updates
Requirements	Vector index & retrieval pipeline, prompt engineering, minimal annotation effort	Pre-trained base model, LoRA/adapters/prefix modules, modest labeled data
Use Cases	Rapidly changing domains (news, regulations), enterprise QA over private corpora	Domain-specialized assistants (legal, medical, brand voice), multi-task adaptation on a budget

PUC Minas

Future Directions

Multimodal Intelligence

- Beyond text—seamlessly combining vision, speech, video, and natural conversation.

Agentic Systems

- Autonomous “agents” that manage long-horizon tasks with coherent, self-correcting workflows.

Ubiquitous & Invisible AI

- Pervasive intelligence embedded in everyday devices and interfaces, working behind the scenes.

Human-Centric Computing

- AI as a seamless collaborator—anticipating needs and streamlining interactions.

Adaptive & On-The-Fly Learning

- Test-time training and continual adaptation for personalized, resilient performance.

THANK YOU