# LLM’S

EXPLORING LLAMA

# Introduction LLM'S

Large Language Models (LLMs) have transformed AI, particularly in Natural Language Processing (NLP), by leveraging advanced architectures like transformers and attention mechanisms, driving innovation across multiple domains.

# LLaMA

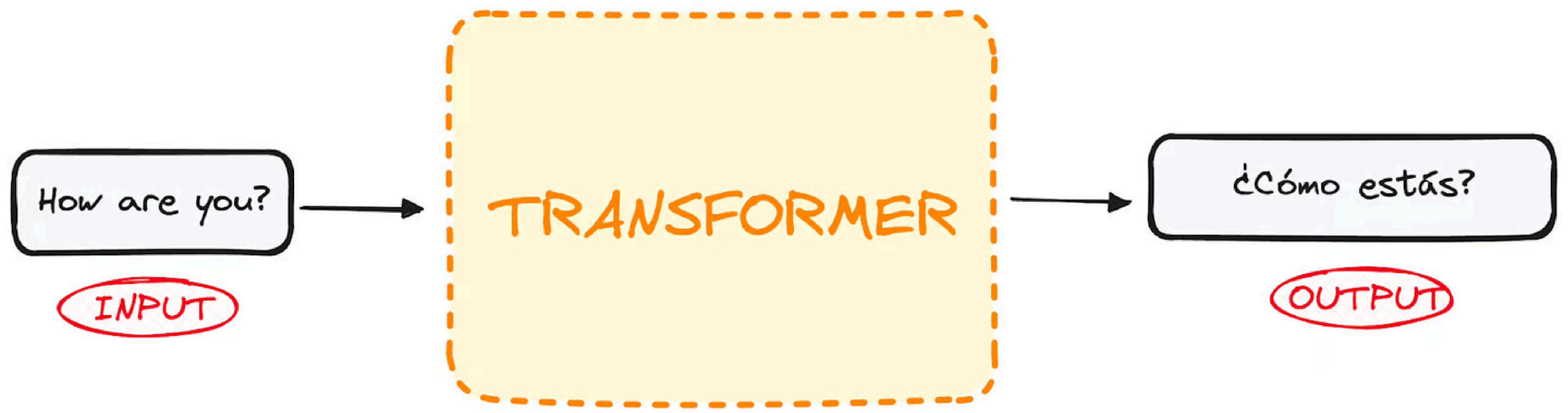
LLaMA (Large Language Model Meta AI) is Meta's open-source contribution to AI, designed to advance research and innovation. It enables customization for specific tasks and industries, offering a family of models ranging from 7B to 65B parameters.

# LLaMA Models

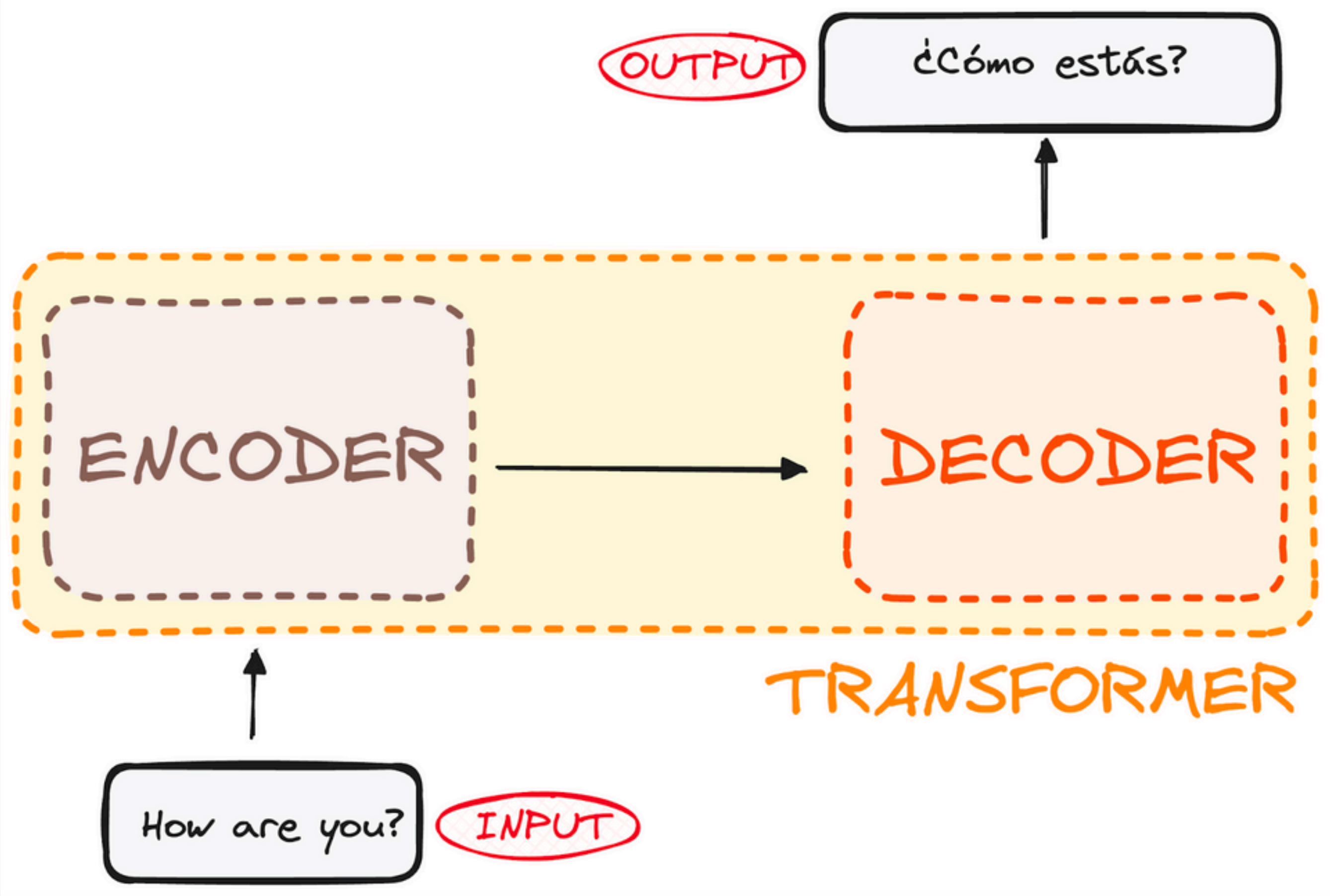
- LLaMA: A strong foundation with LLaMA-13B outperforming GPT-3.
- LLaMA2: Enhanced with open-source availability, longer context handling, and Grouped-Query Attention (GQA).
- LLaMA3: Improved multilingual support, longer token handling, and faster processing.

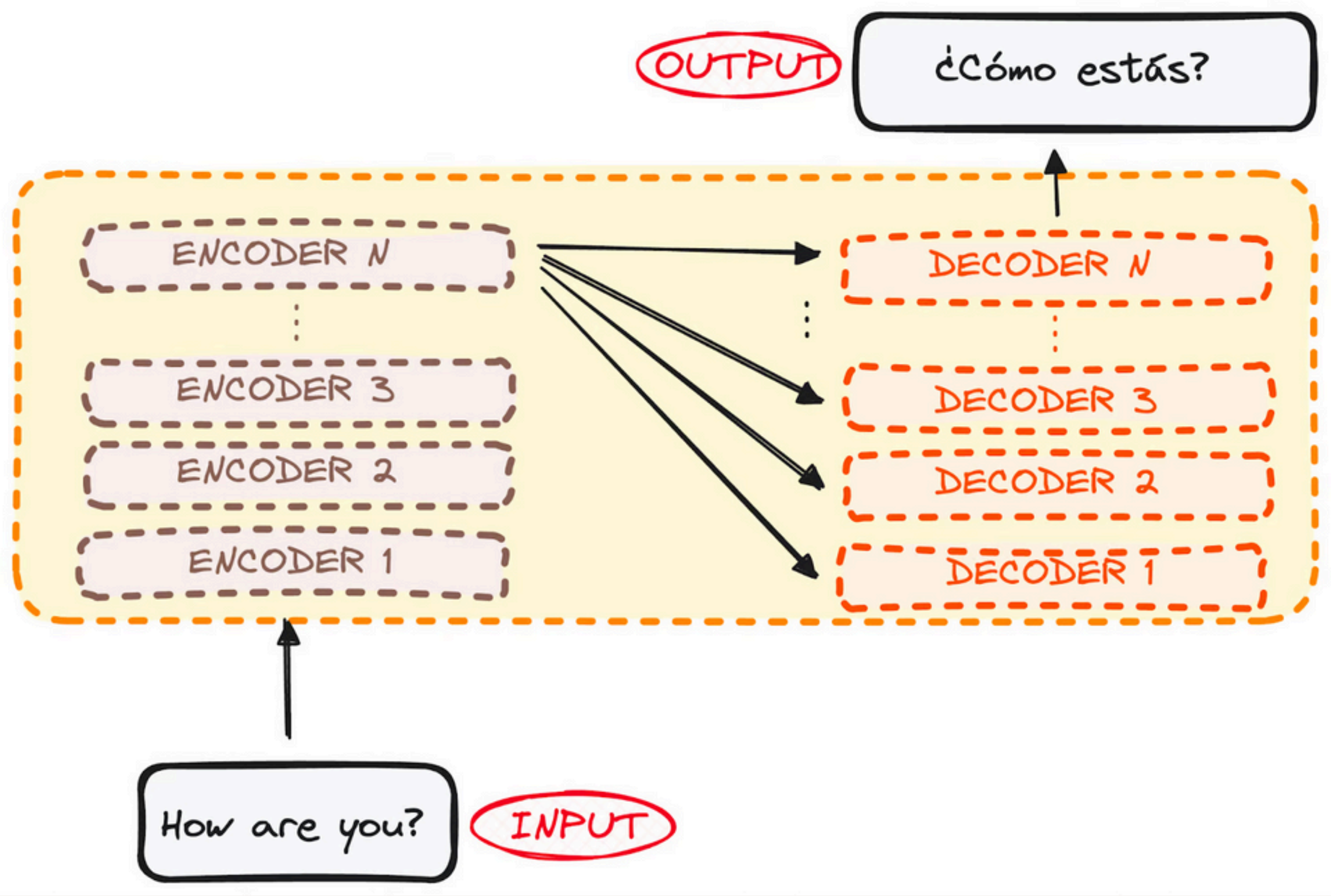
# Attention is all u need !

# Transformers



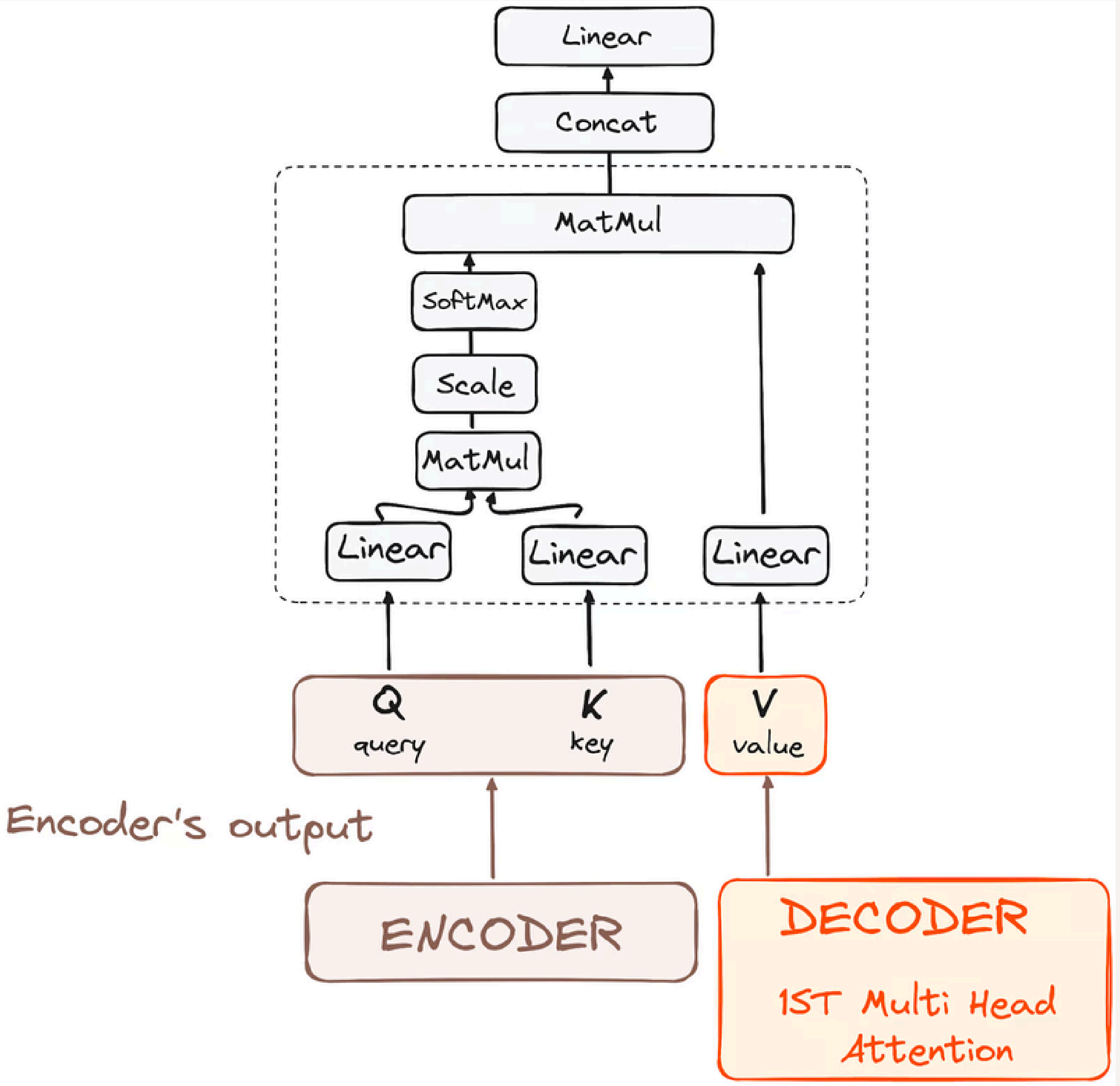
# Transformers

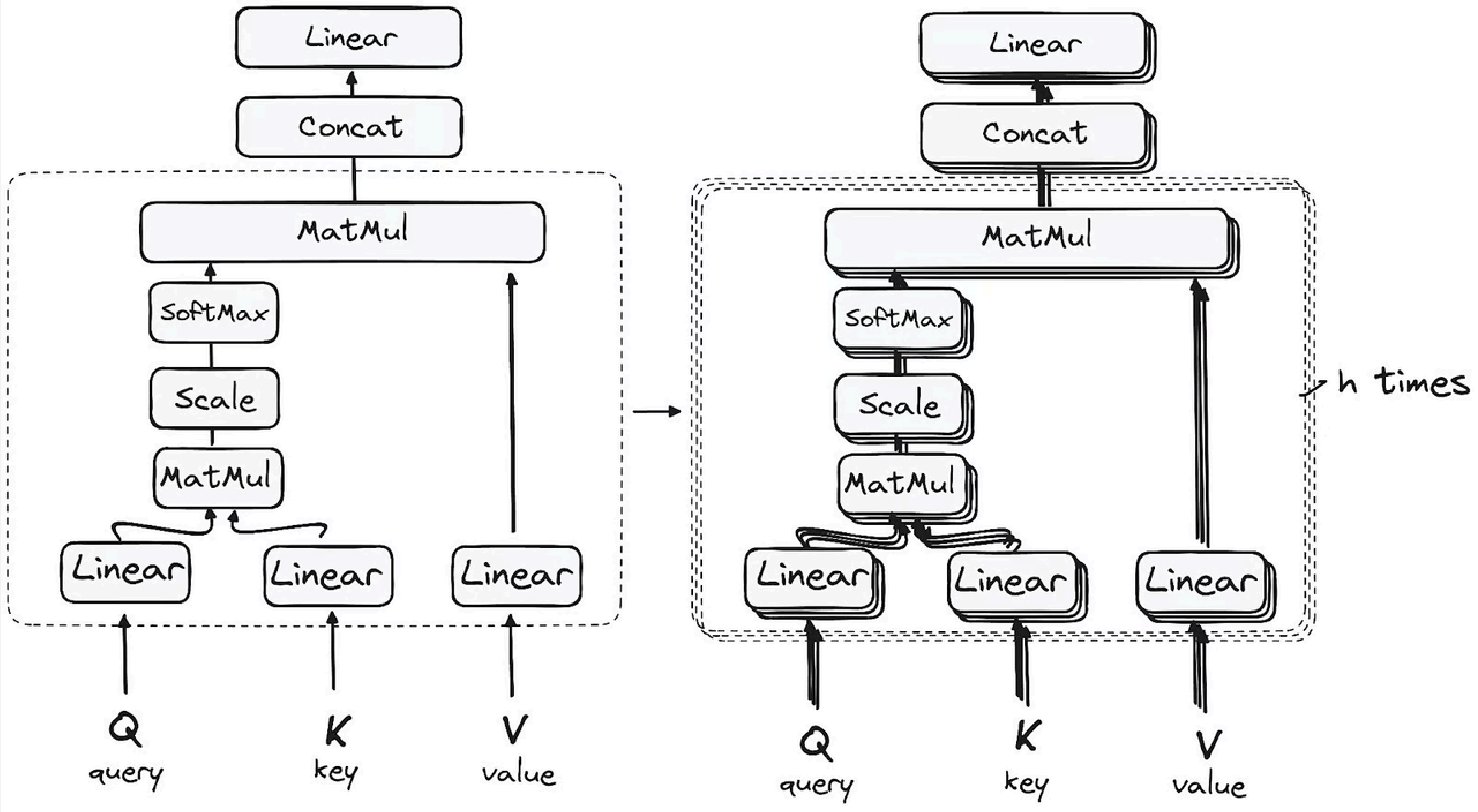


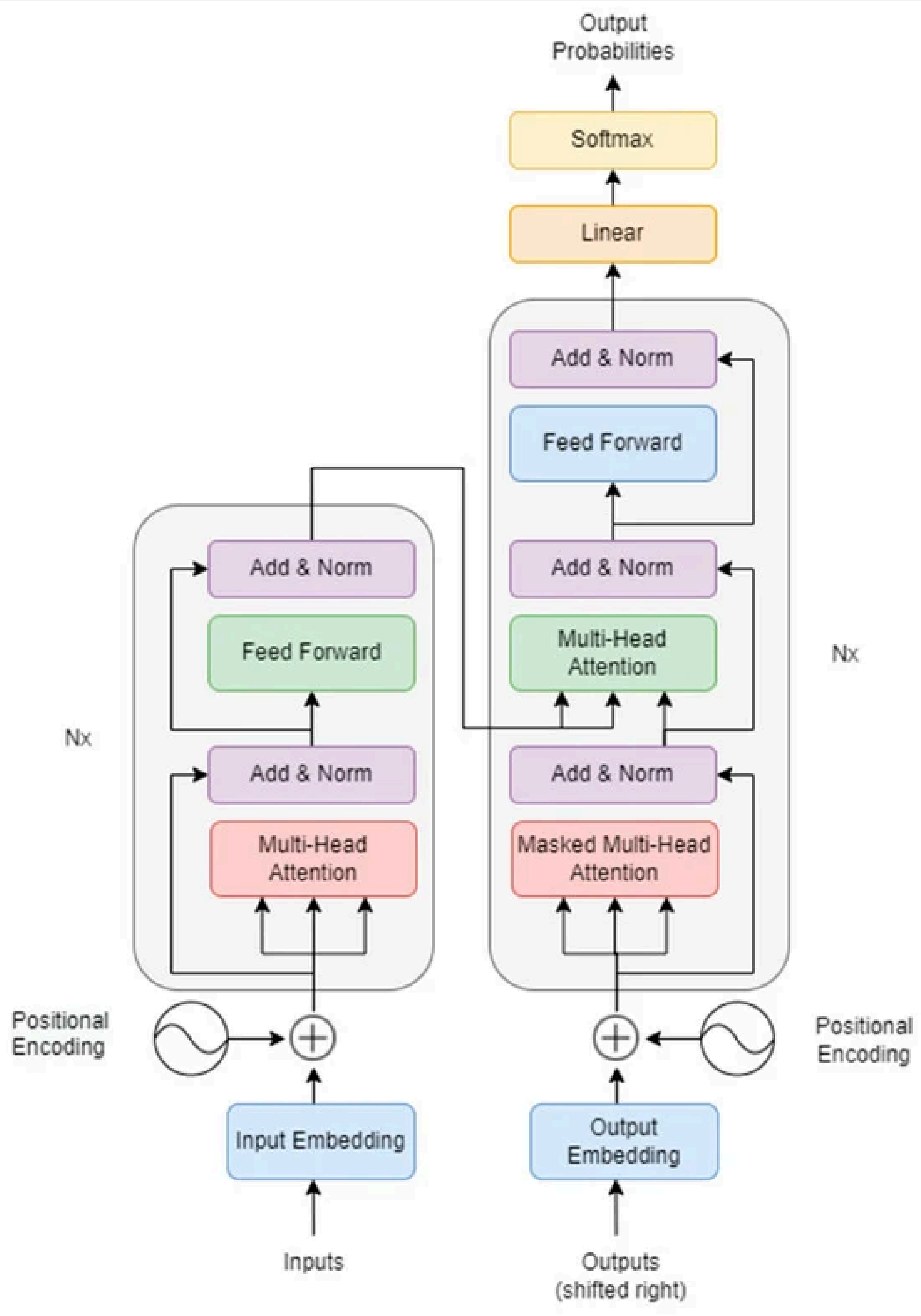


Large Language Models

# MULTI HEADED SELF ATTENTION



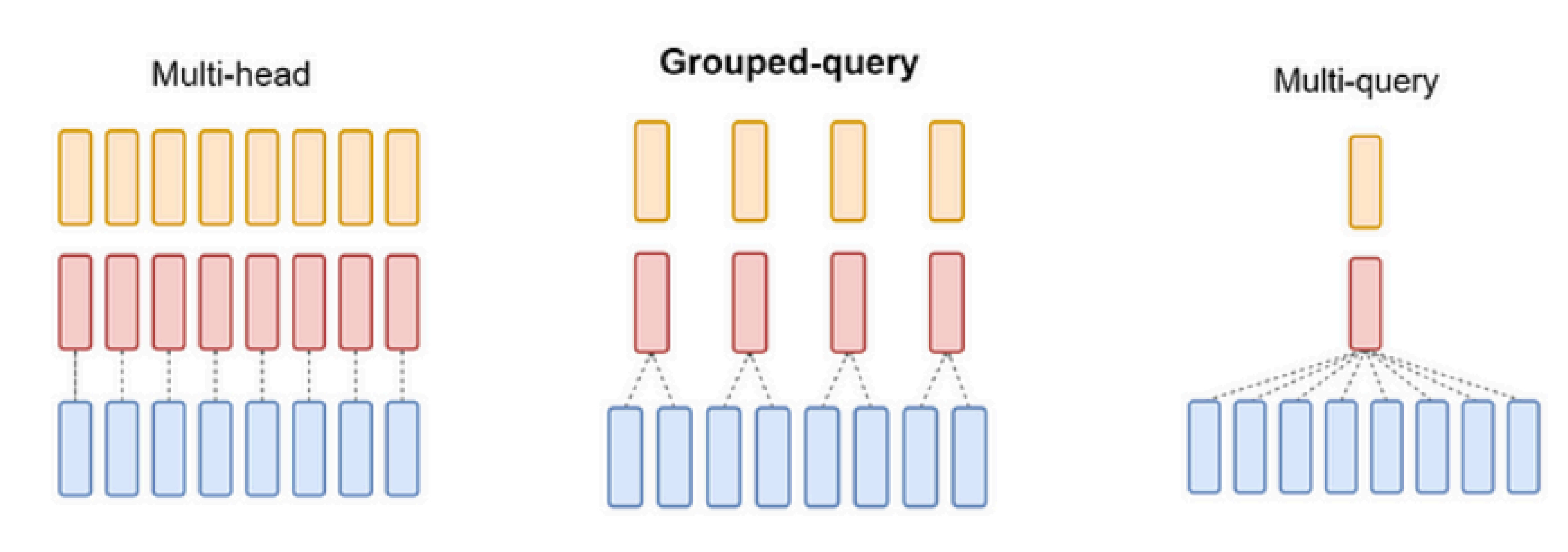


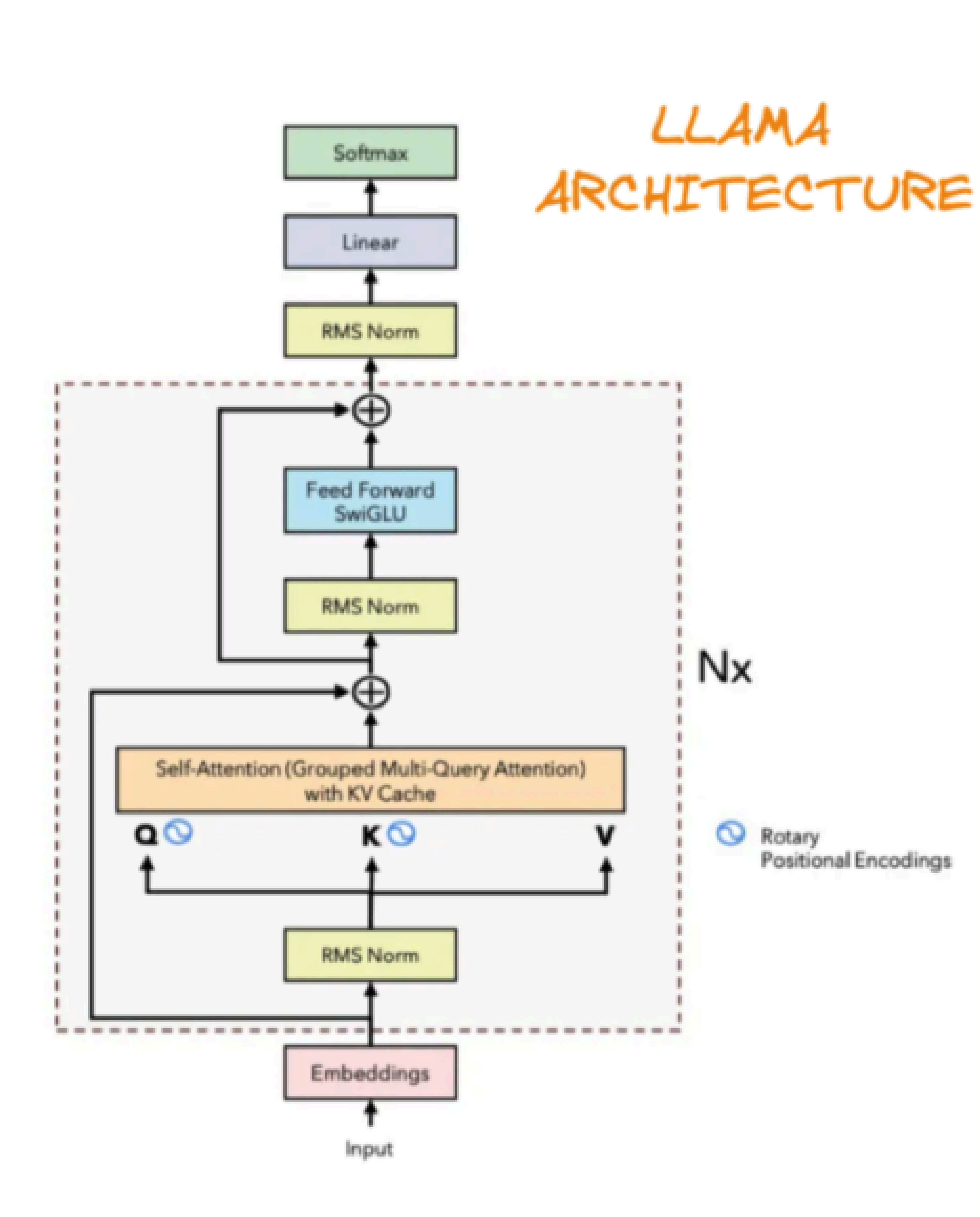

Output
Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked Multi-Head
Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

# Architechture LLaMA 3

Llama 3.3 is an auto-regressive language model that uses an optimized transformer architecture.The tuned versions use SFT and RLHF to align with human preferences for helpfulness and safety.

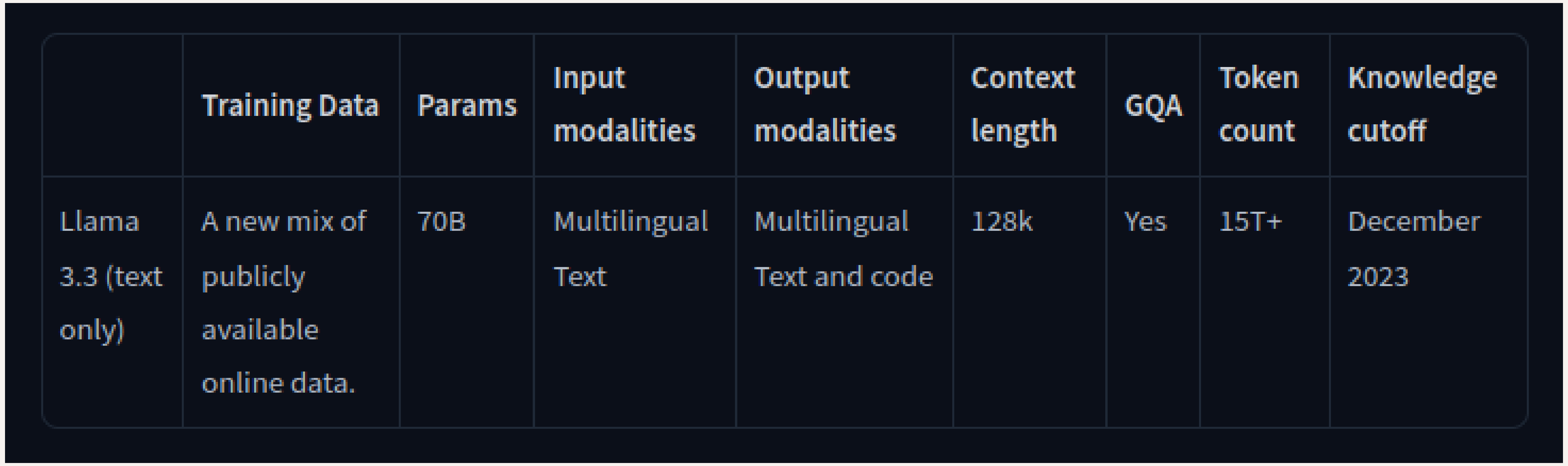
- Input Embeddings
- Rotary Positional Encodings
- RMS Normalization
- **Self-Attention (Grouped Multi-Query Attention with KV Cache)**
- Feed Forward Network (SwigLU)
- RMS Normalization (Post-Attention)
- Layer Stacking (Nx Layers)
- Linear Transformation
- Softmax Output


Multi-head
Grouped-query
Multi-query

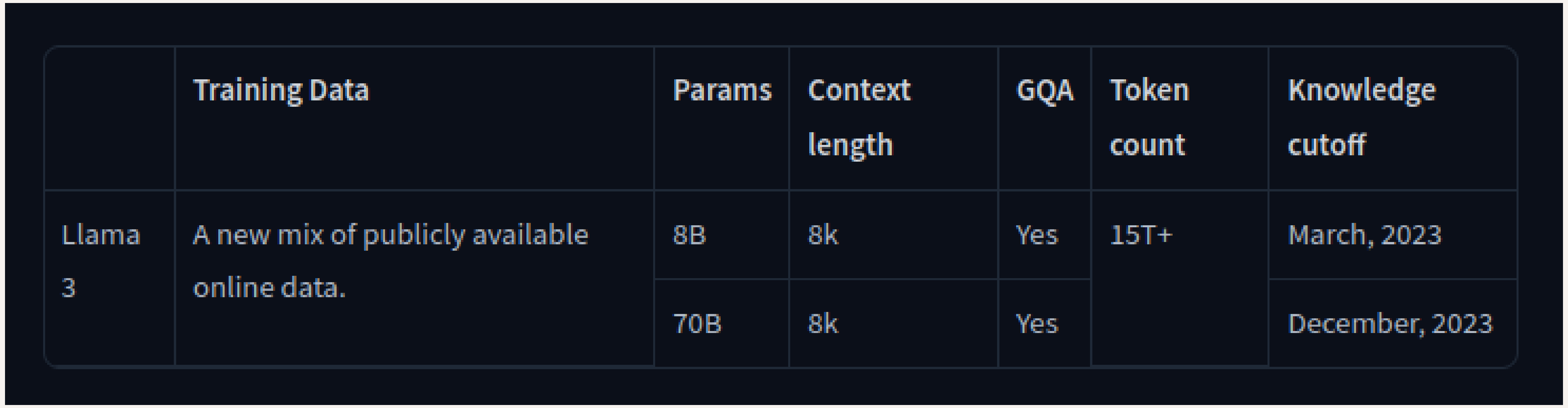


LLAMA ARCHITECTURE

Softmax
Linear
RMS Norm
Feed Forward SwiGLU
RMS Norm
Nx
Self-Attention (Grouped Multi-Query Attention) with KV Cache
Q
K
V
Rotary Positional Encodings
RMS Norm
Embeddings
Input

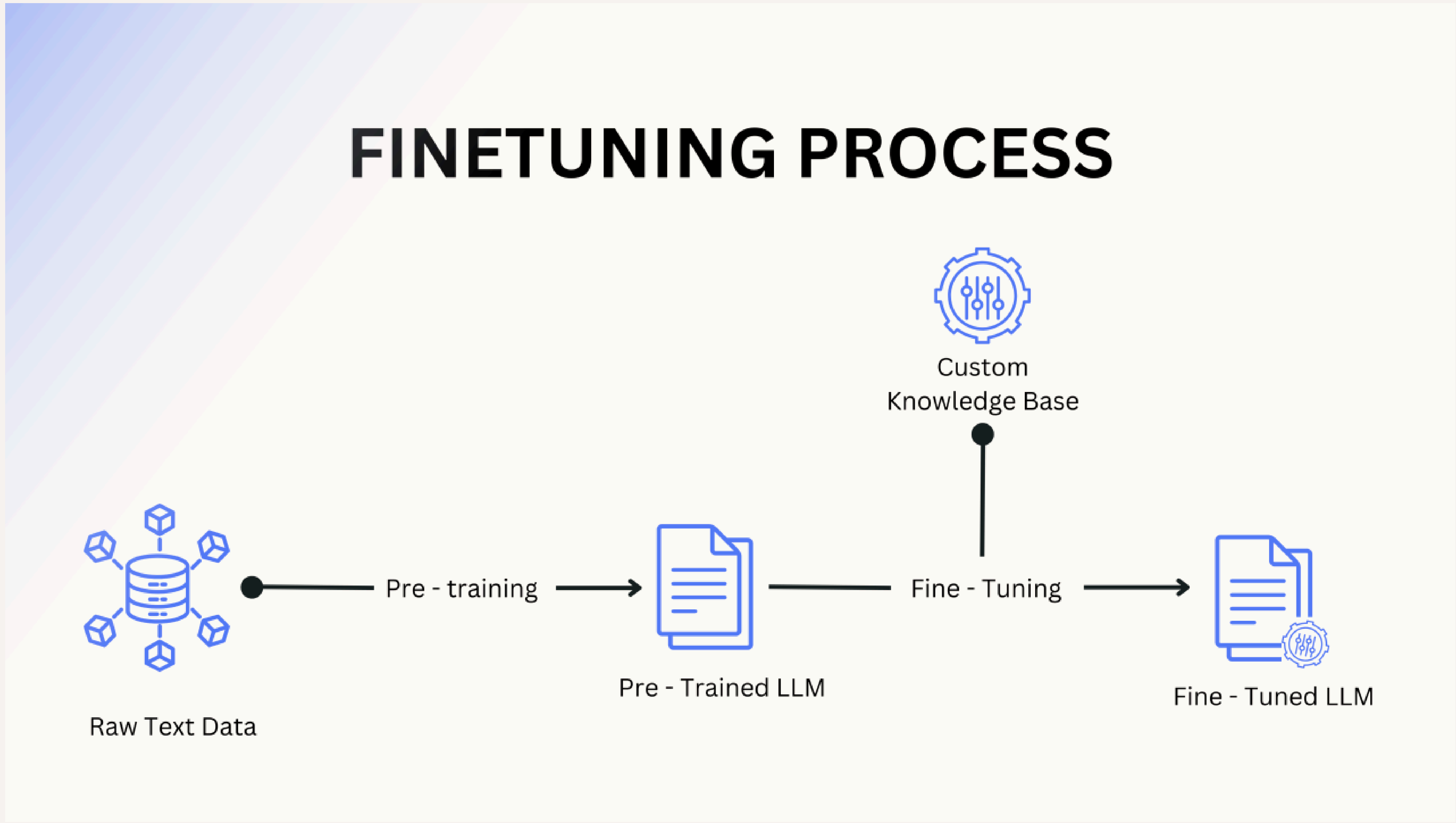


| | Training Data | Params | Input modalities | Output modalities | Context length | GQA | Token count | Knowledge cutoff |
|---|---|---|---|---|---|---|---|---|
| Llama 3.3 (text only) | A new mix of publicly available online data. | 70B | Multilingual Text | Multilingual Text and code | 128k | Yes | 15T+ | December 2023 |



| | Training Data | Params | Context length | GQA | Token count | Knowledge cutoff |
|---|---|---|---|---|---|---|
| Llama 3 | A new mix of publicly available online data. | 8B | 8k | Yes | 15T+ | March, 2023 |
| | | 70B | 8k | Yes | | December, 2023 |

# Fine Tuning Techniques



## Supervised Fine-Tuning (SFT)

- Fine-tune LLaMA 3.3 on high-quality labeled datasets tailored to specific tasks or domains, such as QA pairs or summarization.
- Update model weights using supervised learning to ensure task-relevant and accurate responses.

## Reinforcement Learning with Human Feedback (RLHF)

- Collects model output rankings from human annotators. Trains a reward model based on human preferences.
- Refines the model using Proximal Policy Optimization (PPO).

## Instructive Tuning

- Fine-tune the model using datasets with diverse instruction-output pairs to enhance its ability to follow natural language commands.
- Aligns the model to respond accurately and effectively to human-like instructions.

# Comparision with other llms

| Category | Llama 3.3 | GPT-4o | Mistral 7B |
|---|---|---|---|
| Model SIze | 70 billion parameters | 1.76 trillion parameters | 7 billion parameters |
| Performance | Efficient, good for translation and chatbots | Advanced, better for complex tasks | Optimized for lightweight tasks, fast inference |
| Customization | Open-source, highly customizable | Closed-source, limited customization | Open-source, highly customizable |
| Scalability | Limited scalability, local hardware | Highly scalable, cloud-based | Limited scalability, optimized for small-scale tasks |

# Comparision with other llms

| Category | Llama 3.3 | GPT-4o | Mistral 7B |
|---|---|---|---|
| Multi-Modal Capabilities | No | Yes (text and image inputs) | No |
| Hardware Requirements | Runs on consumer-grade hardware | Requires powerful cloud infrastructure | Runs on consumer-grade hardware |
| Cost Efficiency | More cost-effective | More expensive due to cloud costs | Extremely cost-effective |
| Applications | Content creation, chatbots | coding, Q&A | Lightweight NLP tasks, summarization, coding |

# Lets look into a simple DEMO

# Conclusion

- Significant improvements in natural language understanding and generation.
- Handles complex tasks with reduced computational requirements.
- Paves the way for innovation in AI-driven solutions.

# Thank you

# SHE LAW AI

Empowering Women Groups

Guide: Unnikrishnan Radhakrishnan, PhD