Module 02

Content By: Raghav Bali

Module 02

Building Blocks of Large Language Models

Content By: Raghav Bali

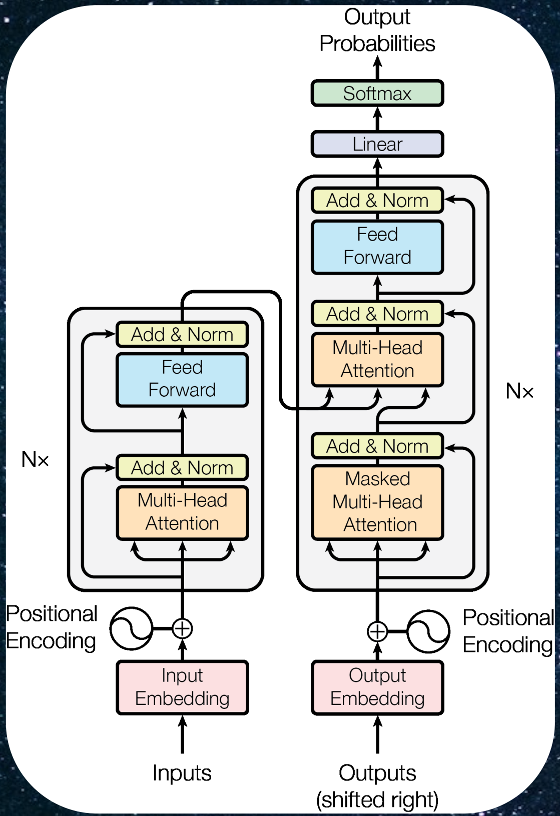
Agenda

- Transformer Architectures
- Evaluation and Benchmarks
- Evolution of LM to LLMs

Content By: Raghav Bali

Quick Recap?

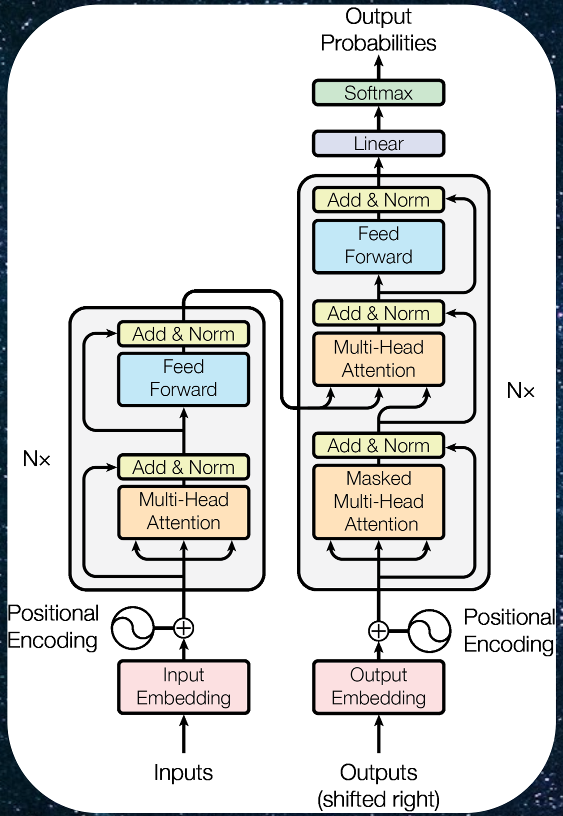
Transformers



Multi-Head Self-Attention

Self-attention mechanism allows the model to weigh the importance of different words in a sentence relative to each other while Multiple-attention heads allow the model to learn multiple features/concepts from different representation subspaces.

Transformers

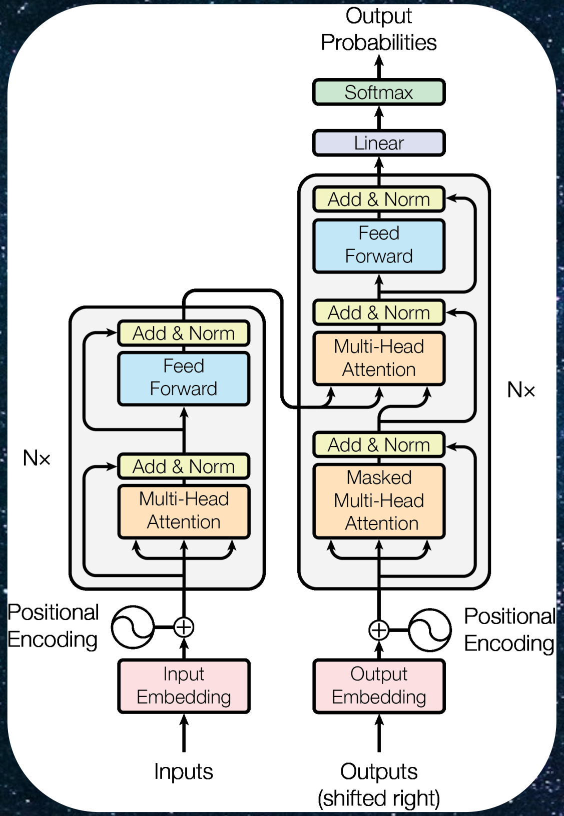


Multi-Head Self-Attention

Positional Encoding

Positional encodings enable the model to maintain sequence information, crucial for tasks where word order matters.

Transformers



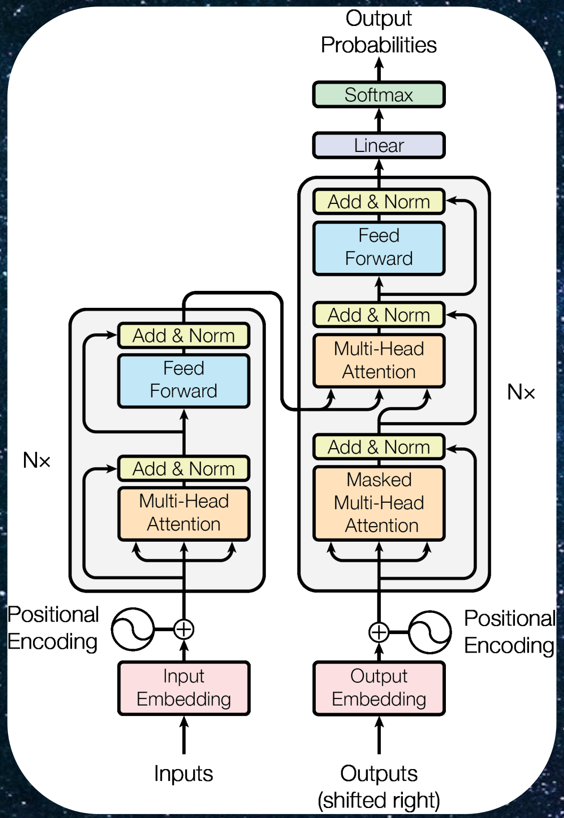
Multi-Head Self-Attention

Positional Encoding

Layer Normalization and Residual Connections

Normalization and Residual Connections were already known effective techniques but the transformer architecture makes use of these concepts within each encoder/decoder block allowing for stable and efficient training.

Transformers



Multi-Head Self-Attention

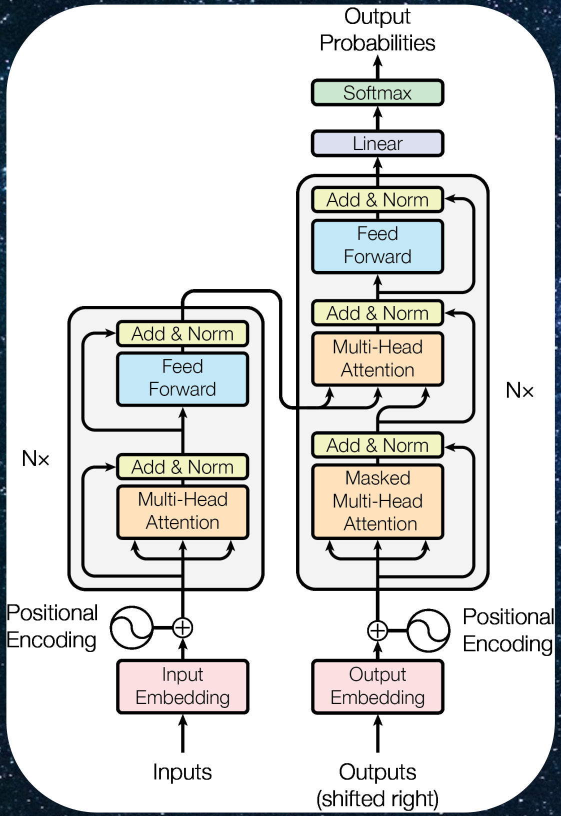
Positional Encoding

Layer Normalization and Residual Connections

Stacked Encoder-Decoder Architecture

The stacked nature of both encoder and decoder components allows transformers to capture and process complex interaction features from the entire input sequence

Transformers



Multi-Head Self-Attention

Positional Encoding

Layer Normalization and Residual Connections

Stacked Encoder-Decoder Architecture

Content By: Raghav Bali

Transformer Architectures

Content By: Raghav Bali

Transformer Architectures

Encoder-Decoder Architectures

- Google T5
- Transformer-XL
- BART

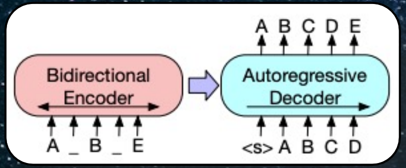
Encoder-Only Architectures

- BERT
- ELECTRA
- ALBERT

Decoder-Only Architectures

- GPT-x
- Chinchilla
- LLaMA

Encoder-Decoder Architectures



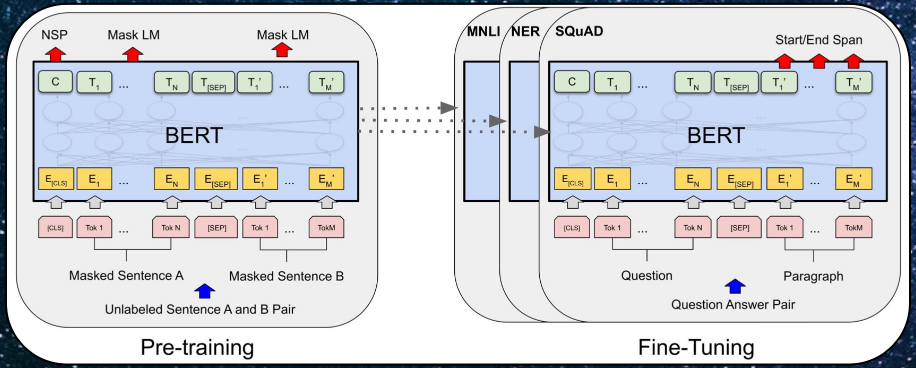
BART

Key Highlights:

- T5 frames all NLP tasks as a text-to-text problem
- Transformer-XL extended context length limitations of earlier models
- BART presents a bi-directional encoder coupled with a autoregressive decoder.
- These models are effective for various NLP tasks

Encoder-Only Architectures

BERT



Key Highlights:

- Designed for NLP tasks involving understanding and representation learning.
- Pre-trained on large datasets the fine-tuned for specific tasks.
- Training objective during pre-training is Masked Language Modeling

Decoder-Only Architectures

LLaMA

GPT

Gemma

Key Contributions:

- Pretrained in unsupervised fashion with autoregressive objective of predicting next token.
- Easily fine-tuned for NLP tasks for classification, translation using different heads.
- Revolutionized the NLP space

LLM Evaluation & Benchmarks

Content By: Raghav Bali

LLM Evaluation Metrics

Traditional Metrics

- F1 Score
- Accuracy

Task Specific Metrics

- Fluency: Perplexity
- Translation/Summarization: BLUE, ROUGE
- Question Answering: Exact Match
- Robustness: Adversarial Testing

LLM Evaluation Metrics

Perplexity

- Well defined for autoregressive models
- Defined as exponentiated average negative log-likelihood of a sequence

  OR
- a measurement of how well a probability model predicts a sample.
- Lower is better, ranges from [0,inf)

Hugging Face is a startup based in New York City and Paris

p(word)

LLM Evaluation Metrics

BLEU & ROUGE

- BLEU: Bilingual Evaluation Understudy
- Evaluate translation quality by comparing generated text to reference
- Calculates precision at different n-gram lengths
- Penalizes shorter translations

- ROUGE: Recall Oriented Understudy for Gisting Evaluation
- Evaluate summary quality by comparing generated text to reference
- Case insensitive metric
- Penalizes shorter translations

LLM Benchmarks

Task Specific Metrics

- GLUE: generalization and understanding capabilities
- SuperGLUE: more challenging tasks for assessing language understanding
- SQuAD: reading comprehension and question answering
- XLNI: multi-lingual language inference
- OpenLLM Leaderboard:
- MTEB: text embedding benchmark
- LMSys Chatbot Arena: human voting based Elo ratings
- LLMPerf: latency and throughput benchmarks

Evolution of LMs to LLMs

Content By: Raghav Bali

Evolution of LMs to LLMs

GPT-2

Content By: Raghav Bali

Evolution of LMs to LLMs

GPT-2

Content By: Raghav Bali

Evolution of LMs to LLMs

1.5 Billion Parameters

40GB Training Data

GPT-2

Evolution of LMs to LLMs

1.5 Billion Parameters

40GB Training Data

GPT-2

Evolution of LMs to LLMs

1.5 Billion Parameters

40GB Training Data

GPT-2

GPT-3

Content By: Raghav Bali

Evolution of LMs to LLMs

1.5 Billion Parameters

40GB Training Data

GPT-2

GPT-3

Content By: Raghav Bali

Evolution of LMs to LLMs

GPT-2

1.5 Billion Parameters

40GB Training Data

GPT-3

175 Billion Parameters

570GB Training Data

[Few,1,0]-Shot Capabilities

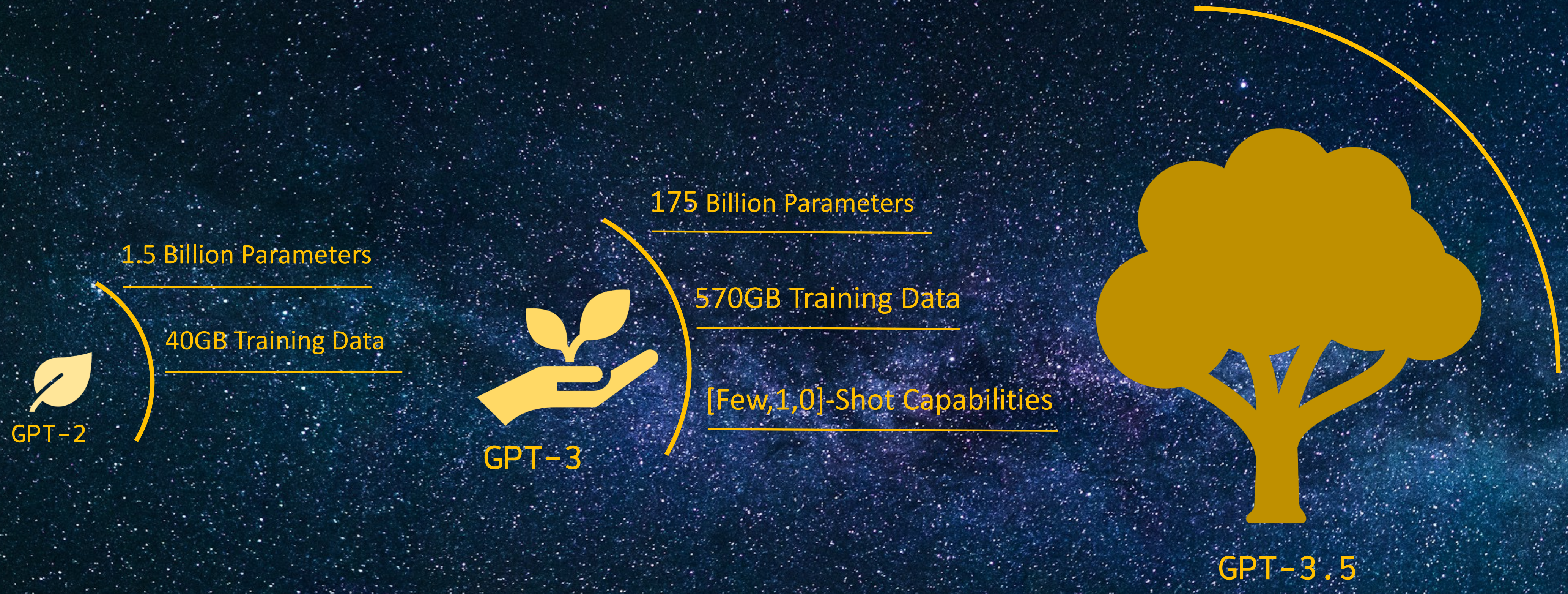
Evolution of LMs to LLMs

GPT-2
- 1.5 Billion Parameters
- 40GB Training Data

GPT-3
- 175 Billion Parameters
- 570GB Training Data
- [Few,1,0]-Shot Capabilities

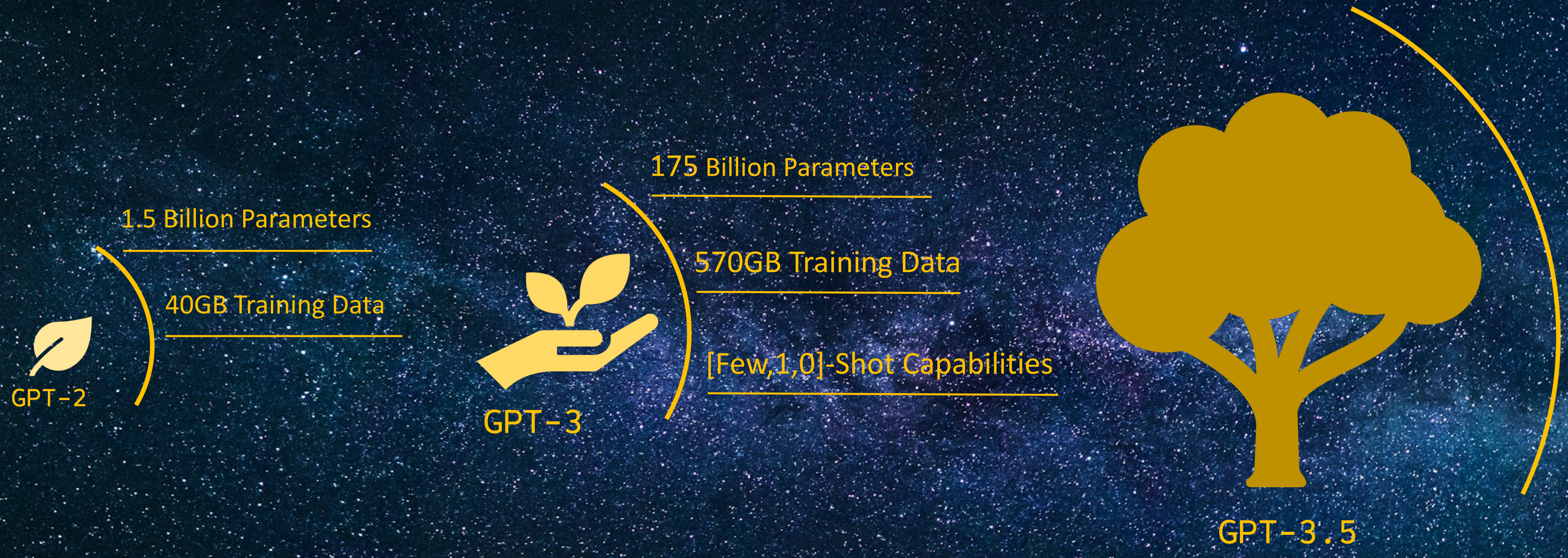
Content By: Raghav Bali

Evolution of LMs to LLMs



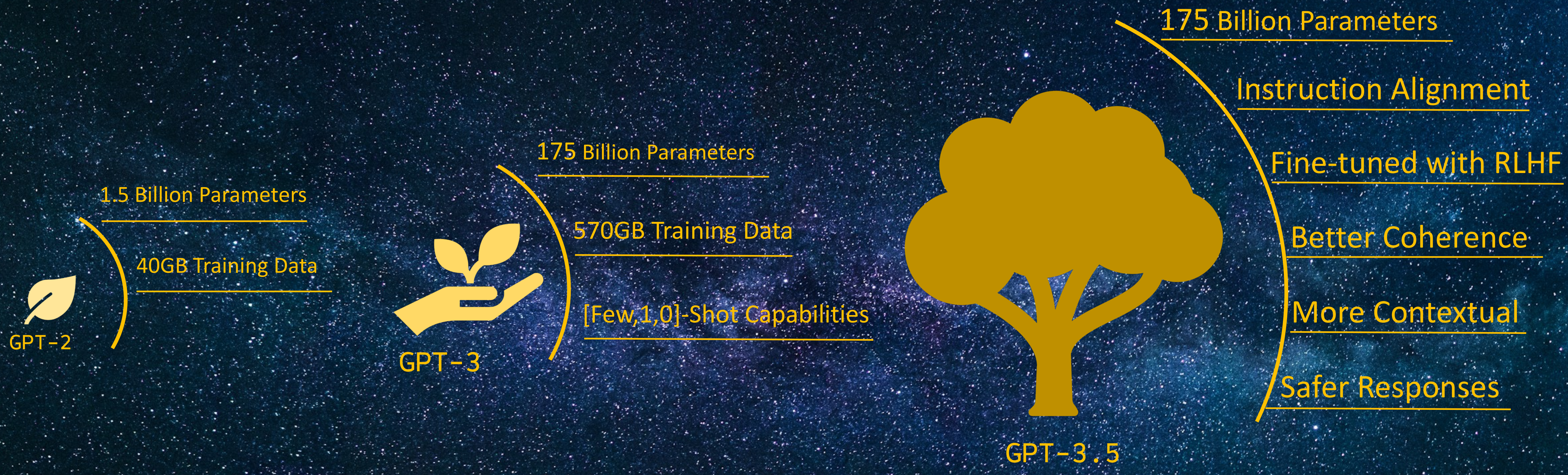
Content By: Raghav Bali

Evolution of LMs to LLMs



Content By: Raghav Bali

Evolution of LMs to LLMs



Content By: Raghav Bali

Evolution of LMs to LLMs

GPT-2

GPT-3

GPT-3.5

Content By: Raghav Bali

Evolution of LMs to LLMs

Pretraining

GPT-2

Large Training Dataset
From internet

GPT-3

GPT-3.5

Extremely Large
Training Dataset
from Internet

Training Objective

Language Modeling

Evolution of LMs to LLMs

	Pretraining	Supervised Fine-Tuning
GPT-2	Large Training Dataset From internet	Task Specific Datasets for fine-tuning
GPT-3	Extremely Large Training Dataset from Internet	Larger Task Specific Datasets for fine-tuning
GPT-3.5		Usual SFT + Ideal Assistant Responses (Prompt, Response)
Training Objective	Language Modeling	Language Modeling

Content By: Raghav Bali

Evolution of LMs to LLMs

	Pretraining	Supervised Fine-Tuning	Reward Modeling/RLHF
GPT-2	Large Training Dataset From internet	Task Specific Datasets for fine-tuning	
GPT-3	Extremely Large Training Dataset from Internet	Larger Task Specific Datasets for fine-tuning	
GPT-3.5		Usual SFT + Ideal Assistant Responses (Prompt, Response)	Response Alignment through response comparison datasets
Training Objective	Language Modeling	Language Modeling	Binary Classification/ Reinforcement Learning

Content By: Raghav Bali

Hands-On

Let Us Tune Some GPT!