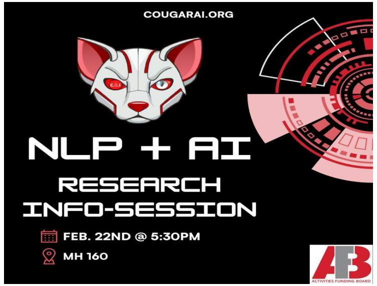
LLM with NLP: The science behind the hype

Sadat Shahriar,
University of Houston



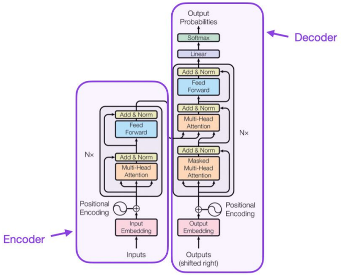
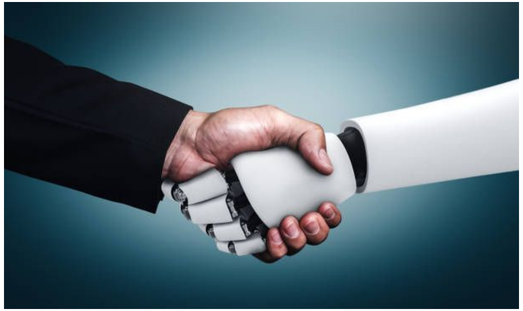
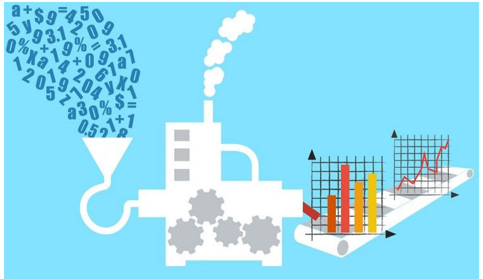


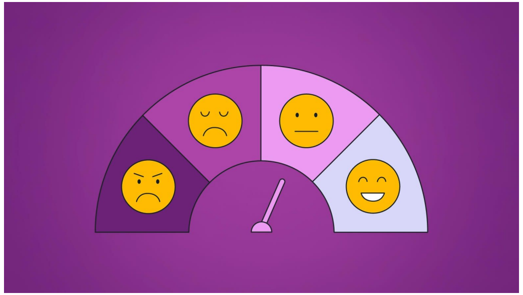
Figure 1: The Transformer - model architecture.

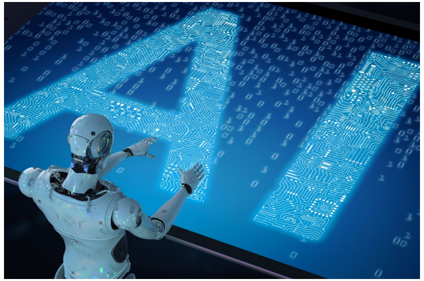
Why Natural Language Processing so important

Why Natural Language Processing so important contd…









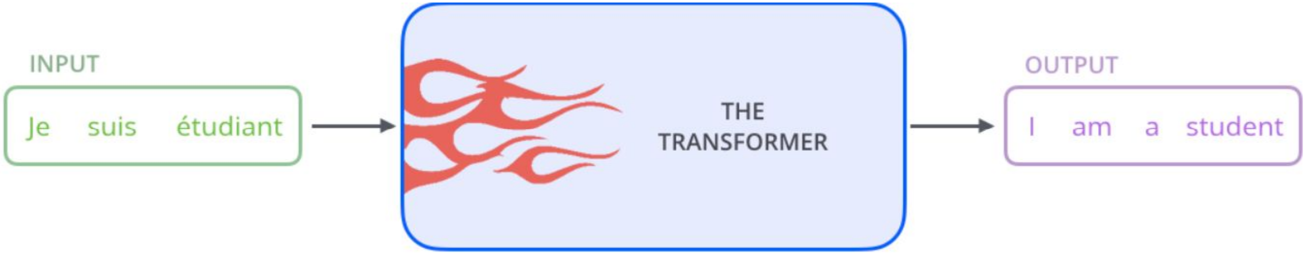
Research in NLP

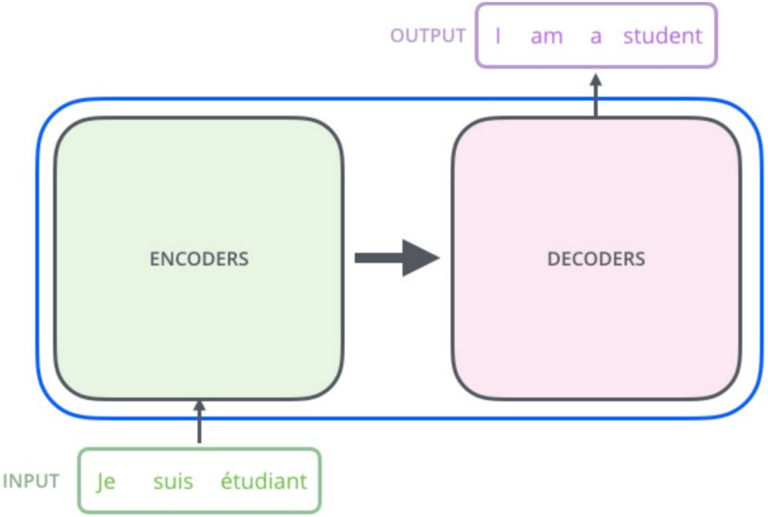
- **Computational Social Science and Cultural Analytics**
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- **Generation**
- **Information Extraction**
- Information Retrieval and Text Mining
- **Interpretability and Analysis of Models for NLP**
- Language Grounding to Vision, Robotics and Beyond
- **Large Language Models**
- Linguistic Diversity
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- **Machine Learning for NLP**

Research in NLP contd…

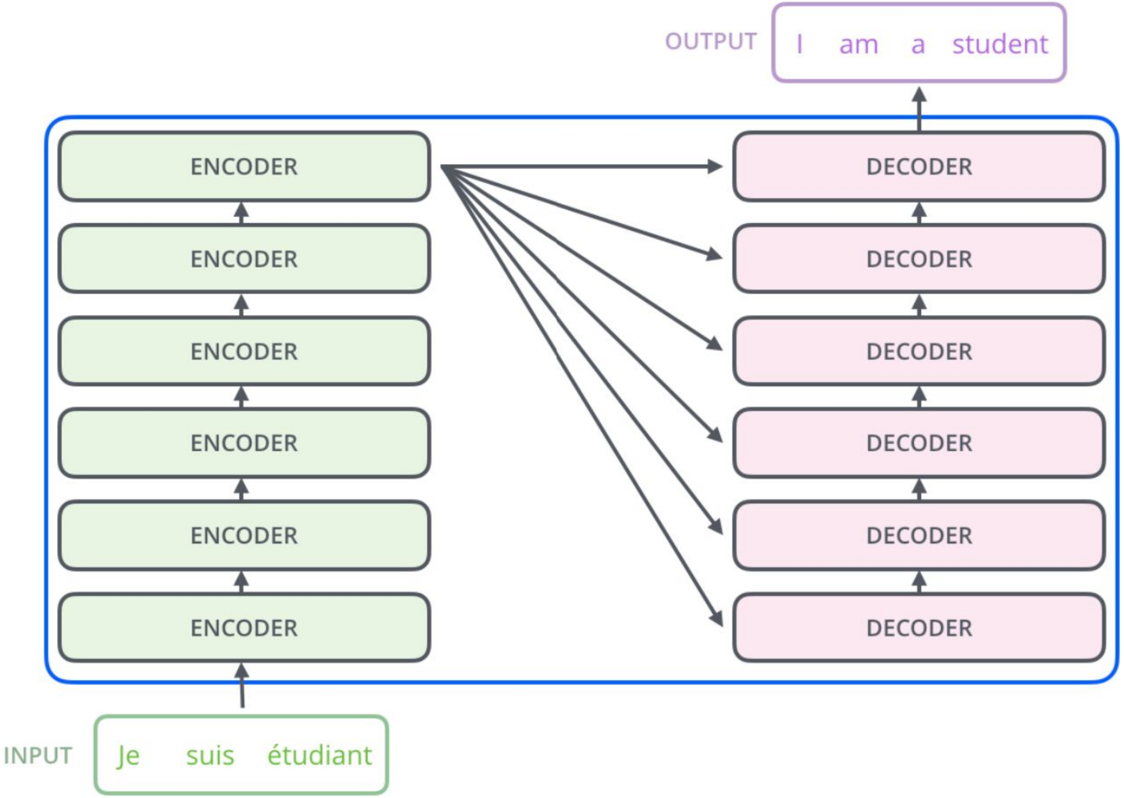
- **Machine Translation**
- **Multilingualism and Cross-Lingual NLP**
- **NLP Applications**
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- **Speech and Multimodality**
- Summarization
- Syntax: Tagging, Chunking and Parsing

Transformer: The architectural father (most slides taken from jalammar blog)

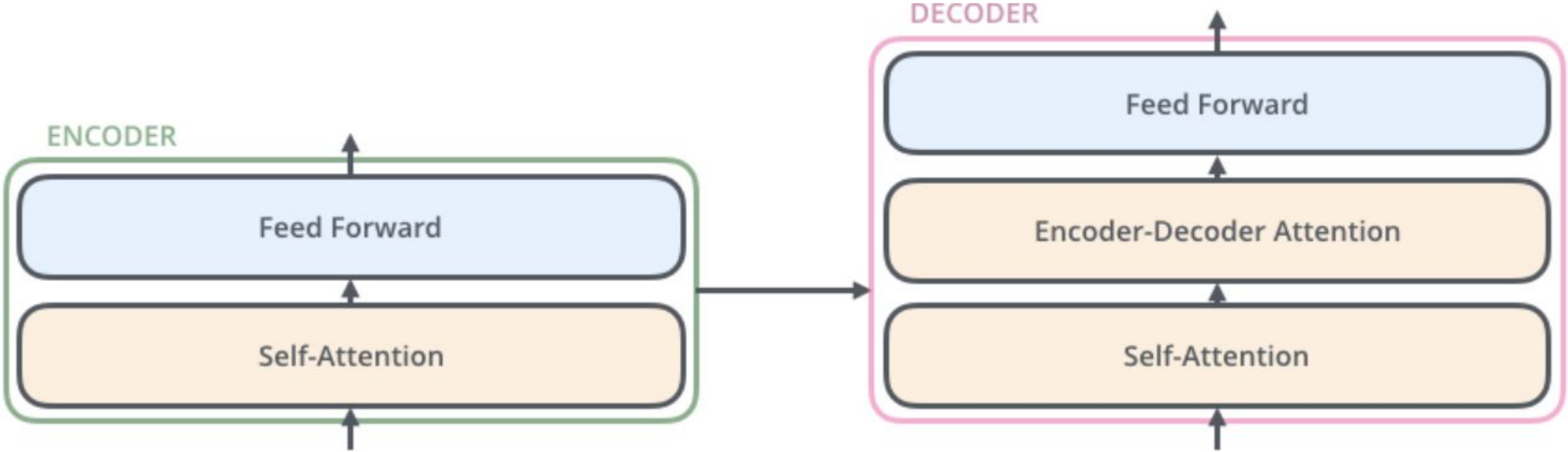




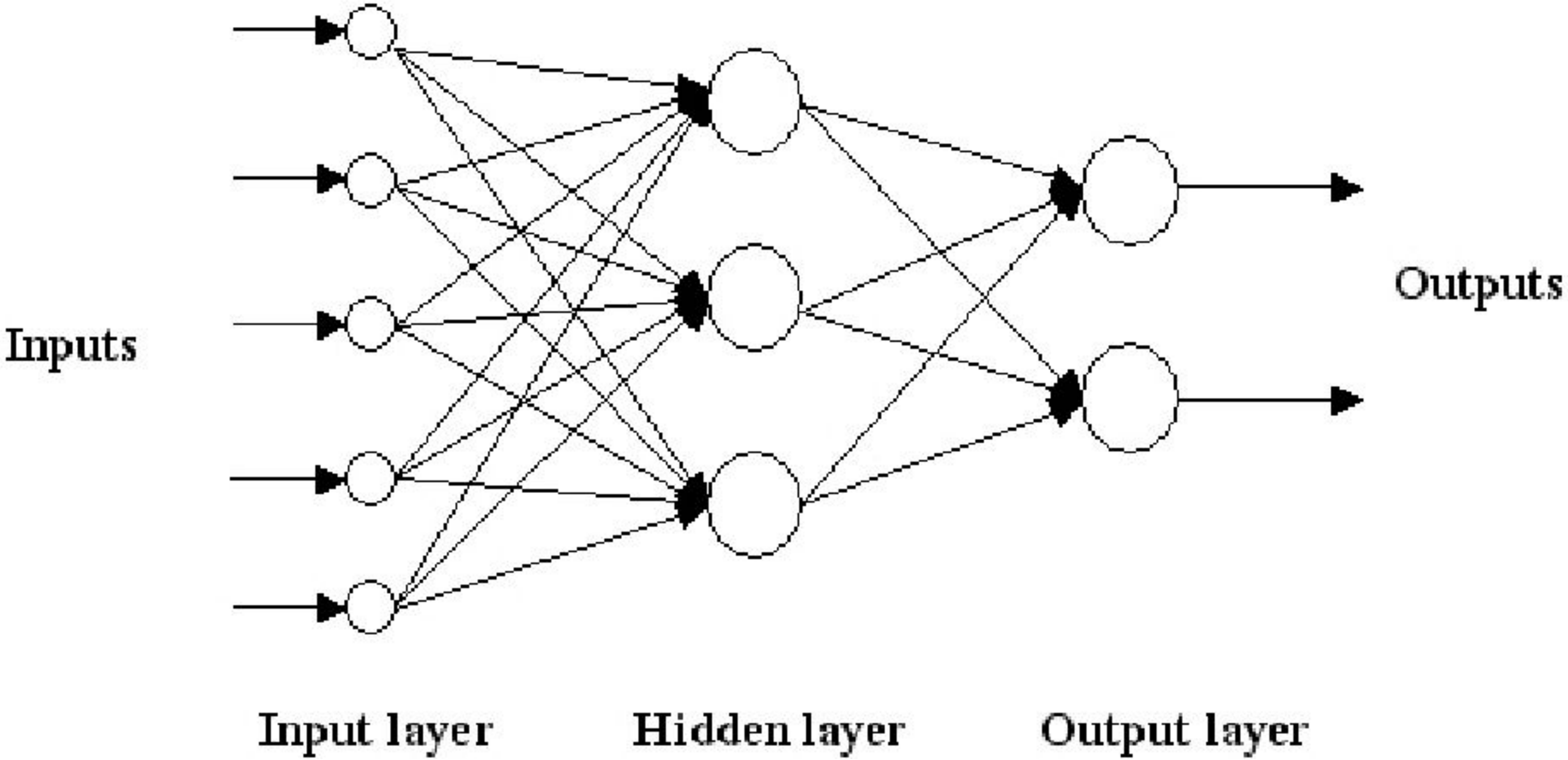
Transformer



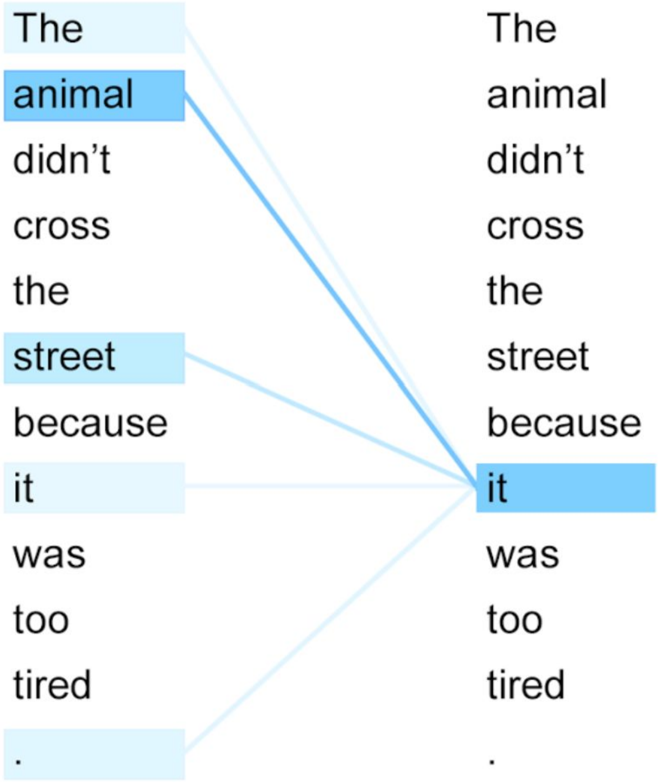
Transformer



Feed Forward neural network

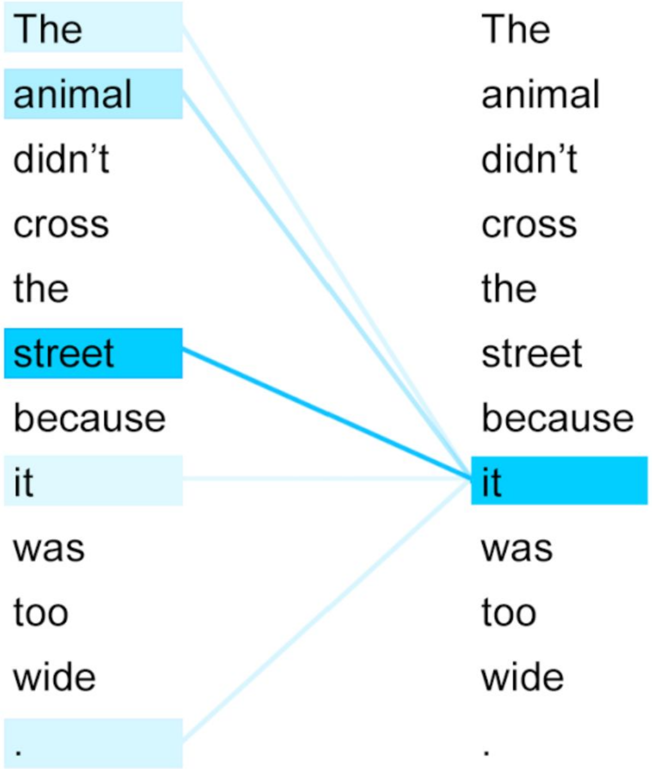


What is an attention?



The
animal
didn’t
cross
the
street
because
it
was
too
tired
.

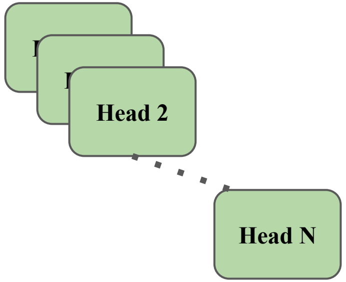
The
animal
didn’t
cross
the
street
because
it
was
too
tired
.



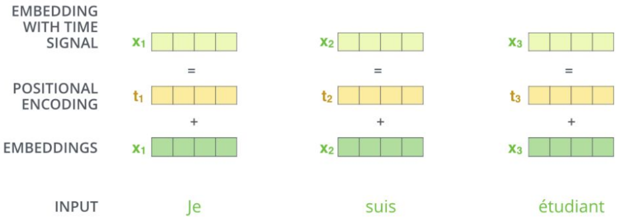
The
animal
didn’t
cross
the
street
because
it
was
too
wide
.

The
animal
didn’t
cross
the
street
because
it
was
too
wide
.

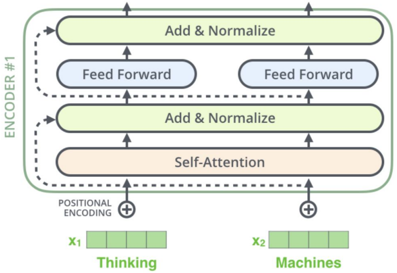
Multi-head



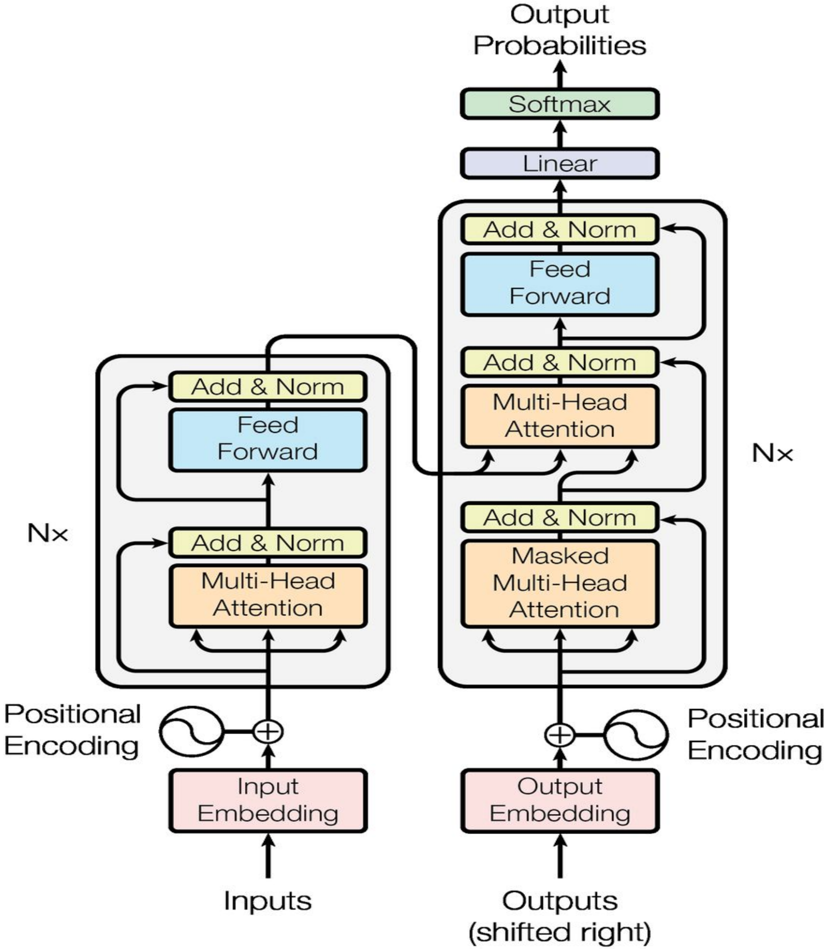
Positional Encoding



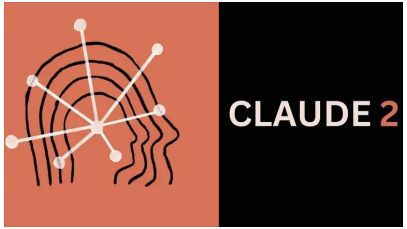
Residual Connection



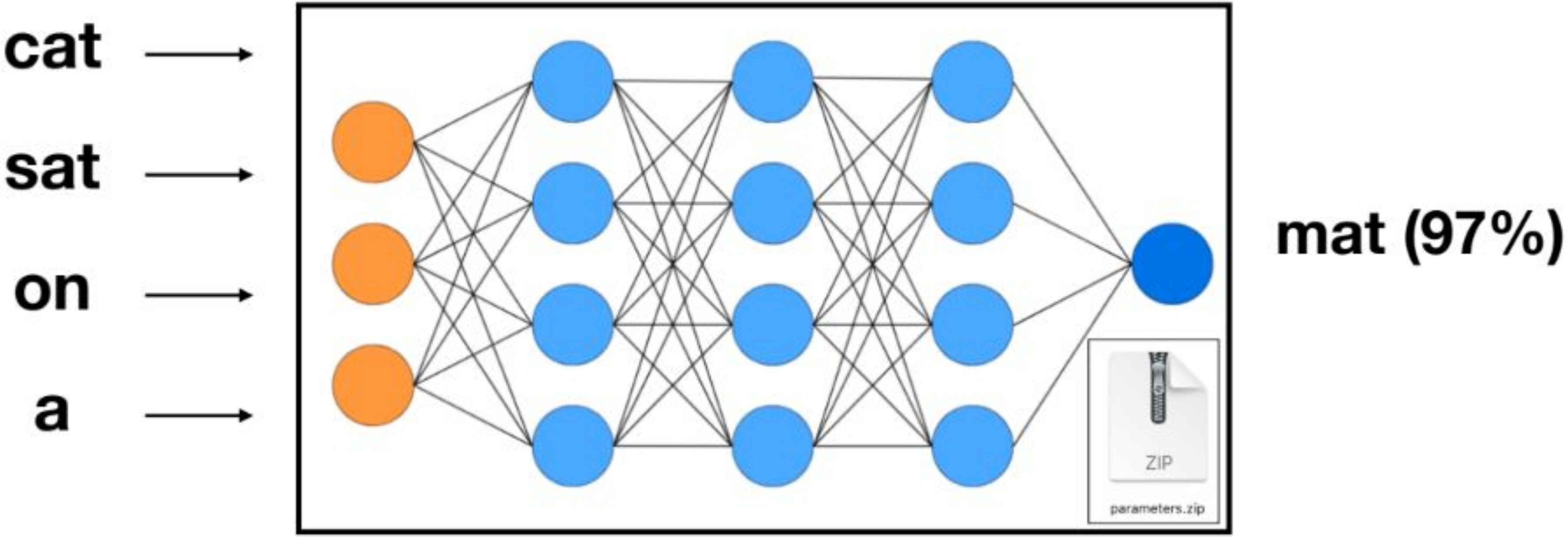
The Transformer



Large Language Models



The Training Process: the next word prediction



cat

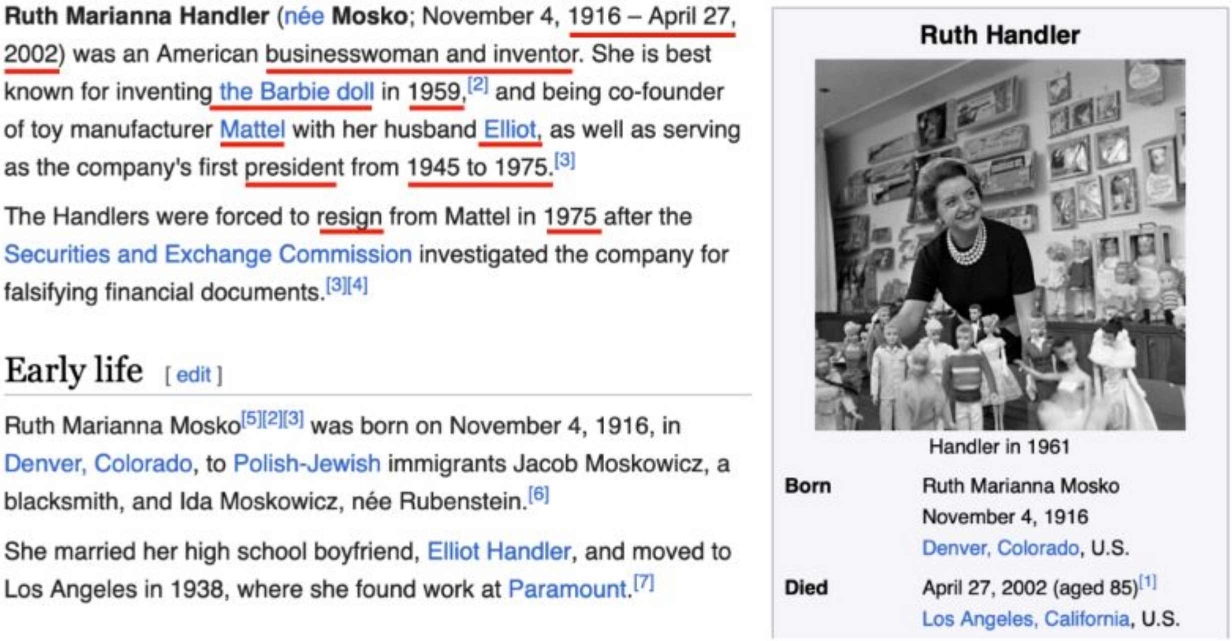
sat

on

a

mat (97%)

Next word prediction forces the neural network to learn a lot about the world:



Ruth Marianna Handler (née **Mosko**; November 4, 1916 – April 27, 2002) was an American businesswoman and inventor. She is best known for inventing the Barbie doll in 1959,[2] and being co-founder of toy manufacturer Mattel with her husband Elliot, as well as serving as the company's first president from 1945 to 1975.[3]

The Handlers were forced to resign from Mattel in 1975 after the Securities and Exchange Commission investigated the company for falsifying financial documents.[3][4]

Early life [edit]

Ruth Marianna Mosko[5][2][3] was born on November 4, 1916, in Denver, Colorado, to Polish-Jewish immigrants Jacob Moskowicz, a blacksmith, and Ida Moskowicz, née Rubenstein.[6]

She married her high school boyfriend, Elliot Handler, and moved to Los Angeles in 1938, where she found work at Paramount.[7]

Ruth Handler

Handler in 1961

Born	Ruth Marianna Mosko November 4, 1916 Denver, Colorado, U.S.
Died	April 27, 2002 (aged 85)[1] Los Angeles, California, U.S.

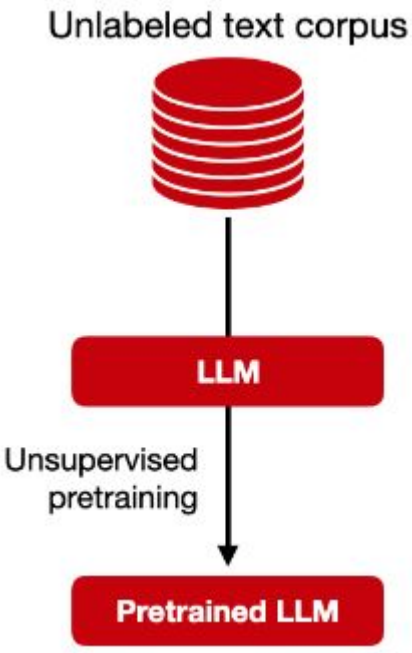
Thanks to Andrej Karpathy

But why is it so?

- We donot fully know
- Billions of parameters collaborate to come up with the next word, how would you visualize that?
- A rough idea: NSP helps in a better contextual understanding, and finding relational dependencies, in turn helping in semantic understanding

The pretraining

What is the capital of France?



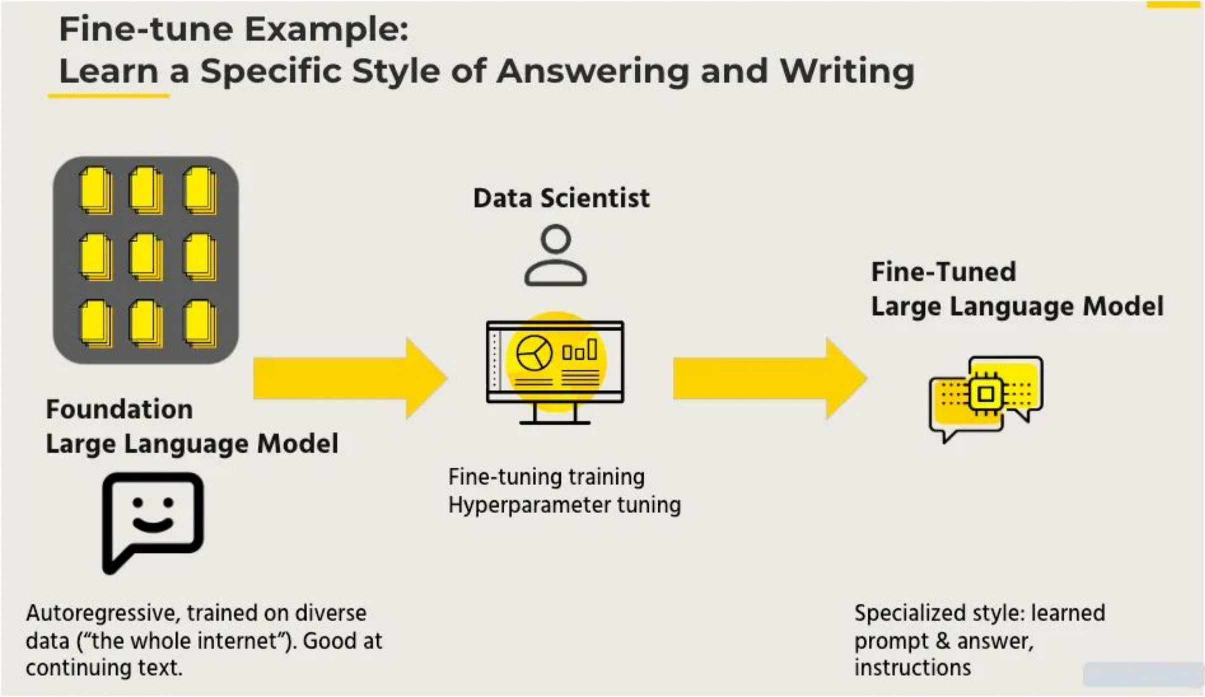
What is the capital of Germany?

What is the capital of Belgium?

Who is the president of France?

Monalisa is in the Louvre museum

Fine Tuning



Fine-tune Example:
Learn a Specific Style of Answering and Writing

Foundation
Large Language Model

Autoregressive, trained on diverse data ("the whole internet"). Good at continuing text.

Data Scientist

Fine-tuning training
Hyperparameter tuning

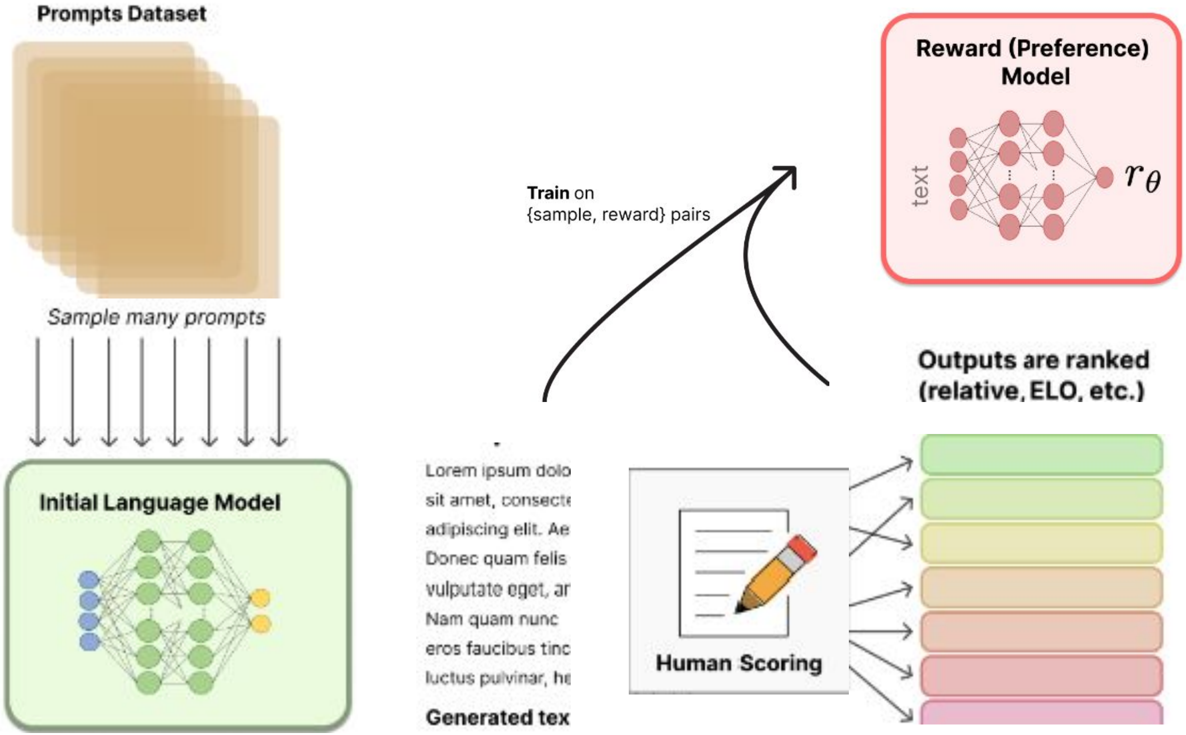
Fine-Tuned
Large Language Model

Specialized style: learned prompt & answer, instructions

Thanks to https://www.labellerr.com/blog/comprehensive-guide-for-fine-tuning-of-llms/

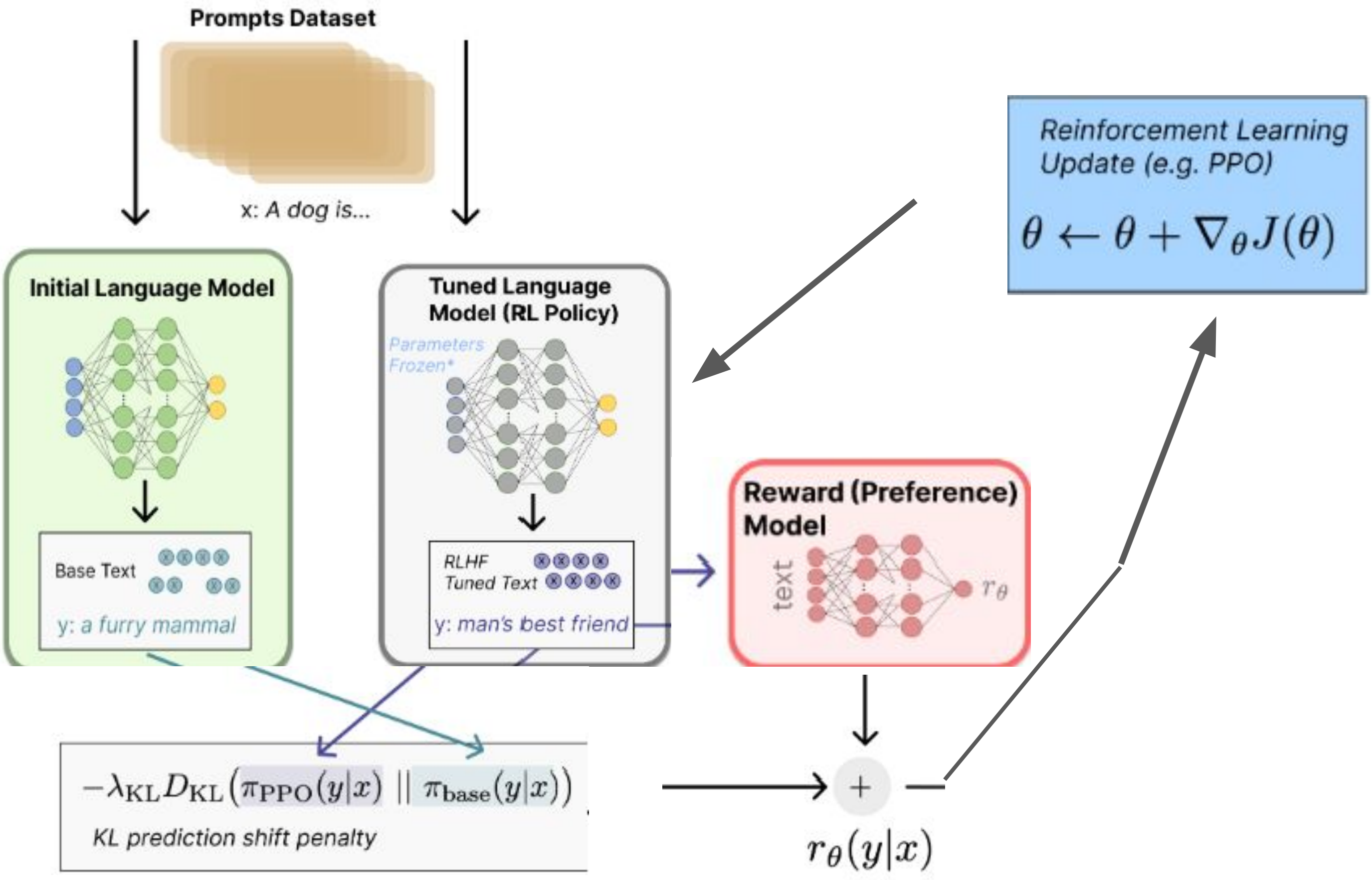
Reinforcement Learning with Human Feedback (RLHF)

Part 1: Reward Modeling

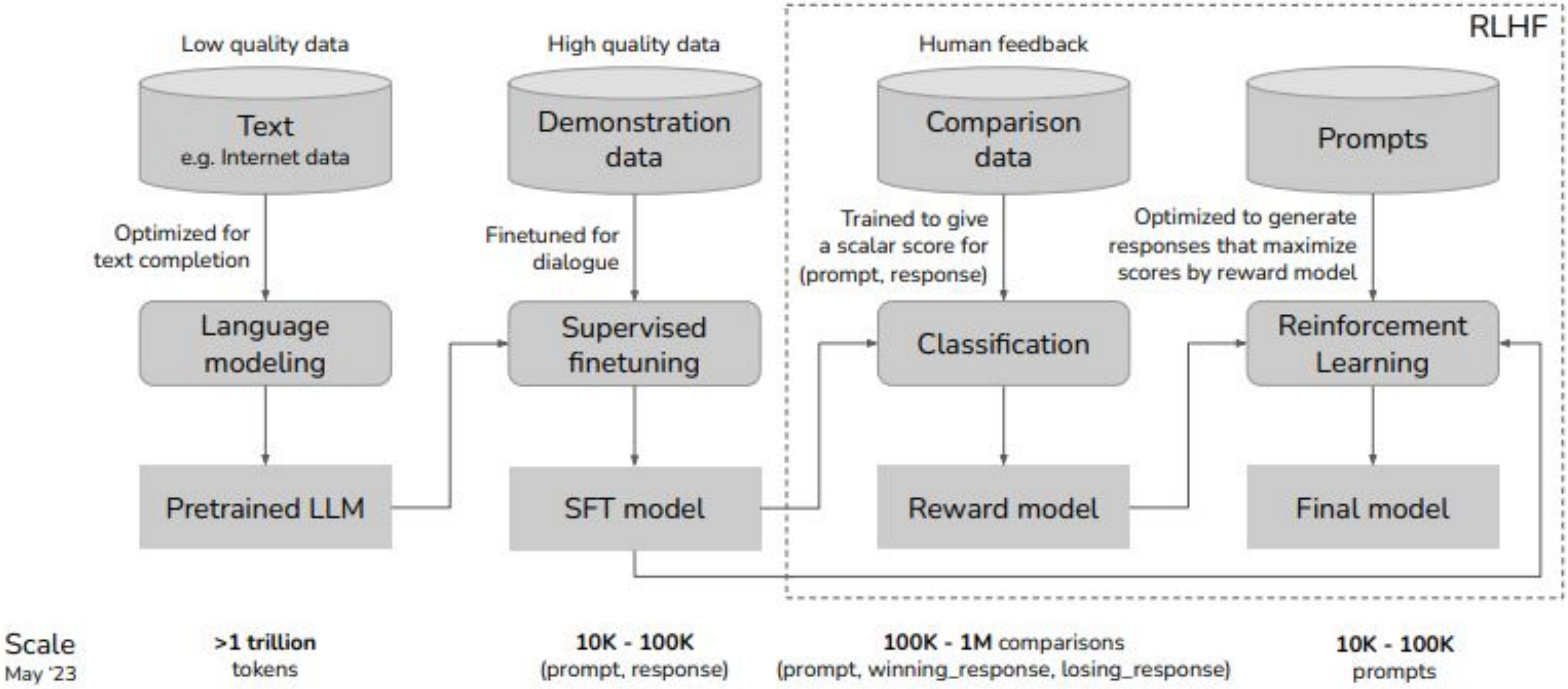


Reinforcement Learning with Human Feedback (RLHF)

Part 2: Policy Optimization


Prompts Dataset
x: A dog is...
Initial Language Model
Base Text
y: a furry mammal
Tuned Language Model (RL Policy)
Parameters Frozen*
RLHF Tuned Text
y: man's best friend
Reward (Preference) Model
text
r_θ
Reinforcement Learning Update (e.g. PPO)
$\theta \leftarrow \theta + \nabla_\theta J(\theta)$
$-\lambda_{\text{KL}} D_{\text{KL}} (\pi_{\text{PPO}}(y|x) \;||\; \pi_{\text{base}}(y|x))$
KL prediction shift penalty
+
$r_\theta(y|x)$

Summary of Training



RLHF

Low quality data

Text
e.g. Internet data

Optimized for text completion

Language modeling

Pretrained LLM

High quality data

Demonstration data

Finetuned for dialogue

Supervised finetuning

SFT model

Human feedback

Comparison data

Trained to give a scalar score for (prompt, response)

Classification

Reward model

Prompts

Optimized to generate responses that maximize scores by reward model

Reinforcement Learning

Final model

Scale May '23	**>1 trillion** tokens	**10K - 100K** (prompt, response)	**100K - 1M** comparisons (prompt, winning_response, losing_response)	**10K - 100K** prompts

The Machine and System Requirement

- GPT-3.5 was trained on **6000 GPUs** for almost a month.
- To run open-sourced model, you definitely need a PC with good GPU

Say, you have the smaller Llama2 model (7b). You need 7*4 =**28 Gb VRAM** just to load the model

- LoRA and QLoRA comes to rescue. Let you load the model in 4 bit quantization. That means, 7b model will take approximately 7*.5 = **3.5 Gb VRAM**
- To fine-tune and/or RLHF, you need more.
- **Google colab** can help you a lot to fit the data.

Practical Tips and…

1. **Prompt Engineering:** LLMs are *less* about programming, much more about *communication*. Learn how to prompt. CoT, few-shot etc. https://www.promptingguide.ai/
2. **Building Apps:** Most applications can be built without any extra training steps. Frameworks like *Langchain* and *Llama_index* are leading the space.
3. **Retrieval Augmented Generation (RAG):** You definitely need the LLM equipped with your "own" data. RAG basically helps you to do that.
4. **Supervised Fine-Tuning (SFT):** If you want to fine-tune, you donot necessarily have to tune all your parameters. Use PEFT from huggingface.
5. **Do I need LLM?** Not necessarily.
6. **Vector Embedding:** A numerical representation of your text that comes handy in numerous applications.

References: Direct

1. https://huggingface.co/blog/rlhf
2. https://arxiv.org/pdf/2204.05862.pdf
3. https://huyenchip.com/2023/05/02/rlhf.html
4. https://www.youtube.com/watch?v=zjkBMFhNj_g&t=73s
5. https://jalammar.github.io/illustrated-transformer/
6. https://2023.aclweb.org/
7. https://medium.com/@shahriarsadat71_26111
8. https://sadat1971.github.io/blogs_and_tutorials.html

References: images and info

1. https://www.google.com/url?sa=i&url=https%3A%2F%2Fsproutsocial.com%2Finsights%2Fsentiment-analysis%2F&psig=AOvVaw3pgQUHmF7O_3hV8SFbBB67&ust=1708502971233000&source=images&cd=vfe&opi=89978449&ved=0CBMQjRxqFwoTCPj56Ne7uYQDFQAAAAAdAAAAABAE
2. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.istockphoto.com%2Fphoto%2F3d-rendering-humanoid-robot-handshake-to-collaborate-future-technology-gm1277975917-377039976&psig=AOvVaw3Kf8zRaLU1_2AZwUAIlDd0&ust=1708502228919000&source=images&cd=vfe&opi=89978449&ved=0CBMQjRxqFwoTCOC-6fW4uYQDFQAAAAAdAAAAABAK
3. https://www.google.com/url?sa=i&url=https%3A%2F%2Facutrans.com%2Ftop-10-most-commonly-spoken-languages-in-the-world%2F&psig=AOvVaw37wMmuRlEnjdErS3t2E9-e&ust=1708503079194000&source=images&cd=vfe&opi=89978449&ved=0CBMQjRxqFwoTCJiLpYu8uYQDFQAAAAAdAAAAABAE
4. https://www.google.com/imgres?imgurl=https%3A%2F%2Fwww.lingualinx.com%2Fhubfs%2Flanguage%2520translator%2520job%2520%25281%2529.jpg&tbnid=ht2fpnX8syLepM&vet=12ahUKEwiZzMiZvLmEAxU478kDHdlmARoQMygBegQIARB6..i&imgrefurl=https%3A%2F%2Fwww.lingualinx.com%2Fblog%2Flanguage-translator-role&docid=dnRygdIJELHHTM&w=809&h=432&q=languae%20translation&ved=2ahUKEwiZzMiZvLmEAxU478kDHdlmARoQMygBegQIARB6
5. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FThe-Transformer-model-architecture_fig1_323904682&psig=AOvVaw0Wwz2Gz4eRknboOp2F-4ST&ust=1708555451005000&source=images&cd=vfe&opi=89978449&ved=0CBMQjRxqFwoTCJjyqpj_uoQDFQAAAAAdAAAAABAE
6. https://blog.research.google/2017/08/transformer-novel-neural-network.html
7. https://news.ycombinator.com/item?id=37067933

Thanks

For any questions, send me a DM in MSTeams (UH people)!

Email: sadat.shrr@gmail.com

LinkedIn: https://www.linkedin.com/in/sadat-shahriar/ (send me a note that you attended my talk !)

Visit my website: https://sadat1971.github.io/

My medium blog: https://medium.com/@shahriarsadat71_26111