LLMs in the wilderness: Leveraging the power of LLMs in our business


Ali Najafi

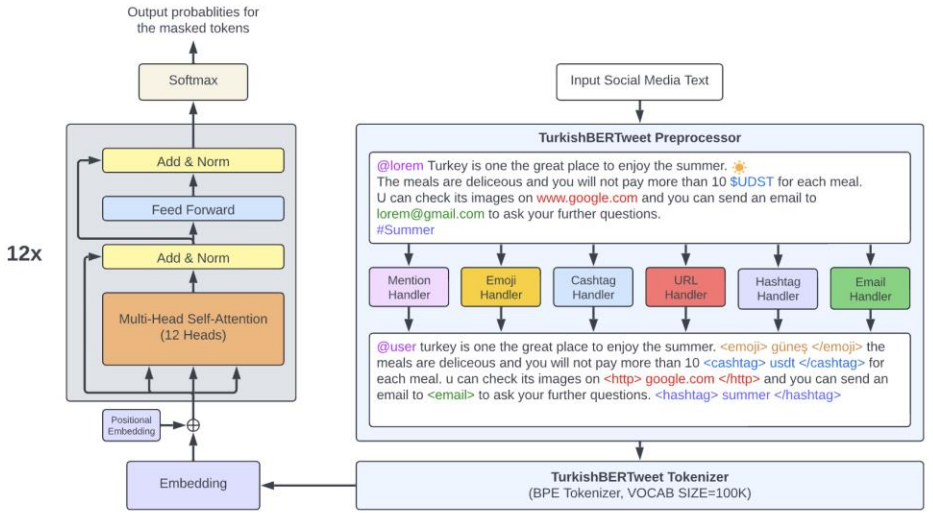
Sabancı Üniversitesi

VBYO 2025

Ali Najafi

PhD in Computer Science, Sabancı University, 2024–Present
MSc in Computer Science, Sabancı University, 2022–2024

🌐 **najafi-ali.com**



Published Works:

🔥 Turkishbertweet: Fast and reliable large language model for social media analysis

🚫 VRLLab at HSD-2Lang 2024: Turkish hate speech detection online with TurkishBERTweet

📂 First public dataset to study 2023 Turkish general election

Google Research

On going Projects:

🧩 Social Media Data Adaptors: Standardization and Anonymization of Social Network Datasets

🔍 Cross-Platform Bot Detection via Modeling of Social Media Account Behavior

📊 Analyzing the Impact of Voting Agency Discrepancies on Social Media Discourse During Turkey's 2023 Election Nights

By the end of this session, You will know

- How to approach an NLP project
- What resources exist
- What the pros and cons of LLMs are
- How to apply models to a real project
- How realistic the LLM-Hype is
- How to finetune generative Language models

Project

Hands on Project

- Amazon company after witnessing the advances in the AI field, decides to develop a Chatbot for its customers to help them purchase shoes!
- They recruited a Machine Learning Engineer to develop it!
- They provided the details of their shoe products and their prices!
- Now, let's say you are the ML guy, how can you build such a system?

Project

Let’s check out the data!

- The dataset is available at amazon-uk-shoes-dataset
- The downloaded data is in JSON format!
- Available information:
  - url: Link to the product
  - title: title of the product
  - asin: Product Unique ID
  - price: Price of the product
  - brand: Brand of the product
  - product_details: Details of product
  - breadcrumbs: a navigational aid that allows users to keep track of their current location on a website or interface
  - images_list: links of product images
  - features: features of product

Project

amazon_uk_shoes_dataset.json

Query | Download | Open in app

	url	title	asin	price	brand	product_details
1	https://www.amazon.co.uk/dp/B08BLP231I	Geox Jr Sandal Strada B Fisherman, Br	B08BLP231K	£50.00	Visit the Geox Store	Package Dimensions : 31.2 x
2	https://www.amazon.co.uk/dp/B08N587YZ	Fila Women's Oakmont Tr Sneaker	B08N587YZ9	£49.57 - £234.95	Fila	Product Dimensions : 32.51 x
3	https://www.amazon.co.uk/dp/B0918K4H1	Gabor Rollingsoft Trainers in Plus Si	B0918K4H1W	No data.	Gabor	Is Discontinued By Manufacturer
4	https://www.amazon.co.uk/dp/B07KMB98C	Merrell Women'S Bare Access Xtr Trail	B07KMB98CG	£67.00 - £182.44	Visit the Merrell Store	Package Dimensions : 28.96 x
5	https://www.amazon.co.uk/dp/B08CN3S1Z	Desigual Women's Shoes_Runner_cmotl S	B08CN3S1ZK	£38.96 - £81.10	Desigual	Package Dimensions : 34.6 x
6	https://www.amazon.co.uk/dp/B07KJX15D	Aquatalia Men's Adamo Dress Calf Penn	B07KJX15DV	£161.63 - £202.04	Aquatalia	Product Dimensions : 24.13 x
7	https://www.amazon.co.uk/dp/B089GYT1H	Womens Convertible Slip-on, Tan, 12	B089GYT1HP	£52.82	Unknown	Package Dimensions : 32 x 18
8	https://www.amazon.co.uk/dp/B01614251	Kappa Delhi Footwear Unisex, Mesh, Wo	B01614251K	£46.18	Visit the Kappa Store	Date First Available : 30 De
9	https://www.amazon.co.uk/dp/B07MJ4DRR	Naturalizer Women's Cairo 4 Sneaker,	B07MJ4DRRZ	£41.66 - £73.51	Naturalizer	Package Dimensions : 32.76 x
10	https://www.amazon.co.uk/dp/B07GX8GPJ	Rockport - Womens Cobb Hill Hattie Hi	B07GX8GPJN	£120.35 - £130.73	Visit the Rockport Store	Package Dimensions : 30.48 x
11	https://www.amazon.co.uk/dp/B012DS5U2	Reebok Men's NPC Ii Fashion Sneaker	B012DS5U2Y	£47.00 - £156.93	Visit the Reebok Store	Is Discontinued By Manufacturer
12	https://www.amazon.co.uk/dp/B08D3F9JQ	adidas Women's Superstar Bold W Gymna	B08D3F9JQY	£85.00 - £93.61	Visit the adidas Store	Package Dimensions : 28.8 x
13	https://www.amazon.co.uk/dp/B01GRUZWA	Geox Women's D Stinge a Fashion Sneak	B01GRUZWAO	£64.95 - £117.00	Visit the Geox Store	Is Discontinued By Manufacturer
14	https://www.amazon.co.uk/dp/B07SQ5MLX	K-Swiss Men's Clean Court Ii CMF Low-	B07SQ5MLXZ	£20.99 - £67.57	K-Swiss	Package Dimensions : 30.48 x
15	https://www.amazon.co.uk/dp/B00LIHRNQ	Volcom Mens Sub Zero Boot Winter	B00LIHRNQI	£65.11 - £238.10	Visit the Volcom Store	Package Dimensions : 36.3 x
16	https://www.amazon.co.uk/dp/B07H3XFPR	ASICS Mens Gel-Ziruss 2 Running Shoes	B07H3XFPR9	£99.79 - £124.99	ASICS	Product Dimensions : 25.25 x
17	https://www.amazon.co.uk/dp/B07KQKYZY	bugatti Women's 432636065969 Low-Top	B07KQKYZYR	£28.77 - £62.87	bugatti	Is Discontinued By Manufacturer
18	https://www.amazon.co.uk/dp/B08HJJ5N7	Pablosky Girl's 495000 Sandal	B08HJJ5N7R	£25.76 - £36.69	Pablosky	Package Dimensions : 23 x 16
19	https://www.amazon.co.uk/dp/B08635QDS	HROYL Girls Dance Shoes Satin Latin/C	B08635QDSF	£25.99 - £28.99	Visit the HROYL Store	Is Discontinued By Manufacturer
20	https://www.amazon.co.uk/dp/B084GB1DS	MILLET Unisex's Hike Up Mid GTX Slouc	B084GB1DS3	£108.49 - £143.40	Visit the MILLET Store	Package Dimensions : 35 x 25
21	https://www.amazon.co.uk/dp/B07T26WQF	adidas Originals Men's Superstar Foun	B07T26WQF5	£27.50 - £214.11	adidas Originals	Product Dimensions : 30 x 35
22	https://www.amazon.co.uk/dp/B0824F5JR	victoria Women's Tenis Piel/Virutas G	B0824F5JRM	£42.49 - £75.49	VICTORIA	Package Dimensions : 29 x 29
23	https://www.amazon.co.uk/dp/B08XJG46Q	ZYLDK Garden Shoes Children's Unisex	B08XJG46Q4	£8.06 - £9.10	ZYLDK	Package Dimensions : 23 x 19

ABOUT THIS FILE

Last Updated	3 years ago
Owner	Crawl Feeds
Created	3 years ago
Size	13.16 MB

Displaying 9 columns, 11,605 rows in table
amazon_uk_shoes_dataset

TABLE COLUMNS

- url
- title
- asin
- price
- brand
- product_details
- breadcrumbs
- images_list
- features

NLP Tasks

Language Translation

- EN -> TR, EN -> FA

Text Classification/Regression

- Sentiment Analysis
- Personality Detection
- Topic Detection
- Hate Speech Detection
- Stance Prediction

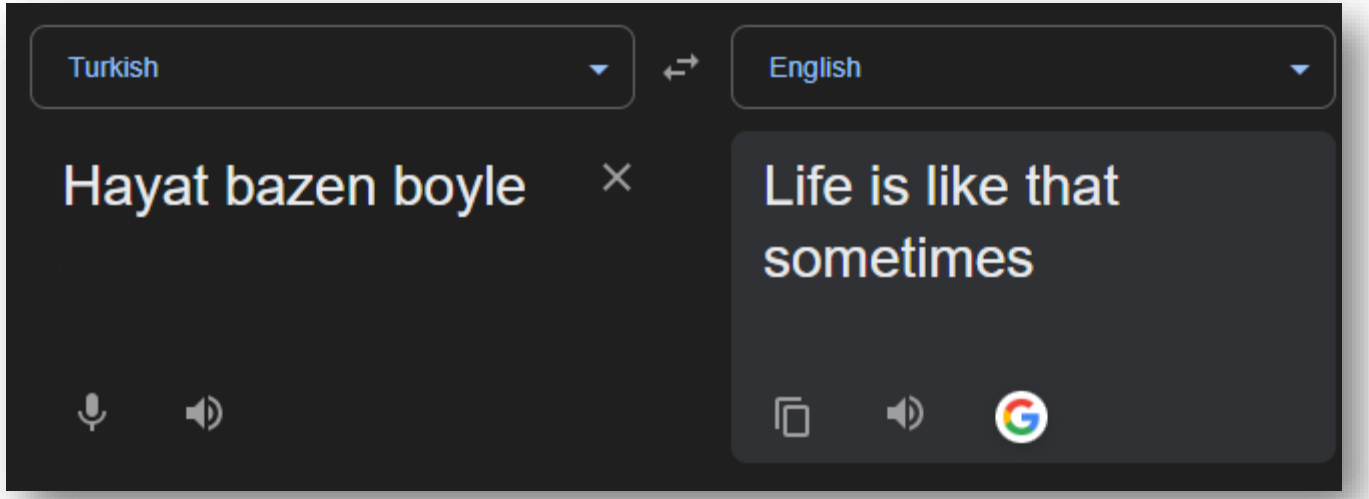
Token Classification

- Name Entity Recognition (NER)

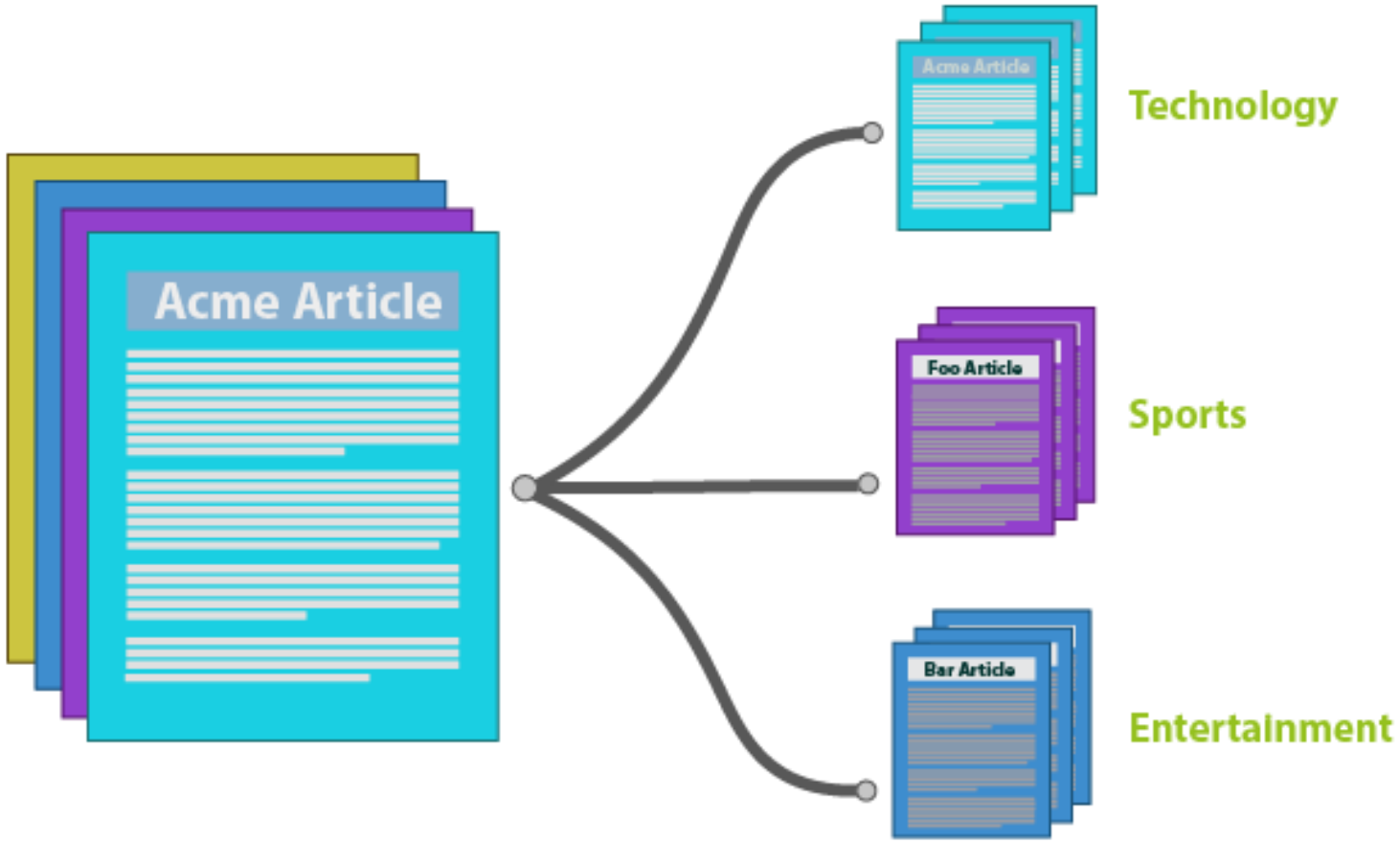
Text Generation

- Question Answering
- Text Summarization

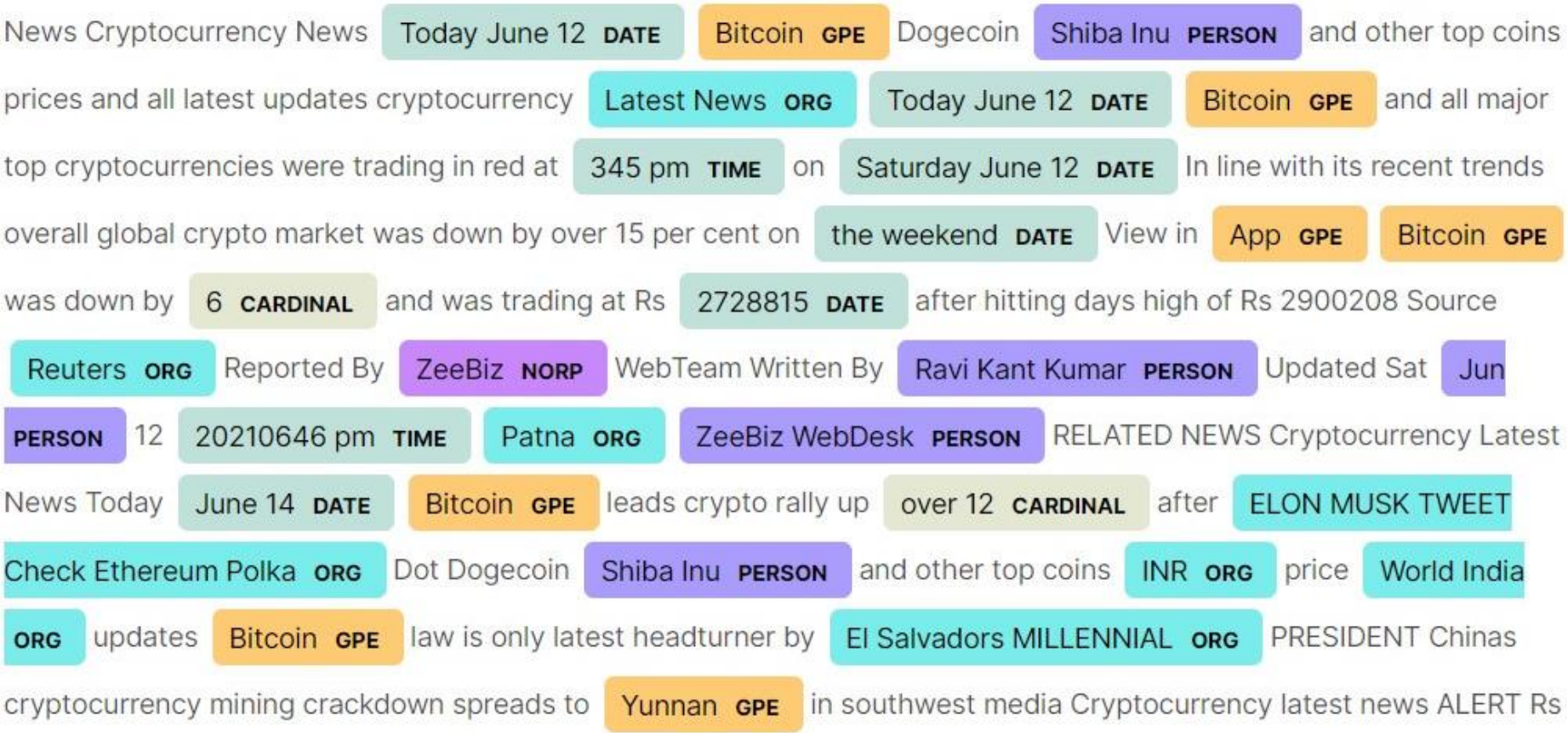
Language Translation



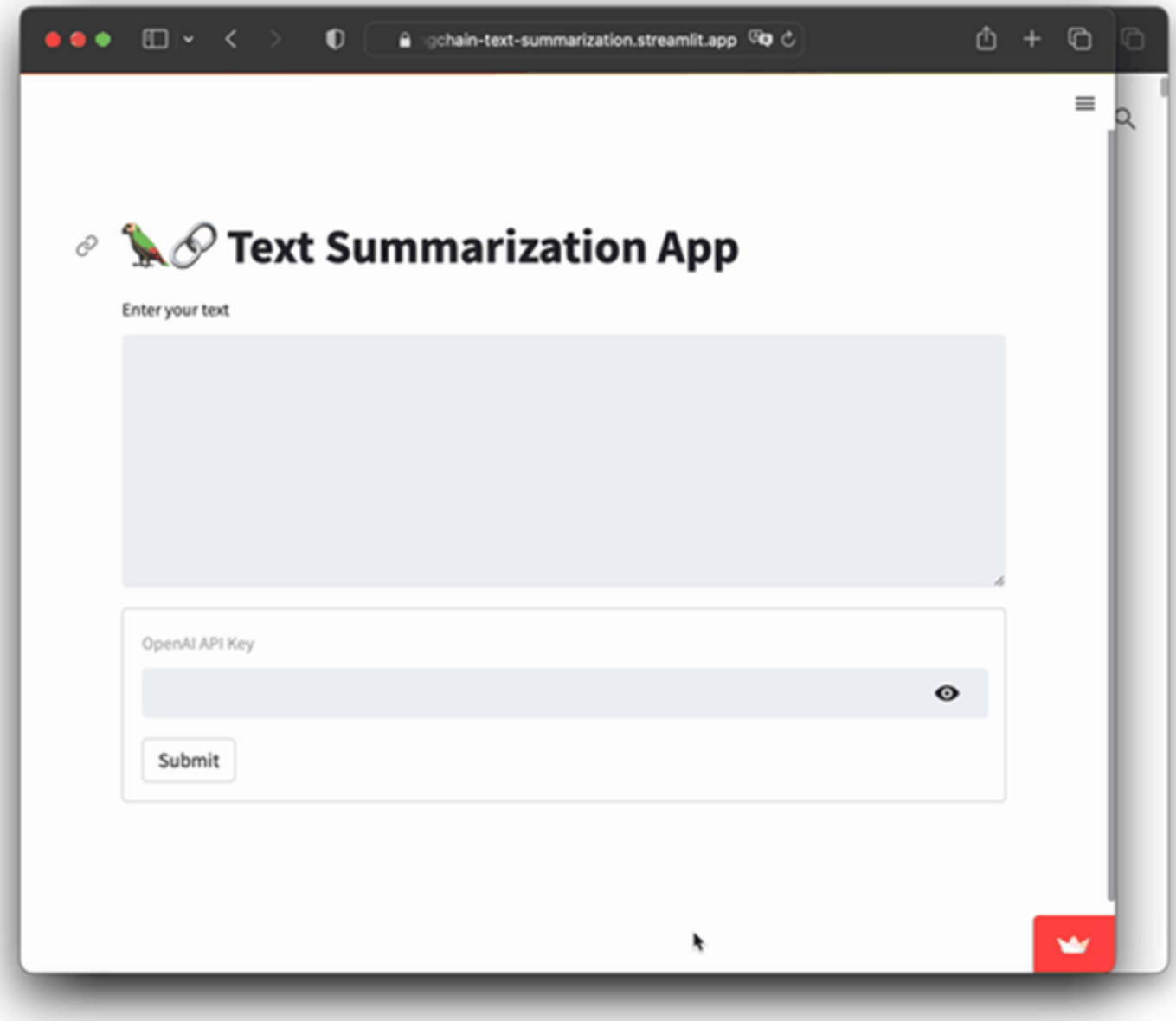
Text Classification



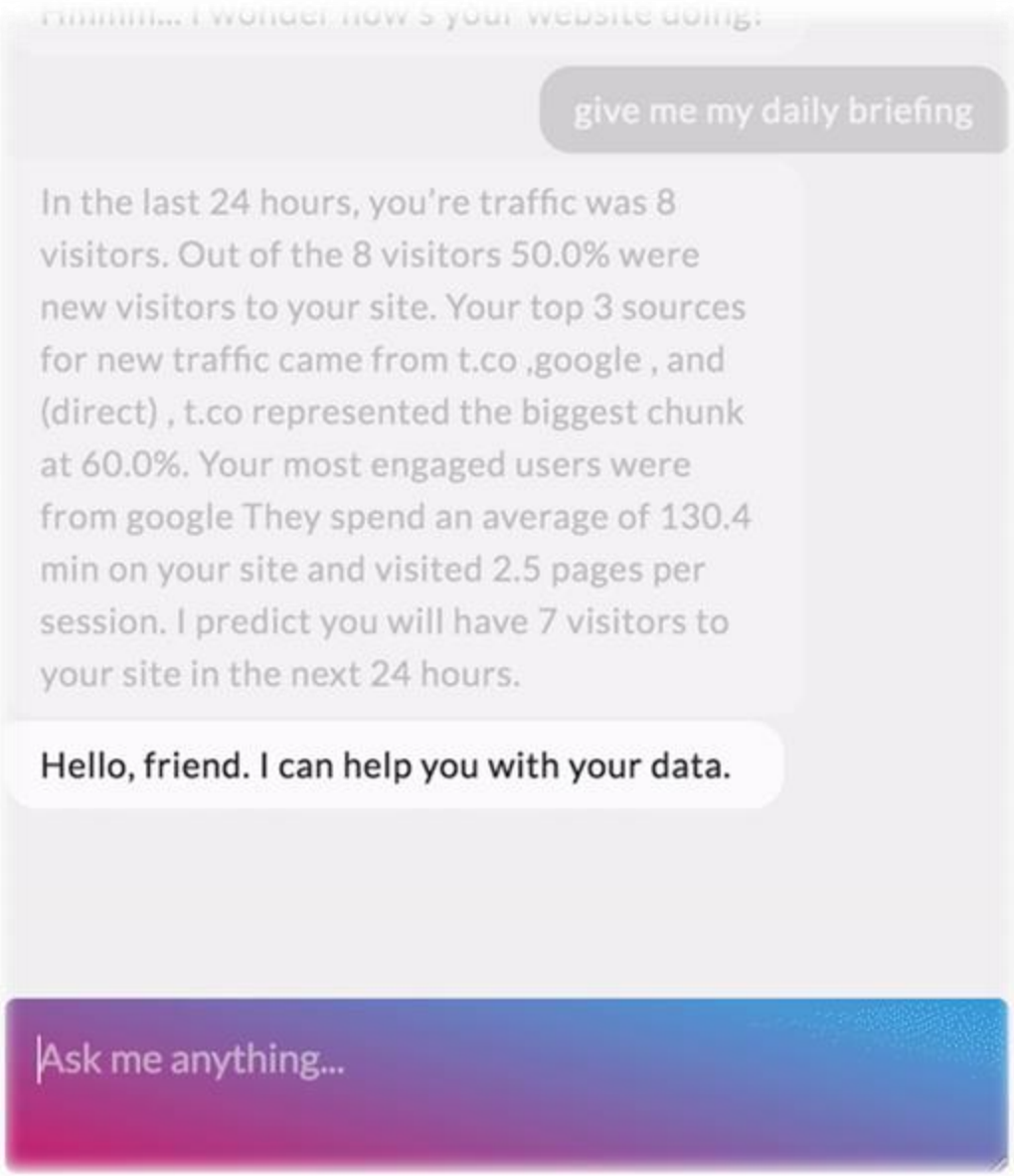
Token Classification



Text Summarization

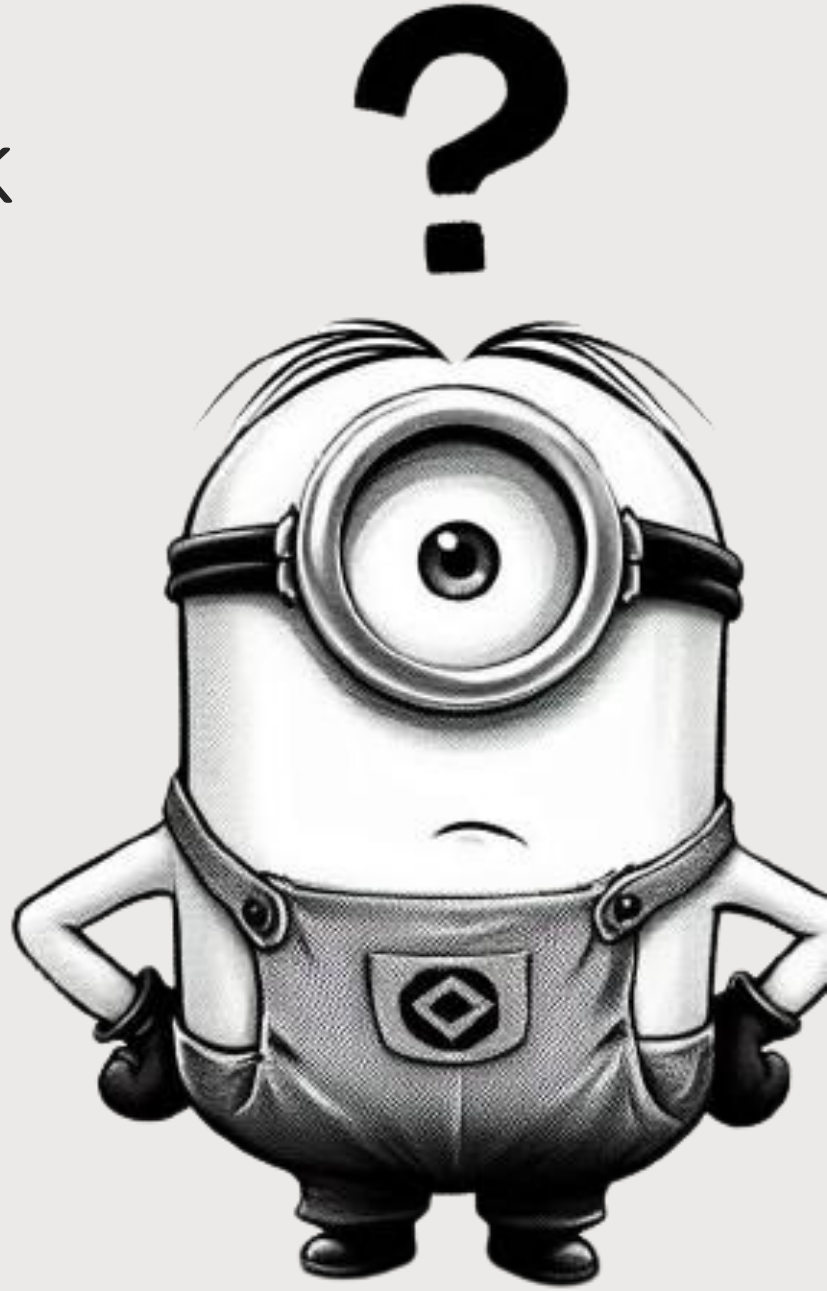


Question Answering



We should ask ourselves

- Do we need LLMs?
- Can we afford them?
- Do we have access to skilled people?
- Is the outcome significant enough to invest in them?



Which model?

There are many already pre-trained models available on Hugging Face

BUT Which model should I choose?

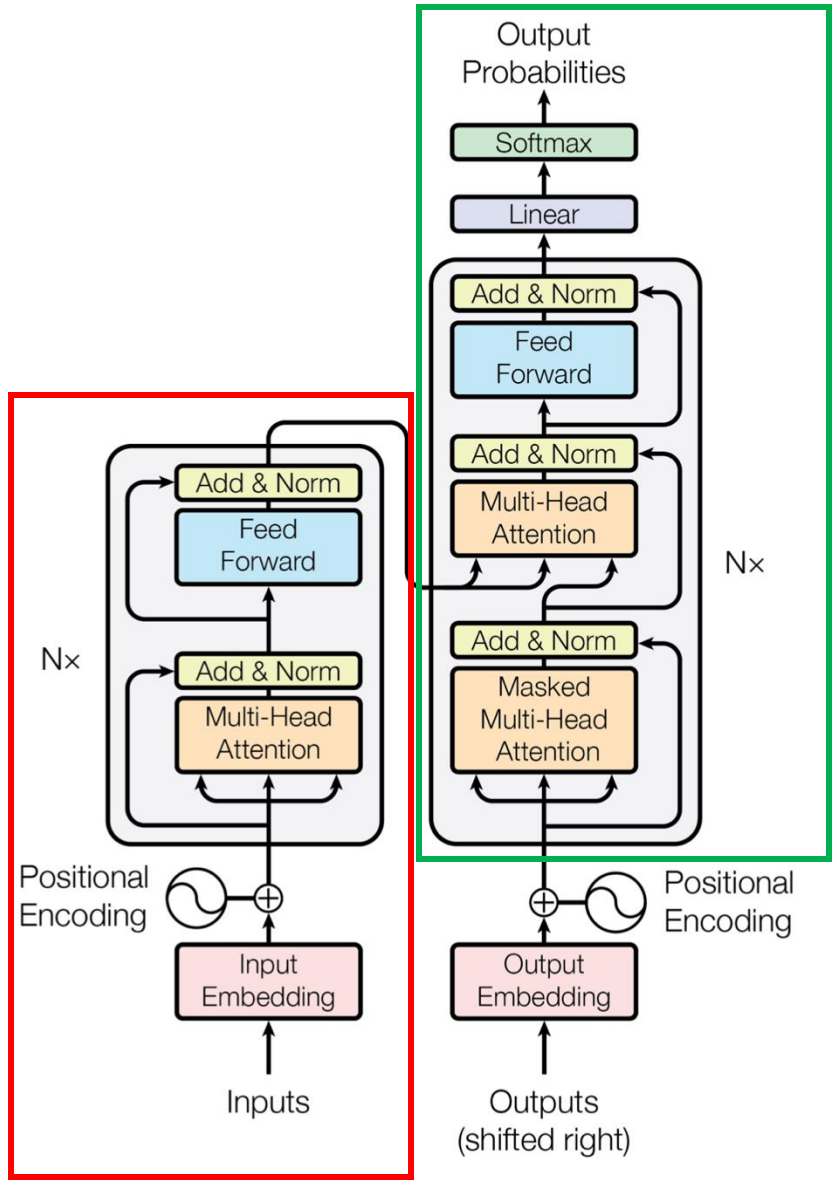
Several factors help us answer this question!

1. NLP Task
2. Computational Resources
3. Domain Requirement
4. LICENSE

NLP Task

- If your task is other than text generation, Encoder models are available!

- If your task contains generation, Encoder-Decoder OR Decoder models are available. Need to mention that you can still use these types of models for non-generative tasks as well!



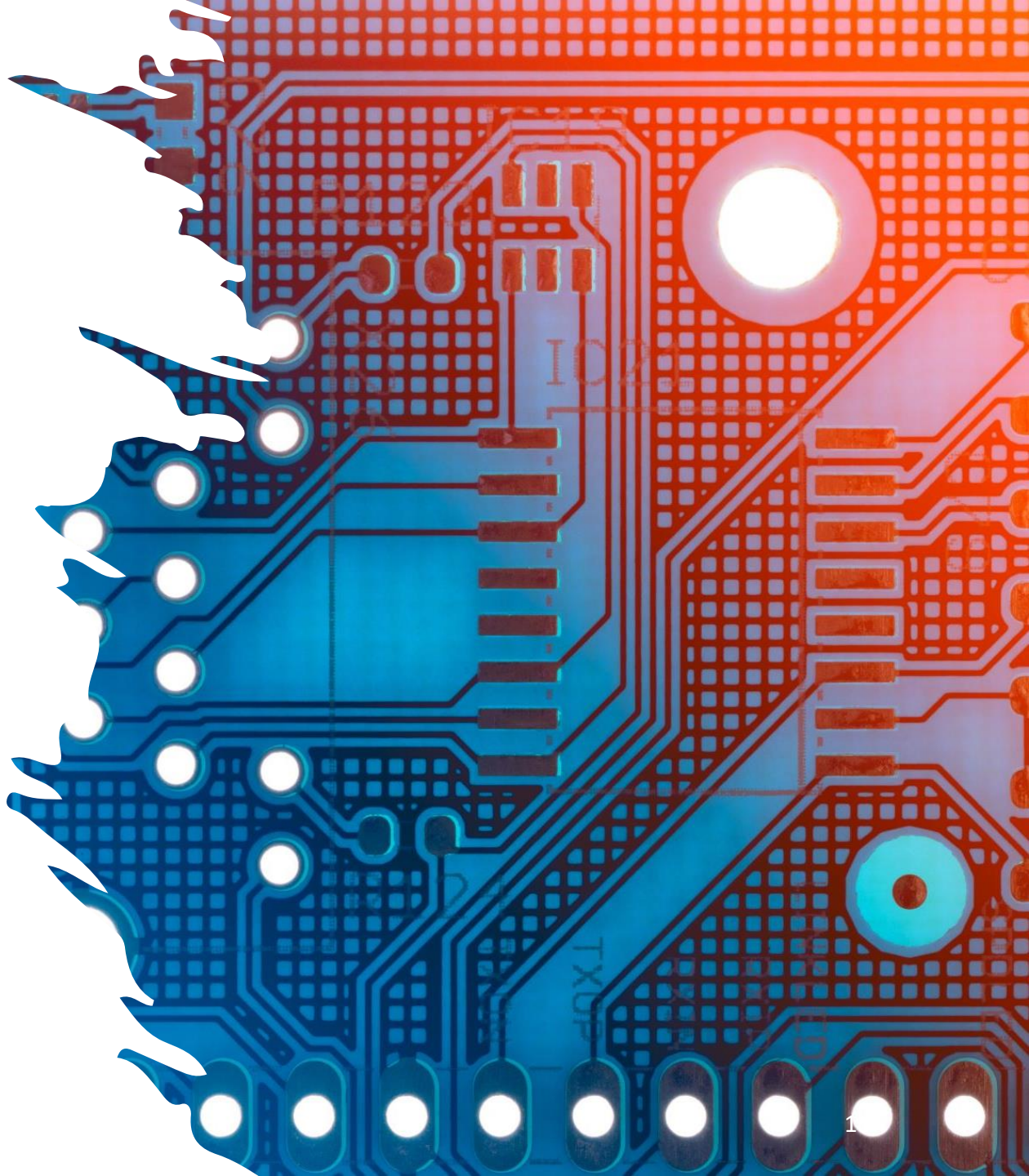
Computational Resources

Computation power plays a great role in model selection.

Computation power means the GPU infrastructure of the company!

There are two main things about GPUs that we are interested in:

- RAM
- Speed



Domain Requirements

- Each model has been trained on specific data for instance:
  - Internet Data
  - Code Data
  - …
- In our project, models may have seen data about shoes, but they have no clue about the available products of Amazon company.
- Language matters: If you are targeting customers of the UK, then the model needs to interact in English. What if you want a chatbot that can handle customers with different languages?

Terminologies in LLM World

- Pre-Training: Training a language model from scratch
- Finetuning: Training the pre-trained model on a domain-specific dataset(s)
- Instruction Finetuned: The model is finetuned on a dataset that contains Question-Answer pairs.
- Zero Shot: Using model without finetuning
- Tokenizer Special tokens: Tokens that the tokenizer maps into a single ID.

Model Architectures and their objectives

_____ models like _____ models can be trained on _____ objectives.

1. Encoder; BERT-based; classification or masked language modeling

2. Encoder-Decoder; T5-based; sequence-to-sequence or masked language modeling

3. Decoder; GPT-based; causal language modeling or masked language modeling

Language Modeling

Causal Language Modeling: predict the next token

- Input: Hard work always pays _____.
- Output: off

Masked Language Modeling: predict the masked token

- Input: Hard work ____ pays off.
- Output: always

Sequence to sequence: predict the corrupted tokens

- Input: Hard _____ pays off.
- Output: work always

Finetuning Approaches

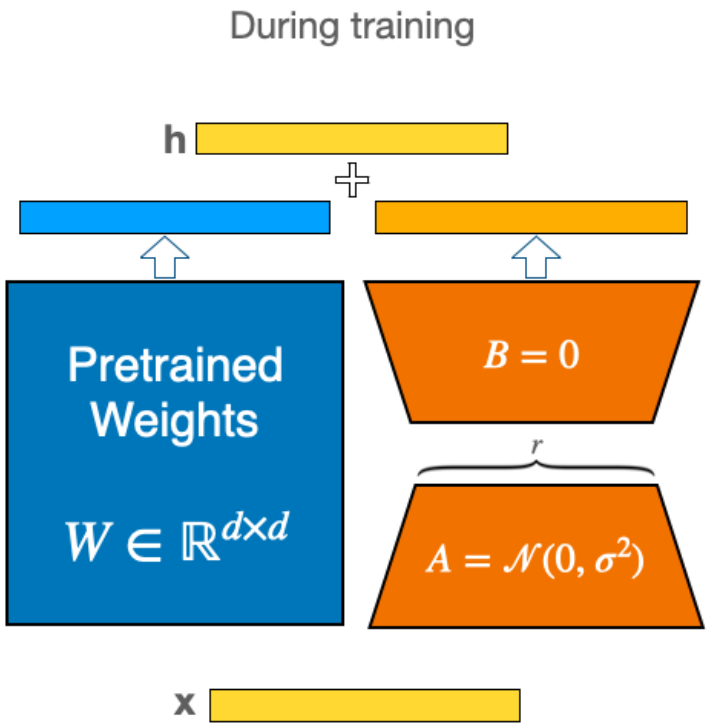
Full-Finetuning: Updating directly the wights of the models

LoRA Finetuning: Just updating the LoRA adapter's weights and keeping the model's weights intact

RAG Finetuning: Finetuning with improved prompts using our knowledge database.

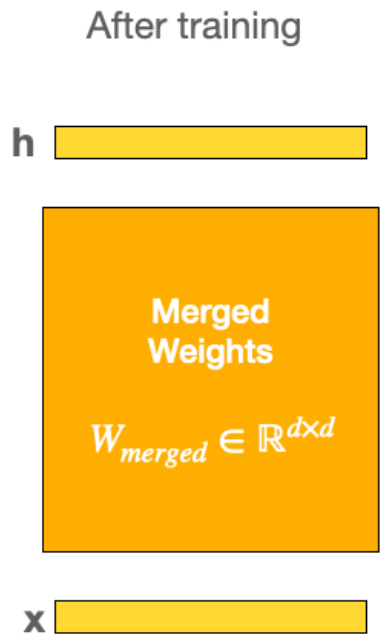
RLHF finetuning: Finetuning the model with a dataset based on human feedbacks using reinforcement learning

LoRA Finetuning



$$h = Wx + BAx$$

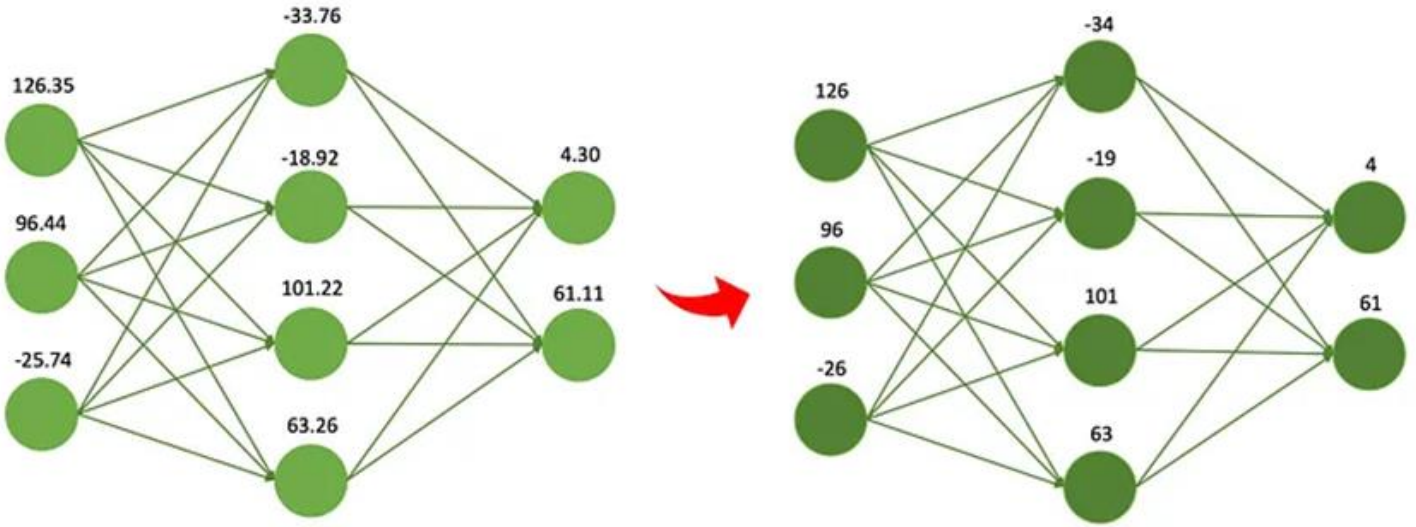
$$h = \underbrace{(W + BA)}_{W_{merged}}x$$



Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Quantization

- It is not simply rounding the numbers!
- It enables us to load huge models on our devices!
- 2bit
- 4bit
- 8bit



Before

32bit

After

8bit

Libraries available for quantization

- Bitsandbytes library of Huggingface
  - https://github.com/bitsandbytes-foundation/bitsandbytes
- AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration
  - https://github.com/mit-han-lab/llm-awq
- SeeDOT:
  - normally used in the IoT field
  - Developed by Microsoft
  - https://github.com/microsoft/EdgeML/tree/master/tools/SeeDot

Now we are ready to go!

Project

What is Amazon expecting?

A Chatbot that can interact with the customer and provide answers based on the provided data!

You may find an open-source chatbot that handles conversations very well but provides irrelevant information!

As a result, we need to customize our model!

Or in other words, finetune our model!

Let’s start!

- Resources:

  https://huggingface.co/spaces/Vokturz/can-it-run-llm

- Model Selection:

  https://huggingface.co/models

- Data Exploration
- Data Preparation
- Finetuning
- Evaluation

Project

```
questions = [
    f"Can you recommend some {category} shoes from {brand_name}?",
    f"What are the best {category} shoes from {brand_name}?",
    f"Can you show me some {category} shoes from {brand_name}?",
    f"Recommend me some {category} shoes from {brand_name}.",
    f"What {category} shoes do you have from {brand_name}?",
    f"Show me some {category} shoes from {brand_name}?",
    f"{category} shoes from {brand_name}",
]
```

```
answers = [
    f"Here are some {category} shoes from {brand_name}:",
    f"Here are some {category} shoes from {brand_name} that you might like:",
    f"Here are some {category} shoes from {brand_name} that you might find interesting:",
    f"Here are some {category} shoes from {brand_name} that you might find appealing:",
    f"Here are some {category} shoes from {brand_name} that you might like:",
]
```

```
rejection_answers = [
    f"Unfortunately, we do not have any {category} shoes from {brand_name} at the moment.",
    f"Sorry, we do not have any {category} shoes from {brand_name}.",
    f"No {category} shoes found for {brand_name}.",
    f"Currently, we do not have any {category} shoes from {brand_name}.",
    f"Sorry, we do not have any {category} shoes from {brand_name} at the moment.",
    f"Unfortunately, we do not have any {category} shoes from {brand_name}.",
    f"Currently there are no {category} shoes available from {brand_name}.",
]
```

Finetuning Data Preparation

So far, we have tabular data. To give a conversational taste to the communication between customers and the chatbot, we should transform our dataset into a textual conversational dataset.

🚀 Deploying the LLMs

- Transformers library is unfortunately slow, but it has a great programming interface and is great for finetuning.
- As a result, for deploying LLMs, it is highly recommended to use libraries like **vLLM** or **TGI**.
- The following web pages contain rich content about these libraries.
  - **vLLM:**
    - **How does vLLM optimize the LLM serving system?**
    - **vLLM Doc**
  - **TGI:**
    - **Deploying Large Language Models With HuggingFace TGI**

- ali.najafi@sabanciuniv.edu
- **www.najafi-ali.com**