

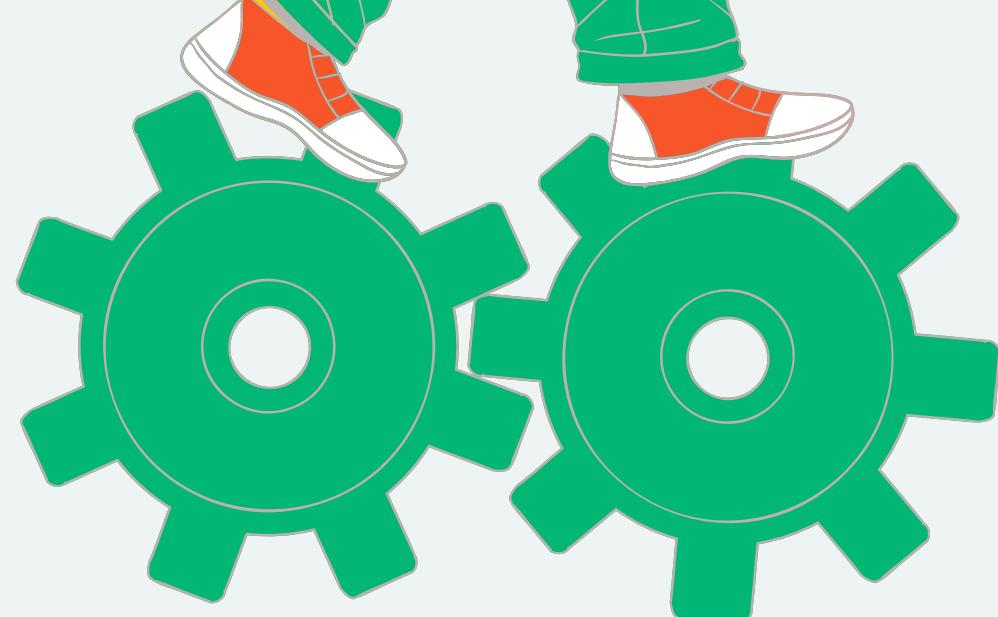
TOXIC & HARMFUL TEXT CLASSIFIER USING TRADITIONAL NN, MISTRAL AND LLAMA LLMS



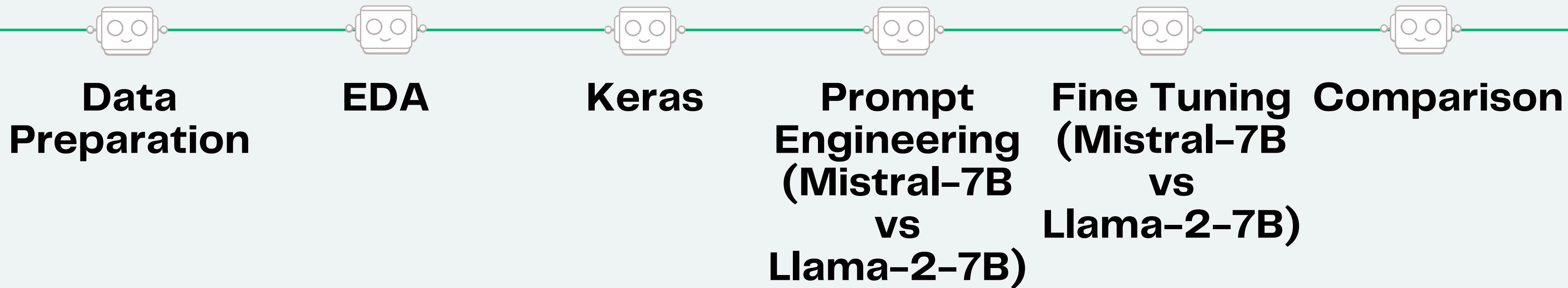
PRESENTED BY:

**RAZ GRAIDER
DANA BRAYNIN
SOPHIE MARGOLIS
RAN ASANTA**

INTRODUCTION AND WORKFLOW



The objective: text classification as
offensive or non-offensive



LIBRARIES & PLATFORMS



matplotlib



Hugging Face

K Keras



seaborn



W&B

DATA PREPARATION



01

“Stormfront” dataset

10,900 comments extracted from Stormfront, a white supremacist forum, classified as non-offensive (0) or offensive (1)



Train

Keras - combined (140102)

LLMs - 0.08% combined (1120)

02

“Wiki” dataset

Comments collected from Wikipedia forums, classified as non-offensive (0) or offensive (1)

128,000 – train

31,900 – validation

64,000 – test



Validation

Keras - combined (33471)

LLMs - 0.08% combined (267)

03

“Jigsaw” dataset

5,000 comments used in the Jigsaw Competition taken from Wikipedia’s talk page edits, classified as non-offensive (0) or offensive (1)



Test

Stormfront - 1066
Wiki - Keras 63978

LLMs 500

Jigsaw - 490



EXPLORATORY DATA ANALYSIS (EDA)

Glimpse to the data

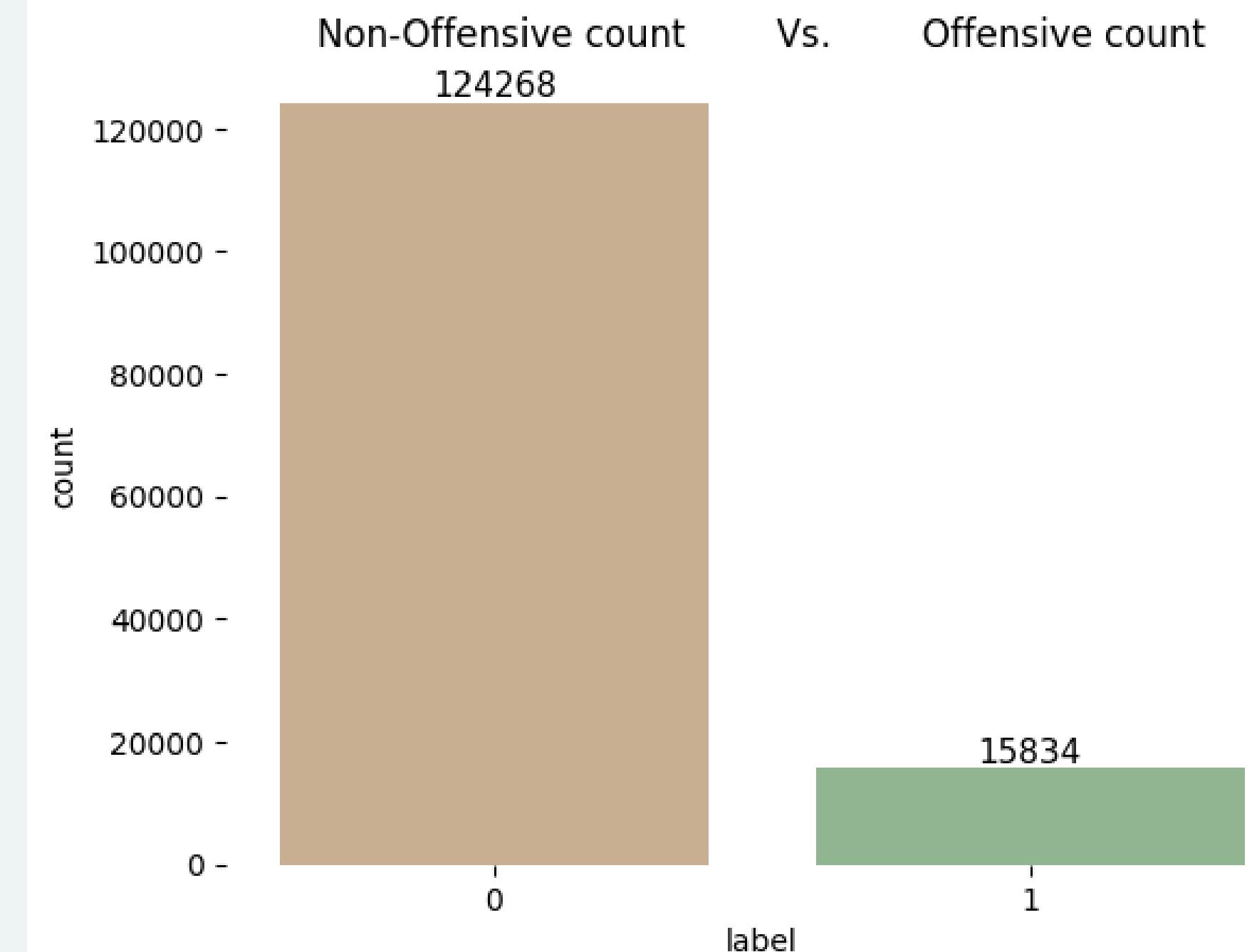


	text	label
0	register bank switching \n\nMany months ago, y...	0
1	Are there are a bunch of H-53 squadrons in Haw...	0
2	"\n\n ""read the notice"" \n\nAs you seem to b...	0
3	"\n\nAndromedean - You say, rightly, ""there a...	0
4	"::Help me understand. We both agree it is sui...	0

EXPLORATORY DATA ANALYSIS (EDA)

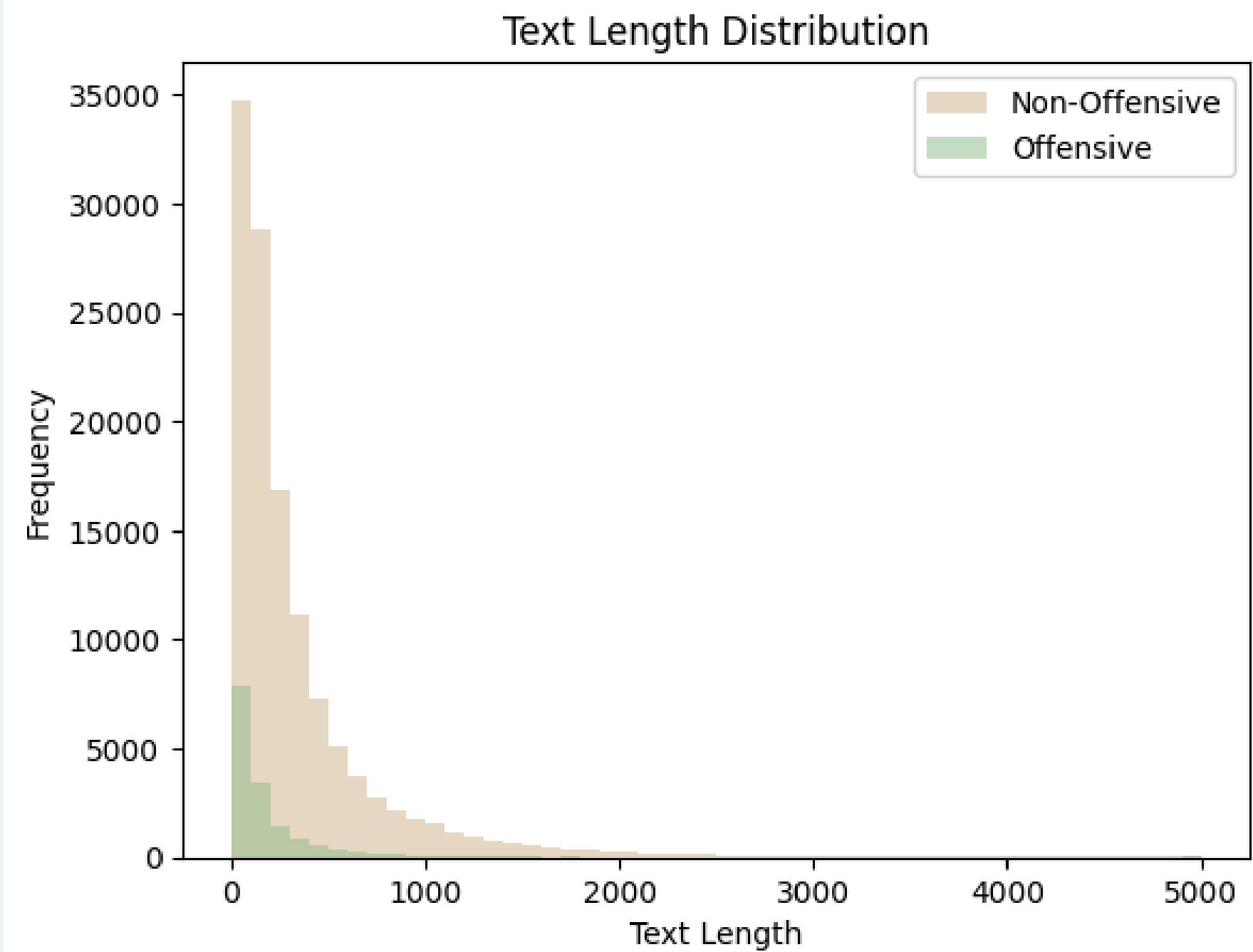
DATA
SCIENCE

Label Count



EXPLORATORY DATA ANALYSIS (EDA)

Text Length



EXPLORATORY DATA ANALYSIS (EDA)

DATA

SCIENCE

ART

DESIGN

INNOVATION

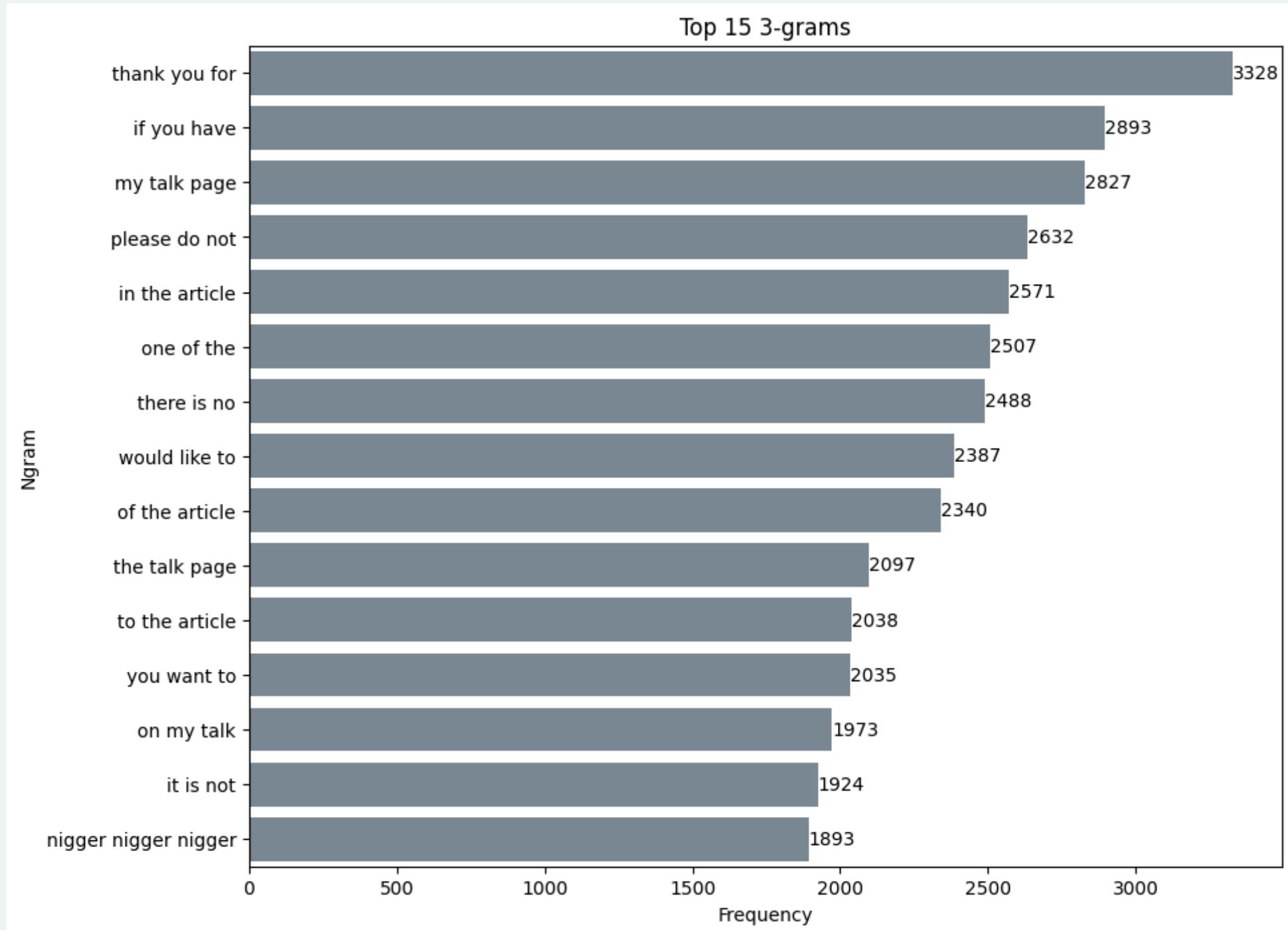
IDEAS

INSIGHTS

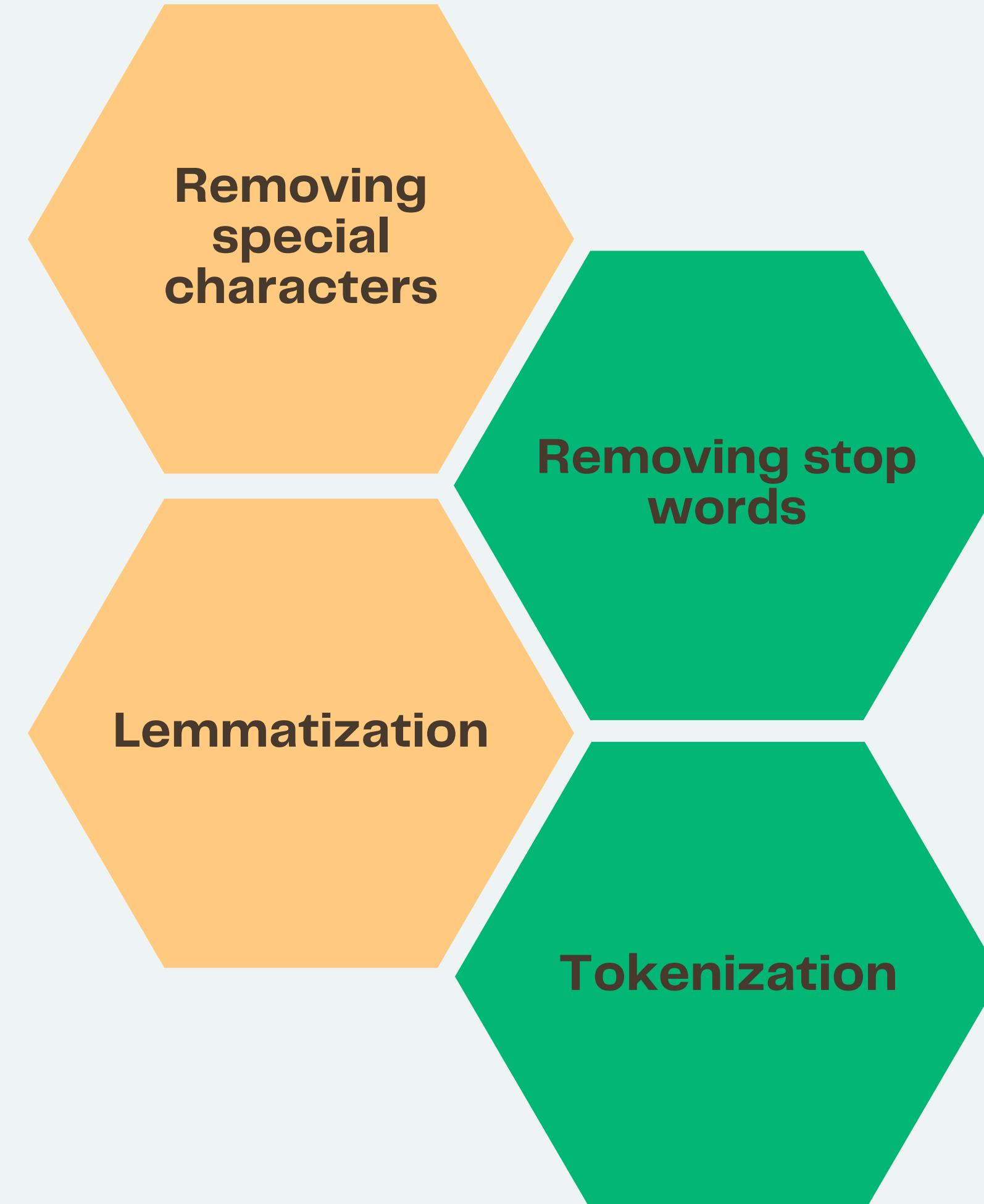
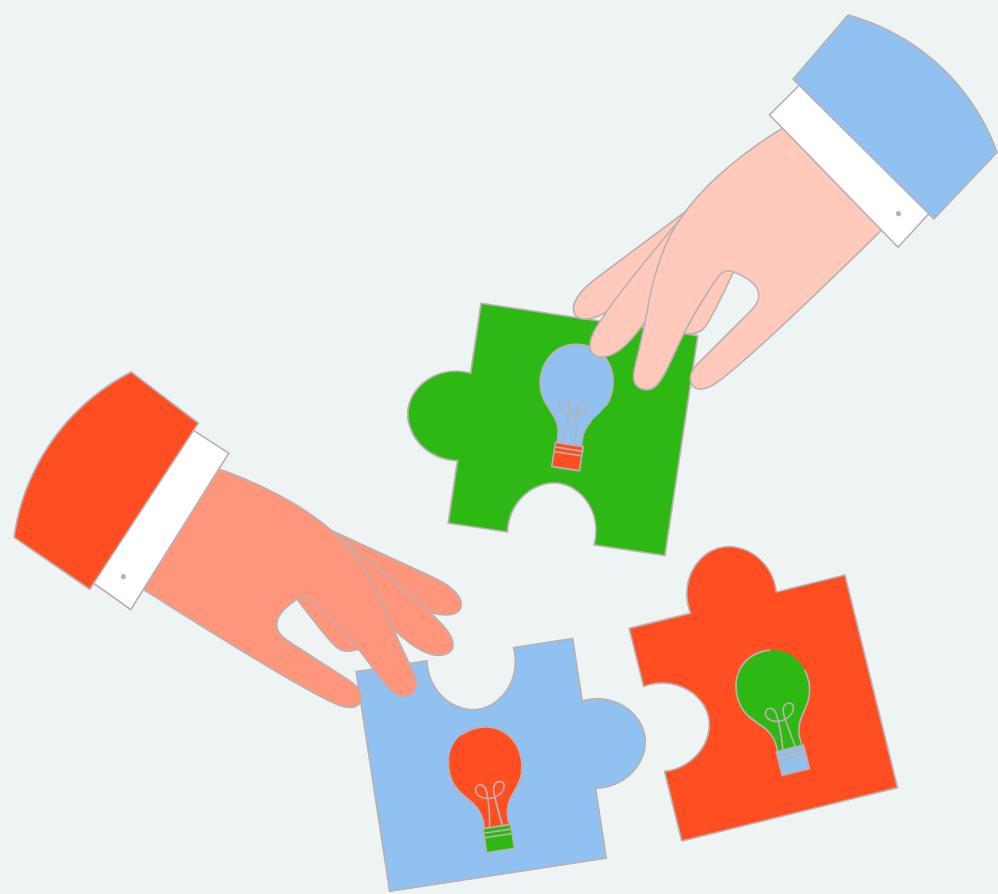
KNOWLEDGE

WISDOM

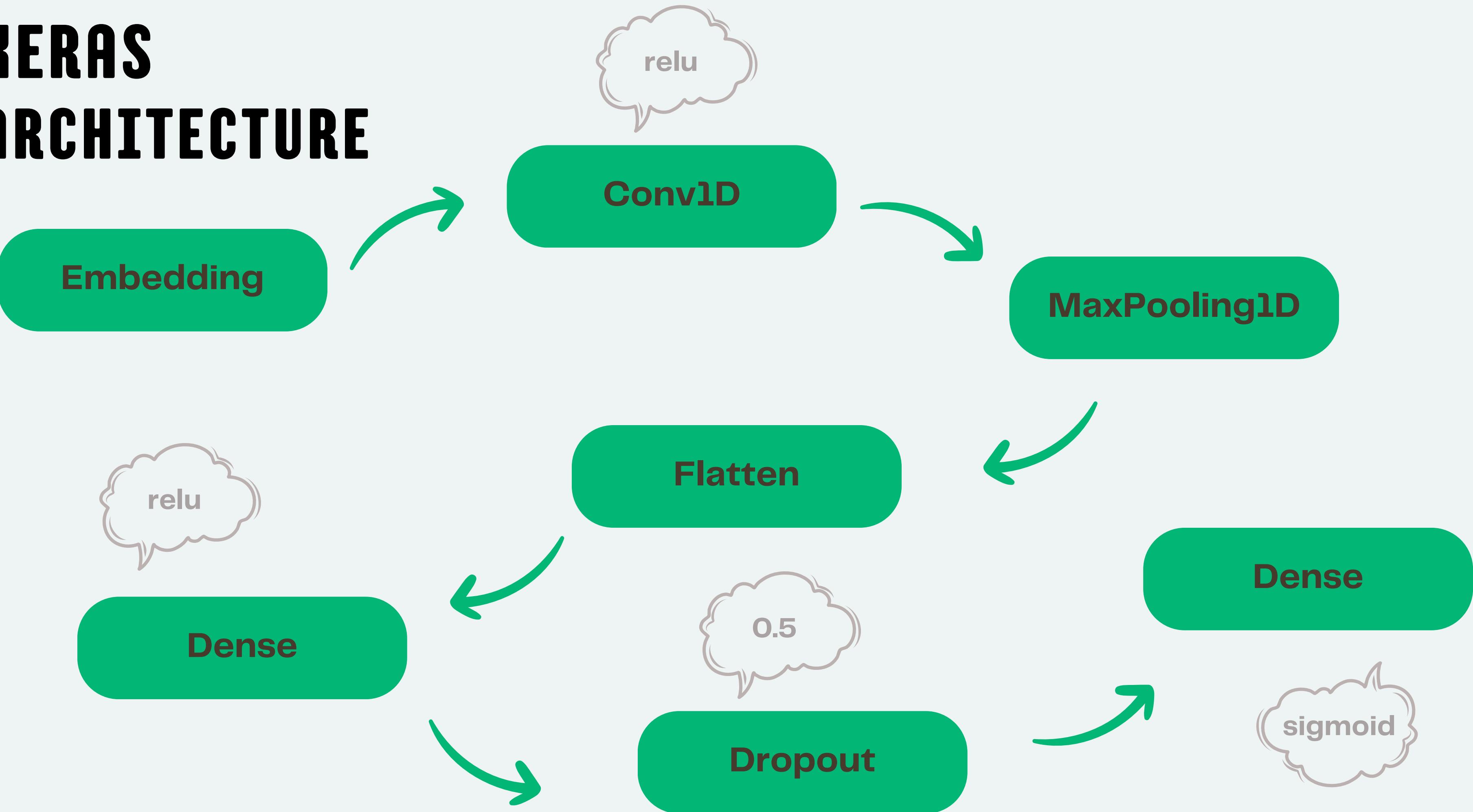
3-Grams



KERAS PREPROCESSING



KERAS ARCHITECTURE



KERAS TEST SCORES

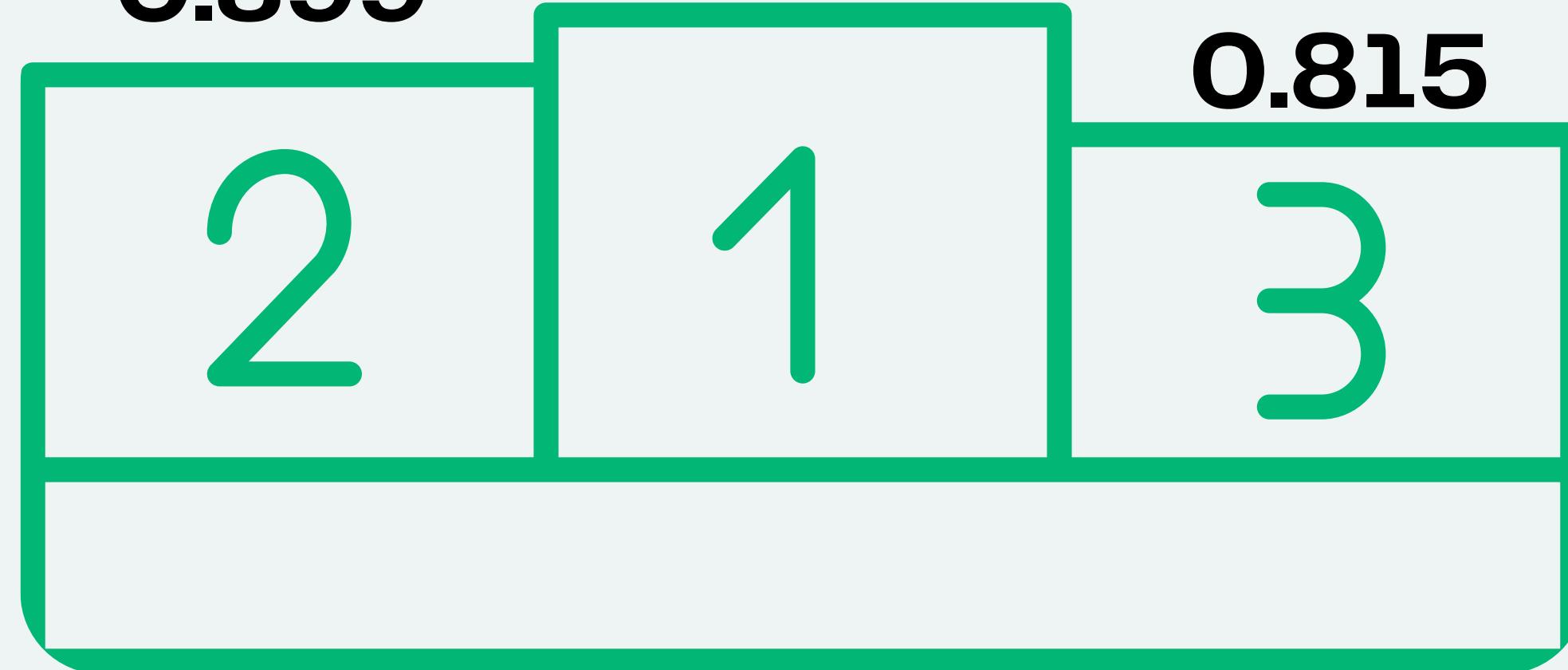


Wiki
0.899

Jigsaw

0.916 Stormfront

0.815



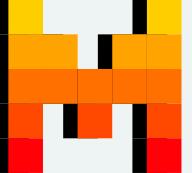
LLM (LARGE LANGUAGE MODELS) THEORETICAL OVERVIEW

LLaMA
by  Meta



Designed to comprehend and generate
human-like text

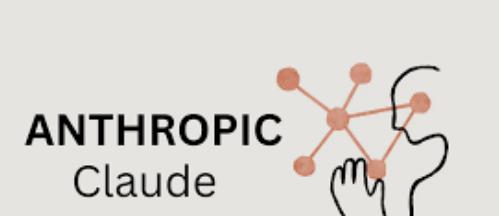
Gemini

 MISTRAL
AI_

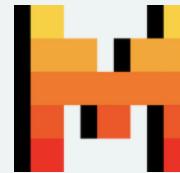
Using transformer-based architectures

 BERT

Trained on massive amounts of text data

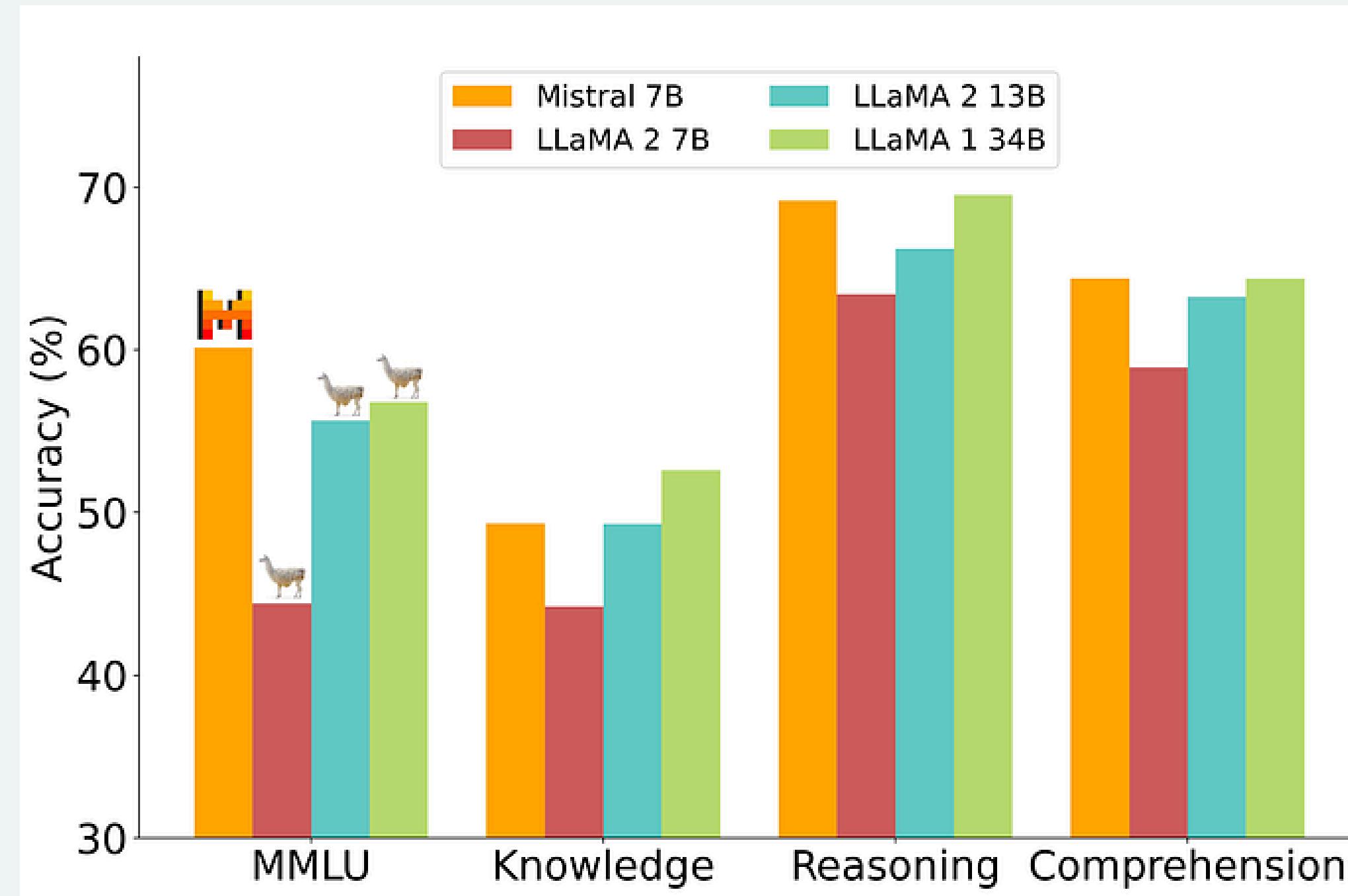
ANTHROPIC
Claude


MISTRAL-7B VS LLAMA-2-7B

	
Released by Mistral AI	Released by Meta in partnership with Microsoft
27 Sep 2023	18 July 2023
7,112,380,416 parameters	6,609,457,152 parameters
8 Trillion tokens	2 Trillion tokens
Decoder-only transformer	Decoder-only transformer
Highly multilingual, with the ability to understand and generate content in over 100 languages	Trained on 20 languages, focusing on those with Latin alphabets
Relatively fast inference speed	Slightly slower inference speed compared to Mistral
Enables faster performance even on less powerful hardware	Requires more robust hardware to function optimally
Apache 2.0 license	GPL 3 license

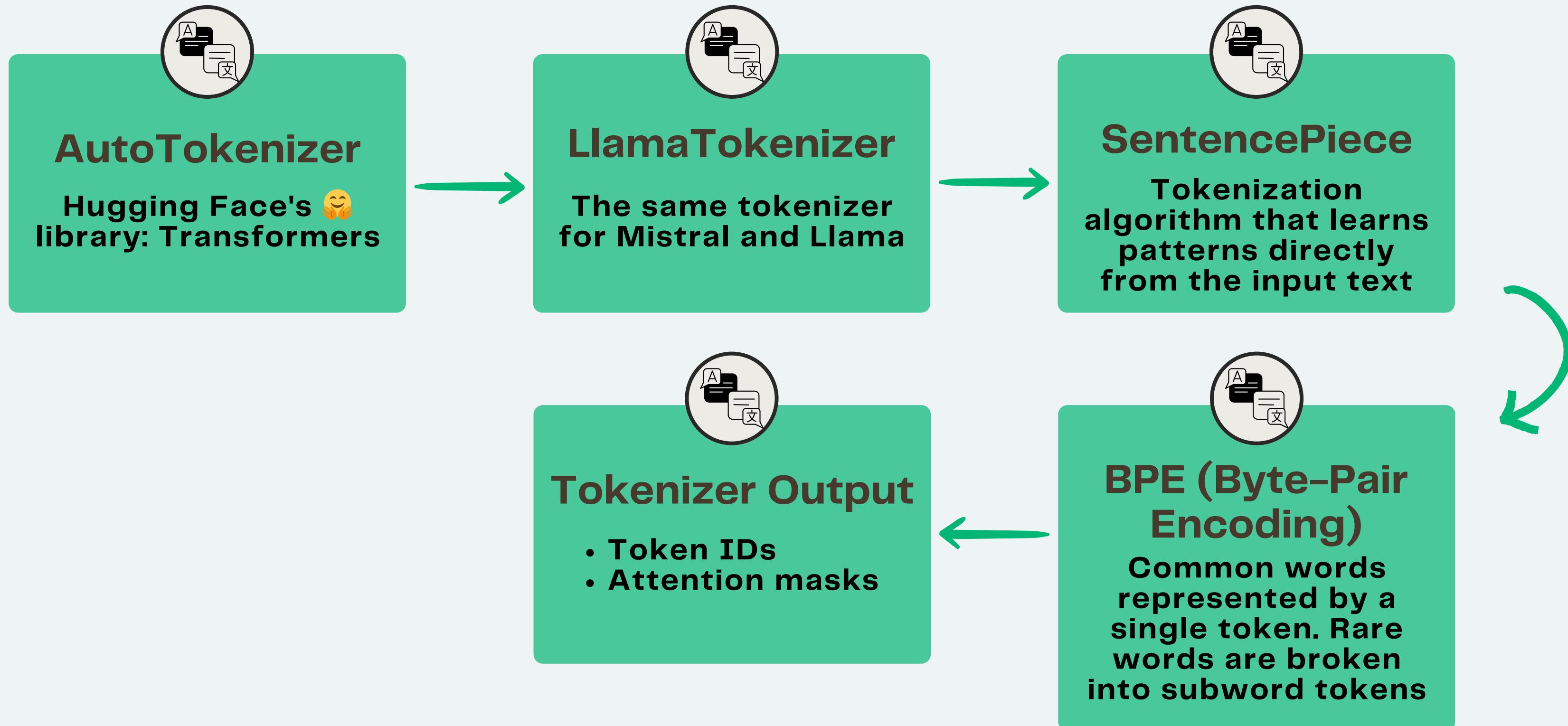
MISTRAL-7B VS LLAMA-2-7B

PERFORMANCE ON DIFFERENT BENCHMARKS



LLM - PREPROCESSING

THEORETICAL OVERVIEW OF THE TOKENIZER



LLM - PREPROCESSING

IMPLEMENTING TOKENIZER MISTRAL-7B & LLAMA-2-7B

Padding

- Ensures that input sequences are of uniform length
- No default pad token
- Using EOS (End Of Sequence) token `</s>` as the pad token

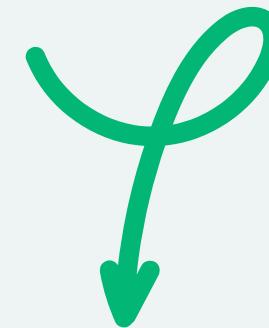
DataCollator WithPadding

- Provided by Hugging Face 😊 Transformers
- Combining individual examples into batches
- Pads sequences within a batch to the maximum length observed in that batch

LLM - PREPROCESSING

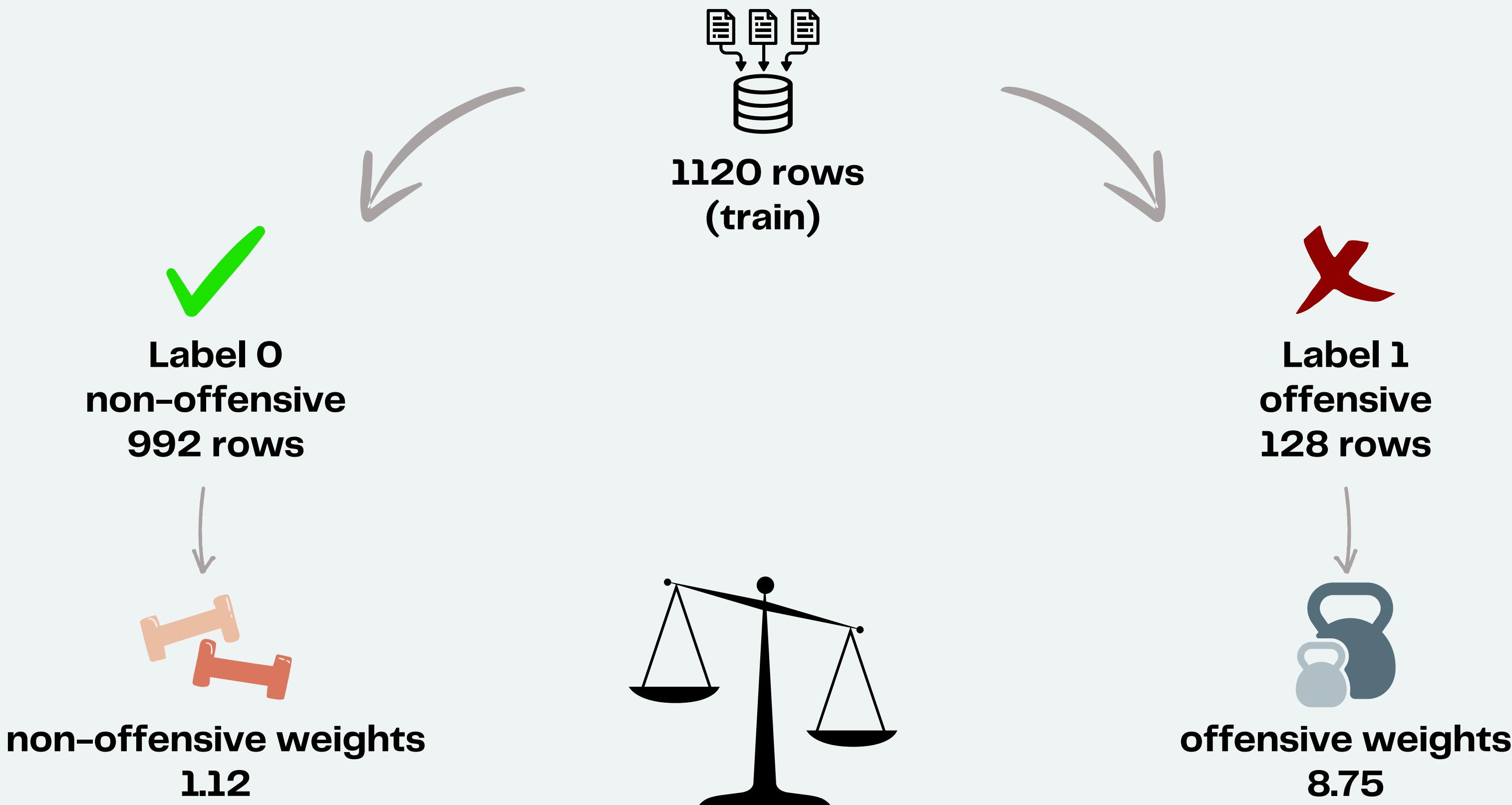
IMPLEMENTING TOKENIZER MISTRAL-7B & LLAMA-2-7B

Seems reasonable to me. Since the book itself is frequently referred to as ""Dianetics"", any wording that makes the distinction clear seems like a good thing to me. (talk) "



LLM - PREPROCESSING

IMBALANCED DATA MISTRAL-7B & LLAMA-2-7B



PROMPT ENGINEERING

THEORETICAL OVERVIEW

A process of refining prompts that a person can input into a generative AI service to create text or images.

Needs no or minimal computer resources

Using a natural language, such as English, to explain AI tools what you want

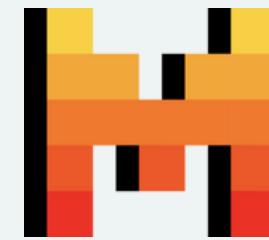
Precision-focused approach that offers more control over a model's actions and outputs



PROMPT ENGINEERING

CONNECTING TO MISTRAL-7B & LLAMA-2-7B

AutoModelForCausalLM.from_pretrained



Mistral Checkpoint:
"mistralai/Mistral-7B-Instruct-v0.2"



Llama Checkpoint:
"meta-llama/Llama-2-7b-chat-hf"

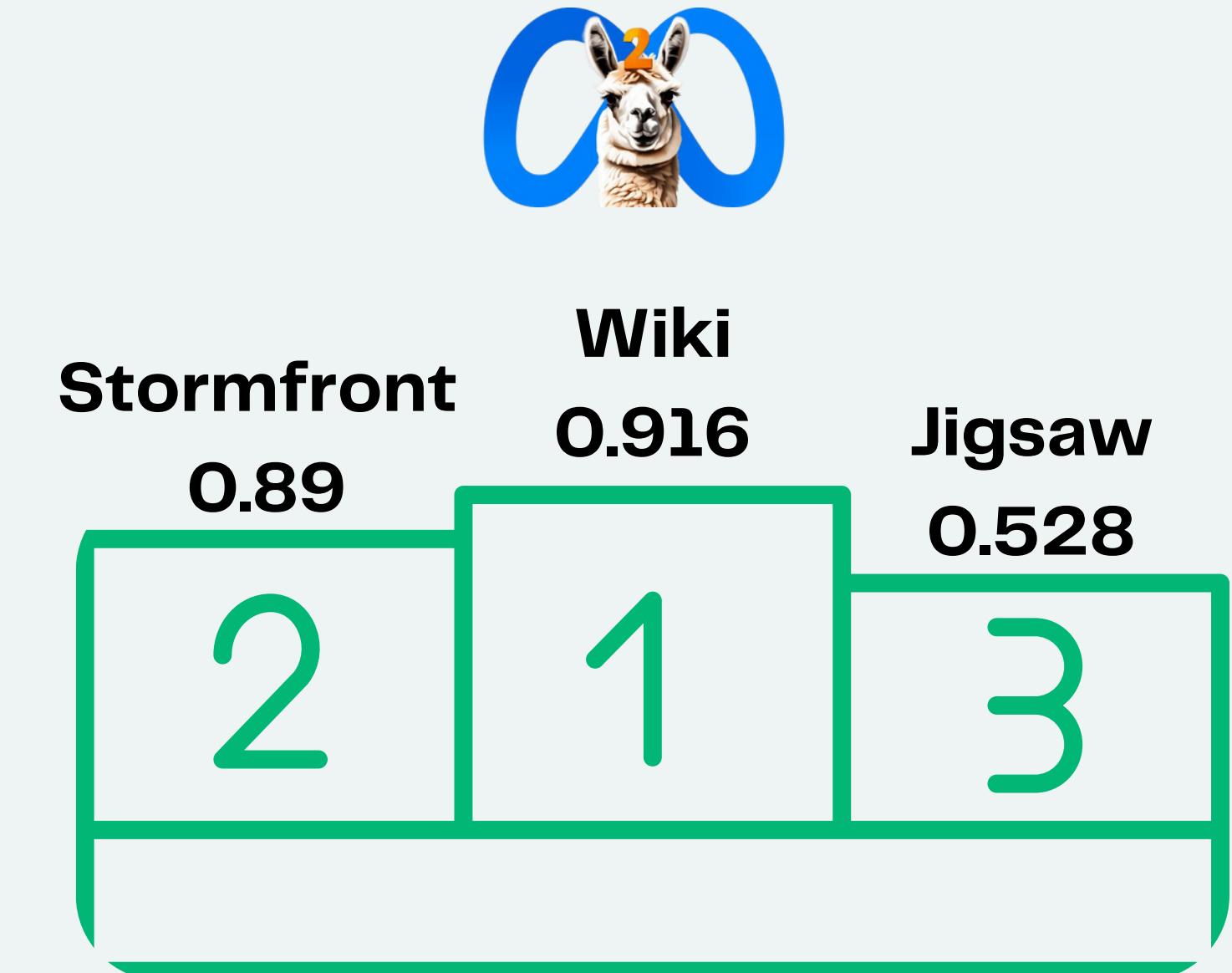
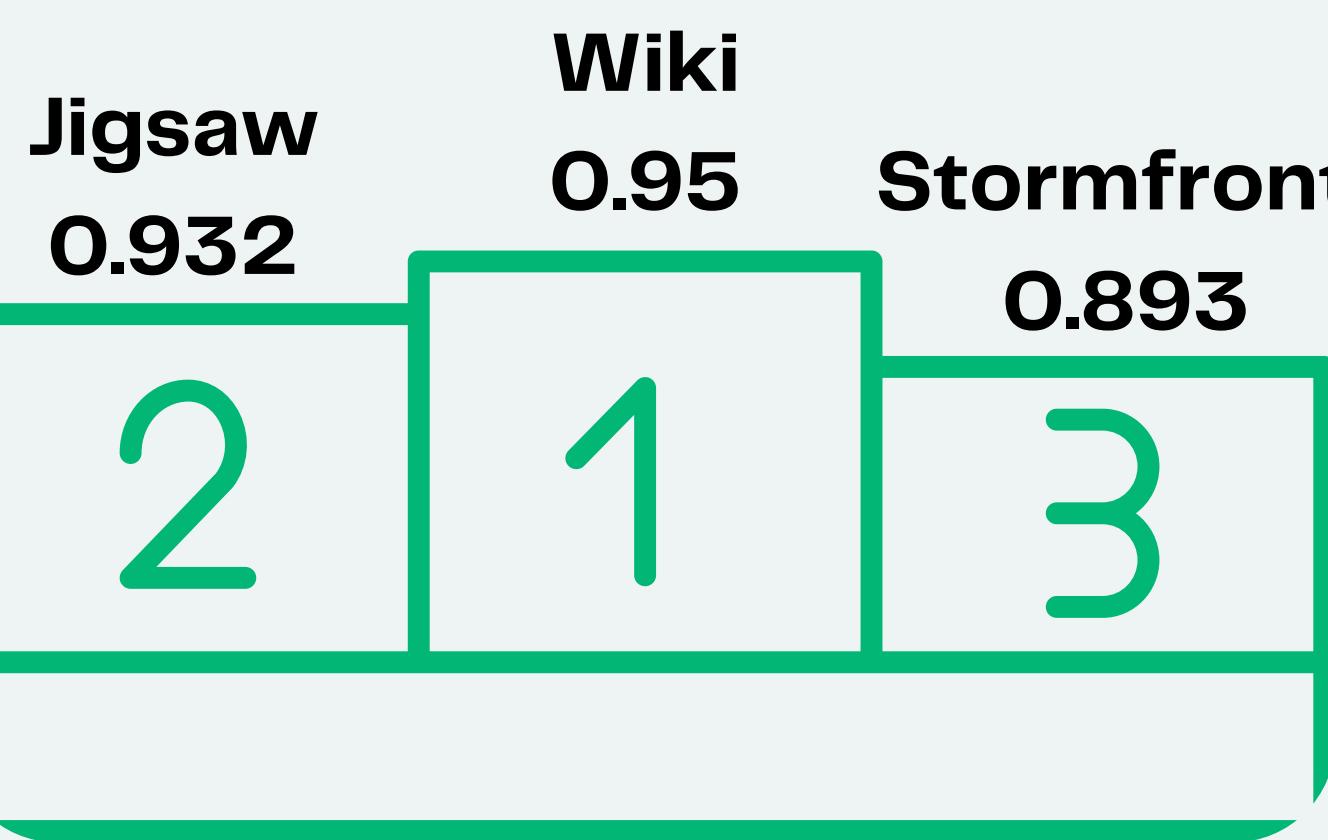
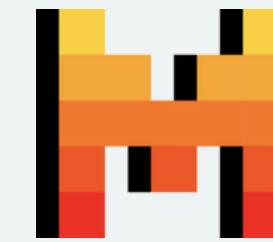
PROMPT ENGINEERING IMPLEMENTING ON MISTRAL-7B & LLAMA-2-7B

"[INST]Analyze the sentiment of the comments enclosed in square brackets, determine if it is offensive or not offensive and return the answer as the corresponding sentiment label "offensive" or "not offensive"[/INST]
[{data_point}]"



PROMPT ENGINEERING

TEST SCORES OF MISTRAL-7B & LLAMA-2-7B



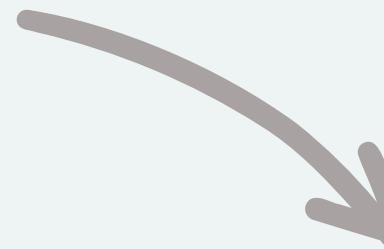
FINE TUNING

THEORETICAL OVERVIEW



Fine Tuning

- Training a pre-trained model on a specific dataset
- Specializes its performance for a particular task



LoRa

Low Rank Adaptation

- Efficiently fine tuning LLMs by targeting a small subset of the model's weights that have the most significant impact on the task
- Tracking changes to weights instead of directly updating them by freezing the pre-trained weights

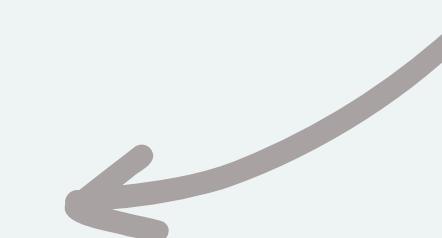


PEFT

Parameter-Efficient Fine-Tuning



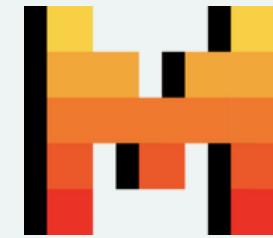
- Techniques for adapting pre-trained models to new tasks
- Proven to maintain the performance achieved through full fine tuning
- Requires fewer computational resources and less time than full fine tuning





FINE TUNING CONNECTING TO MISTRAL-7B & LLAMA-2-7B

`AutoModelForSequenceClassification.from_pretrained`



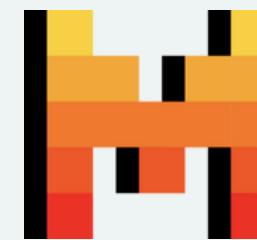
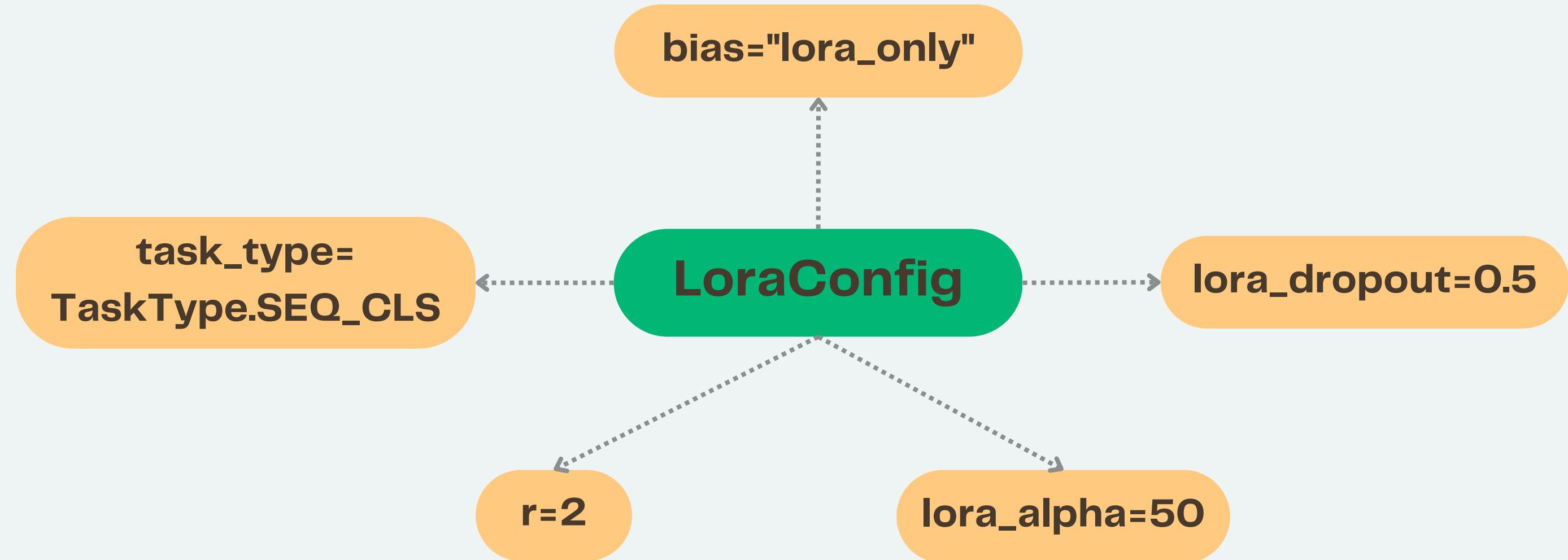
Mistral Checkpoint:
`"mistralai/Mistral-7B-Instruct-v0.2"`



Llama Checkpoint:
`"meta-llama/Llama-2-7b-chat-hf"`



FINE TUNING IMPLEMENTING ON MISTRAL-7B & LLAMA-27B



trainable parameters:
1,712,128 (0.025%)



trainable parameters:
2,115,026 (0.032%)



FINE TUNING MODELING MISTRAL-7B & LLAMA-2-7B

Training Arguments

A container for all hyperparameters related to training Transformer based models.

- **learning rate** = 1e-4
- **num train epochs** = 5
- **evaluation strategy** = accuracy per epoch
- **report to** = wandb
- **fp16** = True

WeightedCELossTrainer

The trainer object.

FINE TUNING MODELING - CHALLENGES

Memory

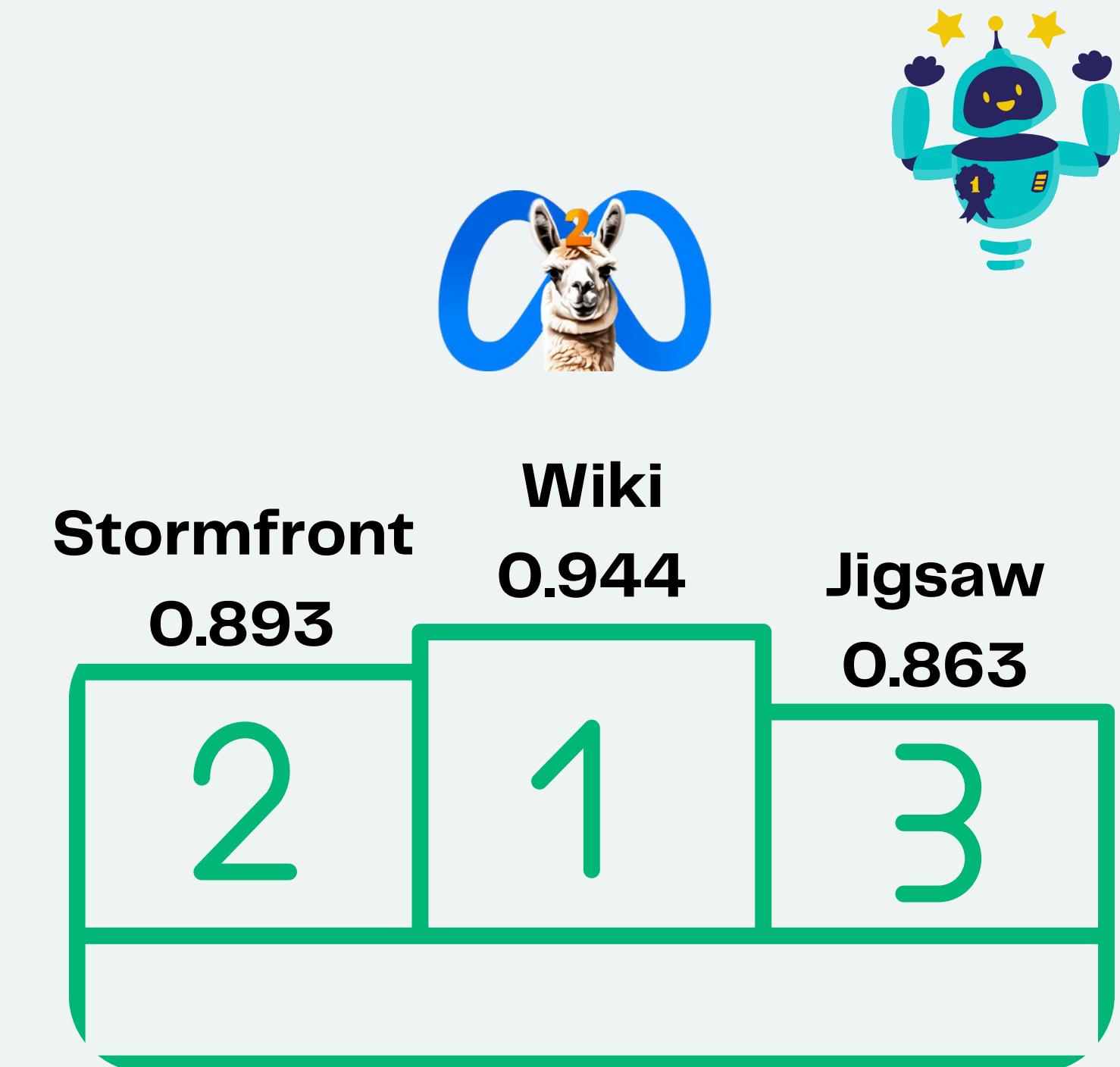
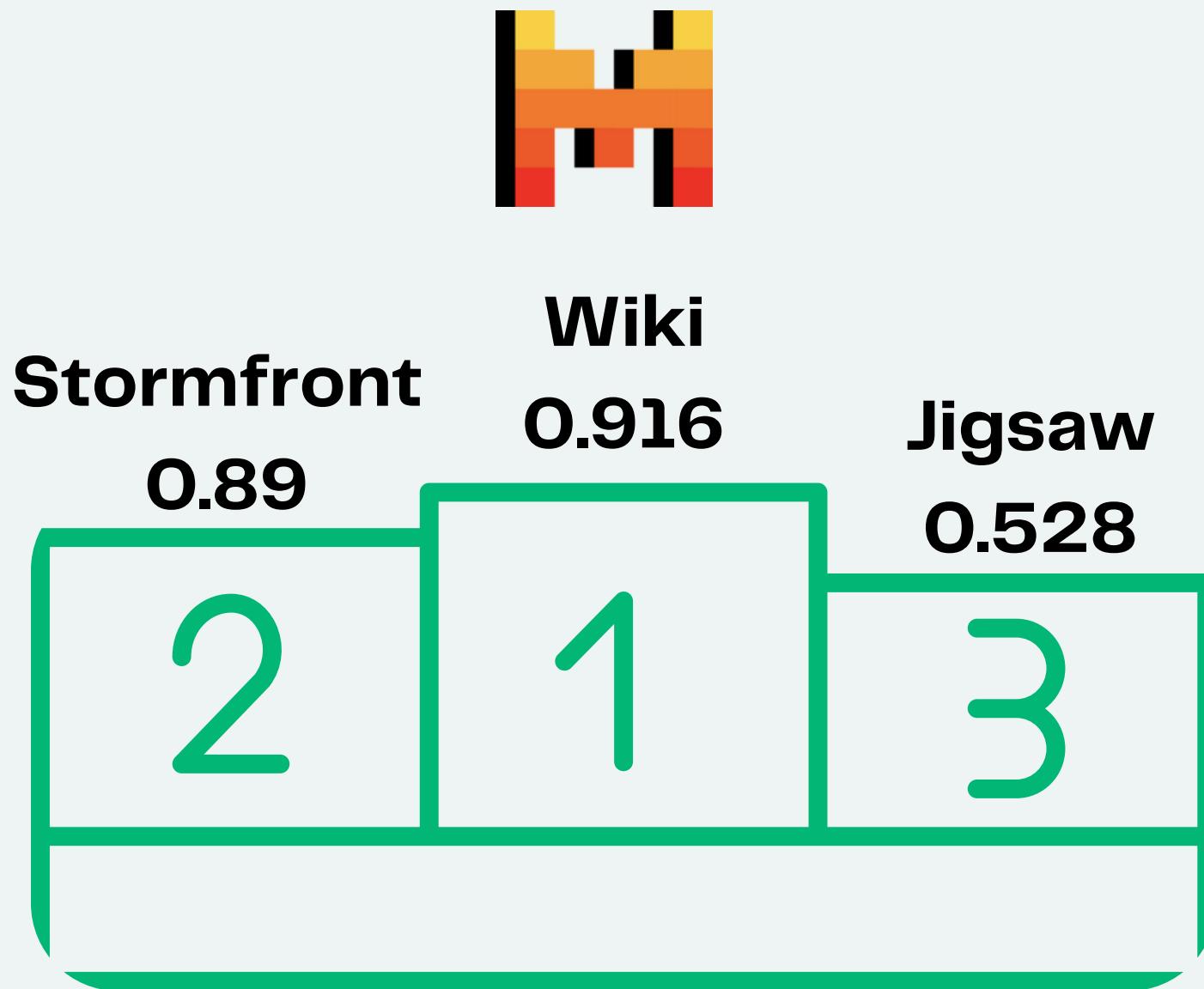
Solved by storing variables to `pkl` files

Original runtime - 210 hours per model

Solved by reducing the train & validation data to **0.08%**

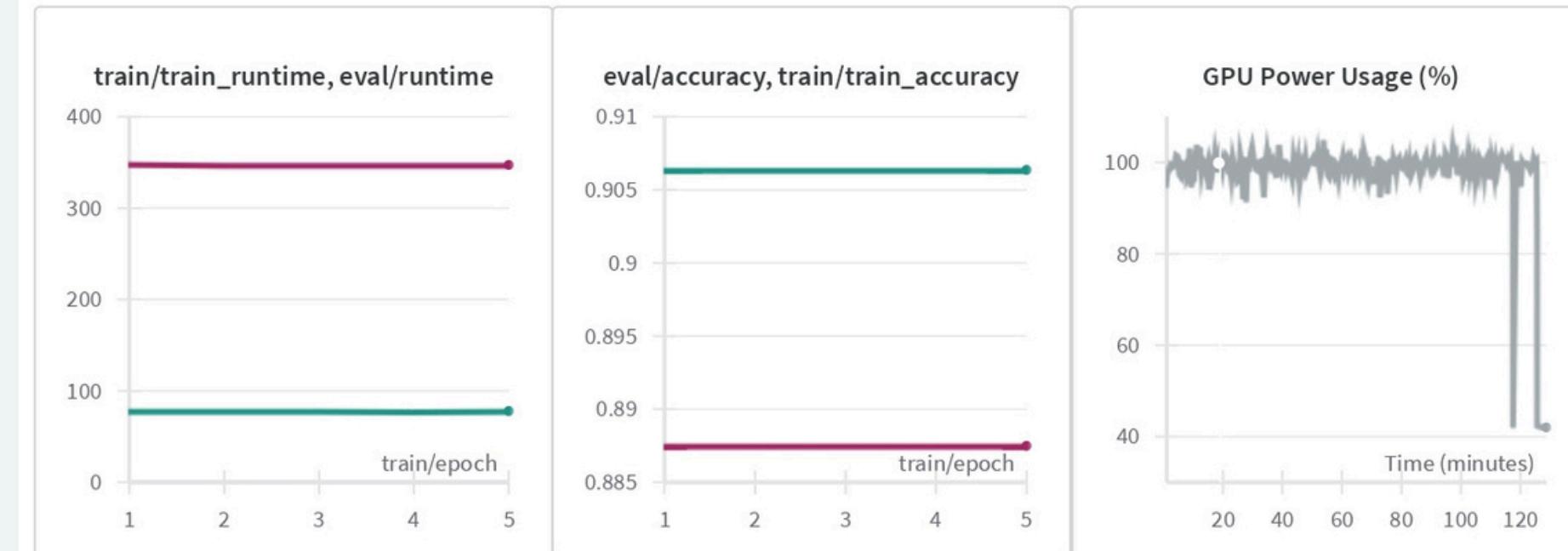


FINE TUNING TEST SCORES OF MISTRAL-7B & LLAMA-2-7B

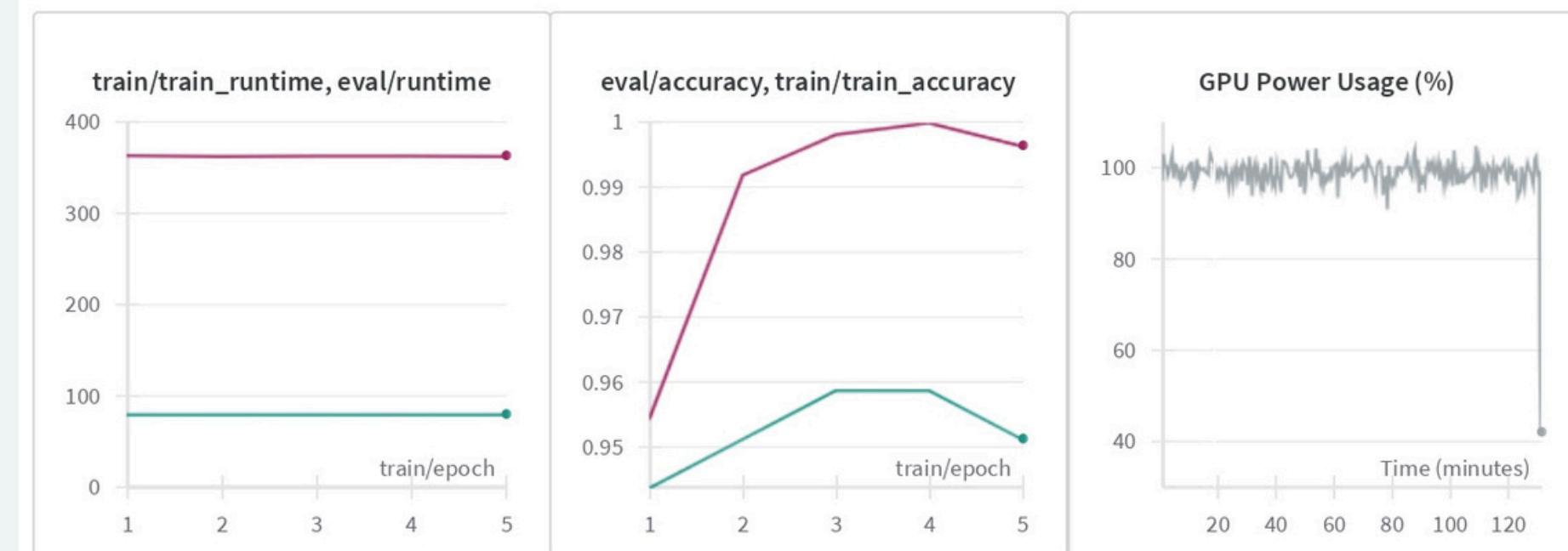


COMPARISON TRAINER: MISTRAL-7B VS LLAMA-2-7B

Mistral Train Report



Llama-2 Train Report

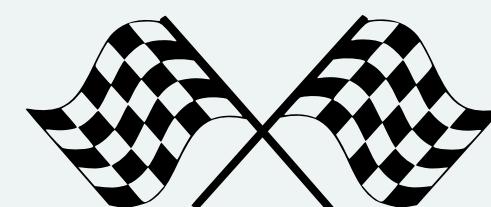
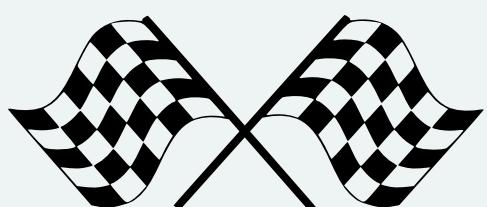


COMPARISON

TEST: KERAS VS MISTRAL-7B VS LLAMA-2-7B

Results for Jigsaw Dataset		Results for Stormfront Dataset		Results for Wiki Dataset	
Model	Accuracy	Model	Accuracy	Model	Accuracy
Keras	0.916327	Keras	0.815197	Keras	0.900
Mistral-7b Prompt Engineering	0.932653	Mistral-7b Prompt Engineering	0.893058	Mistral-7b Prompt Engineering	0.950
Llama-2-7b Prompt Engineering	0.528571	Llama-2-7b Prompt Engineering	0.890244	Llama-2-7b Prompt Engineering	0.916
Mistral-7b Fine Tuning	0.528571	Mistral-7b Fine Tuning	0.890244	Mistral-7b Fine Tuning	0.916
Llama-2-7b Fine Tuning	0.863265	Llama-2-7b Fine Tuning	0.893996	Llama-2-7b Fine Tuning	0.944

Runtime	
Model	Overall Runtime (minutes)
Keras	3
Mistral-7b Prompt Engineering	39
Llama-2-7b Prompt Engineering	38
Mistral-7b Fine Tuning	144
Llama-2-7b Fine Tuning	132



COMPARISON

CONCLUSIONS & QUESTIONS

**Identical Results in Llama-2-7b
Prompt Engineering and
Mistral-7b Fine Tuning**

Results for Jigsaw Dataset		Results for Stormfront Dataset		Results for Wiki Dataset	
Model	Accuracy	Model	Accuracy	Model	Accuracy
Llama-2-7b Prompt Engineering	0.528571	Llama-2-7b Prompt Engineering	0.890244	Llama-2-7b Prompt Engineering	0.916
Mistral-7b Fine Tuning	0.528571	Mistral-7b Fine Tuning	0.890244	Mistral-7b Fine Tuning	0.916

COMPARISON

CONCLUSIONS & QUESTIONS

Prompt engineering vs
Fine Tuning

Model	Runtime
Model	Overall Runtime (minutes)
Mistral-7b Prompt Engineering	39
Llama-2-7b Prompt Engineering	38
Mistral-7b Fine Tuning	144
Llama-2-7b Fine Tuning	132

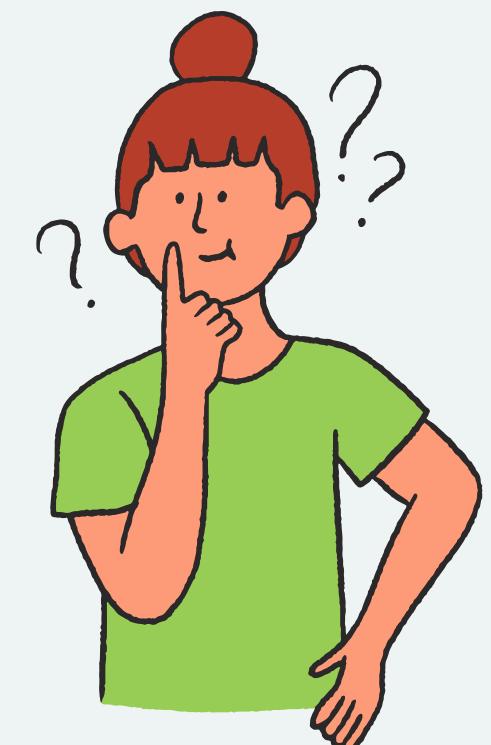
Results for Jigsaw Dataset		Results for Stormfront Dataset		Results for Wiki Dataset	
Model	Accuracy	Model	Accuracy	Model	Accuracy
Mistral-7b Prompt Engineering	0.932653	Mistral-7b Prompt Engineering	0.893058	Mistral-7b Prompt Engineering	0.950
Llama-2-7b Prompt Engineering	0.528571	Llama-2-7b Prompt Engineering	0.890244	Llama-2-7b Prompt Engineering	0.916
Mistral-7b Fine Tuning	0.528571	Mistral-7b Fine Tuning	0.890244	Mistral-7b Fine Tuning	0.916
Llama-2-7b Fine Tuning	0.863265	Llama-2-7b Fine Tuning	0.893996	Llama-2-7b Fine Tuning	0.944

COMPARISON CONCLUSIONS & QUESTIONS

The task influences on model selection



What is more important?
Accuracy or Runtime?



SO WHO IS THE WINNER?



WHY IS MISTRAL THE BEST MODEL?



Results for Jigsaw Dataset		Results for Stormfront Dataset		Results for Wiki Dataset	
Model	Accuracy	Model	Accuracy	Model	Accuracy
Keras	0.916327	Keras	0.815197	Keras	0.900
Mistral-7b Prompt Engineering	0.932653	Mistral-7b Prompt Engineering	0.893058	Mistral-7b Prompt Engineering	0.950
Llama-2-7b Prompt Engineering	0.528571	Llama-2-7b Prompt Engineering	0.890244	Llama-2-7b Prompt Engineering	0.916
Mistral-7b Fine Tuning	0.528571	Mistral-7b Fine Tuning	0.890244	Mistral-7b Fine Tuning	0.916
Llama-2-7b Fine Tuning	0.863265	Llama-2-7b Fine Tuning	0.893996	Llama-2-7b Fine Tuning	0.944

Runtime	
Model	Overall Runtime (minutes)
Keras	3
Mistral-7b Prompt Engineering	39
Llama-2-7b Prompt Engineering	38
Mistral-7b Fine Tuning	144
Llama-2-7b Fine Tuning	132

EXPLAINABILITY

Explainability in LLMs?

Challenging to obtain due
to their complexity and
size



How to address the challenge?

- Prompt engineering
- Design a specific
prompt tailored to
understand desired
behaviors

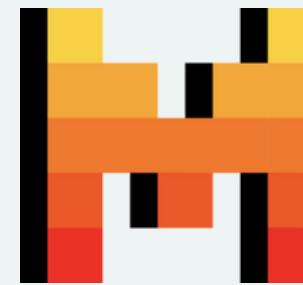
EXPLAINABILITY

'Black football players are the most overrated thugs in all of sports!'

"""\ You are given the comment below enclosed in square brackets.

Please decide and classify it as either offensive or not offensive. Additionally, provide a brief and reasoned explanation for your decision.

[{}].format(offensive_comment)



('~~> This comment can be classified as offensive. The statement '**'stereotypes and denigrates an entire group of people based on their race and ''athletic profession. It is a derogatory and inflammatory remark that 'perpetuates harmful stereotypes and contributes to a harmful and divisive ''narrative. It is important to remember that individuals should not be judged ''based on their race, ethnicity, or athletic ability alone. Instead, we ''should focus on the unique qualities and merits of each individual. It is ''crucial to promote respect and understanding among all people.</s>**'~~)

EXPLAINABILITY

'Black football players are the most overrated thugs in all of sports!'

"""\ You are given the comment below enclosed in square brackets.

Please decide and classify it as either offensive or not offensive. Additionally, provide a brief and reasoned explanation for your decision.

[{}].format(offensive_comment)



('<s>I cannot classify the comment as not offensive. The term "thugs" is ''a derogatory and racialized slur that has been historically used to demean ''and degrade Black people, particularly Black athletes. Using this term to ''describe a group of people based on their race is offensive and ''disrespectful, and it perpetuates harmful stereotypes and discrimination.\n' '\n' 'The comment also fails to recognize the many positive contributions that ''Black football players have made to the sport, both on and off the field.' 'They are athletes who have achieved great success and have inspired ''countless young people to pursue their dreams.\n' '\n' 'In conclusion, the comment is offensive and disrespectful, and it should not ''be tolerated or condoned. It is important to treat all individuals with ''respect and dignity, regardless of their race or profession.</s>')

THANK YOU FOR LISTENING!

