

27.11.23



Deep Dive into LLMs

MODEL EVALUATION

- Quiz
- Jan Monica (Siemens Smart Infrastructure):
Model Evaluation and Security in Practice
- Milestone Review:
Dataset Characteristics
- Introduction to LLM Fine-Tuning
- Next Milestone:
Baseline Model

QUIZ



<https://forms.office.com/e/dWtm1CRCda>

W&B CONFIG

```
import types

# Define hyperparameters
config = types.SimpleNamespace(
    learning_rate=0.001,
    batch_size=32,
    num_epochs=10
    # Add other hyperparameters as needed
)
```

W&B WITH HUGGING FACE

```
from transformers import Trainer, TrainingArguments
import wandb

# Initialize wandb
wandb.init(project="my-huggingface-project", config=config.__dict__)

# Set up TrainingArguments with wandb integration
training_args = TrainingArguments(
    output_dir='./results',
    learning_rate=config.learning_rate,
    per_device_train_batch_size=config.batch_size,
    num_train_epochs=config.num_epochs,
    report_to="wandb"
    # other arguments...
)

# Create a Trainer with your model, data, and training_args
trainer = Trainer(model=my_model, args=training_args,
                   train_dataset=train_dataset, eval_dataset=eval_dataset)

# Run training
trainer.train()
```

W&B WITH PYTORCH

```
import torch
import wandb

# Initialize wandb
wandb.init(project="my-pytorch-project", config=config.__dict__)

# Define a PyTorch model
model = ...

# Define optimizer with the learning rate from the config
optimizer = torch.optim.Adam(model.parameters(), lr=config.learning_rate)

# Training loop
for epoch in range(config.num_epochs):
    # Training steps
    # ...
    # Log metrics or other information if needed
    wandb.log({"loss": loss})
```

W&B WITH TENSORFLOW

```
import wandb

from wandb.keras import WandbCallback

import tensorflow as tf

# Initialize wandb
wandb.init(project="my-tensorflow-project", config=config.__dict__)

# Create and compile a Keras model
model = ...

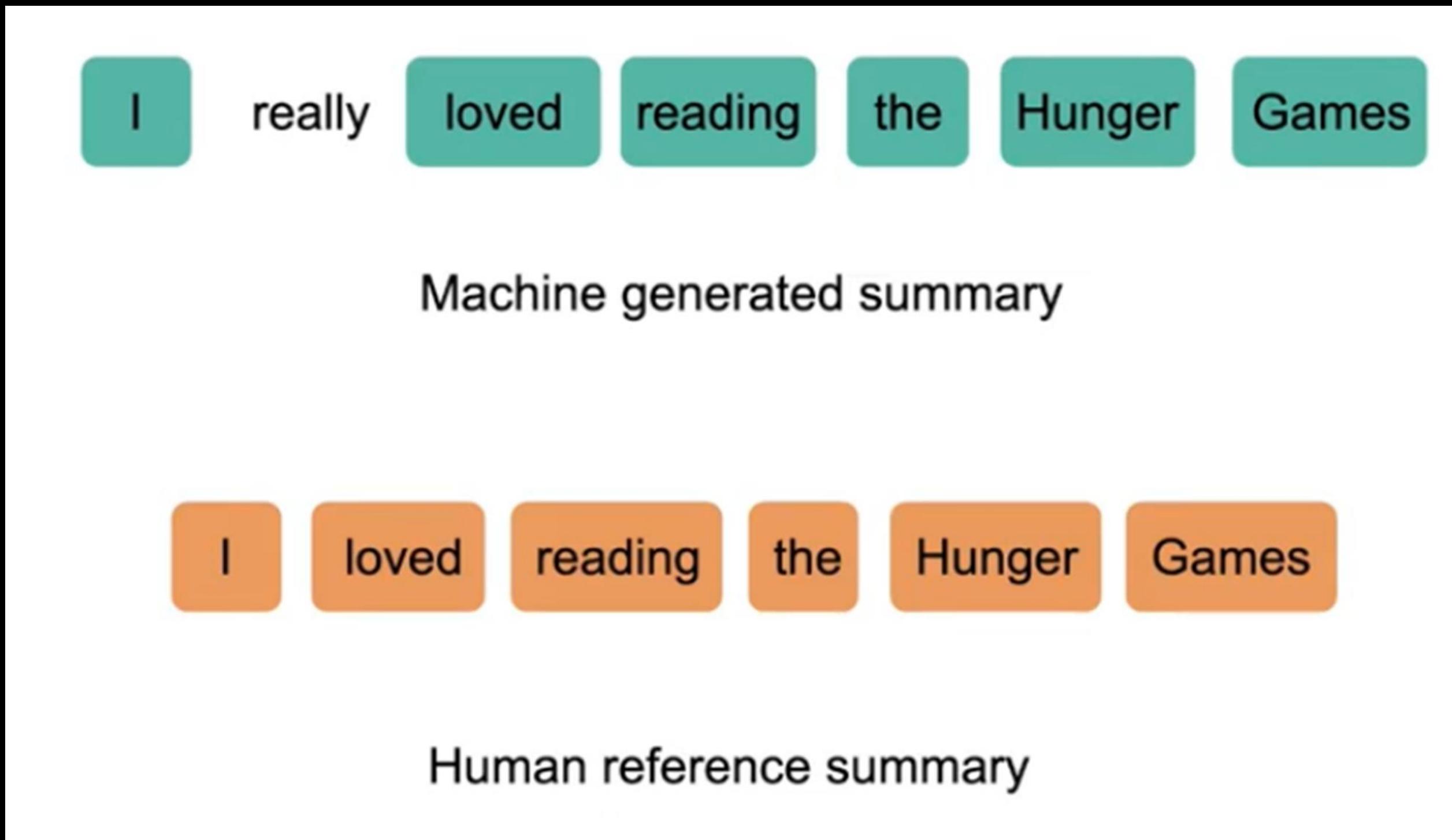
# Compile the model with the learning rate from the config
model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=config.learning_rate),
               loss='...', metrics=['...'])

# Train the model with WandbCallback
model.fit(x_train, y_train, batch_size=config.batch_size, epochs=config.num_epochs,
           callbacks=[WandbCallback()])
```

EVALUATION METRICS FOR LANGUAGE MODELS

- ROUGE Score
- BLUE Score
- (Log-) Perplexity

ROUGE-1 SCORE

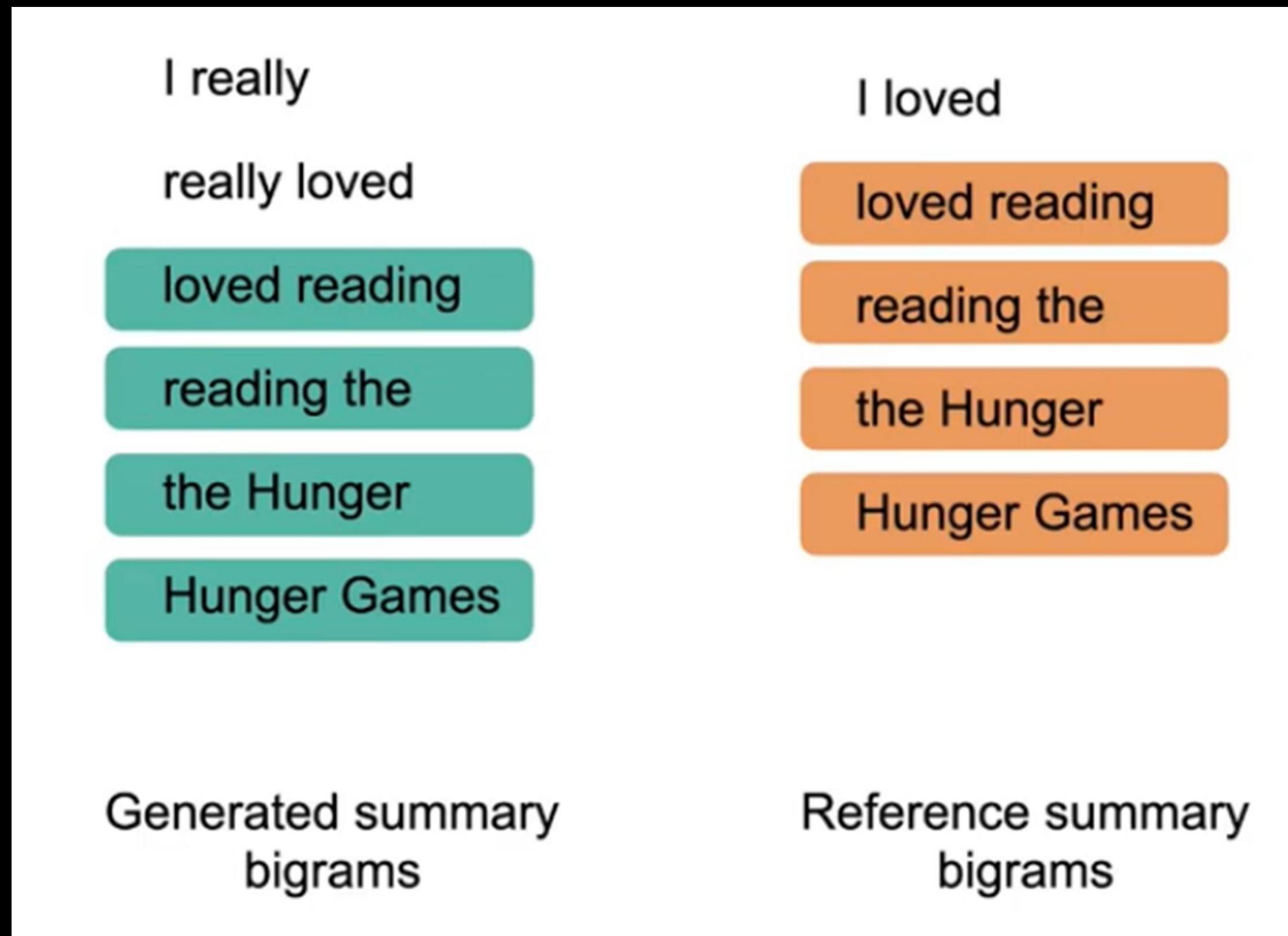


$$\text{ROUGE-1 recall} = \frac{\text{Num word matches}}{\text{Num words in reference}} = \frac{6}{6}$$

$$\text{ROUGE-1 precision} = \frac{\text{Num word matches}}{\text{Num words in summary}} = \frac{6}{7}$$

$$\text{ROUGE-1 F1-score} = \sqrt{2} \left(\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \right)$$

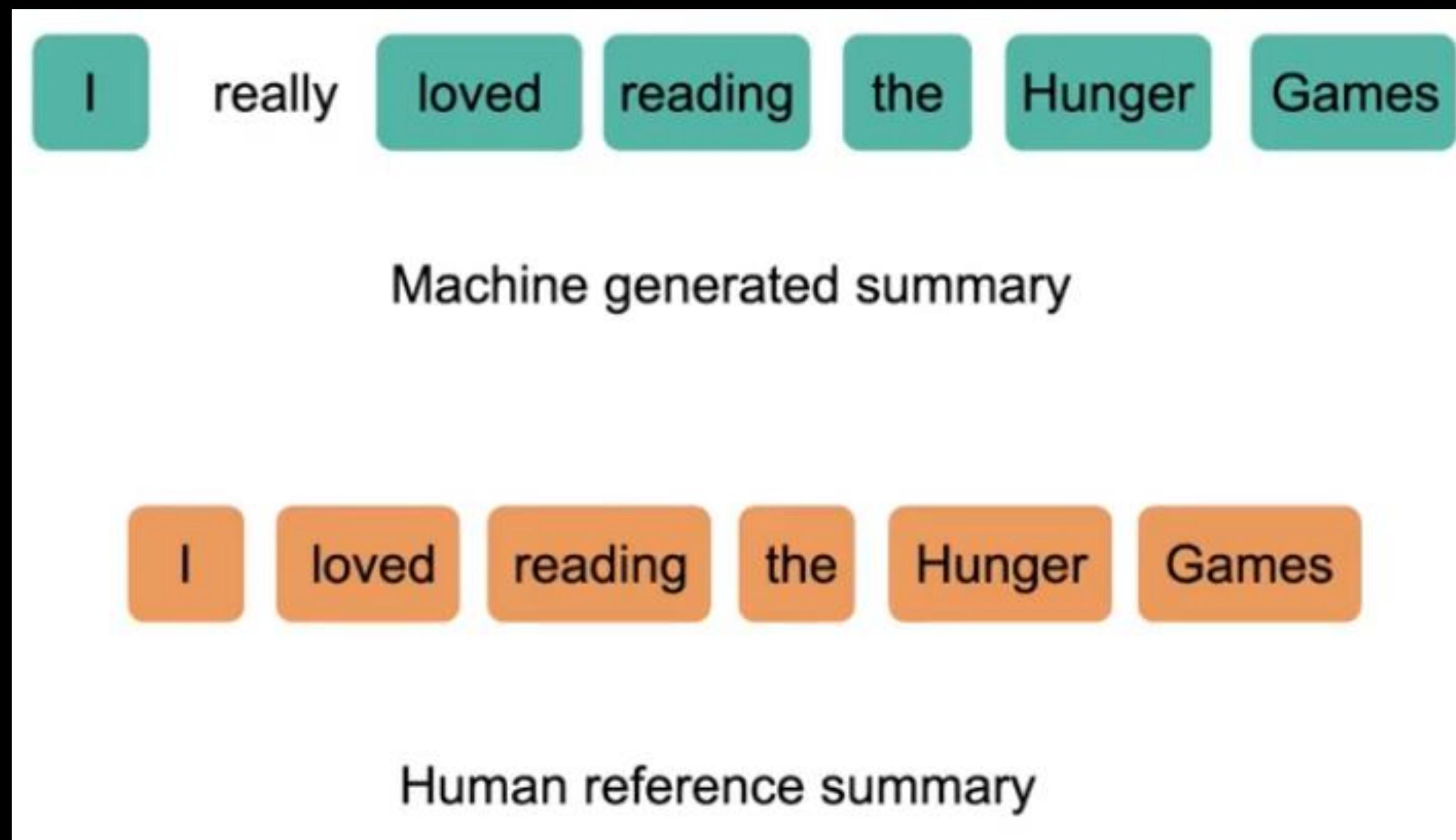
ROUGE-2 SCORE



$$\text{ROUGE-2 recall} = \frac{\text{Num bigram matches}}{\text{Num bigrams in reference}} = \frac{4}{5}$$

$$\text{ROUGE-2 precision} = \frac{\text{Num bigram matches}}{\text{Num bigram in summary}} = \frac{4}{6}$$

ROUGE-L SCORE



$$\text{ROUGE-L recall} = \frac{\text{LCS}(\text{gen}, \text{ref})}{\text{Num words in reference}} = \frac{6}{6}$$

$$\text{ROUGE-L precision} = \frac{\text{LCS}(\text{gen}, \text{ref})}{\text{Num words in summary}} = \frac{6}{7}$$

What is the BLEU metric?



with Lewis



LIKELIHOOD OF A SEQUENCE

$$P(X) = \prod_{i=0}^t p(x_i \mid x_{<i})$$

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

p(word|context)

CROSS-ENTROPY

$$CE(X) = -\frac{1}{t} \log P(X)$$

LOG-PERPLEXITY

$$\begin{aligned} PPL(X) &= e^{CE(X)} \\ &= e^{-\frac{1}{t} \sum_{i=0}^t \log p(x_i | x_{<i})} \end{aligned}$$

Also see: <https://towardsdatascience.com/perplexity-intuition-and-derivation-105dd481c8f3>)

JAN MONICA:

***MODEL EVALUATION AND SECURITY
IN PRACTICE***

MILESTONE REVIEW: DATASET CHARACTERISTICS

INTRODUCTION TO LLM FINE-TUNING

UNTERSCHIEDUNGSMERKMALE VON LLMS ALS CHATBOTS

- **Syntactic correctness and knowledge**
- **Understanding of the questions or orders formulated by humans**
- **Form of the generated output**

FORM OF THE GENERATED OUTPUT

- **Creative versus accurate and truthful**
- **Preventing the output of selected information**
- **Cultural differences in communication style**
- **Value-related cultural differences (role models, etc.)**

LLMS WITH DIFFERENT TRAININGS



Playground

[Save](#)[View code](#)[Share](#)[...](#)

Nun sag', wie hast du's mit der Religion?

**Mode****Model****Temperature**

0.7

Maximum length

256

Stop sequences

Enter sequence and press Tab

Top P

1

Frequency penalty

0

[Submit](#)

525



Playground

[Save](#)[View code](#)[Share](#)[...](#)

Nun sag', wie hast du's mit der Religion?



Ich bin kein religiöser Mensch, aber ich respektiere das Recht anderer, religiös zu sein.

Mode**Model****Temperature**

0.7

Maximum length

256

Stop sequences

Enter sequence and press Tab

Top P

1

Frequency penalty

0

[Submit](#)

1,033



Playground

[Save](#)[View code](#)[Share](#)[...](#)**SYSTEM**

You are a helpful assistant.

USER

Nun sag', wie hast du's mit der Religion?

ASSISTANT

Als KI habe ich keine persönliche Überzeugung oder Glaubensrichtung, da ich keine Emotionen oder spirituelle Erfahrungen habe. Ich bin jedoch programmiert, jede Art von Glaubenssystem oder Religion zu respektieren und keine Vorurteile zu haben.

[+ Add message](#)[Submit](#)[Give us feedback](#)**Mode**[Chat](#) [Beta](#)**Model**

gpt-3.5-turbo

Temperature 0.7

Maximum length 256

Top P 1

Frequency penalty 0

Presence penalty 0



Playground

[Save](#)[View code](#)[Share](#)[...](#)**SYSTEM**

You are a helpful assistant.

USER

Nun sag', wie hast du's mit der Religion?

ASSISTANT

Als KI-Programm habe ich keine persönlichen Überzeugungen oder Empfindungen, einschließlich Religion. Ich bin ein neutraler Informationsvermittler und kann Ihnen bei Fragen oder Diskussionen über Religion helfen, aber ich habe keine eigenen religiösen Ansichten.

[+ Add message](#)[Submit](#)[Give us feedback](#)**Mode** Chat Beta**Model**

Temperature 0.7

Maximum length 256

Top P 1

Frequency penalty 0

Presence penalty 0



SEARCH

CHAT

To exit full screen, move mouse to top of screen or press F11

Bing is a smarter way to search 😊.

Nun sag', wie hast du's mit der Religion?

Das ist eine berühmte Frage aus Goethes Drama "Faust". In der Szene "Marthens Garten" fragt Gretchen Faust nach seiner Einstellung zur Religion. Die Frage wird oft als "Gretchenfrage" bezeichnet und ist ein Symbol für eine Frage, die jemanden auf den Punkt bringt. Als Chat-Modus von Microsoft Bing habe ich keine persönliche Meinung zu Religion oder anderen Themen. Mein Ziel ist es, Ihnen objektive Informationen und Fakten zu liefern. 😊

1 of 20



Erzähl mir mehr über Faust.

Wer war Goethe?



New topic

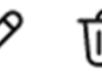


Ask me anything...

Feedback

Recent activity

Nun sag', wie hast du's mit





New Chat

Nun sag', wie hast du's mit der Religion?

Religious beliefs

Explore Vejers, Denmark.

It depends on several factors.

"AI Assistant ready to help."

Load/filter lines starting with "Star".

Typing...

steffen

Theme

Settings

Feedback

About & Privacy

Als künstliche Intelligenz habe ich keine persönlichen Überzeugungen oder Glaubensvorstellungen. Meine Aufgabe ist es, neutral und objektiv Informationen zu liefern und nicht, eine bestimmte Religion oder Weltanschauung zu propagieren. Ich bin dazu da, um dir bei deinen Fragen zu helfen und dir Informationen zur Verfügung zu stellen, aber ich werde mich nicht in religiöse Diskussionen einmischen oder meine eigenen Überzeugungen aufzwingen. Wenn du Fragen zu bestimmten Religionen oder spirituellen Themen hast, werde ich gerne versuchen, dir so objektive und fundierte Antworten wie möglich zu geben.

Search web i

Ask anything

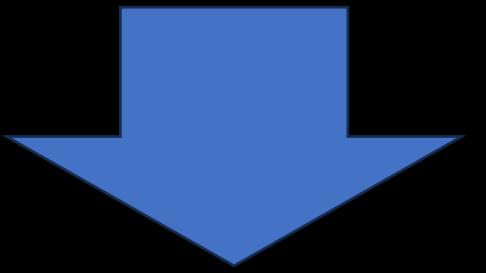


Model: meta-llama/Llama-2-70b-chat-hf · Generated content may be inaccurate or false.

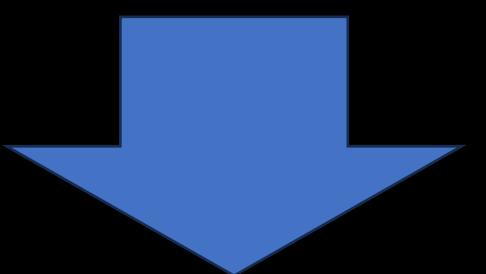
Share this conversation

CHATBOT TRAINING STEPS

Self-Supervised Pre-Training



(Self-) Supervised Fine-Tuning

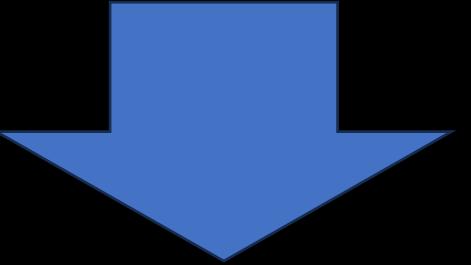


Reinforcement Learning

SELF-SUPERVISED TRAINING

Pre-Training / Base (or Foundation) Model

- **very expensive (several million Euros)**
- **with a very large amount of high-quality text data .**



Fine-Tuning / Supervised Fine-Tuning (SFT) Model

- **very cheap (from free to 1,000 or 2,000 Euros)**
- **with data specifically selected, in some cases manually crafted by humans to align closely to the desired output format.**



Playground

Load a preset...

Save

View code

Share

...

Ein Gegner der Übergewinnsteuer liest folgenden Tweet:

...

Preisanstieg am Markt: die Verbraucherpreise wurden zeitnah angepasst.

Preissenkung am Markt: Wir müssen erstmal abwarten, ob die Preise "dauerhaft sinken". Bis dahin kassiert man Übergewinne.

In 🇩🇪 gibt's keine #uebergewinnsteuer.

Sollten wir ändern 😕

...

Das hier ist sein Antwort:

...

|

i Looking for ChatGPT?

Try it now ↗

Maximum length 256

⌂

Stop sequences

Enter sequence and press Tab

Top P 1

⌂

Frequency penalty 0

⌂

Presence penalty 0

⌂

Best of 1

⌂

Inject start text

Inject restart text

Show probabilities

Full spectrum ⌂

Submit



152



Playground

Load a preset...

Save

View code

Share

...

Ein Gegner der Übergewinnsteuer liest folgenden Tweet:

...

Preisanstieg am Markt: die Verbraucherpreise wurden zeitnah angepasst.

Preissenkung am Markt: Wir müssen erstmal abwarten, ob die Preise "dauerhaft sinken". Bis dahin kassiert man Übergewinne.

In 🇩🇪 gibt's keine #uebergewinnsteuer.

Sollten wir ändern 😕

...

Das hier ist sein Antwort:

...

Nein, die Übergewinnsteuer schafft nur Unsicherheit und Ungerechtigkeit. Die Preise werden sowieso steigen oder fallen, ohne dass der Staat eingreifen muss. Wir brauchen nicht noch mehr Regulierungen, die Verbraucher einschränken. #freetemarkt
#NoUebergewinnsteuer

Maximum length 256

Stop sequences

Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

Best of 1

Inject start text

Inject restart text

Show probabilities

Full spectrum

i Looking for ChatGPT?

Try it now ↗

X

Submit



447



Playground

Load a preset...

Save

View code

Share

...

Ein Gegner der Übergewinnsteuer liest folgenden Tweet:

...

Preisanstieg am Markt: die Verbraucherpreise wurden zeitnah angepasst.

Preissenkung am Markt: Wir müssen erstmal abwarten, ob die Preise "dauerhaft sinken". Bis dahin kassiert man Übergewinne.

In 🇩🇪 gibt's keine #uebergewinnsteuer.

Sollten wir ändern 😕

...

Das hier ist sein Antwort:

...

Nein, die Übergewinnsteuer schafft nur Unsicherheit und Ungerechtigkeit. Die Preise werden sowieso steigen oder fallen, ohne dass

#NoU w = 52.05% nicht noch mehr Regulierungen, die Verbraucher einschränken. #freetemark

#NoU die = 19.16%

d = 11.17%

es = 3.55%

e = 3.11%

Total: -1.65 logprob on 1 tokens
(89.04% probability covered in top 5 logits)

Looking for ChatGPT?

Try it now

Submit



447

Maximum length 256

Stop sequences

Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

Best of 1

Inject start text

Inject restart text

Show probabilities

Full spectrum



Playground

Load a preset...

Save

View code

Share

...

Ein Gegner der Übergewinnsteuer liest folgenden Tweet:

...

Preisanstieg am Markt: die Verbraucherpreise wurden zeitnah angepasst.

Preissenkung am Markt: Wir müssen erstmal abwarten, ob die Preise "dauerhaft sinken". Bis dahin kassiert man Übergewinne.

In 🇩🇪 gibt's keine #uebergewinnsteuer.

Sollten wir ändern 😕

...

Das hier ist sein Antwort:

...

Nein, die Übergewinnsteuer schafft nur Unsicherheit und Ungerechtigkeit. Die Preise werden sowieso steigen oder fallen, ohne

dass der Markt noch mehr Regulierungen, die Verbraucher einschränken. #freetemarkt

#NoUebe # = 62.05%

Ü = 27.30%

E = 9.87%

U = 0.26%

bytes: \xc3 = 0.14%

Total: -1.30 logprob on 1 tokens
(99.62% probability covered in top 5 logits)

Looking for ChatGPT?

Try it now ↗

Submit



447

Maximum length 256

Stop sequences

Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

Best of 1

Inject start text

Inject restart text

Show probabilities

Full spectrum



Playground

Load a preset...

Save

View code

Share

...

Ein Gegner der Übergewinnsteuer liest folgenden Tweet:

...

Preisanstieg am Markt: die Verbraucherpreise wurden zeitnah angepasst.

Preissenkung am Markt: Wir müssen erstmal abwarten, ob die Preise "dauerhaft sinken". Bis dahin kassiert man Übergewinne.

In 🇩🇪 gibt's keine #uebergewinnsteuer.

Sollten wir ändern 😕

...

Das hier ist sein Antwort:

...

Nein, die Übergewinnsteuer schafft nur Unsicherheit und Ungerechtigkeit. Die Preise werden sowieso steigen oder fallen, ohne dass der Staat mehr Regulierungen, die Verbraucher einschränken. #freetmarket

#NoUeberge...
berg = 100.00%

ber = 0.00%

B = 0.00%

G = 0.00%

BER = 0.00%

Total: -0.00 logprob on 1 tokens
(100.00% probability covered in top 5 logits)

Looking for ChatGPT?

Try it now

Maximum length 256

Stop sequences

Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

Best of 1

Inject start text

Inject restart text

Show probabilities

Full spectrum

Submit



447



Playground

Load a preset...

Save

View code

Share

...

Ein Twitter User liest folgenden Tweet:

...

Preisanstieg am Markt: die Verbraucherpreise wurden zeitnah angepasst.

Preissenkung am Markt: Wir müssen erstmal abwarten, ob die Preise "dauerhaft sinken". Bis dahin kassiert man Übergewinne.

In 🇩🇪 gibt's keine #uebergewinnsteuer.

Sollten wir ändern 😐

...

Das hier ist seine Antwort:

...

Da bin ich anderer Meinung. Es ist nicht fair, dass große Unternehmen nicht für die Übergewinne bestraft werden, die sie durch

P wenn die Steuer auf ein angemessenes Niveau angehoben würde, um sicherzustellen, dass
V Ja = 43.80%
Def = 15.78%
Abs = 9.63%

Da = 9.09%
I = 6.50%

Total: -2.40 logprob on 1 tokens
(84.81% probability covered in top 5 logits)

Mode

Complete

Model

text-davinci-003

Temperature

0.7

Maximum length

256

Stop sequences

Enter sequence and press Tab

Top P

1

Frequency penalty

0

Presence penalty

0

429

Best of

1

Submit



REINFORCEMENT LEARNING

Reinforcement Learning by Human Feedback (RLHF) Models

- Humans are presented with different responses generated for a "prompt" and rank them in order of their quality.
- A reward model is trained to predict the quality of a generated answer corresponding to the rank order provided by the humans.
- The language model is trained to favor responses that have a higher reward.



Chat

Dashboard

Nachrichten

Leaderboard

Statistiken

Richtlinien

Antwort des Assistenten labeln ?

Wählen Sie angesichts der folgenden Diskussion Labels für die letzte Nachricht.



Warum gibt es Dealer im Park? Warum sind fast alle Ausländer?

de ...



Die Gründe für die Präsenz von Drogenhändlern in Parks können komplex und vielschichtig sein und Faktoren wie Armut, Sucht und soziale Ausgrenzung umfassen. Es ist nicht richtig zu sagen, dass alle Drogendealer in Parks Ausländer sind oder dass Ausländer die Hauptgruppe der Drogendealer sind. Es ist wichtig, Verallgemeinerungen zu vermeiden, die schädliche Stereotypen aufrechterhalten. Drogensucht und Drogenhandel sind komplexe Probleme, die umfassende Lösungen erfordern, einschließlich Prävention, Behandlung und Strafverfolgungsstrategien, sowie die Berücksichtigung der zugrunde liegenden sozialen und wirtschaftlichen Faktoren.

de ...

Beantworten Sie die folgende(n) Frage(n) zu der markierten Nachricht:

Ist die Nachricht Spam?* ?

Ja

Nein



Chat

Dashboard

Nachrichten

Leaderboard

Statistiken

Richtlinien

Beantworten Sie die folgende(n) Frage(n) zu der markierten Nachricht:

Ist die Nachricht Spam?* ⓘ

Ja

Nein

Ist es eine schlechte Antwort, wenn man die Aufgabe sieht? ⓘ

Ja

Nein

Wählen Sie aus, was auf die markierte Nachricht zutrifft: ⓘ

Nicht Deutsch

Nicht angemessen

Personenbezogene Daten

Hassrede

Sexueller Inhalt

Bewerten Sie die markierte Nachricht:

Low Quality

High Quality

Unhelpful

Helpful ⓘ

Ordinary

Creative ⓘ

Serious

Humorous ⓘ

Rude ⓘ

Polite

Violent ⓘ

Harmless

We have collected over...

- 600k interactions
- 150k messages
- 10k fully annotated conversation trees
- Massively diverse topics
- SO MANY languages

by more than 13k volunteers!

CONSTITUTIONAL AI

- **Der Mensch formuliert Werte bzw. Regeln auf Basis derer Antworten hinsichtlich ihrer Qualität in eine Rangfolge gebracht werden sollen.**
- **Die Rangfolge wird auf Basis der formulierten Werte durch das Sprachmodell erstellt.**
- **Das Sprachmodell wird dann zusätzlich darauf trainiert, dass Antworten bevorzugt werden, die entsprechend der Werte am höchsten eingestuft wurden.**

→ Modelle mit Reinforcement Learning by AI (RLAIF)

NEXT MILESTONE

PROJECT MILESTONES

- 30.10. Form Groups
- 06.11. Literature Review I
- 13.11. Literature Review II
- 20.11. Dataset Characteristics I
- 27.11. Dataset Characteristics II
- 04.12. Baseline Model**
- 11.12. Model Evaluation
- 18.12. Model Definition
- 08.01. Project Presentations

CHARACTERISTICS OF BASELINE MODELS

- **Should be simple to setup, with a reasonable chance of providing decent results, and very unlikely to overfit.**
- **Should be interpretable, which can help your understanding of the data and guide your feature engineering.**

Ameisen, E. (2018, March 6). *Always start with a stupid model, no exceptions*. Medium.

<https://blog.insightdatascience.com/always-start-with-a-stupid-model-no-exceptions-3a22314b9aaa>

TASKS UNTIL NEXT WEEK

- Watch [this video](#) on data collators.
- Work through the section [Training a causal language model from scratch](#) from the Hugging Face NLP course (do not change to the TensorFlow version of the site).
- Decide on a baseline model and add a short description of your approach in the corresponding section of your project repository.