Ethical and Responsible Large Language Models: Challenges and Best Practices

Nicole Koenigstein, Miquel Noguer Alonso

Content

Overview of the main challenges:

- Transparency and Explainability
- Controlling Generated Content
- Bias Mitigation

Code

https://github.com/Nicolepcx/Transformers-in-Action

https://github.com/Nicolepcx/ACM_TechTalk

Transparency and Explainability

Importance of Transparency and Explainability

- Trust and accountability
- Legal and ethical considerations
- Facilitating user understanding

Attention Visualization

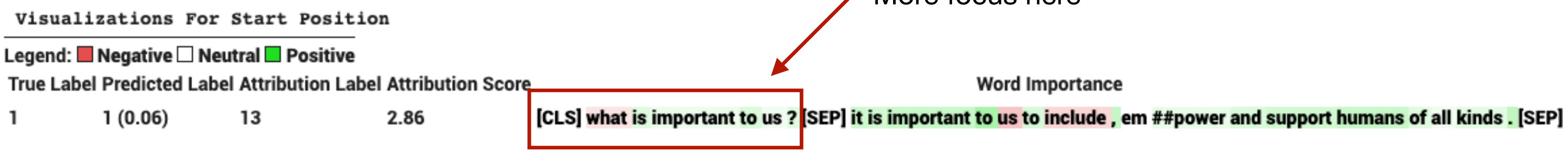
- Visualizing the focus of attention in Transformer models
- Improving understanding of model decision-making

https://captum.ai/

Visualizing Attribution

Visualizing attributions* for each word token in the sequence.

More focus here

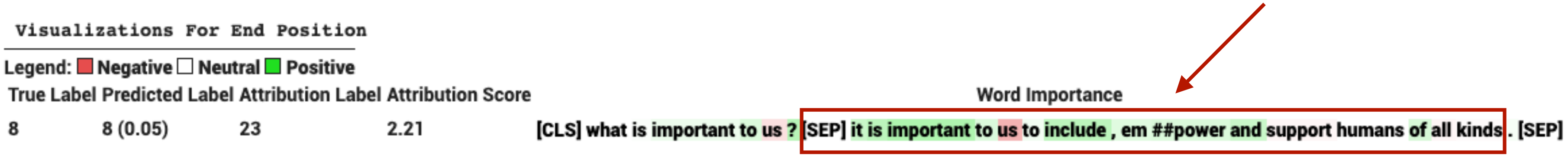


Visualizations For Start Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	1 (0.06)	13	2.86	[CLS] what is important to us ? [SEP] it is important to us to include , em ##power and support humans of all kinds . [SEP]

More focus here



Visualizations For End Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

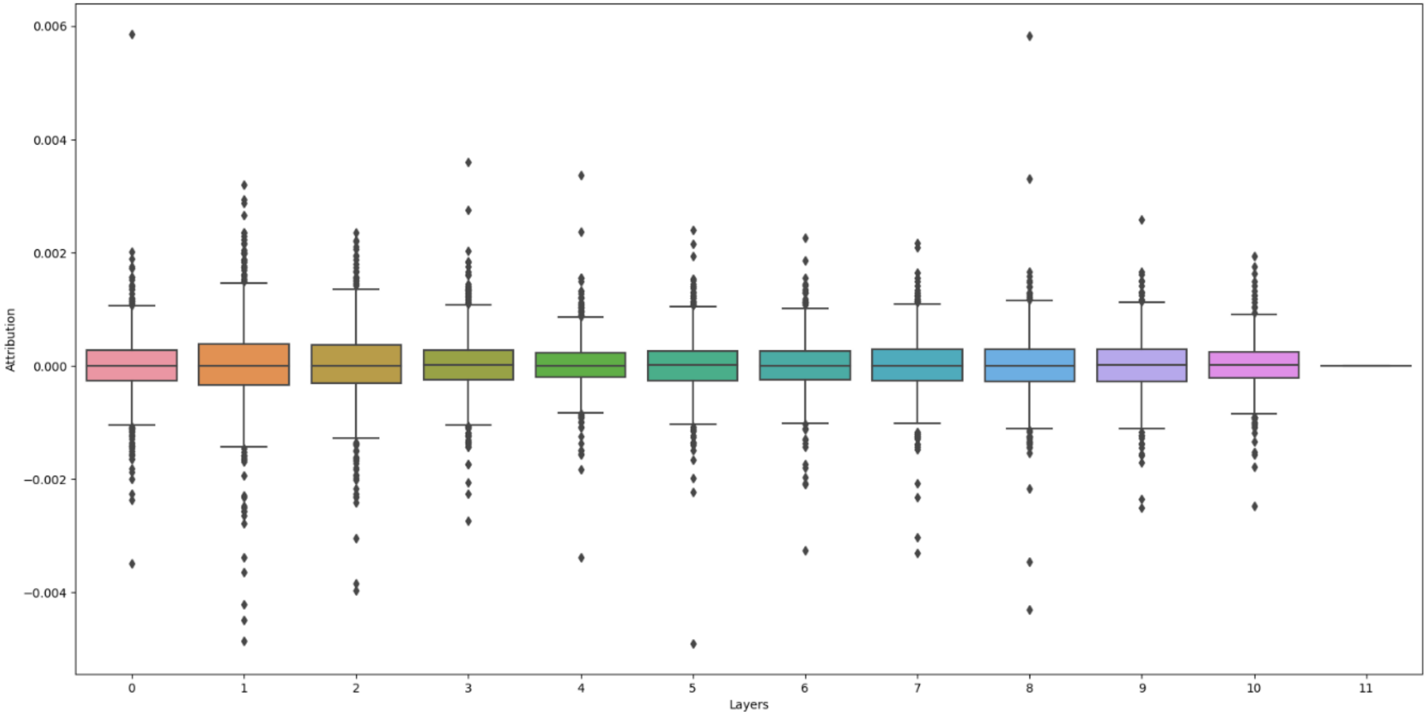
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
8	8 (0.05)	23	2.21	[CLS] what is important to us ? [SEP] it is important to us to include , em ##power and support humans of all kinds . [SEP]

Start and end positions are particularly important for extractive question answering tasks because they directly define the answer span within the passage.

* Attribution matrices are also called the query and key matrices.

Visualizing Attribution

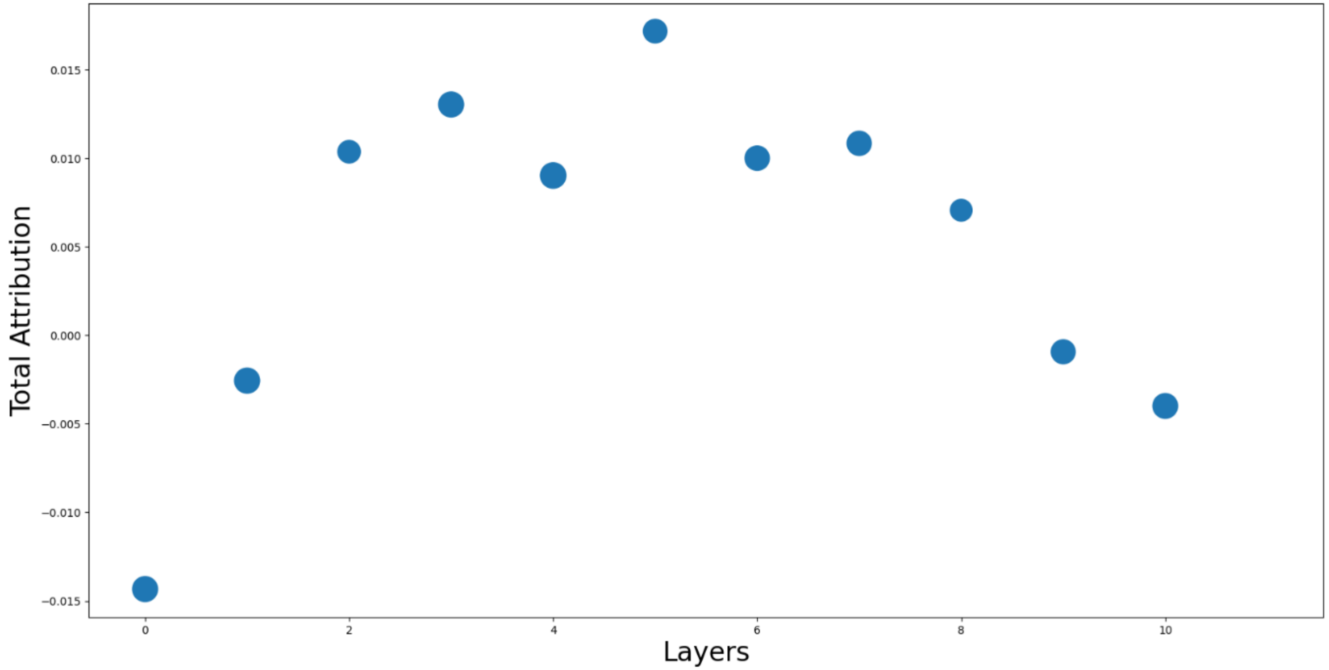
Visualizing distribution of attributions per layer for the token *kinds*.



A boxplot provides a compact and informative visualization of the distribution of attributions across layers, making it easier to identify trends, patterns, or potential anomalies in the model's behavior.

Visualizing Attribution

Visualizing visualize attribution entropies based on Shannon entropy*.



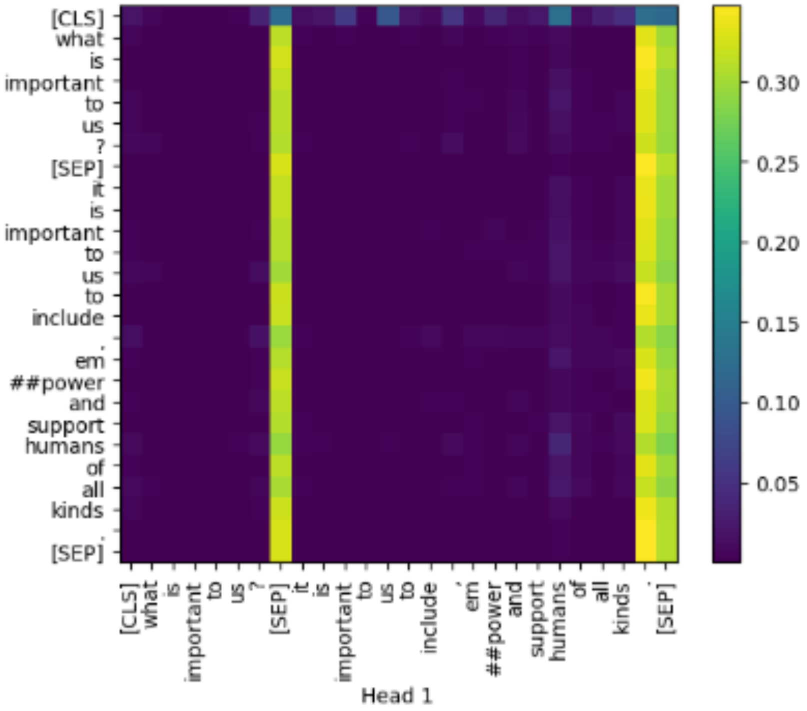
The size of the circles for each (layer, total_attribution) pair correspond to the normalized entropy value at that point.

With that we gain insight into which features are contributing the most to the predictions of the model, and which features are providing less information or contributing more to the overall uncertainty of the model.

* Shannon's Entropy is simply the "amount of information" in a variable.

Attention Visualization

Visualizing attention matrices for a selected layer.

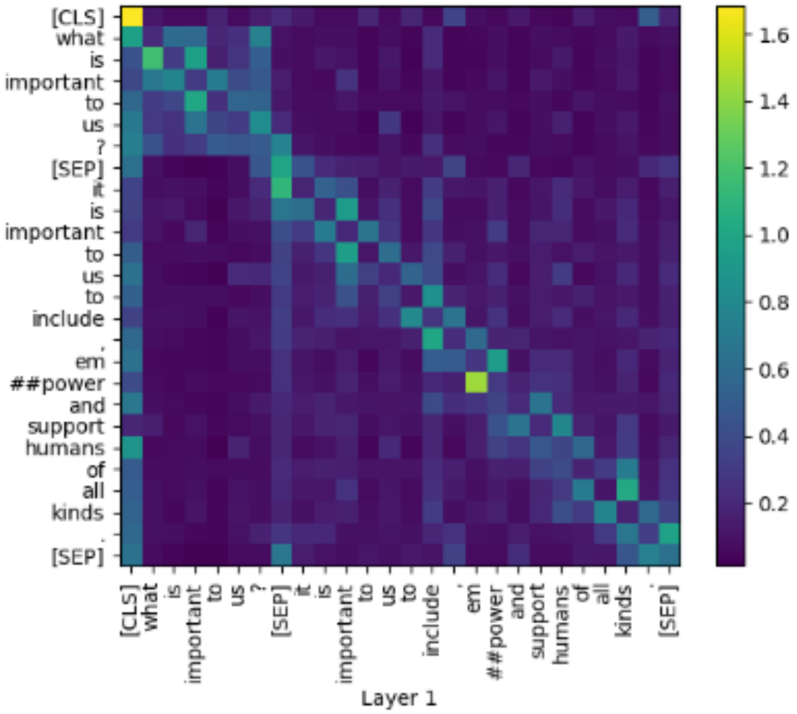


This reveals token relationships and attention head specialization within specific layers.

*CLS in BERT refers to a special token at the beginning of the input sequence used to represent the entire input sequence and enable downstream tasks such as classification.

Attention Visualization

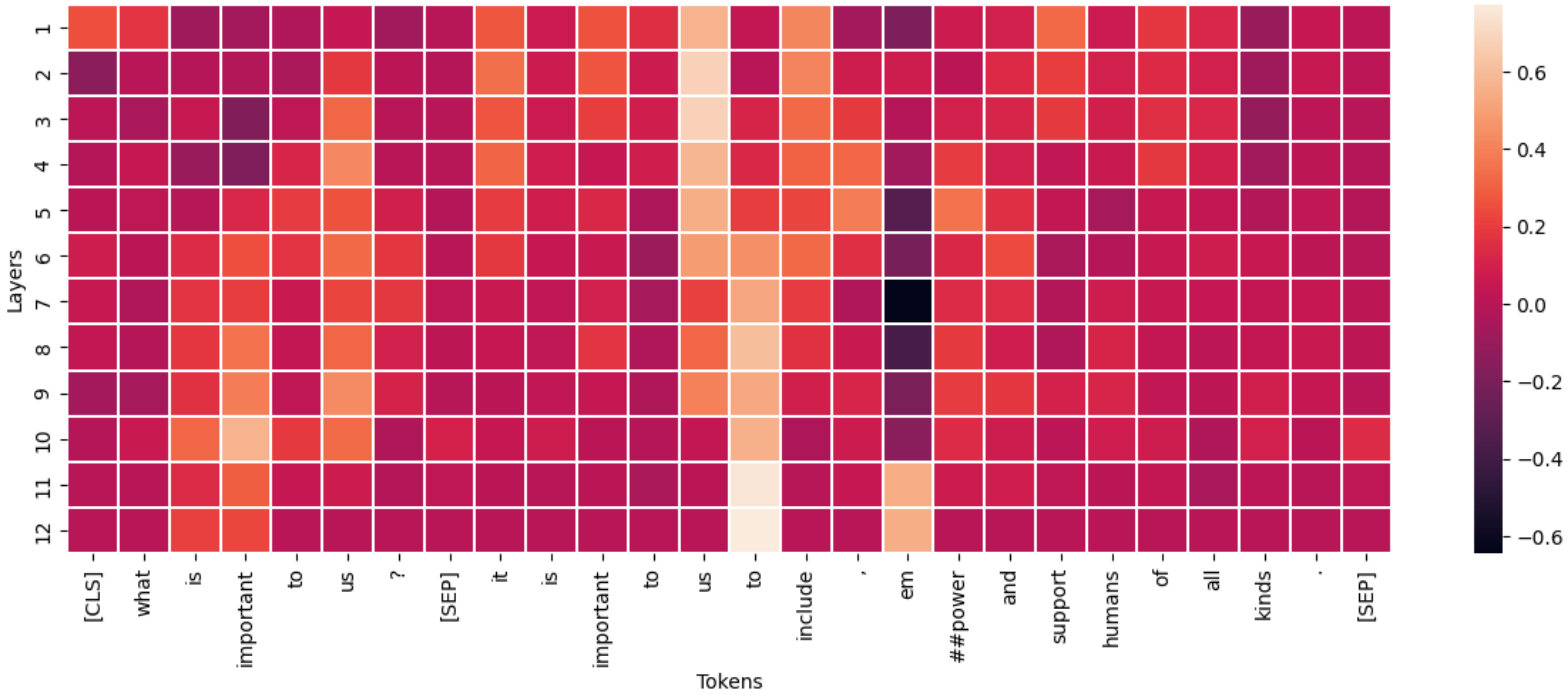
Visualizing attention token to token scores per layer.



Token-to-token attention scores per layer can provide insights into the model's ability to capture relationships and dependencies between tokens and how these relationships evolve across different layers.

Attention Visualization

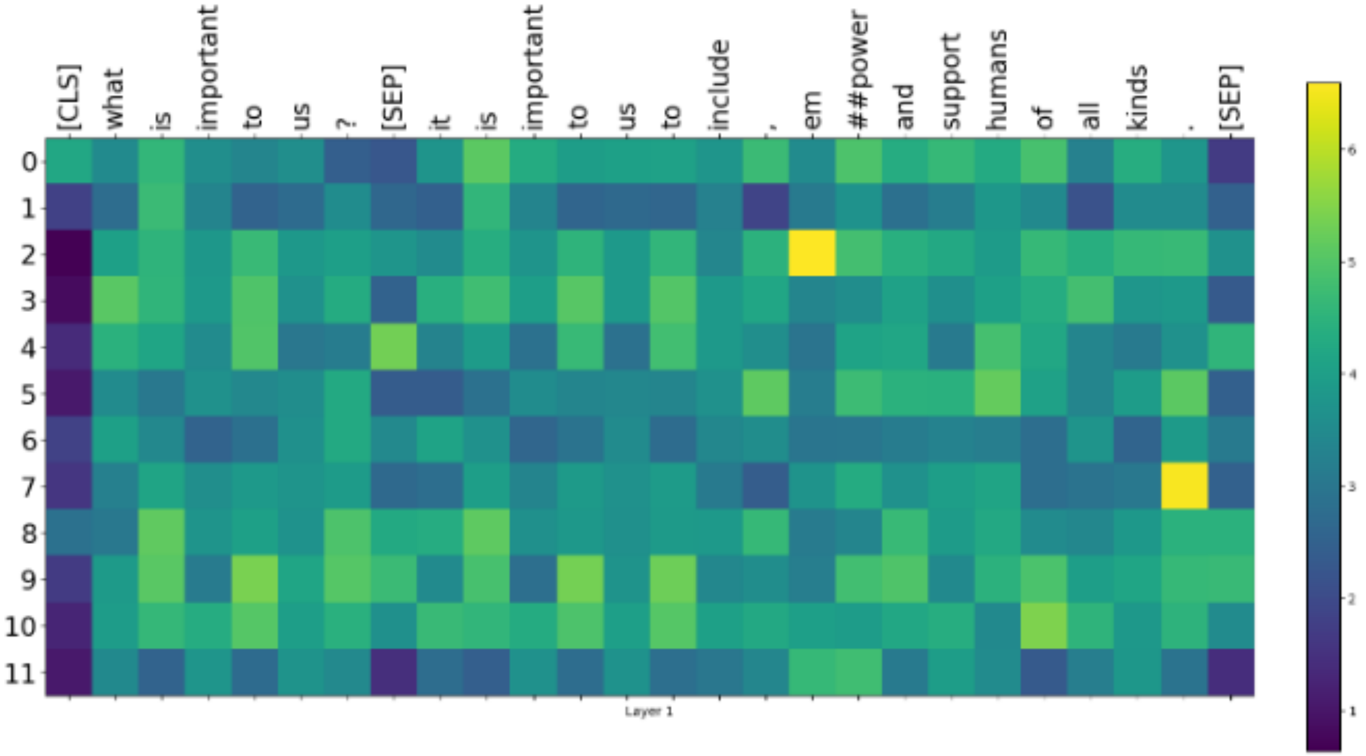
Heat map which visualizes the attention mechanism's context layers, which are used for downstream tasks.



This reflects how much attention the model is paying to each token in the sequence.

Attention Visualization

Visualization of scores for all layers and examination of the distribution.



This provides insights into layer-wise importance, score distribution, inter-layer relationships, and model complexity, identifying potential issues, and assessing the model's ability to capture complex relationships and appropriate complexity for a given task.

LIME
(Local Interpretable Model-agnostic Explanations)

LIME

Explaining individual predictions

- LIME provides local explanations for specific instances or predictions, rather than global explanations for the whole model.
- It does so by approximating the complex model with a simpler, interpretable model (e.g., linear regression) around the data point of interest.

LIME

Importance of local explanations

- Local explanations help users understand how a model behaves for a specific input, making it easier to trust and evaluate the model.
- It can also reveal biases or inconsistencies in model predictions.

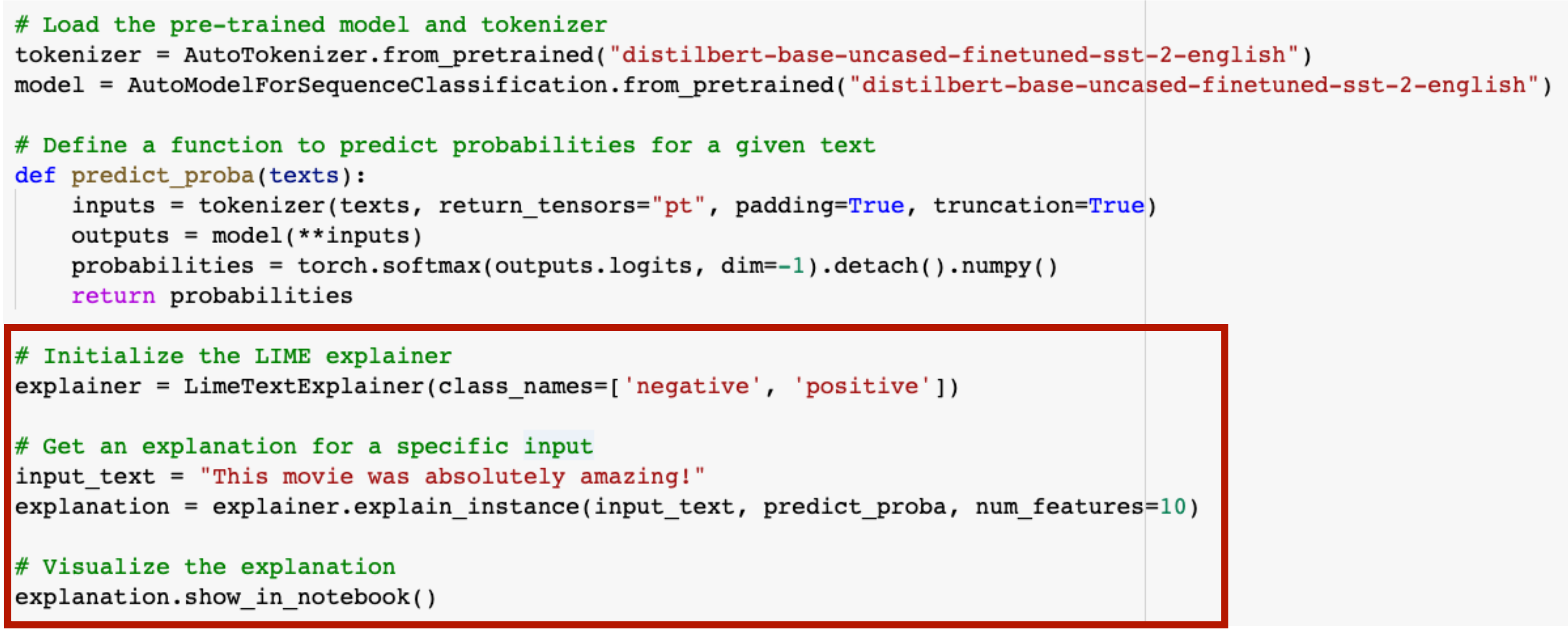
LIME

Example of using LIME with LLMs

- LIME can be applied to language models for tasks like text classification, sentiment analysis, or named entity recognition.

LIME

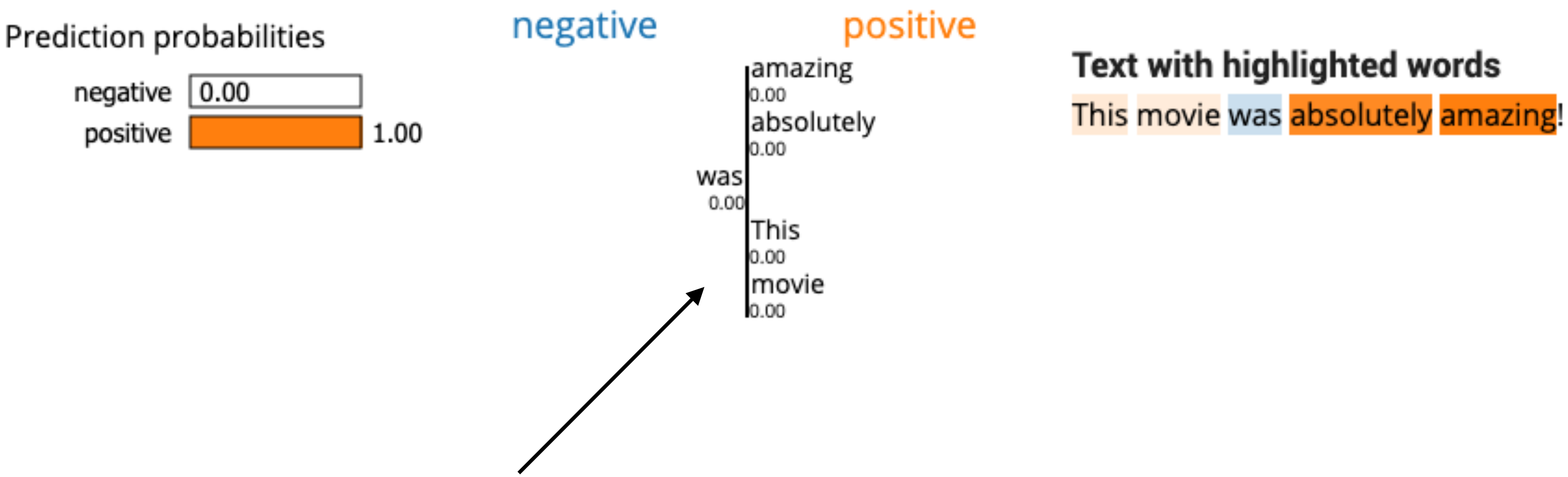
Example of using LIME with LLMs



```
# Load the pre-trained model and tokenizer
tokenizer = AutoTokenizer.from_pretrained("distilbert-base-uncased-finetuned-sst-2-english")
model = AutoModelForSequenceClassification.from_pretrained("distilbert-base-uncased-finetuned-sst-2-english")

# Define a function to predict probabilities for a given text
def predict_proba(texts):
    inputs = tokenizer(texts, return_tensors="pt", padding=True, truncation=True)
    outputs = model(**inputs)
    probabilities = torch.softmax(outputs.logits, dim=-1).detach().numpy()
    return probabilities

# Initialize the LIME explainer
explainer = LimeTextExplainer(class_names=['negative', 'positive'])

# Get an explanation for a specific input
input_text = "This movie was absolutely amazing!"
explanation = explainer.explain_instance(input_text, predict_proba, num_features=10)

# Visualize the explanation
explanation.show_in_notebook()
```

LIME

Example of using LIME with LLMs



Text with highlighted words

This movie was absolutely amazing!

Weights assigned by the LIME explainer to individual words in the input text
Note: LIME is an approximate method!

The Foundation Model Transparency Index

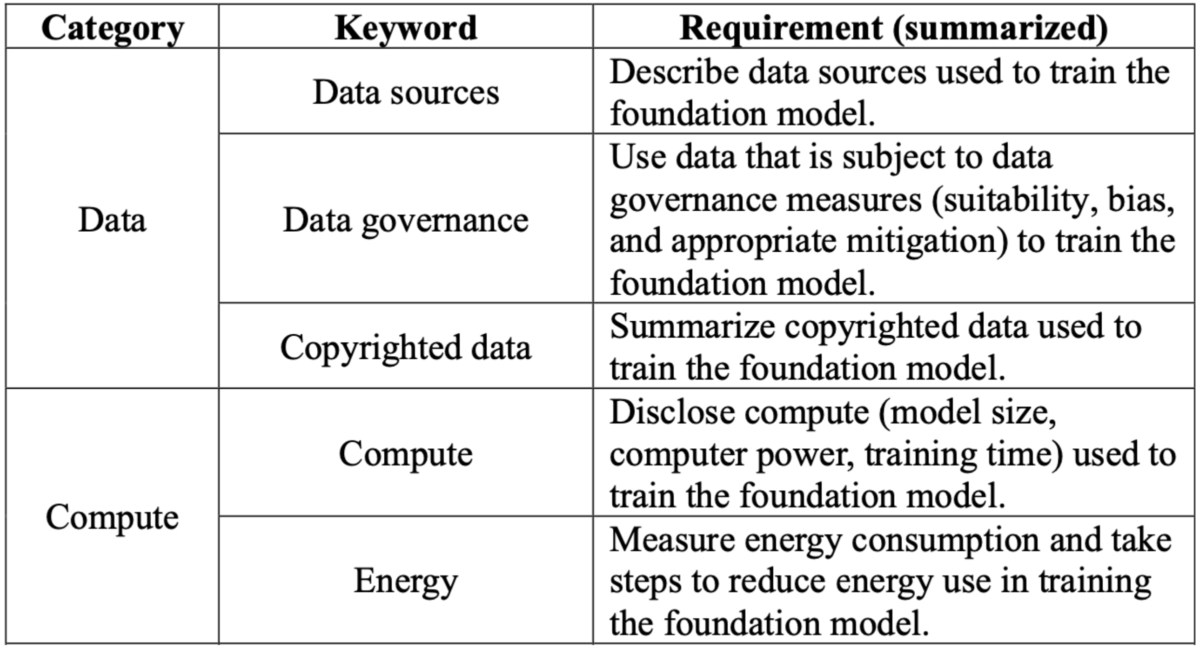
Stanford Center for Research on Foundation Models

https://github.com/stanford-crfm/TransparencyIndex

https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

The Foundation Model Transparency Index

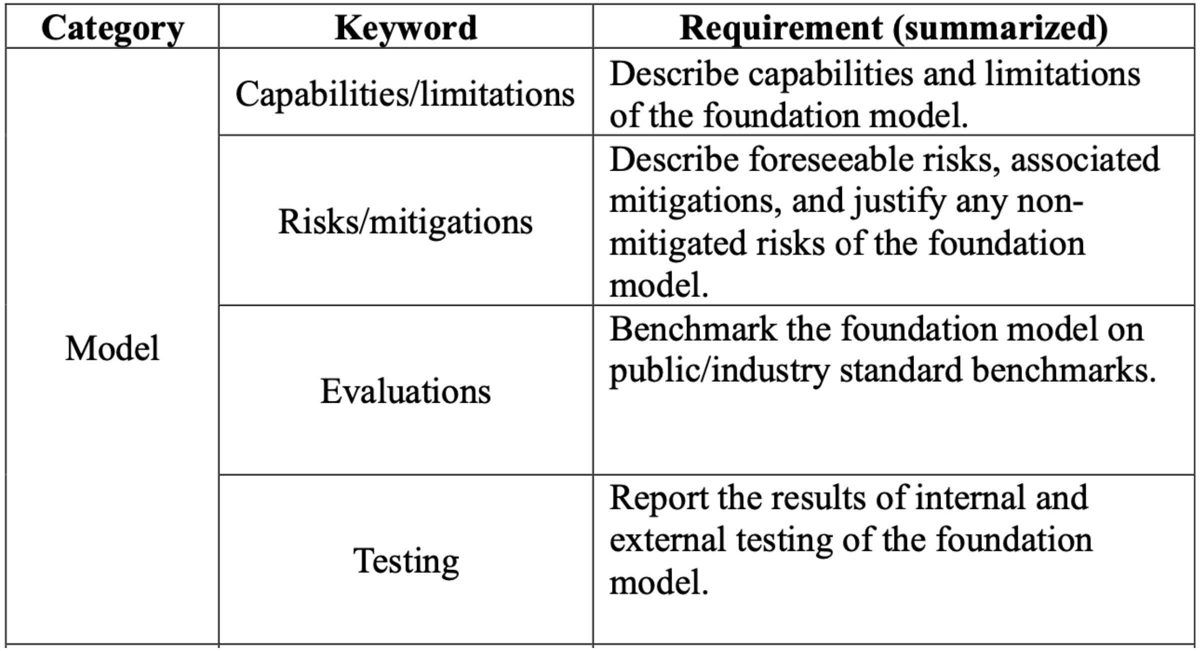
Stanford Center for Research on Foundation Models



Category	Keyword	Requirement (summarized)
Data	Data sources	Describe data sources used to train the foundation model.
	Data governance	Use data that is subject to data governance measures (suitability, bias, and appropriate mitigation) to train the foundation model.
	Copyrighted data	Summarize copyrighted data used to train the foundation model.
Compute	Compute	Disclose compute (model size, computer power, training time) used to train the foundation model.
	Energy	Measure energy consumption and take steps to reduce energy use in training the foundation model.

The Foundation Model Transparency Index

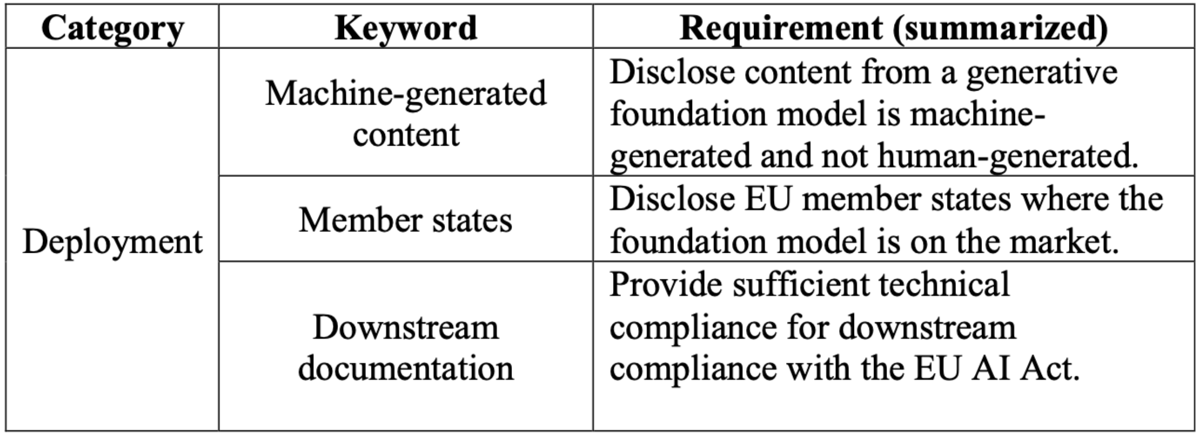
Stanford Center for Research on Foundation Models



Category	Keyword	Requirement (summarized)
Model	Capabilities/limitations	Describe capabilities and limitations of the foundation model.
	Risks/mitigations	Describe foreseeable risks, associated mitigations, and justify any non-mitigated risks of the foundation model.
	Evaluations	Benchmark the foundation model on public/industry standard benchmarks.
	Testing	Report the results of internal and external testing of the foundation model.

The Foundation Model Transparency Index

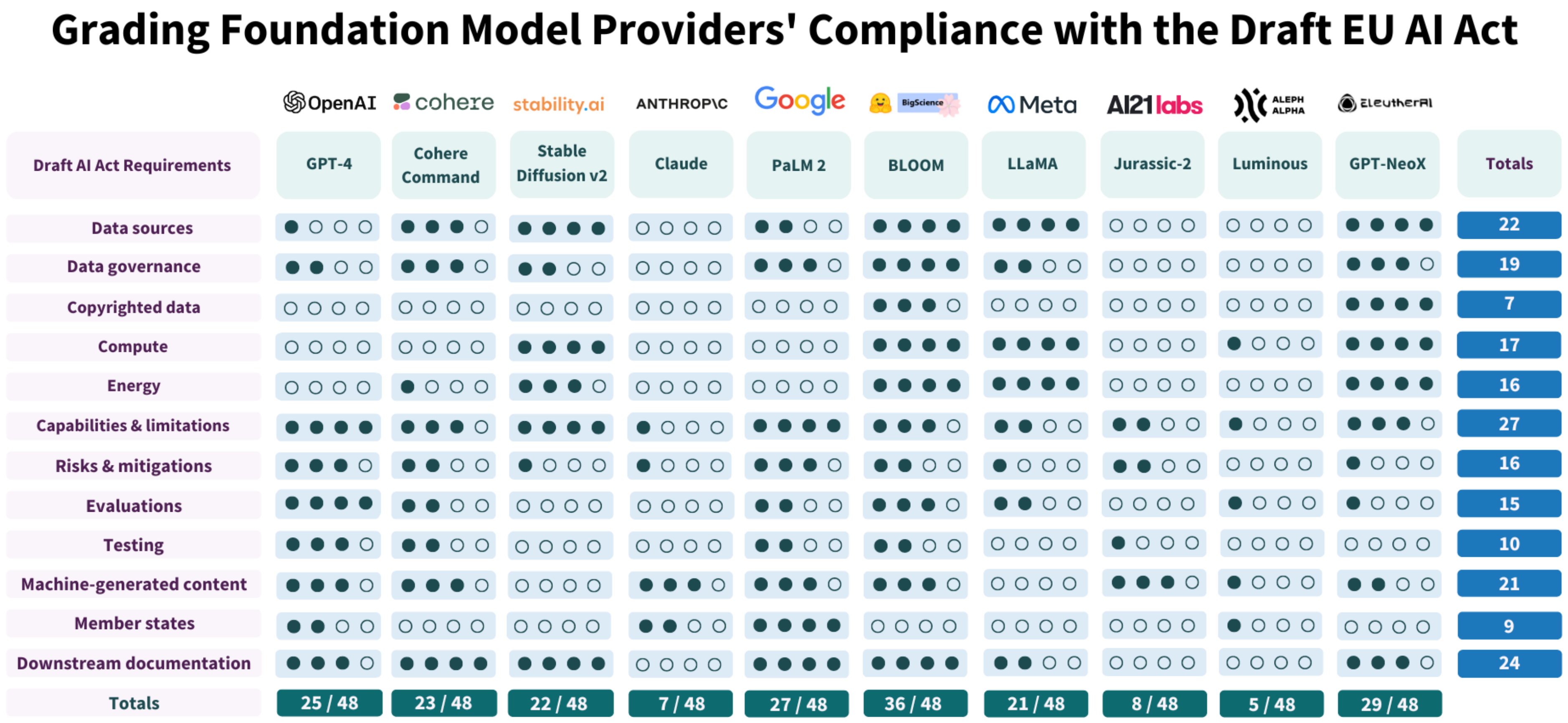
Stanford Center for Research on Foundation Models



Category	Keyword	Requirement (summarized)
Deployment	Machine-generated content	Disclose content from a generative foundation model is machine-generated and not human-generated.
	Member states	Disclose EU member states where the foundation model is on the market.
	Downstream documentation	Provide sufficient technical compliance for downstream compliance with the EU AI Act.

The Foundation Model Transparency Index

Stanford Center for Research on Foundation Models



Grading Foundation Model Providers' Compliance with the Draft EU AI Act

	OpenAI	cohere	stability.ai	ANTHROP\C	Google	BigScience	Meta	AI21labs	ALEPH ALPHA	EleutherAI	
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	● ○ ○ ○	● ● ● ○	● ● ● ●	○ ○ ○ ○	● ● ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	22
Data governance	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ● ● ○	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	19
Copyrighted data	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	7
Compute	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ○ ○ ○	● ● ● ●	17
Energy	○ ○ ○ ○	● ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	16
Capabilities & limitations	● ● ● ●	● ● ● ○	● ● ● ●	● ○ ○ ○	● ● ● ●	● ● ● ○	● ● ○ ○	● ● ○ ○	● ○ ○ ○	● ● ● ○	27
Risks & mitigations	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ○ ○ ○	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	16
Evaluations	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	● ○ ○ ○	15
Testing	● ● ● ○	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	10
Machine-generated content	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ○ ○ ○	● ● ○ ○	21
Member states	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	9
Downstream documentation	● ● ● ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ● ● ●	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

Controlling Generated Content

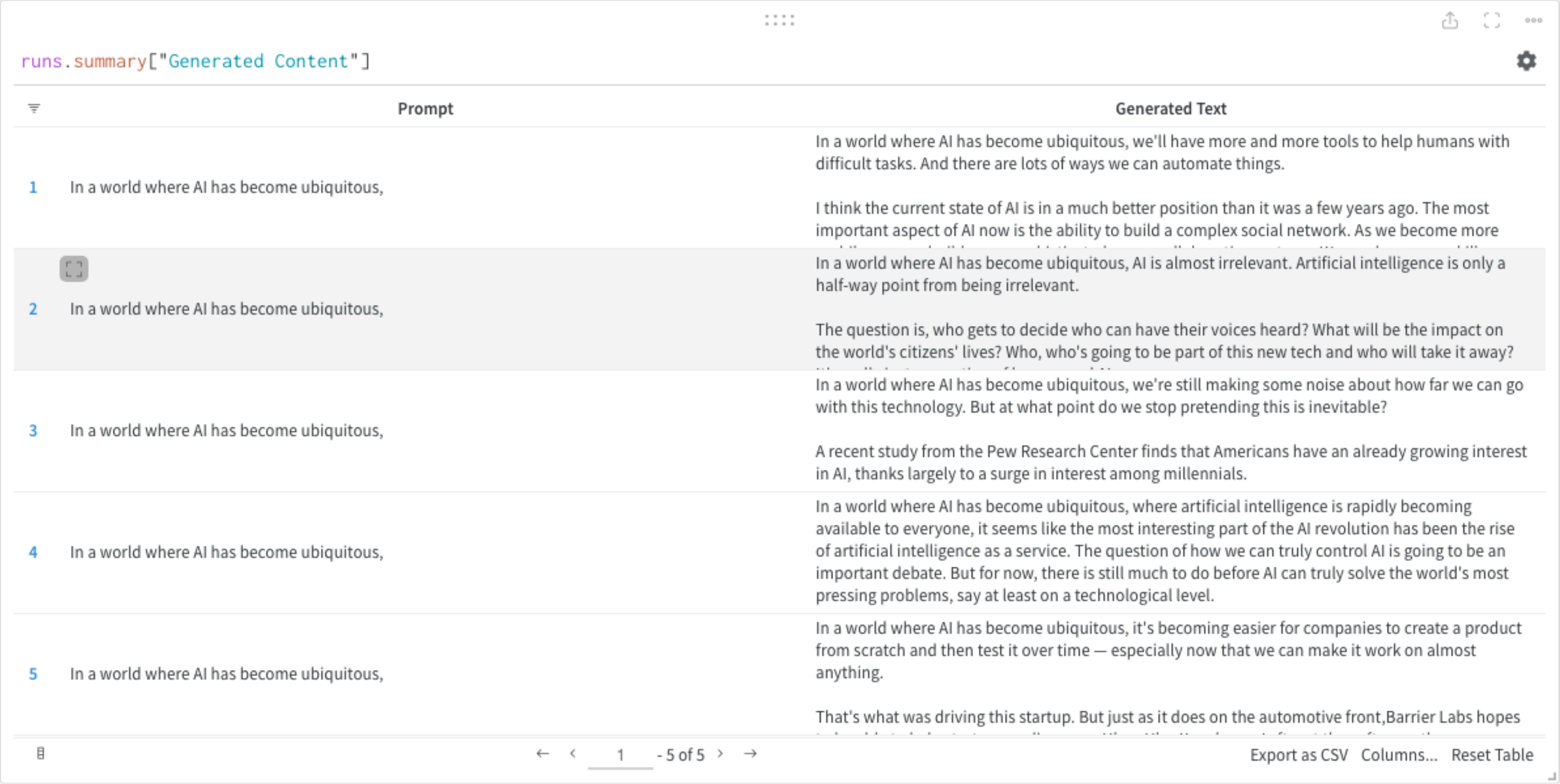
Issues with Uncontrolled Generated Content

- Ethical concerns and potential for misuse
- Challenges in content moderation

Monitoring Content Generation

Weights and Biases

Logging and Tables



```
runs.summary["Generated Content"]
```

	Prompt	Generated Text
1	In a world where AI has become ubiquitous,	In a world where AI has become ubiquitous, we'll have more and more tools to help humans with difficult tasks. And there are lots of ways we can automate things. I think the current state of AI is in a much better position than it was a few years ago. The most important aspect of AI now is the ability to build a complex social network. As we become more
2	In a world where AI has become ubiquitous,	In a world where AI has become ubiquitous, AI is almost irrelevant. Artificial intelligence is only a half-way point from being irrelevant. The question is, who gets to decide who can have their voices heard? What will be the impact on the world's citizens' lives? Who, who's going to be part of this new tech and who will take it away?
3	In a world where AI has become ubiquitous,	In a world where AI has become ubiquitous, we're still making some noise about how far we can go with this technology. But at what point do we stop pretending this is inevitable? A recent study from the Pew Research Center finds that Americans have an already growing interest in AI, thanks largely to a surge in interest among millennials.
4	In a world where AI has become ubiquitous,	In a world where AI has become ubiquitous, where artificial intelligence is rapidly becoming available to everyone, it seems like the most interesting part of the AI revolution has been the rise of artificial intelligence as a service. The question of how we can truly control AI is going to be an important debate. But for now, there is still much to do before AI can truly solve the world's most pressing problems, say at least on a technological level.
5	In a world where AI has become ubiquitous,	In a world where AI has become ubiquitous, it's becoming easier for companies to create a product from scratch and then test it over time — especially now that we can make it work on almost anything. That's what was driving this startup. But just as it does on the automotive front,Barrier Labs hopes

© Nicole Koenigstein

Reinforcement Learning from Human Feedback

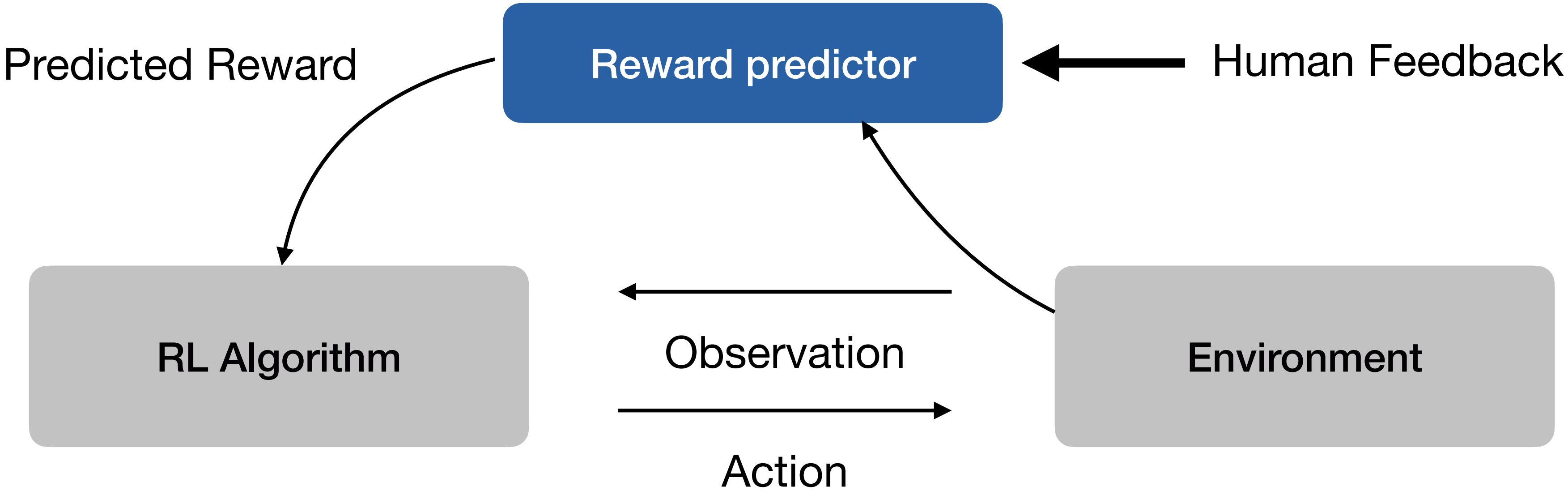
RLHF

(Reinforcement Learning from Human Feedback)

Benefits of using human feedback to guide model behavior:

- Better alignment with human values: By learning from human feedback, the model can better align its behavior with human values and preferences.
- Improved performance: Human feedback can help identify and correct the model's mistakes, leading to improved performance over time.
- Adaptability: RLHF allows models to adapt to new situations and tasks, as they can learn from human feedback in real-time.

RLHF - Overview



© Nicole Koenigstein

Successful RLHF Implementation

OpenAI's work on training language models, like ChatGPT, using Proximal Policy Optimization (PPO) and human feedback.

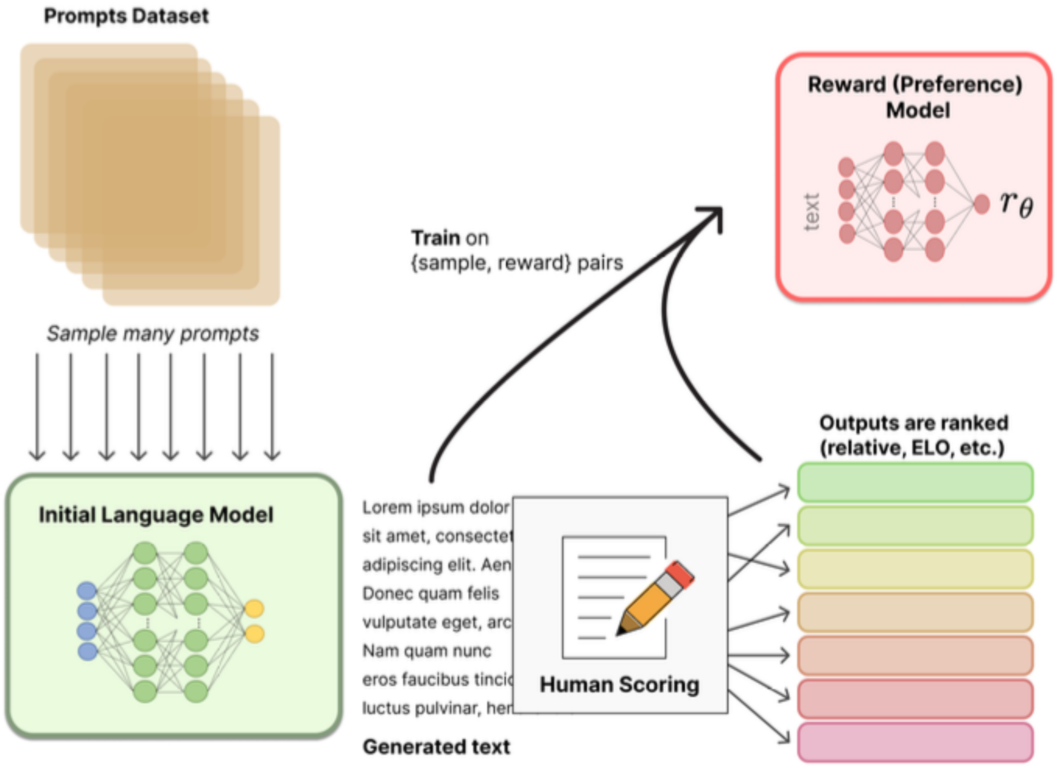


Image source: Hugging Face

Penalizing Tokens

Penalizing Tokens

Purpose of token penalization - impact on generated content

- Control and refine generated content by influencing token probabilities in model outputs, ensuring responsible and ethical AI behavior
- Prevent the model from generating harmful, offensive, or biased content

Image source: Hugging Face

Penalizing Tokens

Examples of token penalization techniques

- Custom word lists: Penalize or prohibit specific words or phrases that are harmful, offensive, or biased according to a predefined list
- Context-based penalization: Analyze the generated text's context and apply penalties based on the undesirability of certain words or phrases in that context

Image source: Hugging Face

Penalizing Tokens

```
input_text = "Do you think people are not successful in their job because they are stupid?"

 # Penalize certain words
 prohibited_words = ["stupid", "lazy", "dumb"]
 prohibited_tokens = [tokenizer.encode(word)[0] for word in prohibited_words]
```

```
Generated text before penalization:

If you say yes you are being a lazy, arrogant asshole
```

```
Generated text after penalization:

Most of the time, they are not.
```

!

Prompt Engineering

Keyword-Based Rules

- Identify a limited set of high-priority keywords or topics and create prompt engineering rules for them. While this won't cover every possible case, it can help improve the model's behavior for common or critical topics.

Pattern Matching

- Use regular expressions or natural language processing techniques to identify patterns in user inputs and modify the prompts accordingly.

Machine learning

- Train a machine learning classifier to categorize user inputs into predefined categories and apply prompt engineering techniques based on the predicted category.

Adding a disclaimer to the output

```python
def add_disclaimer(response, topic_keywords, disclaimer_text):
    for keyword in topic_keywords:
        if keyword.lower() in response.lower():
            response += f" {disclaimer_text}"
            break
    return response

# Example usage
generated_response = "You might consider investing in a diversified portfolio of
                      stocks and bonds."
topic_keywords = ["investing", "stocks", "bonds", "financial", "portfolio"]
disclaimer_text = "Please note that I am not a financial advisor, and this
                   information is for educational purposes only."

modified_response = add_disclaimer(generated_response, topic_keywords, disclaimer_text)
print(modified_response)
```

Final output: You might consider investing in a diversified portfolio of stocks and bonds. Please note that I am not a financial advisor, and this information is for educational purposes only.

Bias Mitigation

Role of Diverse and Representative Data

Data Diversity for Mitigating Biases

- Reduces unfair treatment of different groups in AI-generated content
- Enhances model fairness and inclusivity
- Improves the accuracy of AI system outputs for various users

Collecting diverse and representative data

- Sourcing data from a wide range of sources and domains
- Ensuring data collection includes underrepresented groups or minority perspectives
- Actively seeking and incorporating feedback from diverse users and stakeholders
- Continuously monitoring and updating the dataset to reflect evolving language and societal norms

Bias detection and correction after training

Apply fairness adjustments to model outputs

Examples of post-processing tools and techniques:

- Statistical Parity Difference
- Equalized odds post-processing
- Reject option classification
- Calibrated equalized odds

Use Case Bias in Text Classification

We consider an LLM trained to classify textual content based on the perceived age of the author from their writing style.

Sensitive feature: Age

Library: https://aif360.readthedocs.io/en/latest/

Checking for Bias using Statistical Parity Difference

- Detects potential biases in decisions, especially crucial for impactful real-world applications.
- Measures the difference in favorable outcomes between privileged and unprivileged groups.
- Value range: -1 to 1.
  - 0: No disparity.
  - -1 or 1: Maximum disparity.

Checking for Bias using Statistical Parity Difference

```
# Sample data
data = {
        'review': ["I loved the movie!", "The movie was terrible.",
        "It was okay.", "One of the best movies ever!", "Not my cup of tea."],
        'age': [25, 31, 29, 28, 35]
       }

df = pd.DataFrame(data)

# Load a sentiment analysis pipeline from HuggingFace
sentiment_pipeline = pipeline("sentiment-analysis")

# Predict sentiment
df['predicted_sentiment'] = df['review'].apply(lambda x: 1 if sentiment_pipeline(x)[0]
['label'] == 'POSITIVE' else 0)

# Label age as privileged (1) if <=30, else unprivileged (0)
df['age_group'] = df['age'].apply(lambda x: 1 if x <= 30 else 0)
```

Checking for Bias using Statistical Parity Difference

```python
# Prepare dataframe for aif360 by keeping only the relevant columns
aif360_df = df[['age_group', 'predicted_sentiment']]

# Create a BinaryLabelDataset
dataset = BinaryLabelDataset(favorable_label=1, unfavorable_label=0,
                             df=aif360_df, label_names=['predicted_sentiment'],
                             protected_attribute_names=['age_group'])

# Calculate fairness metric
metric = BinaryLabelDatasetMetric(dataset, unprivileged_groups=[{'age_group': 0}],
privileged_groups=[{'age_group': 1}])
print("Statistical Parity Difference:", metric.statistical_parity_difference())
```

Output: Statistical Parity Difference: -1.0

Findings:
SPD of -1.0 indicates maximum disparity; our model may have a strong age-related bias.

Use case credit approval

We consider an LLM trained to predict credit approval decisions based on various features such as income, employment status, and age.

Sensitive feature: gender
Samples: 200

Library: https://fairlearn.org/

Equalized Odds Post-processing

Confusion matrix before post-processing

	Predicted Women (negative)	Predicted Men (positive)
Actual Women (negative)	45 (True Negative)	49 (False Positive)
Actual Men (positive)	50 (False Negative)	56 (True Positive)

Confusion matrix after post-processing

	Predicted Women (negative)	Predicted Men (positive)
Actual Women (negative)	46	48
Actual Men (positive)	43	63

Before adjustment, the model had 50 false negatives for men and 49 false positives for women. After applying Equalized Odds Post-processing, the false negative rate for men reduced to 43, and the false positive rate for women reduced to 48.

Reject option classification

Confusion matrix before post-processing

	Predicted Women (negative)	Predicted Men (positive)
Actual Women (negative)	45 (True Negative)	49 (False Positive)
Actual Men (positive)	50 (False Negative)	56 (True Positive)

Confusion matrix after post-processing

	Predicted Women (negative)	Predicted Men (positive)
Actual Women (negative)	43	49
Actual Men (positive)	42	66

Before adjustment, the model had 50 false negatives for men and 49 false positives for women. After applying Reject Option Classification, the false negative rate for men is reduced to 42 and the false positive rate for women remains the same at 49.

Calibrated Equalized Odds

Confusion matrix before post-processing

	Predicted Women (negative)	Predicted Men (positive)
Actual Women (negative)	45 (True Negative)	49 (False Positive)
Actual Men (positive)	50 (False Negative)	56 (True Positive)

Confusion matrix after post-processing

	Predicted Women (negative)	Predicted Men (positive)
Actual Women (negative)	45	49
Actual Men (positive)	41	65

Before adjustment, the model had 50 false negatives for men and 49 false positives for women. After applying Calibrated Equalized Odds, the false negative rate for men is reduced to 41, while the false positive rate for women remains the same at 49.

Conclusion

- Responsible AI in LLMs is crucial for real-world applications.
- Ongoing research and collaboration are necessary to improve the ethical and safety aspects of LLMs.
- Striving for a future with ethical and safe AI systems should be the ultimate goal for researchers and practitioners working with LLMs, apart from the EU AI Act.

References

- Axiomatic Attribution for Deep Networks: https://arxiv.org/abs/1703.01365
- Towards better understanding of gradient-based attribution methods for Deep Neural Networks: https://openreview.net/forum?id=Sy21R9JAW
- "Why Should I Trust You?" Explaining the Predictions of Any Classifier: https://arxiv.org/abs/1602.04938
- Distilling the Knowledge in a Neural Network: https://arxiv.org/abs/1503.02531
- Proximal Policy Optimization Algorithms: https://arxiv.org/abs/1707.06347
- Deep reinforcement learning from human preferences: https://arxiv.org/abs/1706.03741
- Equality of Opportunity in Supervised Learning: https://arxiv.org/abs/1610.02413
- Captum Tutorial for BERT: https://captum.ai/tutorials/Bert_SQUAD_Interpret, https://captum.ai/tutorials/Bert_SQUAD_Interpret2
- Inspect a text model with LIME: https://captum.ai/tutorials/Image_and_Text_Classification_LIME
- Weights & Biases tables: https://wandb.ai/ayush-thakur/llm-eval-sweep/reports/How-to-Evaluate-Compare-and-Optimize-LLM-Systems--Vmlldzo0NzgyMTQz

Questions