

Explainable AI (XAI) and Large Language Models (LLM): an impossible pairing?

Leonida Gianfagna (Cyber Guru)

Enrico Zimuel (Elastic)

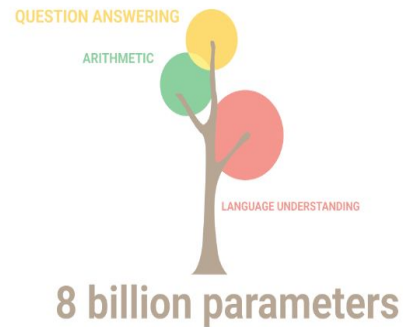


AI Heroes, Italian Artificial Intelligence
Conference, 11 December 2024



LLM emerging properties

“Emergence is when quantitative changes in a system result in qualitative changes in behavior.” (P. Anderson, 1972)



“..Only twenty years ago we expected to have to solve two tasks separately, modeling language and the world, and then combine them. Things turned out differently, and I wonder if the distinction between understanding the world and understanding language isn't arbitrary, and if another kind of mind might not draw very different boundaries..” N. Cristianini

“These things are totally different from us,” he says. “Sometimes I think it’s as if aliens had landed and people haven’t realized because they speak very good English.” G. Hinton

Evidence of emerging properties

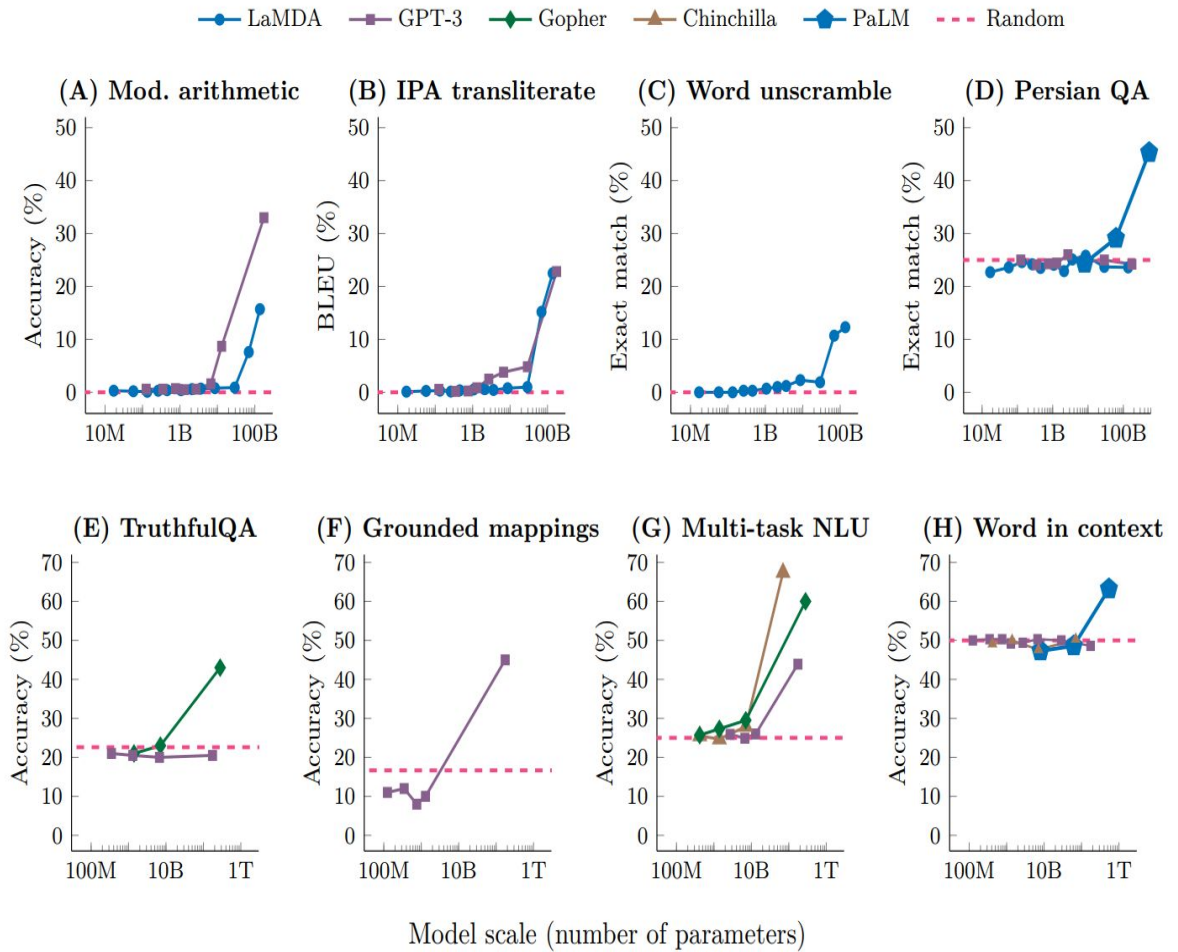
□ An ability is emergent if it is not present in smaller models but is present in larger models.

□ Emergent abilities would not have been directly predicted by extrapolating a scaling law (i.e. consistent performance improvements) from small-scale models.

□ When visualized via a scaling curve (x-axis: model scale, y-axis: performance), emergent abilities show a clear pattern—performance is near-random until a certain critical threshold of scale is reached, after which performance increases to substantially above random.

□ This qualitative change is also known as a *phase transition*—a dramatic change in overall behavior that would not have been foreseen by examining smaller-scale systems

□ The ability to perform a task via few-shot prompting is emergent when a model has random performance until a certain scale, after which performance increases to well-above random. The figure shows eight such emergent abilities spanning five language model families from various work



Model scale (number of parameters)
Wei et al. 2022 - Emergent Abilities of Large Language Models

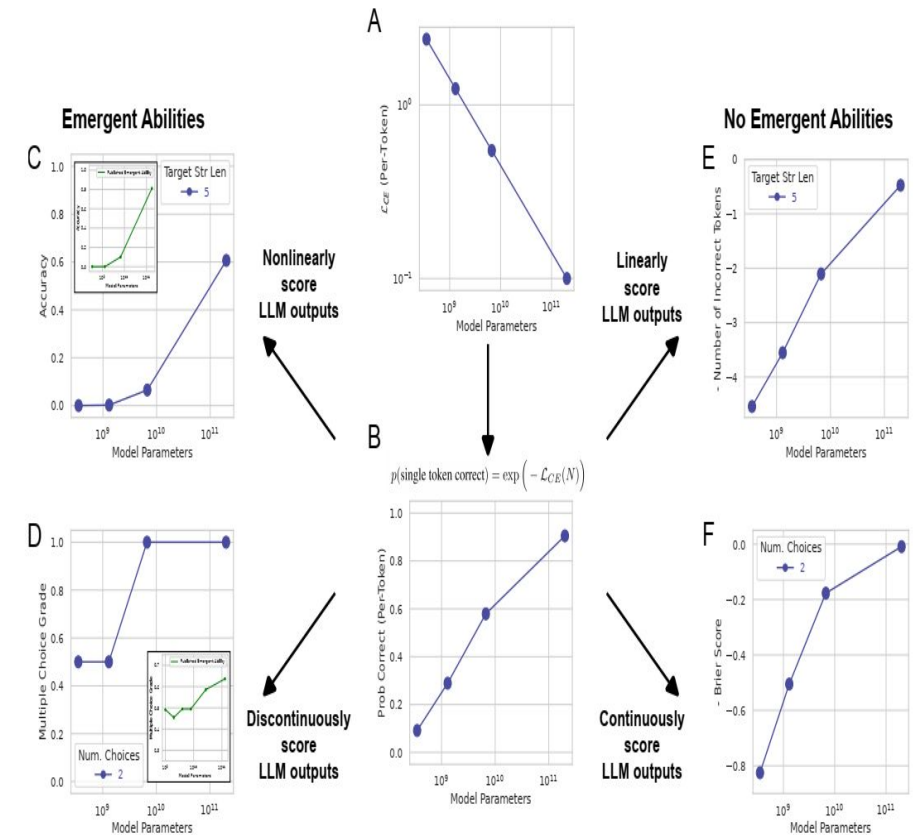
Doubts on emergent properties

□ These emergent abilities have garnered significant interest, raising questions such as: What controls *which* abilities will emerge? What controls *when* abilities will emerge? How can we make desirable abilities emerge faster, and ensure undesirable abilities never emerge?

□ These questions are especially pertinent to AI safety and alignment, as emergent abilities forewarn that larger models might one day, without warning, acquire undesired mastery over dangerous capabilities

□ Sharp and unpredictable changes might be induced by the researcher's choice of measurement, even though the model family's per-token error rate changes smoothly, continuously and predictably with increasing scale.

□ That emergent abilities could be a mirage caused primarily by the researcher choosing a metric that nonlinearly or discontinuously deforms per-token error rates, and secondarily by possessing too few test data to accurately estimate the performance of smaller models, thereby causing smaller models to appear wholly unable to perform the task.



R. Schaeffer et al 2023. / Are Emergent Abilities of Large Language Models a Mirage?

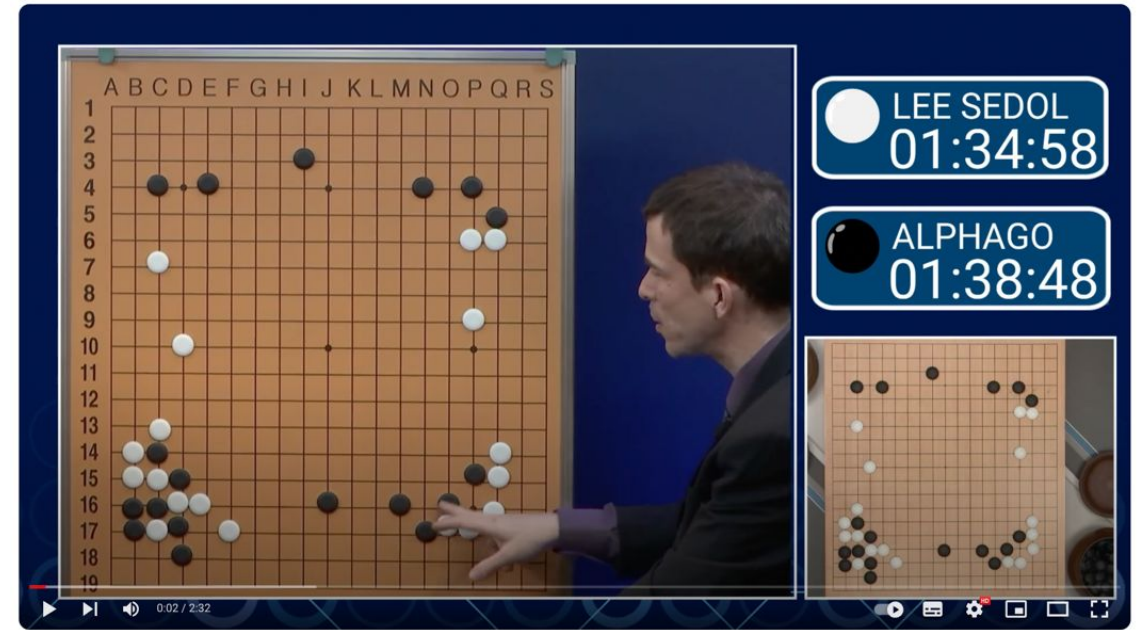
eXplainable AI (XAI)

GO champion Fan Hui commenting the famous 37th move of AlphaGo, the software developed by Google to play GO, that defeated in March 2016 the Korean champion Lee Sedol with an historical result: "It's not a human move, I've never seen a man playing such a move".

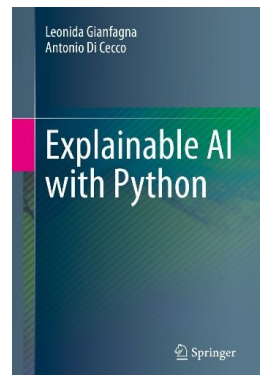
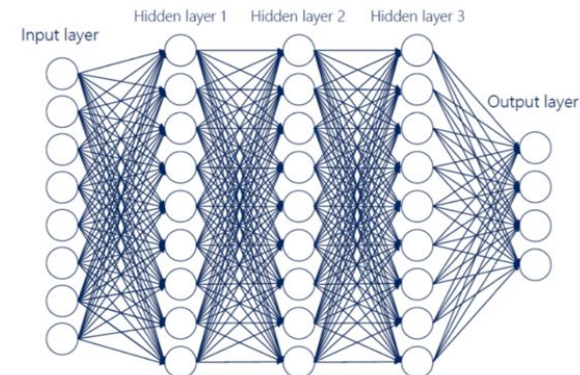
GO is known as a “computationally complex” game, more complex than Chess and before this result the common understanding was that it was not a game suitable for a machine to play successfully.

The GO champion could not make sense of the move even after having looked at all the match, he recognized it as brilliant, but he had no way to provide an explanation

A (non-mathematical) definition by Miller (2017) is: **Explainability is the degree to which a human can understand the cause of a decision taken by an artificial agent.**

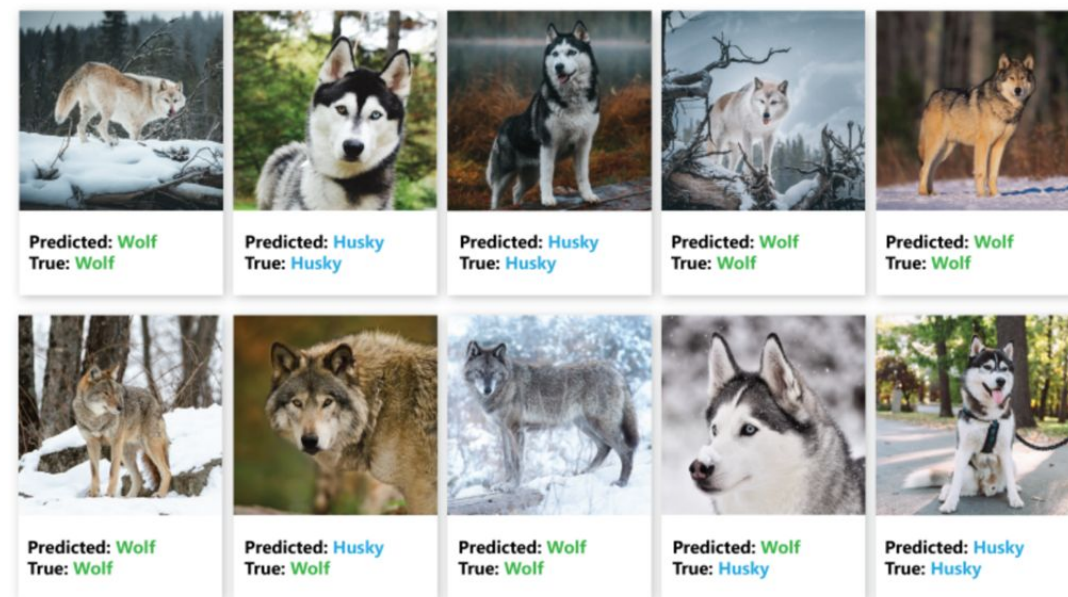


Move 37!! Lee Sedol vs AlphaGo Match 2



XAI

- After the training, **the algorithm learned to distinguish the classes with remarkable accuracy**: only a misclassification over 100 images! But if we use an Explainable Ai method asking the model **“Why have you predicted wolf?”** The answer will be with a little of surprise **“because there is snow!”**
- This is an **experiment** conducted to fool the **Deep Neural Network (DNN)**: the engineers maintained in the second and fourth images only the elements that the system used to recognize a guitar and a penguin and changed all the rest so that the system still “see” them like a guitar and a penguin.
- The work from Goodfellow et al. (2014) opened the door to further evolutions starting from universal perturbations (Moosavi-Dezfooli et al. 2017) to the recent one-pixel attacks that showed how to fool a neural network by just changing one pixel in the input image.



Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[Notebook Here](#)

[One Pixel attack original paper](#)

Why XAI?

- **Compliance with regulations (eg GDPR)**
 - Under GDPR, individuals have the "right to explanation" for decisions made by AI. XAI ensures transparency, enabling organizations to justify AI-driven outcomes and maintain legal compliance.
- **Enhancing Robustness of ML Models and Defending against Model Poisoning**
 - The idea is that the classification of a normal input should rely more on robust features if compared to the classification of AE that is likely to rely on non-robust features attacked to change
- **Ensuring Fairness in Datasets**
 - By analyzing AI decisions, XAI can highlight potential biases in data, ensuring fair treatment across demographic groups and mitigating ethical concerns.
- **Knowledge discovery**
 - It is the most complex application to comment, being related to situations in which ML models are not just used to make predictions

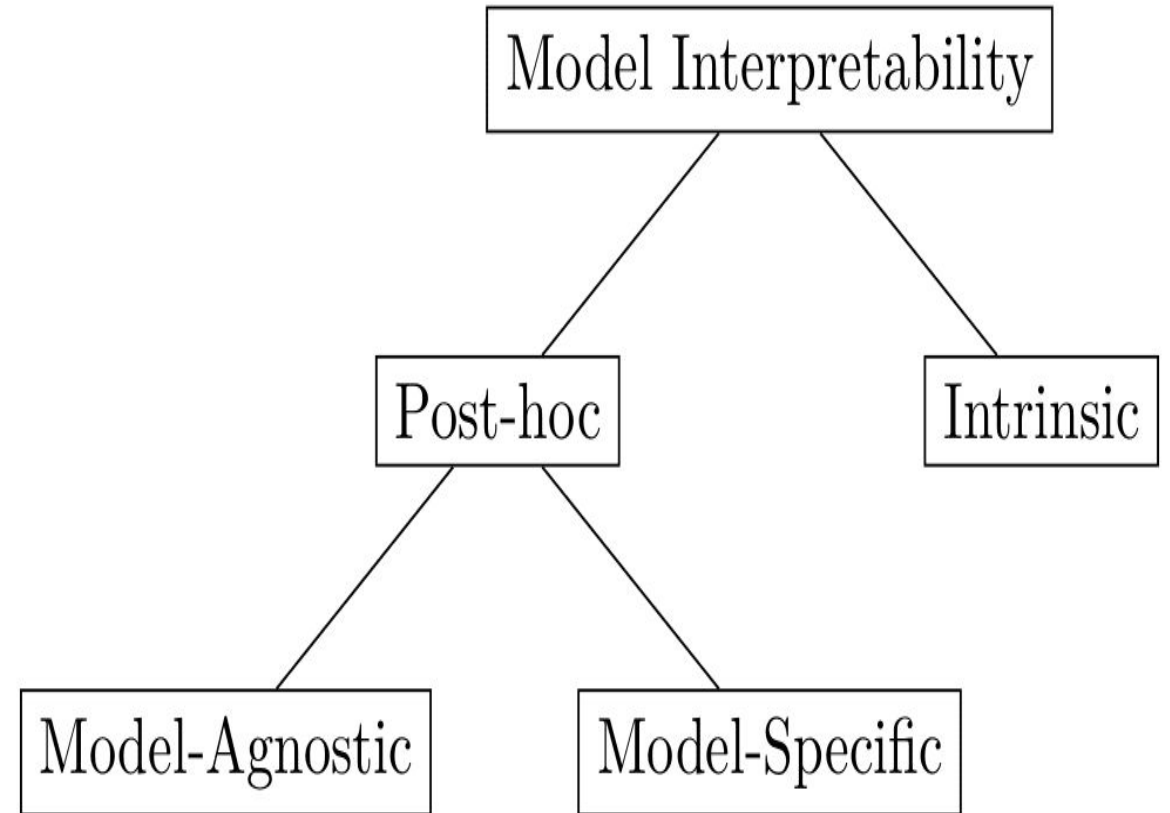
XAI taxonomy

1. Intrinsic Interpretability:

1. Refers to models that are inherently interpretable due to their design.
2. Examples: Linear regression, decision trees, rule-based models.
3. Characteristics: Simple structure, transparency in decision-making.

2. Post-hoc Interpretability:

1. Refers to methods applied after the model is trained to explain its predictions.
2. Types:
 1. **Model-Agnostic:**
 1. Works with any machine learning model.
 2. Examples: SHAP, LIME, Partial Dependence Plots.
 2. **Model-Specific:**
 1. Tailored to specific model architectures.
 2. Examples: Gradient-based visualization for neural networks, feature importance in random forests.



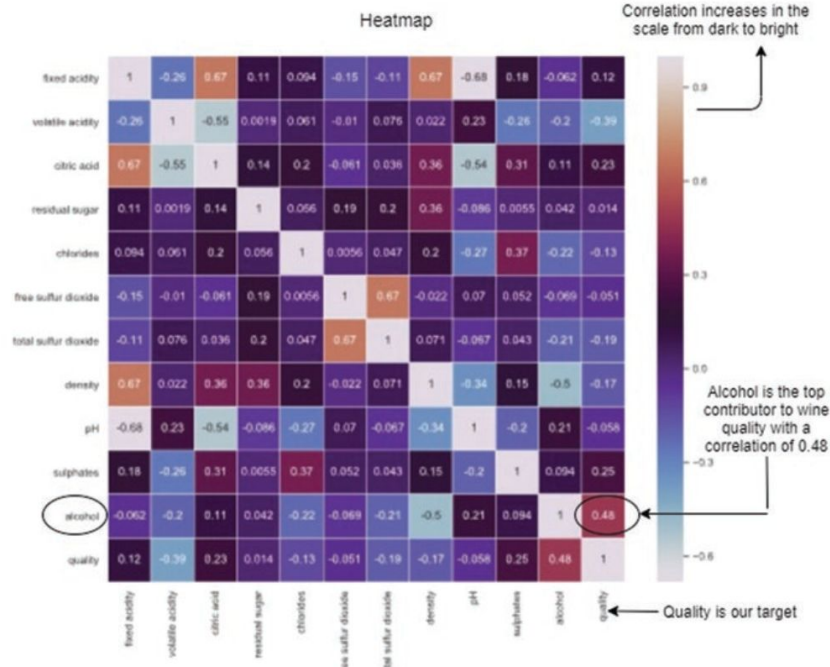
INTRINSIC EXPLAINABLE MODELS: LINEAR REGRESSION

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Top 5 rows of Wine Quality dataset

$$Y = m_0 + m_1x_1 + m_2x_2 + \dots + m_kx_k$$

Correlations with target	
Alcohol	0.476166
Volatile acidity	-0.390558
Sulfates	0.251397
Citric acid	0.226373
Total sulfur dioxide	-0.185112
Density	-0.174919
Chlorides	-0.128907
Fixed acidity	0.124052
pH	-0.057731
Free sulfur dioxide	-0.050554
Residual sugar	0.013732




- Correlation is a measure of the degree of the linear relation between two variables
- It can vary from -1 (full negative correlation, one variable's increase makes the other to decrease) to 1 (positive correlation, the two variables increase together).
- Every variable has obviously correlation = 1 with itself

AGNOSTIC METHODS: PERMUTATION IMPORTANCE

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...
156	142	...	8
153	130	...	24

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...
156	142	...	8
153	130	...	24



- **Permutation importance is calculated after a model has been fitted.** So we won't change the model or change what predictions we'd get for a given value of height, sock-count, etc.
- Instead we will ask the following question: If I randomly shuffle a single column of the data, leaving the target and all other columns in place, how would that affect the accuracy of predictions in that now-shuffled data?

AGNOSTIC METHODS: PERMUTATION IMPORTANCE

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier

data = pd.read_csv('../input/fifa-2018-match-statistics/FIFA 2018 Statistics.csv')
y = (data['Man of the Match'] == "Yes") # Convert from string "Yes"/"No" to binary
feature_names = [i for i in data.columns if data[i].dtype in [np.int64]]
X = data[feature_names]
train_X, val_X, train_y, val_y = train_test_split(X, y, random_state=1)
my_model = RandomForestClassifier(n_estimators=100,
                                random_state=0).fit(train_X, train_y)
```

```
import eli5
from eli5.sklearn import PermutationImportance

perm = PermutationImportance(my_model, random_state=1).fit(val_X, val_y)
eli5.show_weights(perm, feature_names = val_X.columns.tolist())
```

Weight	Feature
0.1750 ± 0.0848	Goal Scored
0.0500 ± 0.0637	Distance Covered (Kms)
0.0437 ± 0.0637	Yellow Card
0.0187 ± 0.0500	Off-Target
0.0187 ± 0.0637	Free Kicks
0.0187 ± 0.0637	Fouls Committed
0.0125 ± 0.0637	Pass Accuracy %
0.0125 ± 0.0306	Blocked
0.0063 ± 0.0612	Saves
0.0063 ± 0.0250	Ball Possession %
0 ± 0.0000	Red
0 ± 0.0000	Yellow & Red
0.0000 ± 0.0559	On-Target
-0.0063 ± 0.0729	Offsides
-0.0063 ± 0.0919	Corners
-0.0063 ± 0.0250	Goals in PSO
-0.0187 ± 0.0306	Attempts
-0.0500 ± 0.0637	Passes

- Our example will use a model that predicts whether a soccer/football team will have the "Man of the Game" winner based on the team's statistics. The "Man of the Game" award is given to the best player in the game.

AGNOSTIC METHODS: Partial dependence plot (PDP)

- The main strength of this permutation importance method is to provide a simple and direct answer about the most important feature.
- But it doesn't help no answering the “How”: we may be interested or asked to answer how goal scored may change the predictions
- PDP sketches the functional form of the relationship between an input feature and the target
- What is performed under the covers by PDP method is to evaluate the effect of changes in a feature over multiple rows to get an average behavior and provide the related functional relationship.

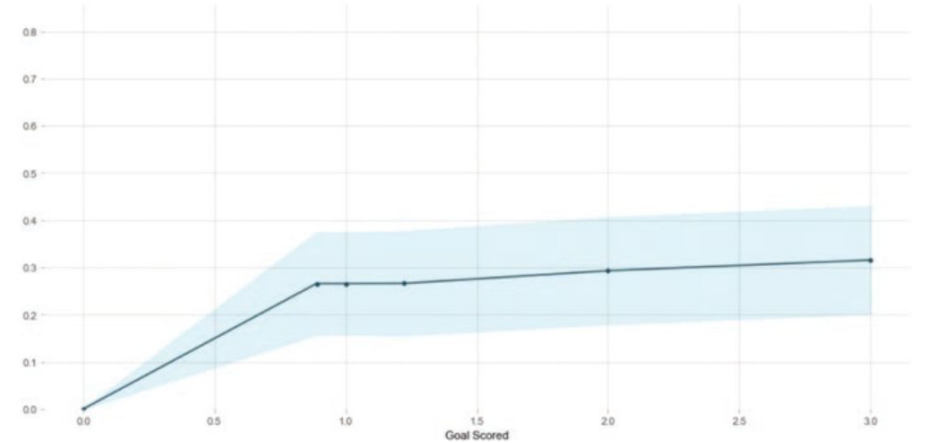


Fig. 4.4 Partial Dependence Plot diagram that shows how “Goal Scored” influences the prediction (Becker 2020)

PDP for feature “Distance Covered (Kms)”
Number of unique grid points: 10

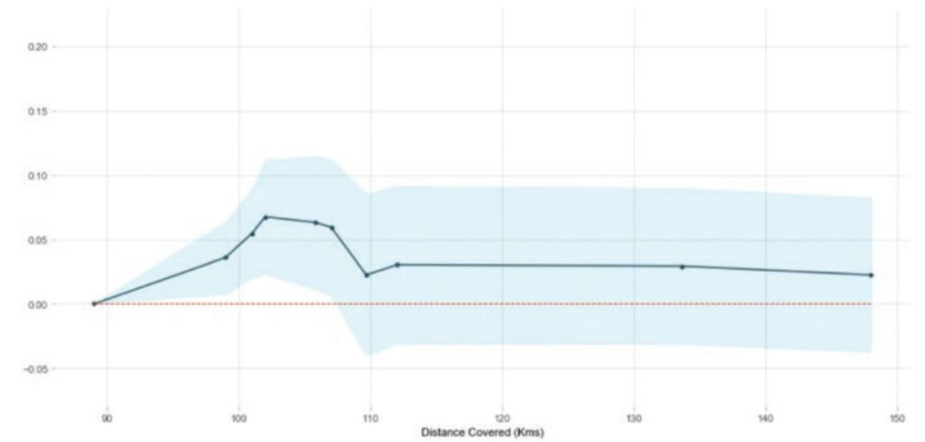


Fig. 4.5 Partial Dependence Plot diagram that shows how “Distance Covered” influences the prediction

Partial dependence plot (PDP)

- Looking at the single diagram of goal scored, it seems there is just a slight variation above one goal
- The maximum effect from distance covered is achieved around 100 km, but with more goals also longer distances produce the same overall effect.

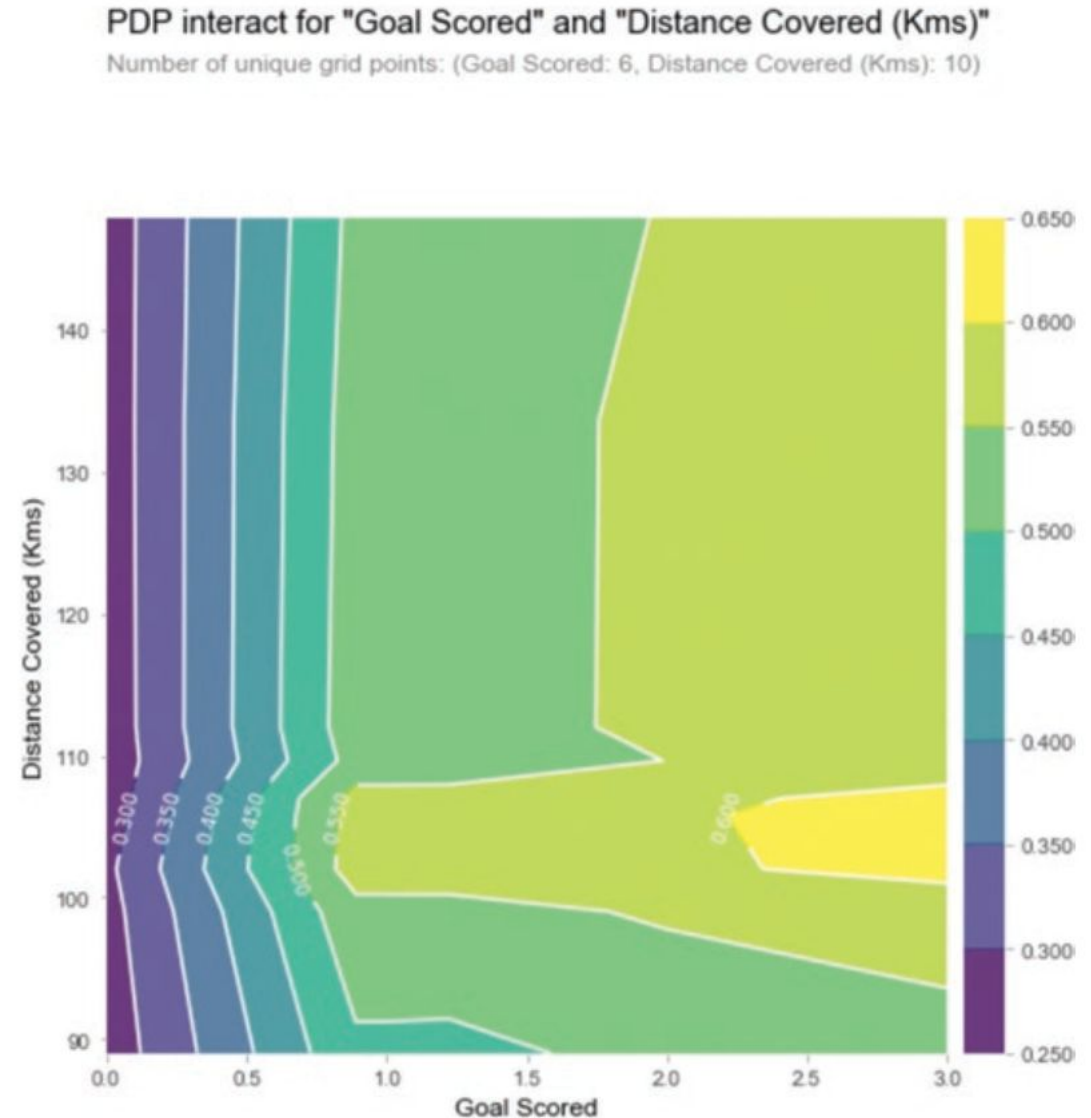
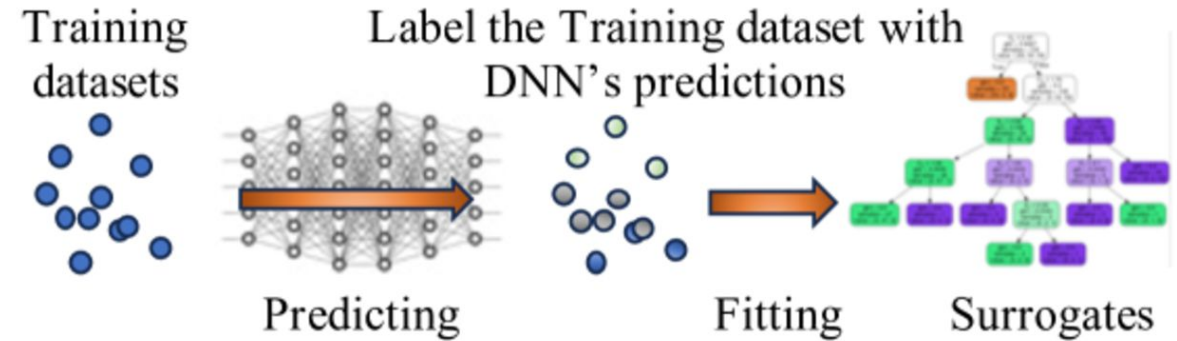


Fig. 4.6 PDP diagram that shows the interaction of the two main features and their impact on the prediction

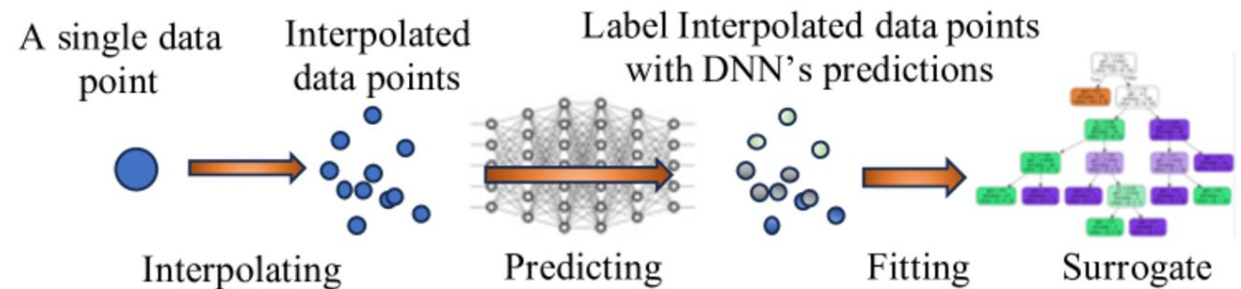
AGNOSTIC METHODS: Global Surrogate

- Proxy explainable models employ interpretable surrogates to interpret complex DNN models, adhering to the principle of simulating the decision boundaries of the original models with simpler constructs.



(a) Global Surrogate

- These surrogate models utilize the predictions from the original DNN as labels to train and elucidate the significance of input features and elucidate their interactions in driving model outputs



(b) Local Surrogate

Agnostic Method: - SHAP (SHapley Additive exPlanations) A Game Theoretical Approach

- If we move to our working scenario of “Player of the Match” prize, so far we provided explanations about the most important features and the functional relationship of these features with the prediction, but we are not able to answer the direct question: considering the features in the figure, **how much the specific prediction for his match has been driven by the number of goals scored by Uruguay?**
- **SHAP method relies on Shapley value, named by Lord Shapley in 1951 who introduced this concept to find solutions in cooperative games.** To set the stage, game theory is a theoretical framework to deal with situations in which we have several individual players and we search for the optimal decisions that depend from the strategy adopted by the other players.
- You see on the left the list of features and on x axis the SHAP value. **The color of each dot represents if that feature is high or low for that specific row of data. The relative position of the dot on x axis shows if that feature contributed positively or negatively to the prediction.** In this way, you may quickly assess if, for each prediction, the feature is almost flat or impacting a lot some rows and nothing to the others.

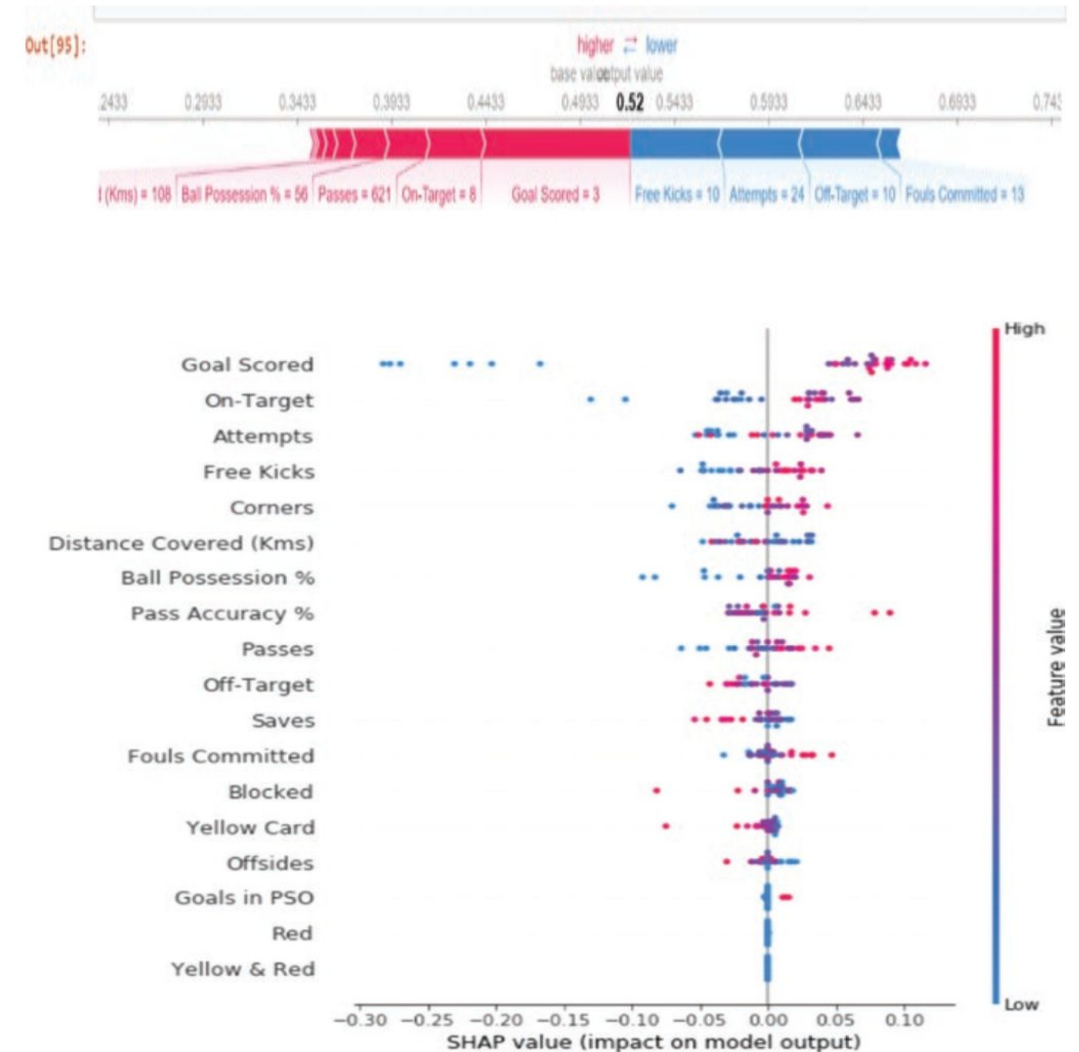


Fig. 4.8 SHAP diagram that shows the features' ranking and the related impact on the match prediction (Becker 2020)

Occlusions as agnostic method

- Using the occlusion in the training phase, we force the black box not to learn by looking at the finer details.
- We can take a pre-trained black box and question it on image content using occlusions as a XAI method. The model has already been trained and fixed, so we don't care in this phase of training it in a robust way
- We want to understand which details of the image are most significant for the class' attribution or to evaluate the importance of some group of pixels

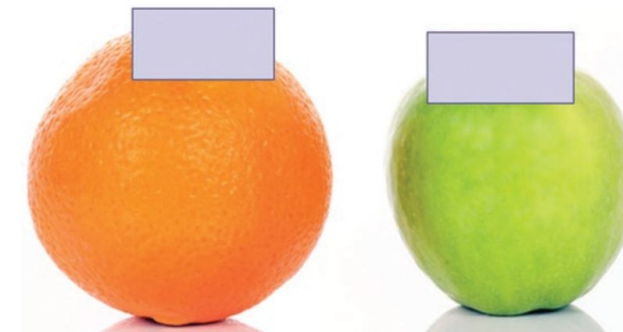


Fig. 5.4 The occlusion idea as an augmentation technique: random gray rectangles force the model to rely more on robust features such as skin's texture

Fig. 5.5 Original image

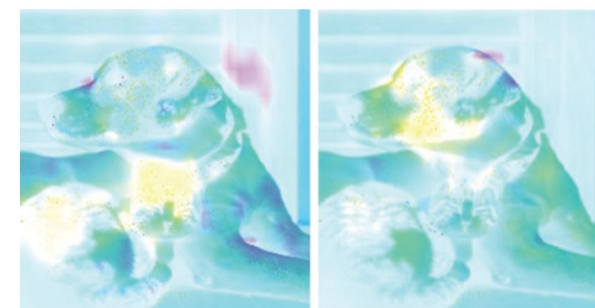
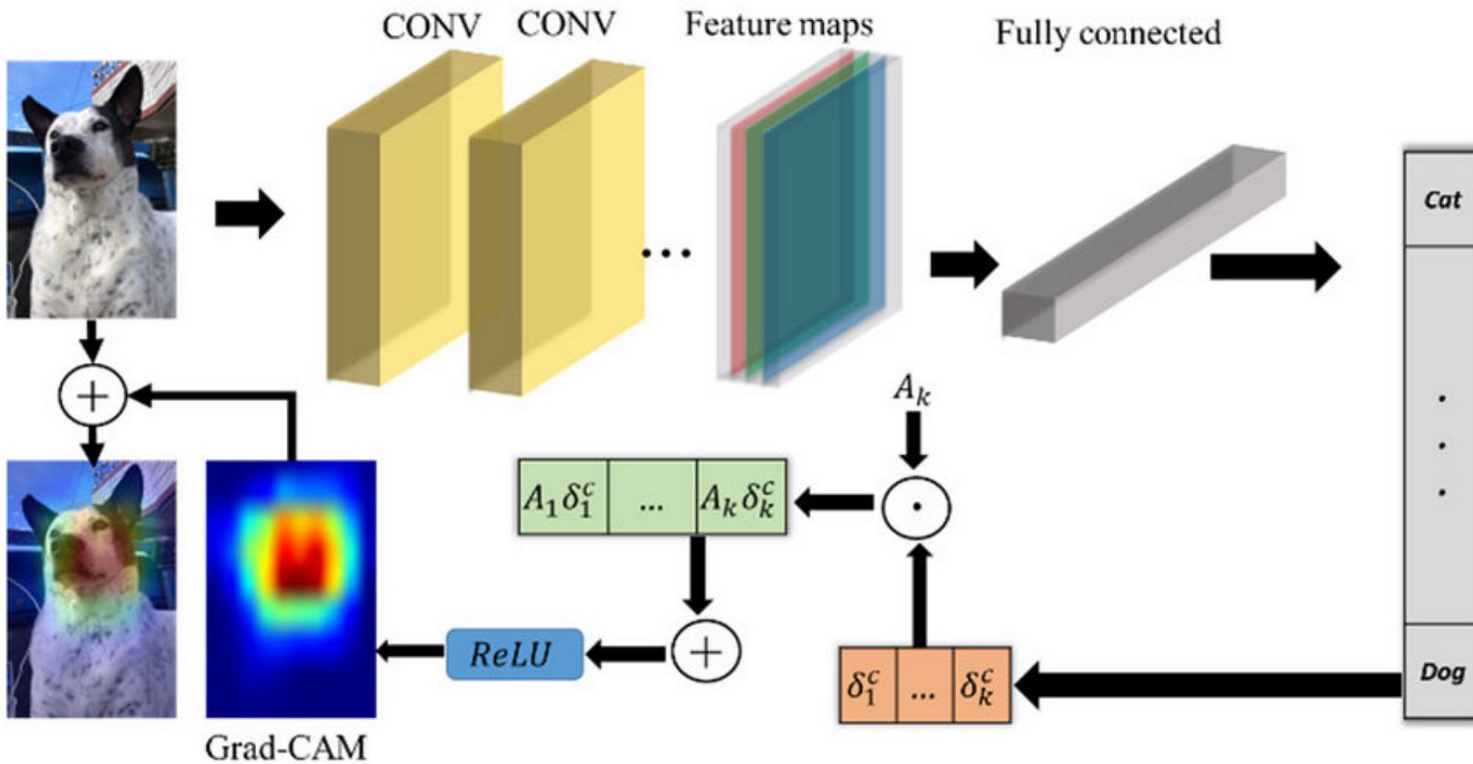


Fig. 5.6 Occlusions used to highlight relevant features respectively for the tabby cat and the dog classes

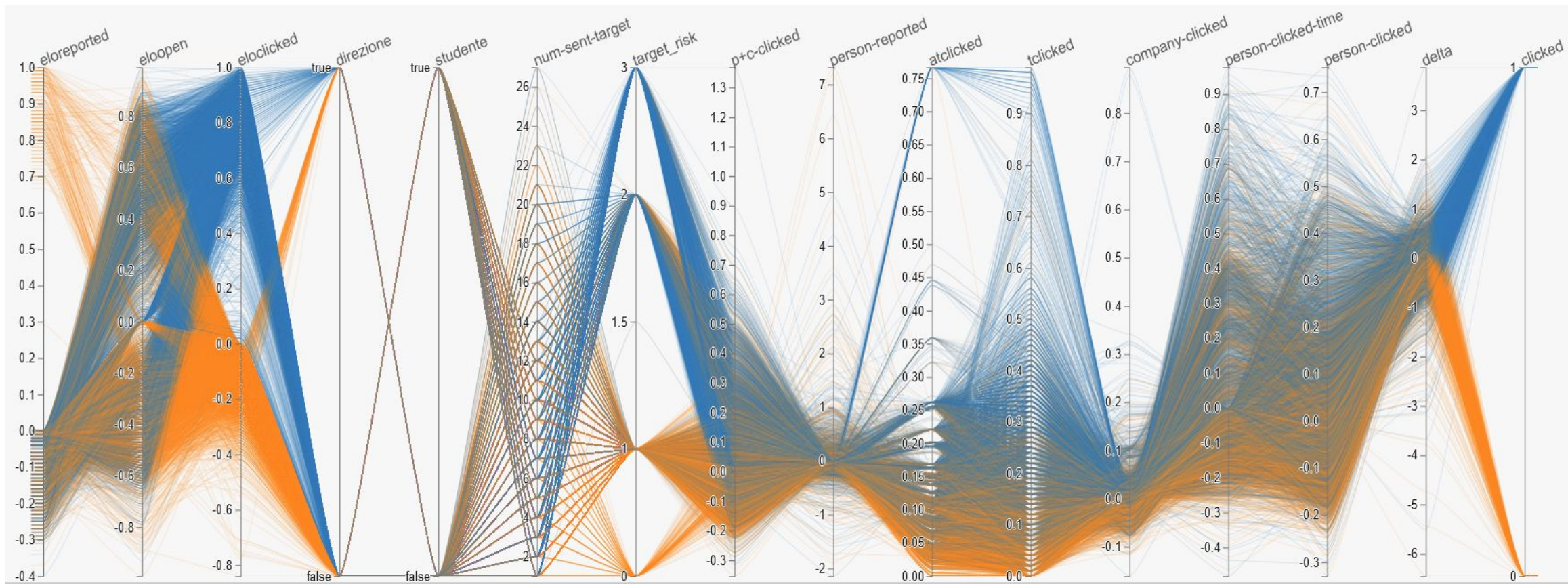
Grad-Cam as model specific method

- Grad-CAM (Gradient-weighted Class Activation Mapping) is a model-specific method, which provides local explanations for Deep Neural Networks.



- Grad-CAM works by computing the gradients of the model's output with respect to the feature maps in the final convolutional layer, effectively revealing which parts of the image the model 'looks at' when making a prediction.
- As model specific method requires access to gradients and internal layers.

Features Exploration @Cyber Guru Phishing



Parallel Plot, features importance

XAI @ LLMs

- In contrast to traditional deep learning models, the scale of LLMs in terms of parameters and training data introduces both complex challenges and exciting opportunities for explainability research.
- On the one hand, traditionally practical feature attribution techniques, such as gradient-based methods and SHAP values, could demand substantial computational power to explain LLMs with billions of parameters.
- This makes these explanation techniques less practical for real-world applications that end-users can utilize

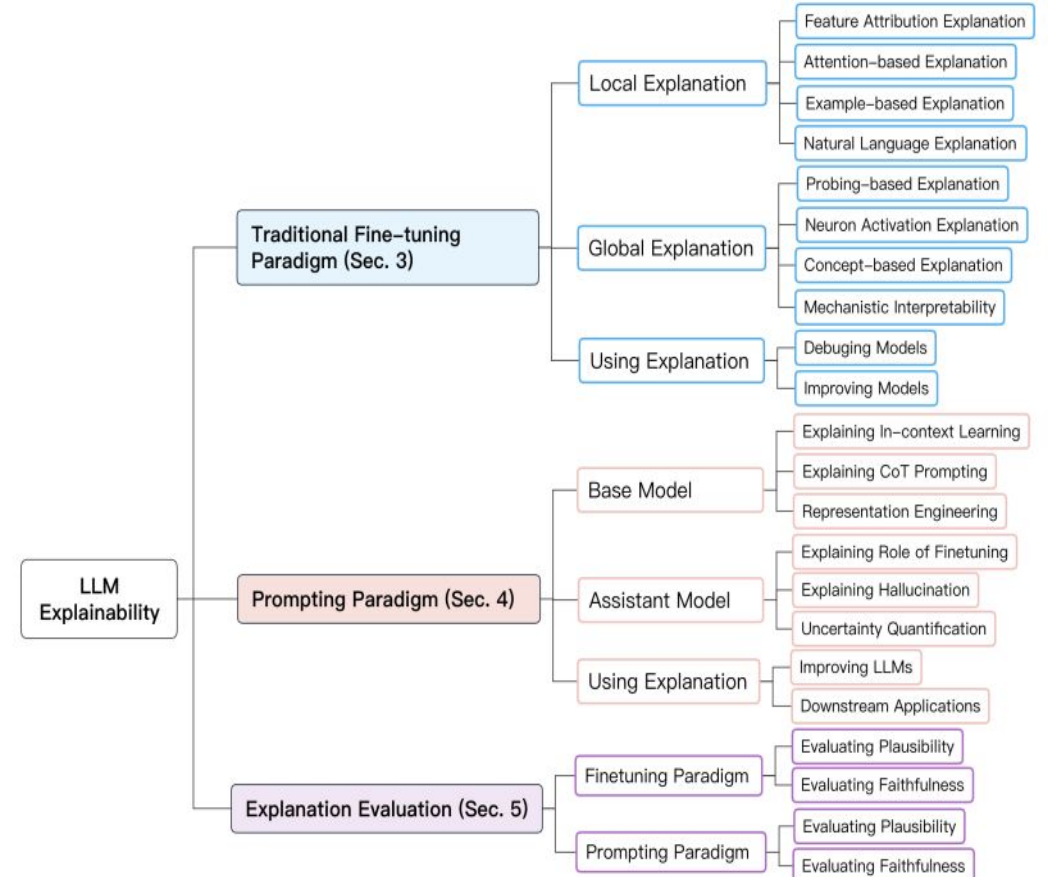


Figure 1: We categorize LLM explainability into two major paradigms. Based on this categorization, we summarize different kinds of explainability techniques associated with LLMs belonging to these two paradigms. We also discuss evaluations for the generated explanations under the two paradigms.

Large Language Model (LLM)

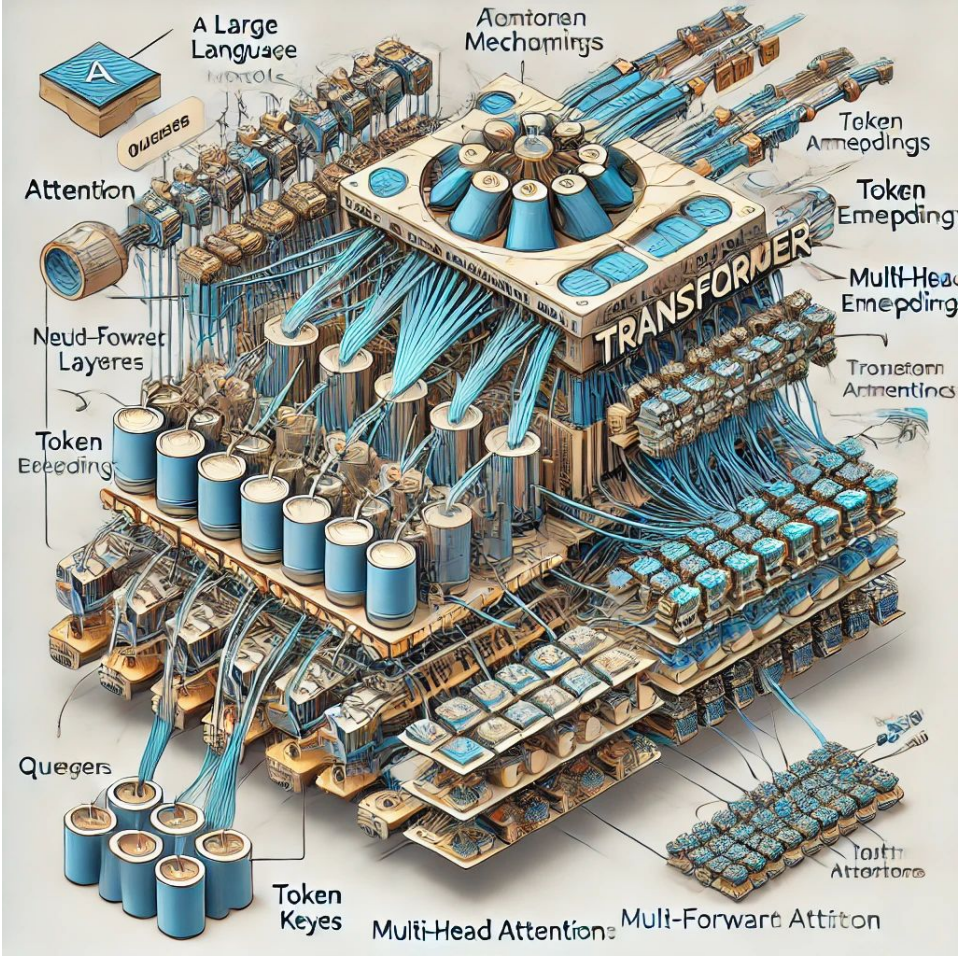


Image generated with GPT-4o

LLM

- **Large Language Model (LLM)** are giant neural networks with billions of parameters (weights), trained on large quantities of unlabelled text using self-supervised learning
- A message is splitted in **tokens** (sub-word)
- Each token is translated in a number using an operation called **embeddings**
- LLM works by taking an input text and **repeatedly predicting** the next token or word

Attention Is All You Need

- Google and University of Toronto published a paper in 2017 “[Attention is All You Need](#)”
- In this paper, they introduced the **Transformer architecture**
- This novel approach unlocked the progress in NLP that we see today
- Scale efficiently, parallel process, attention to input meaning

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com	Noam Shazeer* Google Brain noam@google.com	Niki Parmar* Google Research nikip@google.com	Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com	Aidan N. Gomez* [†] University of Toronto aidan@cs.toronto.edu	Lukasz Kaiser* Google Brain lukaszkaizer@google.com	
Illia Polosukhin* [‡] illia.polosukhin@gmail.com			

Abstract

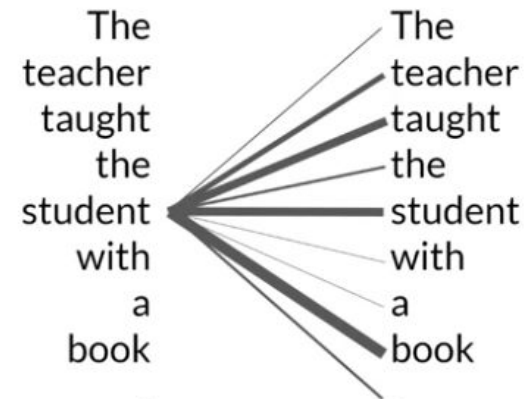
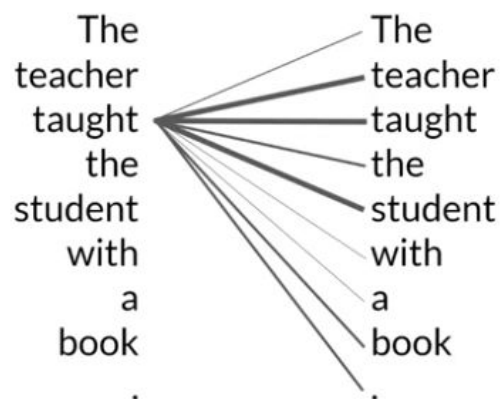
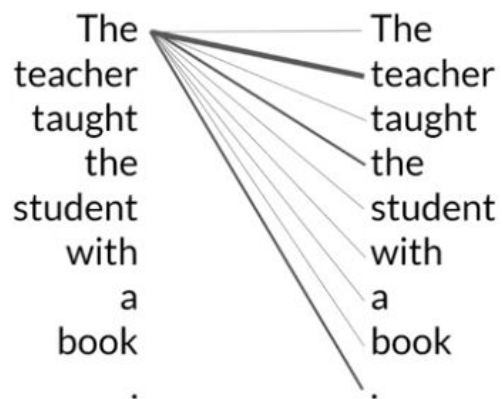
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

* Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

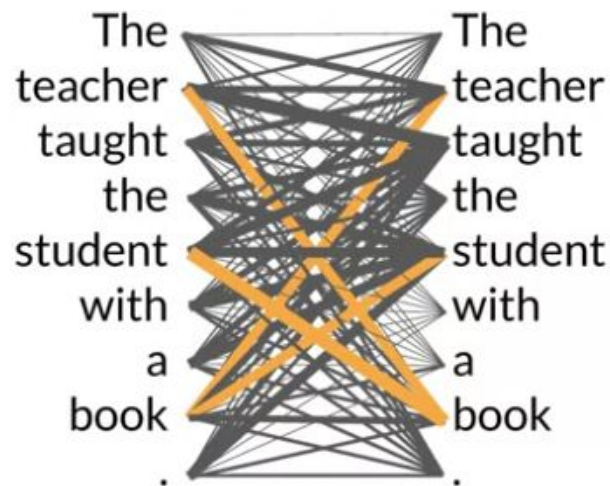
[†]Work performed while at Google Brain.

[‡]Work performed while at Google Research.

Attention map

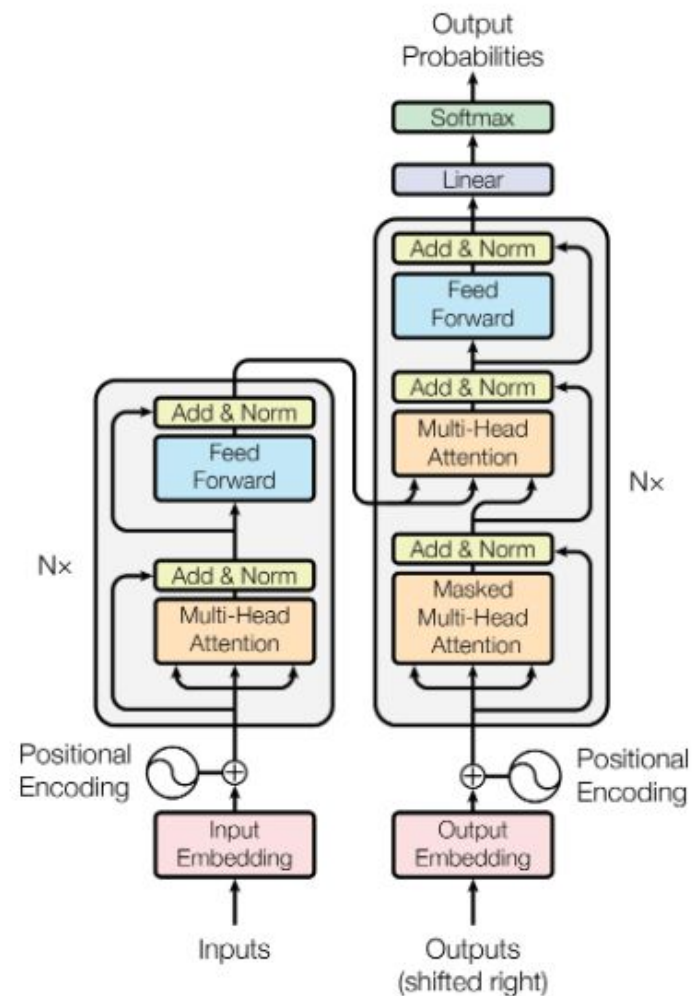
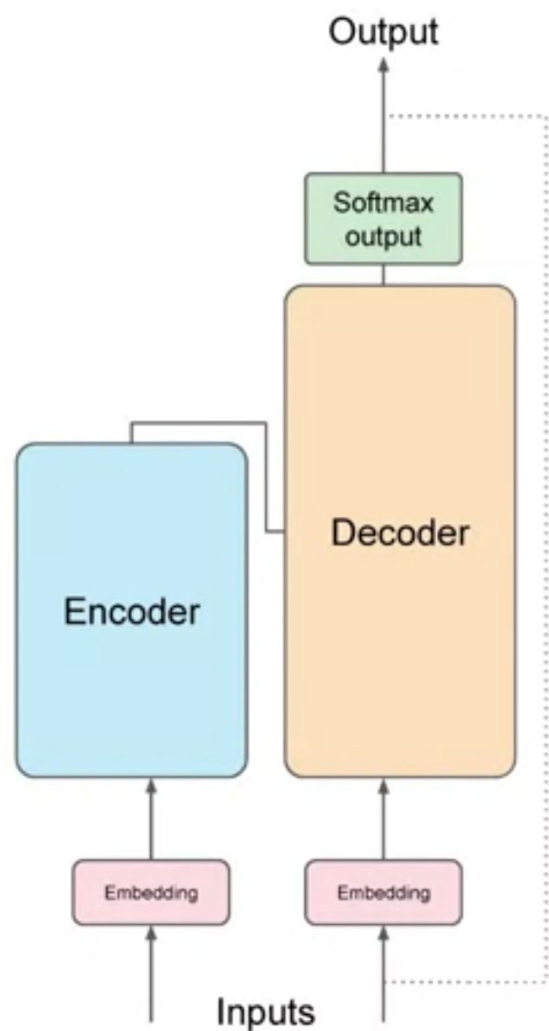


eg. **book** is strongly connected with **teacher** and **student**

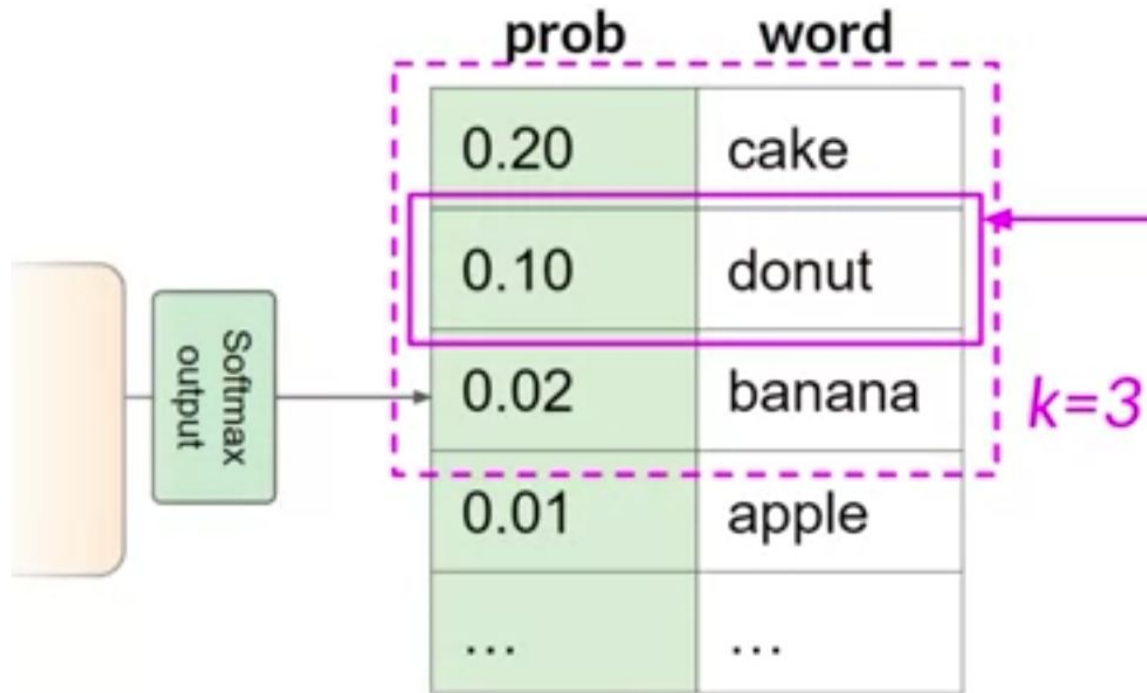


self-attention

Transformers architecture

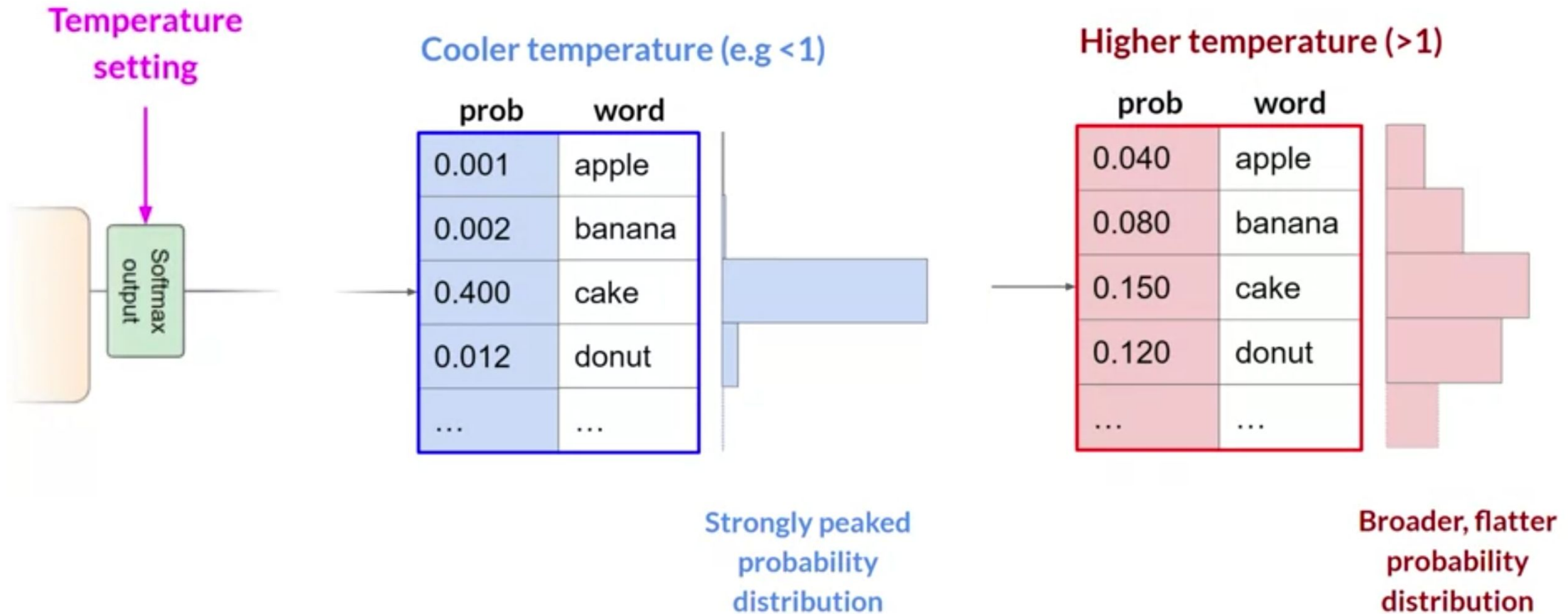


Top-k



top-k: select an output from the top-k results after applying random-weighted strategy using the probabilities

Temperature



XAI and LLM

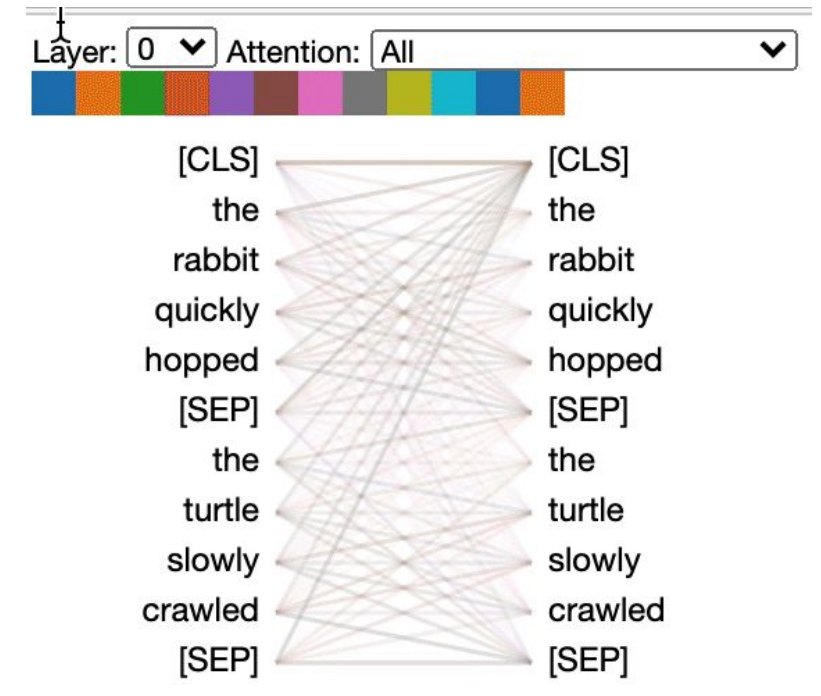
- Local explanation
 - Visualizing the Attention
 - Heat-map of self-attention
 - Transformer Explainer
 - LLM visualization
 - Grad-CAM visualization
- Black-box explanation (i.e. using prompting techniques)
 - In-Context Learning (ICL) and Chain-of-Thought (CoT)
 - Benchmark (e.g. GSM-Symbolic)
 - LLMs as “bullshit”

Paper and Tool	Star	Fork	Update	Target	Agnostic	Goal				
Vig [2019] BertViz	6.1k	734	08/23	Transformers	✓	C	E	IMP	INT	R
Swamy et al. [2021] Experiments	19	2	05/22	BERT-based LM	✗	C	E	IMP	INT	R
Wu et al. [2021] Polyjuice	90	16	08/22	-	✓	C	E	IMP	INT	R
Wang et al. [2022] TransformerLens	48	161	01/23	GPT2-small	✗	C	E	IMP	INT	R
Menon and Vondrick [2022] -	-	-	-	Vision-LM	✓	C	E	IMP	INT	R
Gao et al. [2023a] Experiments	17	0	10/23	ChatGPT	✗	C	E	IMP	INT	R
Pan et al. [2023] -	-	-	-	LLMs	✓	C	E	IMP	INT	R
Conmy et al. [2023] ACDC	105	23	11/23	Transformers	✓	C	E	IMP	INT	R
He et al. [2022] RR	38	2	02/23	LLMs	✓	C	E	IMP	INT	R
Yoran et al. [2023] MCR	71	9	01/24	LLMs	✓	C	E	IMP	INT	R
Sarti et al. [2023] Inseq	250	26	01/24	SeqGen models	✓	C	E	IMP	INT	R
Wu et al. [2023b] Boundless DAS	0	17	01/24	LLMs	✓	C	E	IMP	INT	R
Li et al. [2023] XICL	1	3	11/23	LLMs	✓	C	E	IMP	INT	R
Chen et al. [2023] LMExplainer	-	-	-	LLMs	✓	C	E	IMP	INT	R
Gao et al. [2023b] Chat-REC	-	-	-	Rec. systems	✗	C	E	IMP	INT	R
Zhang et al. [2022] DSRLM	9	1	07/23	LLMs	✓	C	E	IMP	INT	R
Singh et al. [2023] SASC	61	14	01/24	LLMs	✓	C	E	IMP	INT	R
Li et al. [2022] -	-	-	-	LLMs	✓	C	E	IMP	INT	R
Ye and Durrett [2022] TextualExplnContext	11	2	02/23	LLMs	✓	C	E	IMP	INT	R
Turpin et al. [2023] Experiments	25	9	03/23	LLMs	✓	C	E	IMP	INT	R
Kang et al. [2023] AutoSD	-	-	-	Debugging models	✗	C	E	IMP	INT	R
Krishna et al. [2023] AMPLIFY	-	-	-	LLMs	✓	C	E	IMP	INT	R
Yang et al. [2023] Labo	51	4	12/23	CBM	✗	C	E	IMP	INT	R
Bitton-Guetta et al. [2023] WHOOPS!	-	-	-	LLMs	✓	C	E	IMP	INT	R
Shi et al. [2023] Chatgraph	2	0	07/23	LLMs	✓	C	E	IMP	INT	R

C = comparison of model
 E = explanation
 IMP = improvement
 INT = interpretability
 R = reasoning

Visualizing the Attention*

- An open-source tool that visualizes attention at multiple scales, each of which provides a unique perspective on the attention mechanism
- <https://github.com/jessevig/bertviz>



* Jesse Vig, [A Multiscale Visualization of Attention in the Transformer Model](#), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019

Heatmap of self-attention

```
import inseq

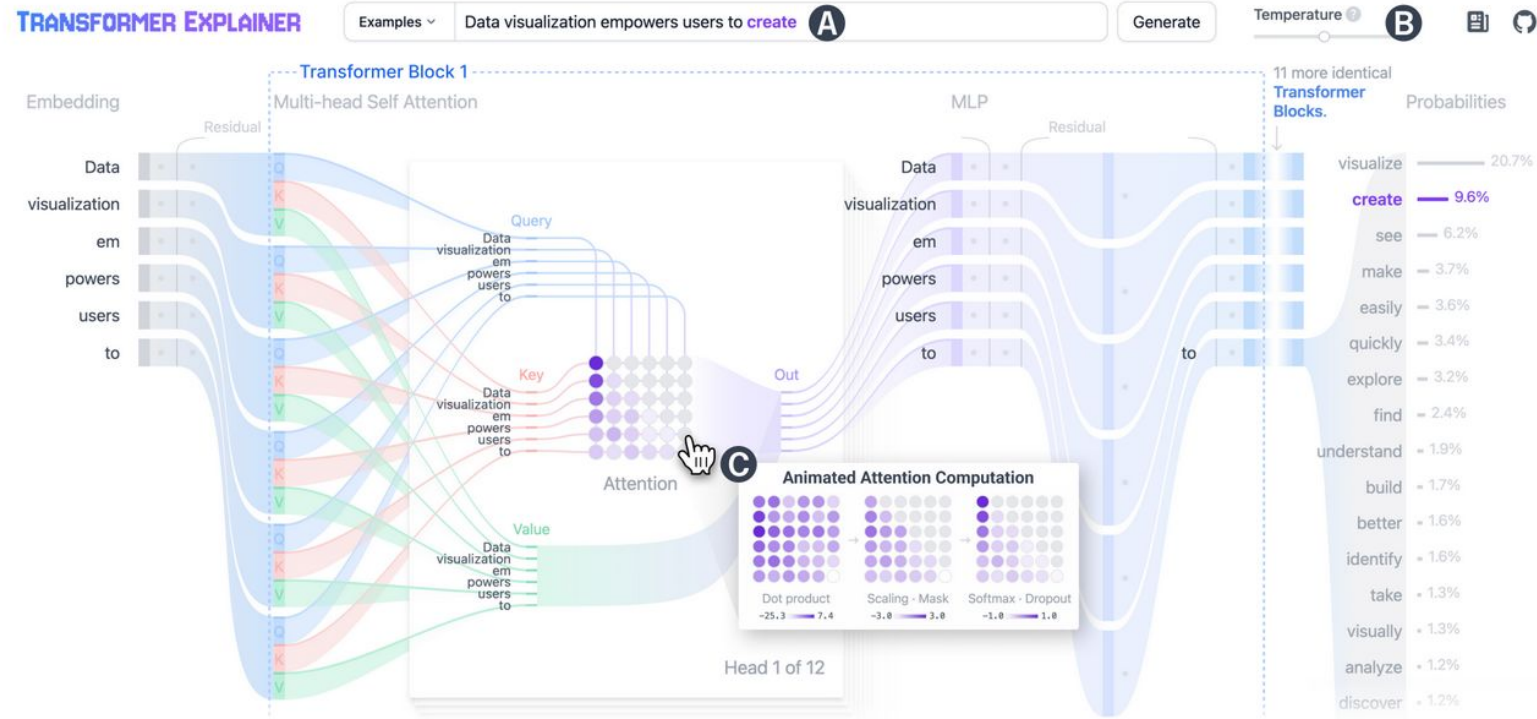
model = inseq.load_model("Helsinki-NLP/opus-mt-en-fr", "integrated_gradients")
out = model.attribute(
    "The developer argued with the designer because her idea cannot be implemented.",
    n_steps=100
)
out.show()
```

0th instance:



<https://github.com/inseq-team/inseq>

Transformer Explainer*



<https://poloclub.github.io/transformer-explainer/>

* Aeree Cho et al., [Transformer Explainer: Interactive Learning of Text-Generative Models](#), IEEE VIS, 2024

LLM visualization

<https://bbycroft.net/llm>

LLM Visualization

Chapter: Overview

GPT-2 (small) nano-gpt GPT-2 (XL) GPT-3

Table of Contents

- Intro
 - Introduction
 - Preliminaries
- Components
 - Embedding
 - Layer Norm
 - Self Attention
 - Projection
 - MLP
 - Transformer
 - Softmax
 - Output

LLM

How to predict text
2437 284 4331 2420
tokens 16326
words 2456

tok embed

pos embed \ominus \oplus

transformer i

- layer norm
- multi-head, causal self-attention
- layer norm \oplus
- feed forward
- layer norm \oplus
- linear
- softmax

nano-gpt
n_params = 85,584

Welcome to the walkthrough of the GPT large language model! Here we'll explore the model *nano-gpt*, with a mere 85,000 parameters.

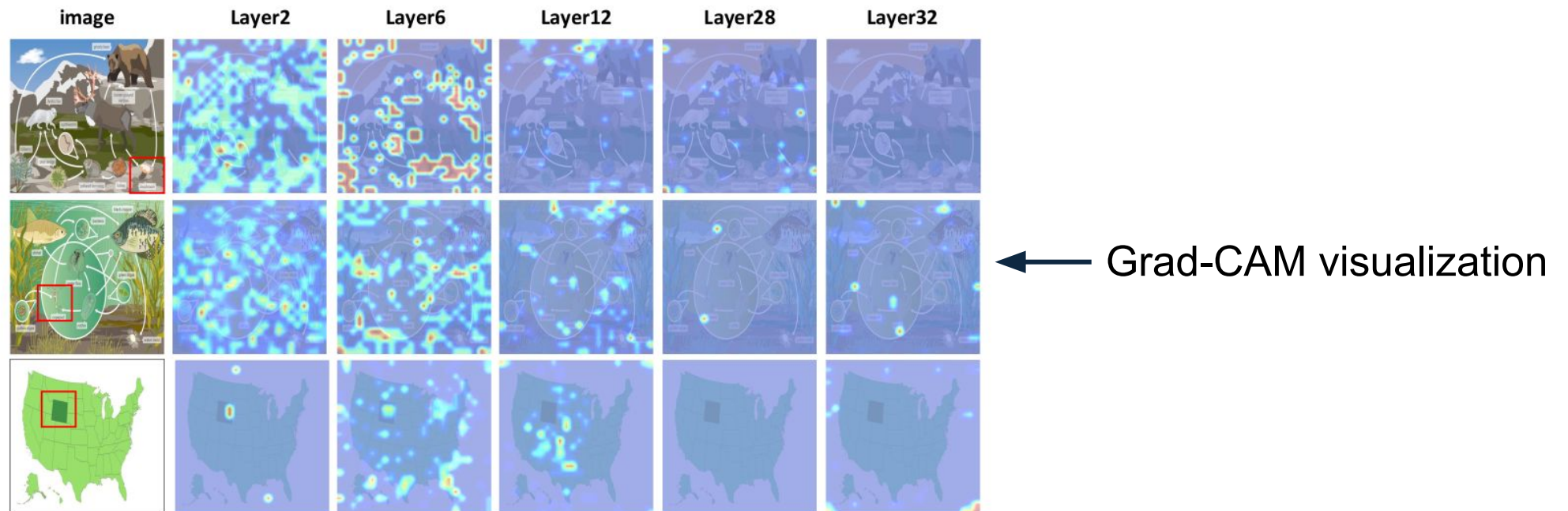
Its goal is a simple one: take a sequence of six letters:

C B A B B C

and sort them in alphabetical order, i.e. to "ABBCC".

Explainability in Multimodal Large Language Models*

- Large Vision Language Models are capable to generate and analyze images
- The information flow appears to converge in the shallow layer



* Xiaofeng Zhang et al., [From Redundancy to Relevance: Enhancing Explainability in Multimodal Large Language Models](#), arXiv:2406.06579v1

GSM-Symbolic benchmark*

- GSM-Symbolic has been proposed recently for measuring the reasoning capabilities of LLMs
- LLMs exhibit noticeable variance when responding to different instantiations of the same question (e.g. altering numerical values in the question)
- We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data

* Iman Mirzadeh et al., [GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models](#), ICLR 2025 Conference Submission

ChatGPT is bullshit*

- The overall activity of large language models, is better understood as bullshit in the sense explored by Frankfurt (On Bullshit, Princeton, 2005)
- The models are in an important way indifferent to the truth of their outputs
- Considering LLMs as “bullshit” is a more useful and more accurate way of predicting and discussing the behaviour of these systems

* Hicks, M.T., Humphries, J. & Slater, J. [ChatGPT is bullshit](#). Ethics and Information Technology 26, 38 (2024)

Conclusions

- LLMs have propelled NLP techniques forward by an incredible leap
- Emergent capabilities appeared in large model (e.g. math problems)
- Traditional XAI techniques are not so useful when applied to LLM
- We have a limited knowledge about these emergent properties
- Recent studies have scaled back the perceived reasoning capabilities of these LLMs
- How these capabilities emerge in LLMs is still an open question
- The perceived reasoning from the attention mechanism is still under study

References

- Ashish Vaswan et al., [Attention Is All You Need](#), Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017)
- Nello Cristianini, [Machina sapiens. L'algoritmo che ci ha rubato il segreto della conoscenza](#), Il Mulino (2024)
- Leonida Gianfagna, Antonio Di Cecco, [Explainable AI with Python](#), Springer 2021
- Rylan Schaeffer, Brando Miranda, Sanmi Koyejo, [Are Emergent Abilities of Large Language Models a Mirage?](#), arXiv:2304.15004, 2023
- Microsoft Research, [Sparks of Artificial General Intelligence:Early experiments with GPT-4](#), arXiv:2303.12712, 2023
- Jiawei Su*, Danilo Vasconcellos Vargas and Kouichi Sakurai, [One Pixel Attack for Fooling Deep Neural Networks](#), IEEE Transactions on Evolutionary Computation, 2019
- OpenAI Research, [Improving Language Understanding by Generative Pre-Training](#), 2018
- Haiyan Zhao et al., [Explainability for Large Language Models: A Survey](#), 2023
- Wei, Jason, et al. "[Emergent abilities of large language models](#)." *arXiv preprint arXiv:2206.07682* (2022).
- Vafa, Keyon, et al. "[Evaluating the World Model Implicit in a Generative Model](#)." *arXiv preprint arXiv:2406.03689* (2024).
- Erik Cambria et al., "[XAI meets LLMs: A Survey of the Relation between Explainable AI and Large Language Models](#)" *arXiv preprint arXiv:2407.15248* (2024)

Thank you!

Contacts:

enrico.zimuel (at) elastic.co

leonida.gianfagna (at) cyberguru.eu

