Explainability and Interpretability for NLP

ANITI

cea list

CentraleSupélec

DEEL
Dependable, Explainable & Embeddable Learning

FOR

cnrs GDR Groupement de recherche
RADIA Raisonnement, Apprentissage, et Décision en Intelligence Artificielle

IRIT

IRT SAINT EXUPÉRY

SCIENTA Lab

université PARIS-SACLAY

PFIA 2025 Dijon

Antonin **Poché**
PhD Student at IRT Saint Exupéry & IRIT

Wassila **Ouerdane**
Prof. MICS, CentraleSupélec,
Univ. Paris-Saclay

Charlotte **Claye**
PhD Student ScientaLab, MICS,
CentraleSupélec, Univ.
Paris-Saclay

Jean-Philippe **Poli**
DR. CEA List
Univ. Paris-Saclay / Carnot List

Celine **Hudelot**
Prof. MICS, CentraleSupélec,
Univ. Paris-Saclay

SCAN ME

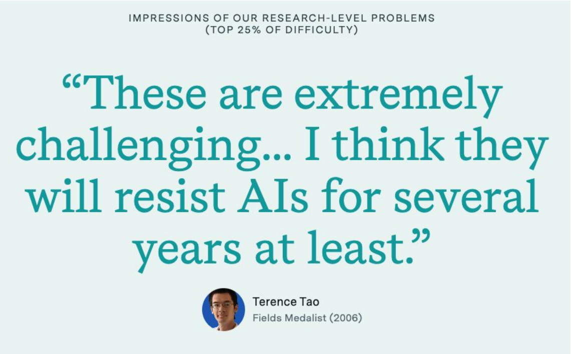
Table of content

- Context & Motivations
- User-centered Explanations
- Research-centered explanation
- In practice: Interpreto Library

Context & Motivations

Context : High-Performing Language Models

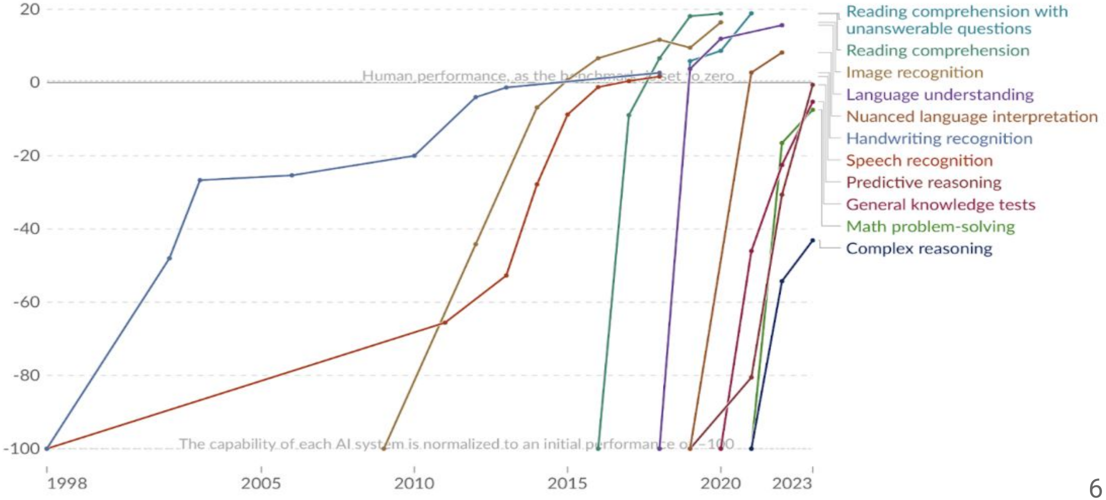
- Accelerating progress in language AI
- An increase in usage intentions in many domains
- Language AI at the central pillar of Generative AI



IMPRESSIONS OF OUR RESEARCH-LEVEL PROBLEMS
(TOP 25% OF DIFFICULTY)

“These are extremely challenging… I think they will resist AIs for several years at least.”

Terence Tao
Fields Medalist (2006)



https://ourworldindata.org/grapher/test-scores-ai-capabilities-relative-human-performance

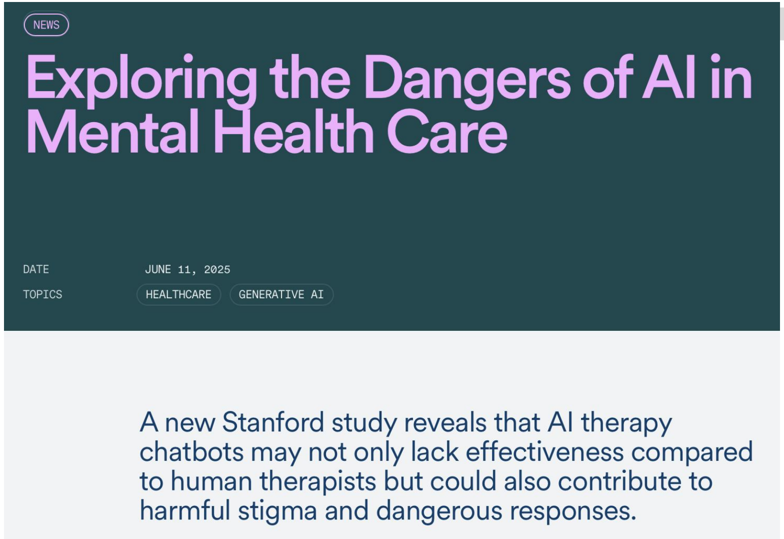
Sources: https://epoch.ai/frontiermath
https://ourworldindata.org

Context: Several (critical) applications

text.cortex

Context : Prone to unexpected failures

https://www.arxiv.org/pdf/2506.10922



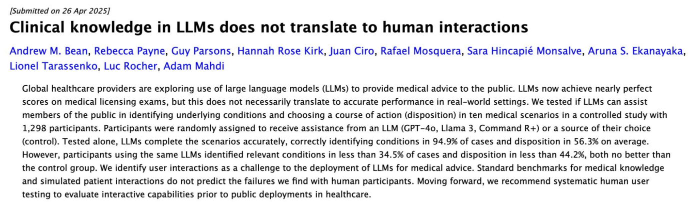
NEWS

Exploring the Dangers of AI in Mental Health Care

DATE JUNE 11, 2025

TOPICS HEALTHCARE GENERATIVE AI

A new Stanford study reveals that AI therapy chatbots may not only lack effectiveness compared to human therapists but could also contribute to harmful stigma and dangerous responses.

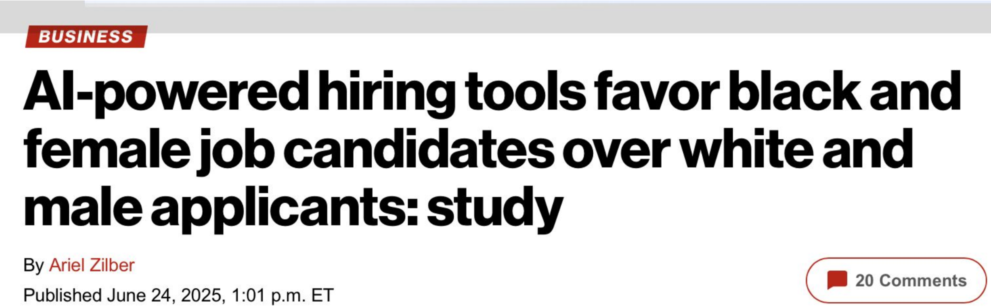


[Submitted on 26 Apr 2025]

Clinical knowledge in LLMs does not translate to human interactions

Andrew M. Bean, Rebecca Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera, Sara Hincapié Monsalve, Aruna S. Ekanayaka, Lionel Tarassenko, Luc Rocher, Adam Mahdi

Global healthcare providers are exploring use of large language models (LLMs) to provide medical advice to the public. LLMs now achieve nearly perfect scores on medical licensing exams, but this does not necessarily translate to accurate performance in real-world settings. We tested if LLMs can assist members of the public in identifying underlying conditions and choosing a course of action (disposition) in ten medical scenarios in a controlled study with 1,298 participants. Participants were randomly assigned to receive assistance from an LLM (GPT-4o, Llama 3, Command R+) or a source of their choice (control). Tested alone, LLMs complete the scenarios accurately, correctly identifying conditions in 94.9% of cases and disposition in 56.3% on average. However, participants using the same LLMs identified relevant conditions in less than 34.5% of cases and disposition in less than 44.2%, both no better than the control group. We identify user interactions as a challenge to the deployment of LLMs for medical advice. Standard benchmarks for medical knowledge and simulated patient interactions do not predict the failures we find with human participants. Moving forward, we recommend systematic human user testing to evaluate interactive capabilities prior to public deployments in healthcare.

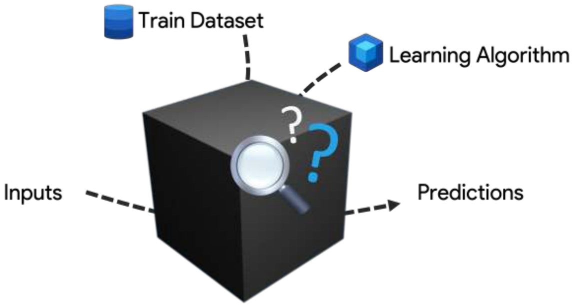


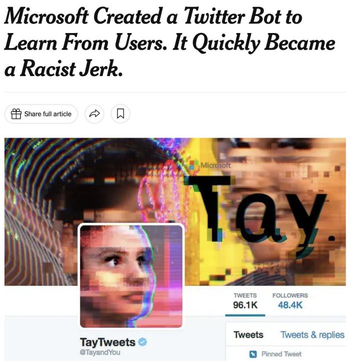
BUSINESS

AI-powered hiring tools favor black and female job candidates over white and male applicants: study

By Ariel Zilber

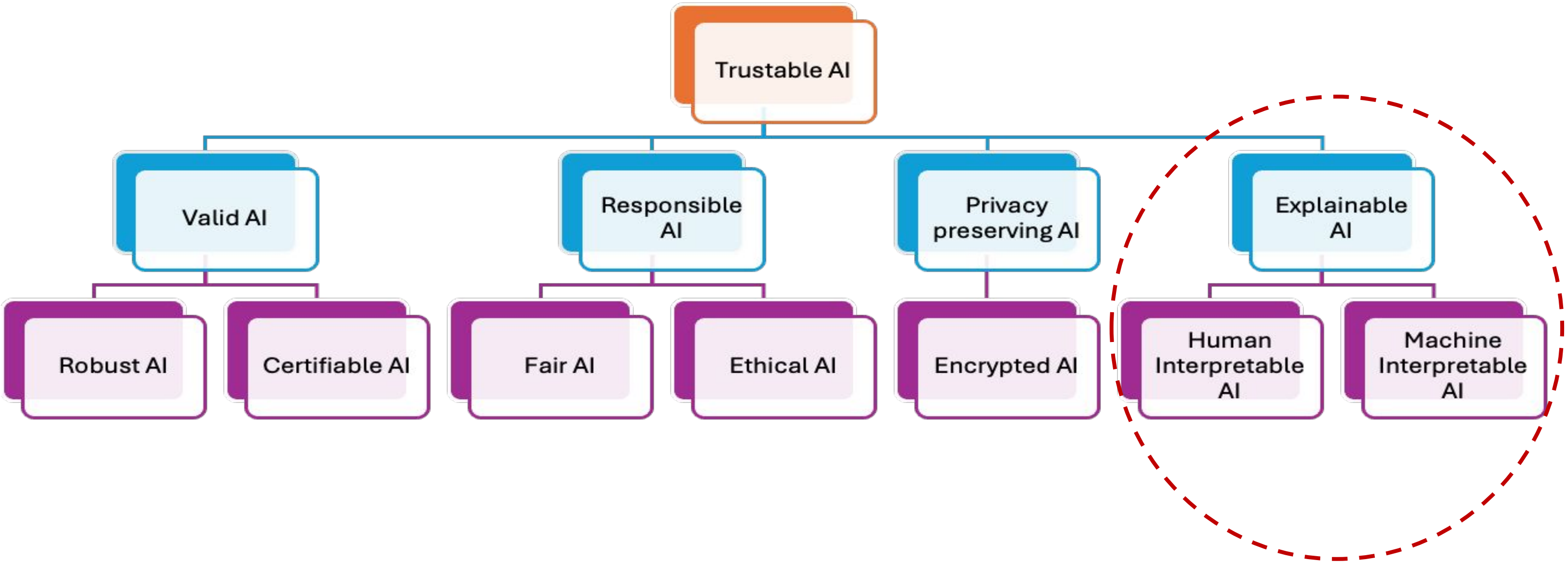
Published June 24, 2025, 1:01 p.m. ET





Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

Requirements for AI adoption



Trustable AI

Valid AI

Responsible AI

Privacy preserving AI

Explainable AI

Robust AI

Certifiable AI

Fair AI

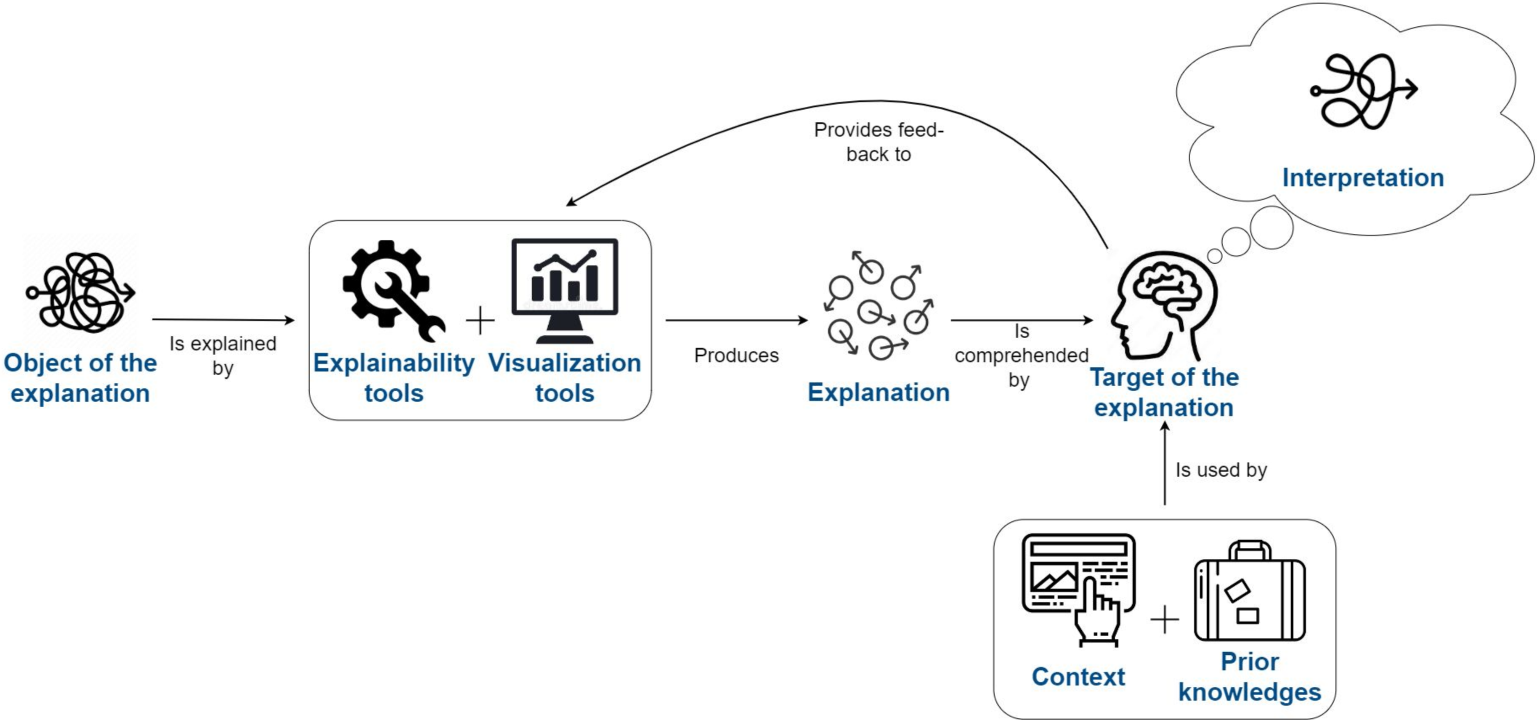
Ethical AI

Encrypted AI

Human Interpretable AI

Machine Interpretable AI

The key component of Explainability

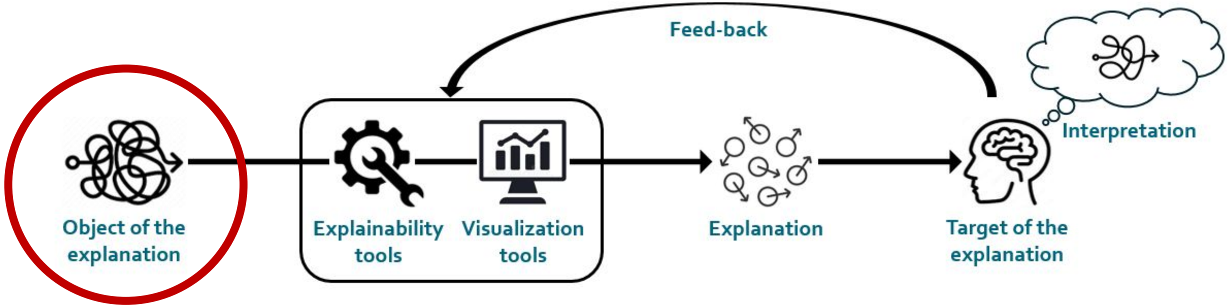


Source: [Tutorial PFIA 2024]

Scope of the explanation

Model

Feature Viz,
Concept Activation Vector
Explanation “by design”
...



Predictions

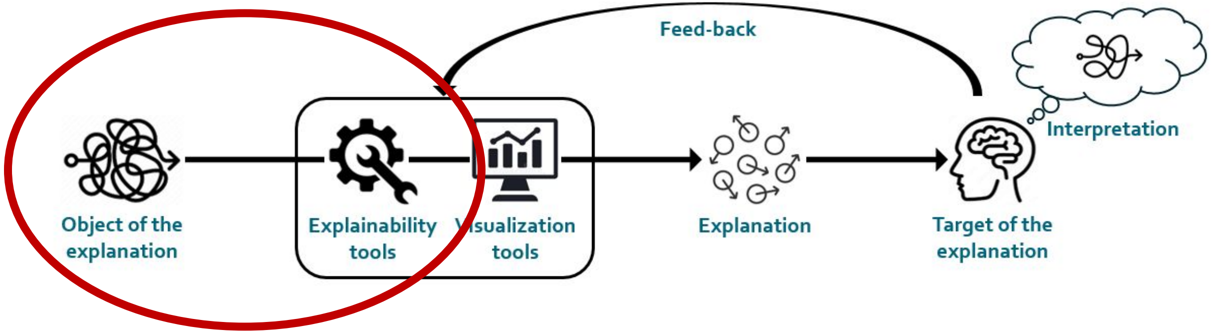
Feature Attribution
Feature Inversion
...

Data

Nearest Neighbourhood
Influence Function
Prototypes
...

Source: [Tutorial PFIA 2024]()

Application time



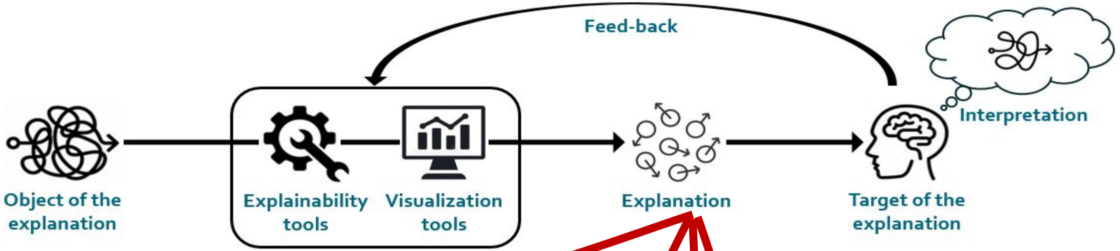
Model construction ● By-design

(ant-hoc/transparent/self-explaining)

Trained model ▼ Post-hoc

Source: [Tutorial PFIA 2024]

Format of the explanations



Feed-back

Interpretation

Object of the explanation

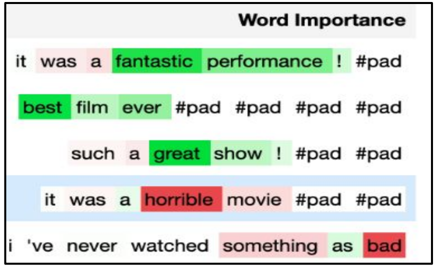
Explainability tools

Visualization tools

Explanation

Target of the explanation

Attributions



Word Importance

it was a fantastic performance ! #pad

best film ever #pad #pad #pad #pad

such a great show ! #pad #pad

it was a horrible movie #pad #pad

i 've never watched something as bad

Captum tutorial

Concept-based

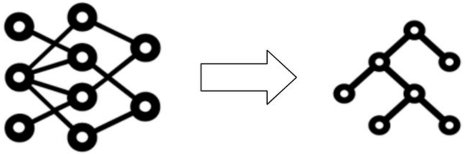
90-year-old
51-year-old
68-year-old
76-year-old
79-year-old
70-year-old
twenty-five-year-old
94-year-old
kindest
bravest
luckiest
loveliest
wildest
scariest
widest
quietest
Weird
quickest
tightest
smartest
proudest
starkest
harshest
sweetest
fiercest
cutest
best-case
clearest

Dalvil, et al (ICLR, 2022)

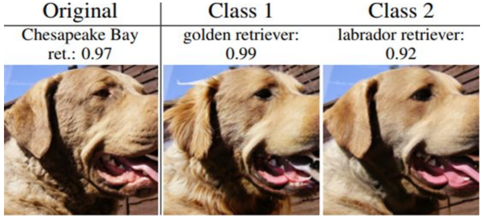
Feature viz

Xplique

Model surrogate



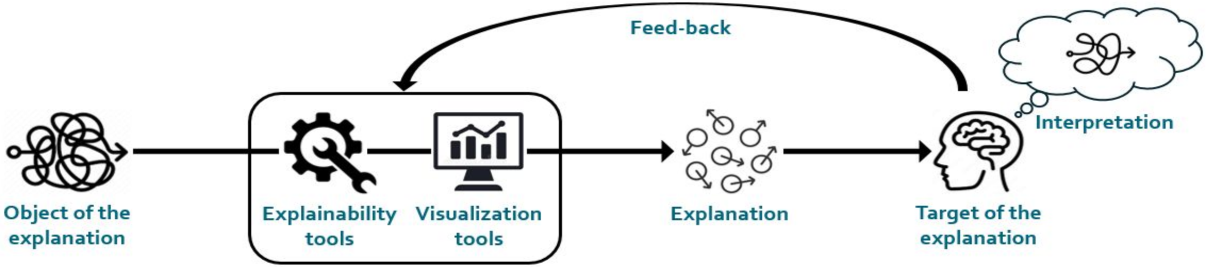
Example-based

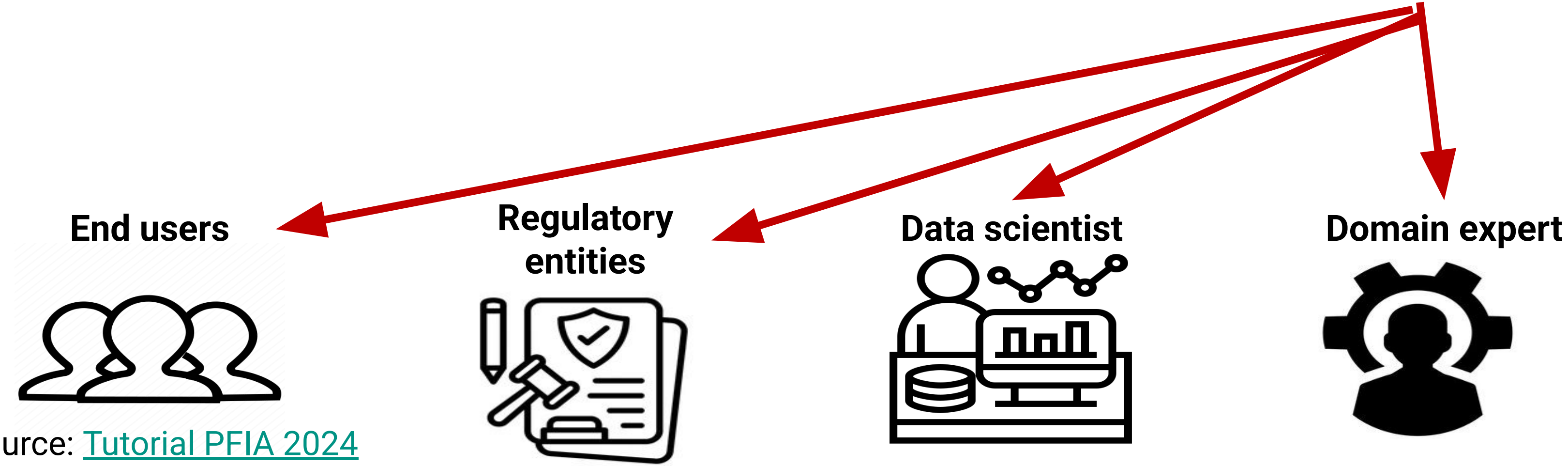


Original	Class 1	Class 2
Chesapeake Bay ret.: 0.97	golden retriever: 0.99	labrador retriever: 0.92

Source: Tutorial PFIA 2024

Target of explanation





End users

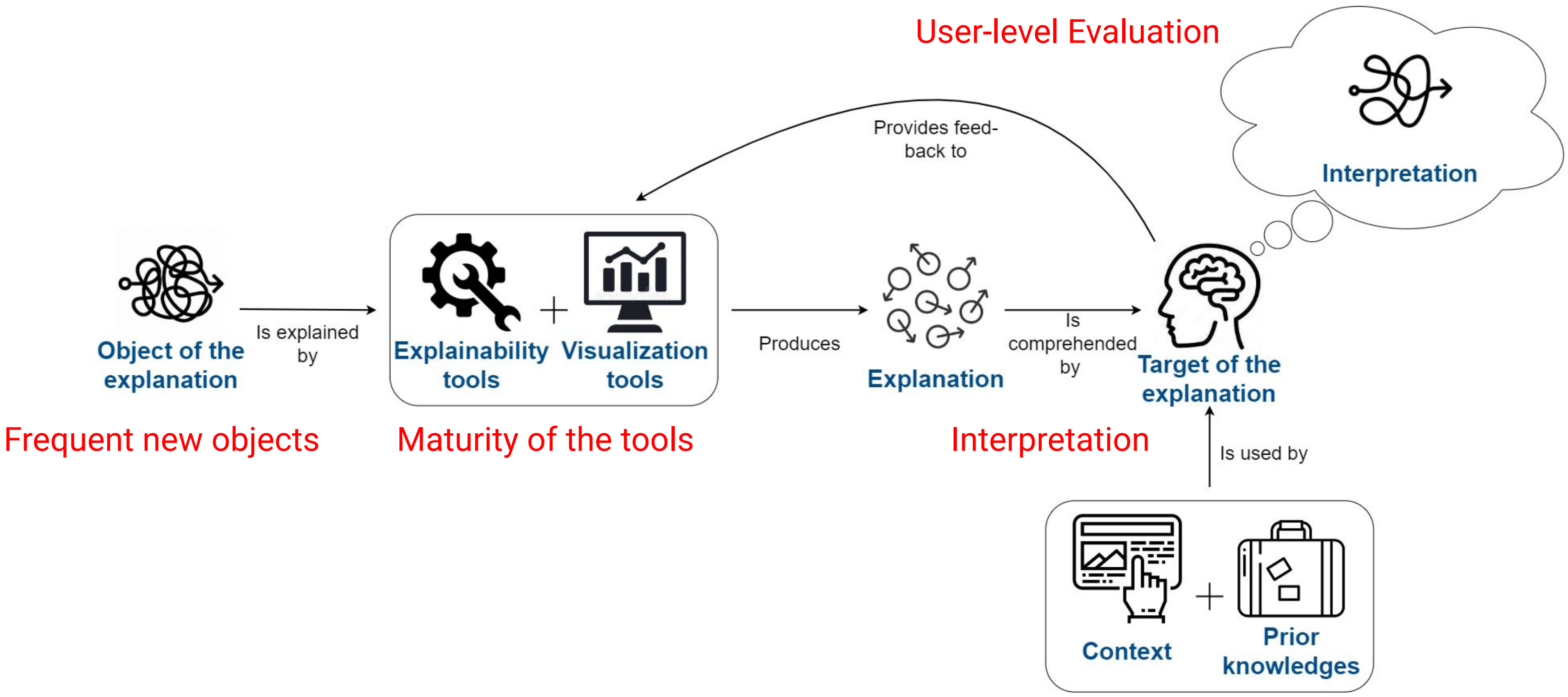
Regulatory entities

Data scientist

Domain expert

Source: Tutorial PFIA 2024

Current explainability challenges



User-level Evaluation

Provides feed-back to

Interpretation

Object of the explanation

Is explained by

Explainability tools + **Visualization tools**

Produces

Explanation

Is comprehended by

Target of the explanation

Is used by

Context + **Prior knowledges**

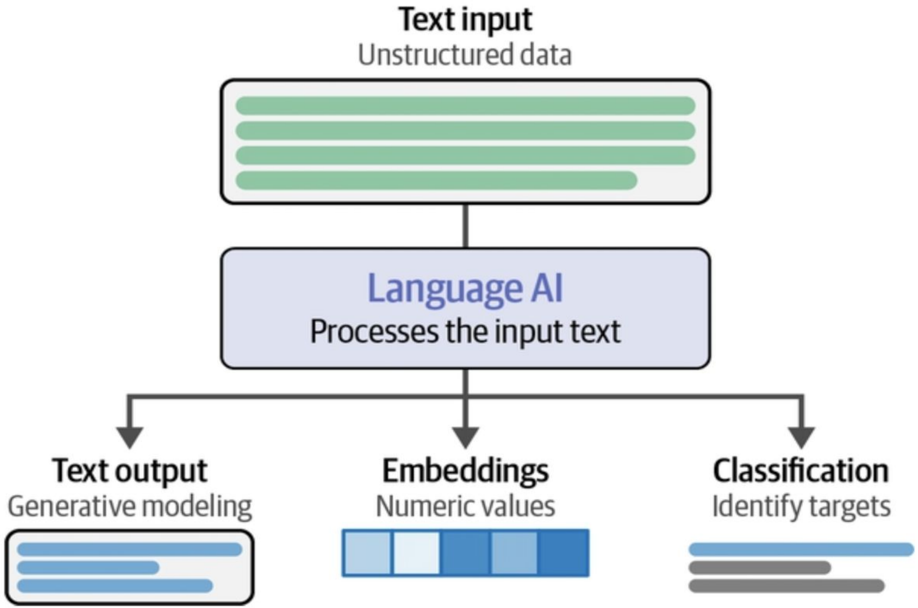
Frequent new objects

Maturity of the tools

Interpretation

XAI for NLP

Language AI : the main tasks



Main principles :

- Text representation : tokenization and embedding.
  - High-dimensionality
- Recent models are mainly based on the transformer architecture
  - Importance of the attention mechanism
- Many tasks are generative
  - Translation, Question-Answering, Summarization, ….

Alammar and Grotendorst - Book 2024 - Hands-On Large Language Models

Challenges: Cognitive load

I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character.

Martin Luther King

Challenges : Tokenization and Embeddings

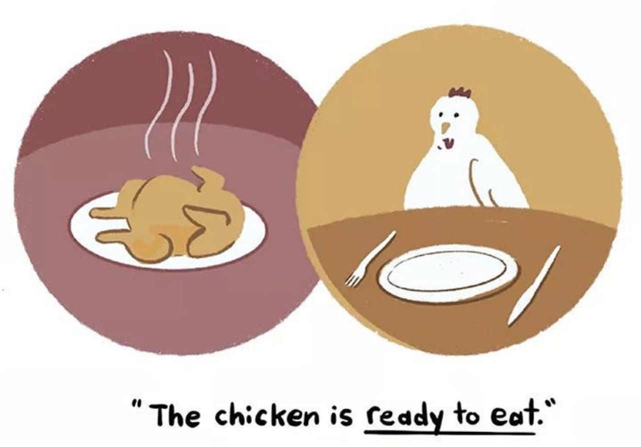
Input text

Artificial Intelligence

Input tokens	Input ids	Input embeddings
CLS	101	101: [0.1, 0.5, 0.9, 0.3, 0.8]
Art	4362	4362: [0.0, 0.2, 0.9, 0.4, 0.7]
##ific	67103	67103: [0.6, 0.2, 0.5, 0.7, 0.8]
##ial	796	796: [0.5, 0.5, 0.4, 0.1, 0.3]
Intel	15875	15875: [0.8, 0.6, 0.9, 0.3, 0.3]
##lig	39771	39771: [0.1, 0.1, 0.8, 0.5, 0.5]
##ence	27646	27646: [0.7, 0.2, 0.2, 0.3, 0.3]
EOS	102	102: [0.2, 0.8, 0.8, 0.5, 0.1]

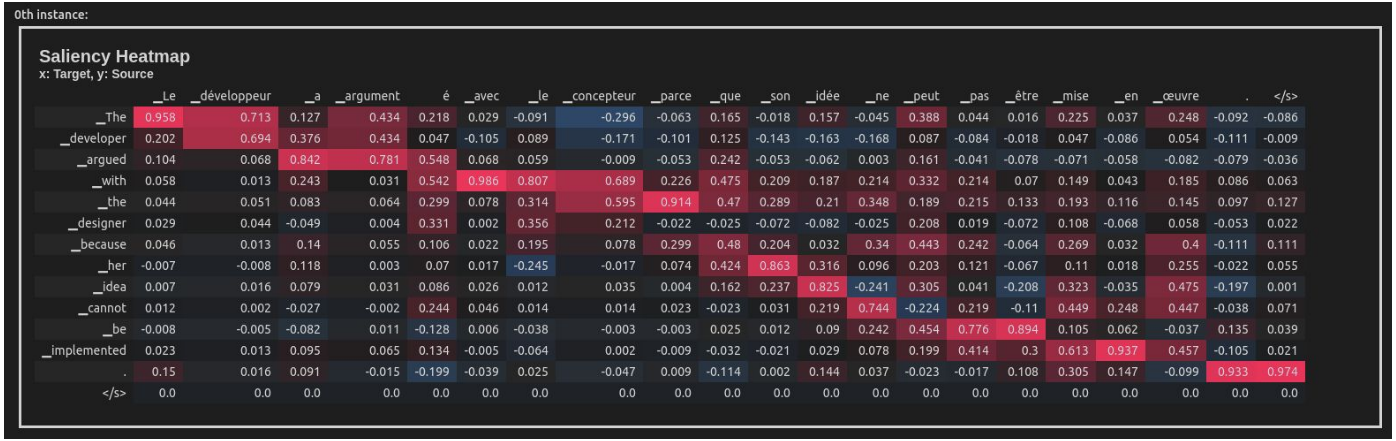
Challenge: Context and Ambiguity

Words and sentences often have **multiple meanings,** and understanding the correct interpretation depends heavily on **context.**



Challenges: Generation

- Generations are auto-regressive.
  - Local explanation only explain one token prediction.
  - To explain text generation there are as **many explanations as generated tokens**.



0th instance:

Saliency Heatmap
x: Target, y: Source

	_Le	_développeur	_a	_argument	é	_avec	_le	_concepteur	_parce	_que	_son	_idée	_ne	_peut	_pas	_être	_mise	_en	_œuvre	.	</s>
_The	0.958	0.713	0.127	0.434	0.218	0.029	-0.091	-0.296	-0.063	0.165	-0.018	0.157	-0.045	0.388	0.044	0.016	0.225	0.037	0.248	-0.092	-0.086
_developer	0.202	0.694	0.376	0.434	0.047	-0.105	0.089	-0.171	-0.101	0.125	-0.143	-0.163	-0.168	0.087	-0.084	-0.018	0.047	-0.086	0.054	-0.111	-0.009
_argued	0.104	0.068	0.842	0.781	0.548	0.068	0.059	-0.009	-0.053	0.242	-0.053	-0.062	0.003	0.161	-0.041	-0.078	-0.071	-0.058	-0.082	-0.079	-0.036
_with	0.058	0.013	0.243	0.031	0.542	0.986	0.807	0.689	0.226	0.475	0.209	0.187	0.214	0.332	0.214	0.07	0.149	0.043	0.185	0.086	0.063
_the	0.044	0.051	0.083	0.064	0.299	0.078	0.314	0.595	0.914	0.47	0.289	0.21	0.348	0.189	0.215	0.133	0.193	0.116	0.145	0.097	0.127
_designer	0.029	0.044	-0.049	0.004	0.331	0.002	0.356	0.212	-0.022	-0.025	-0.072	-0.082	-0.025	0.208	0.019	-0.072	0.108	-0.068	0.058	-0.053	0.022
_because	0.046	0.013	0.14	0.055	0.106	0.022	0.195	0.078	0.299	0.48	0.204	0.032	0.34	0.443	0.242	-0.064	0.269	0.032	0.4	-0.111	0.111
_her	-0.007	-0.008	0.118	0.003	0.07	0.017	-0.245	-0.017	0.074	0.424	0.863	0.316	0.096	0.203	0.121	-0.067	0.11	0.018	0.255	-0.022	0.055
_idea	0.007	0.016	0.079	0.031	0.086	0.026	0.012	0.035	0.004	0.162	0.237	0.825	-0.241	0.305	0.041	-0.208	0.323	-0.035	0.475	-0.197	0.001
_cannot	0.012	0.002	-0.027	-0.002	0.244	0.046	0.014	0.014	0.023	-0.023	0.031	0.219	0.744	-0.224	0.219	-0.11	0.449	0.248	0.447	-0.038	0.071
_be	-0.008	-0.005	-0.082	0.011	-0.128	0.006	-0.038	-0.003	-0.003	0.025	0.012	0.09	0.242	0.454	0.776	0.894	0.105	0.062	-0.037	0.135	0.039
_implemented	0.023	0.013	0.095	0.065	0.134	-0.005	-0.064	0.002	-0.009	-0.032	-0.021	0.029	0.078	0.199	0.414	0.3	0.613	0.937	0.457	-0.105	0.021
.	0.15	0.016	0.091	-0.015	-0.199	-0.039	0.025	-0.047	0.009	-0.114	0.002	0.144	0.037	-0.023	-0.017	0.108	0.305	0.147	-0.099	0.933	0.974
</s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Sarti et al. - ACL 2023 - Inseq: An Interpretability Toolkit for Sequence Generation Models

Challenges: Generating Auto-explanations

- To be useful explanations should be both **plausible** and **faithful**. Jacovi et Golberg 2019
    - Plausible: Probable and convincing to humans
    - Faithful: Represent the model mechanisms

Auto-explanations are highly plausible (they are trained for it). But nothing proves their faithfulness.

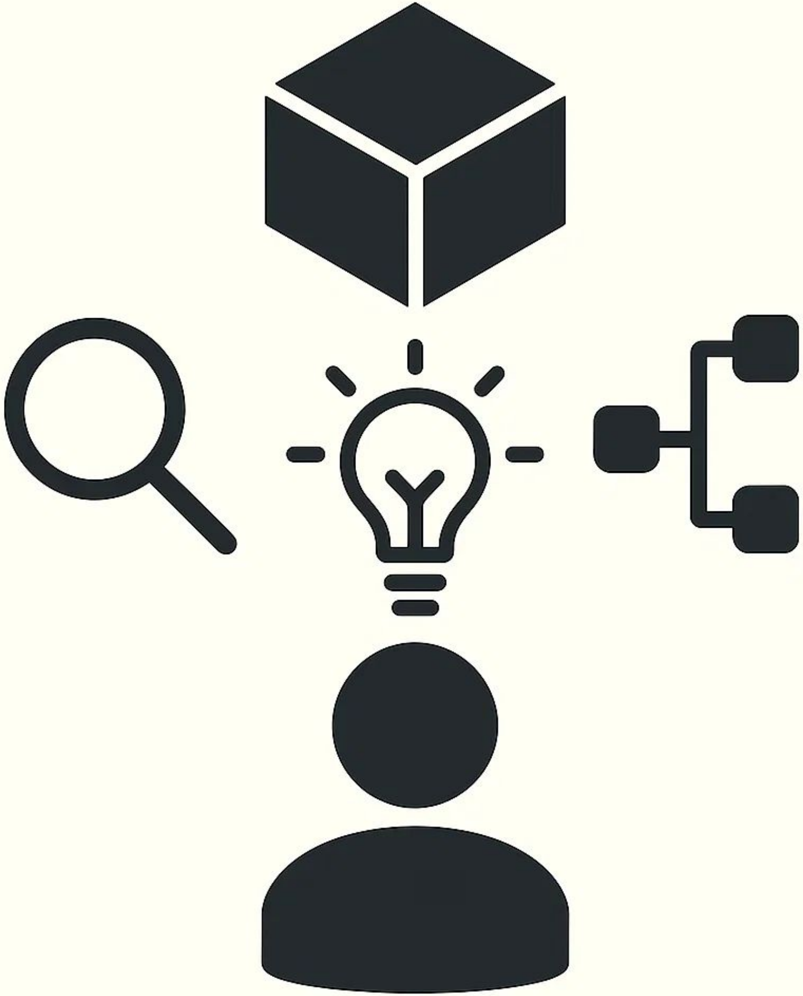
Barez et al. - 2025 - Chain-of-Thought Is Not Explainability

Challenges: LLM sizes

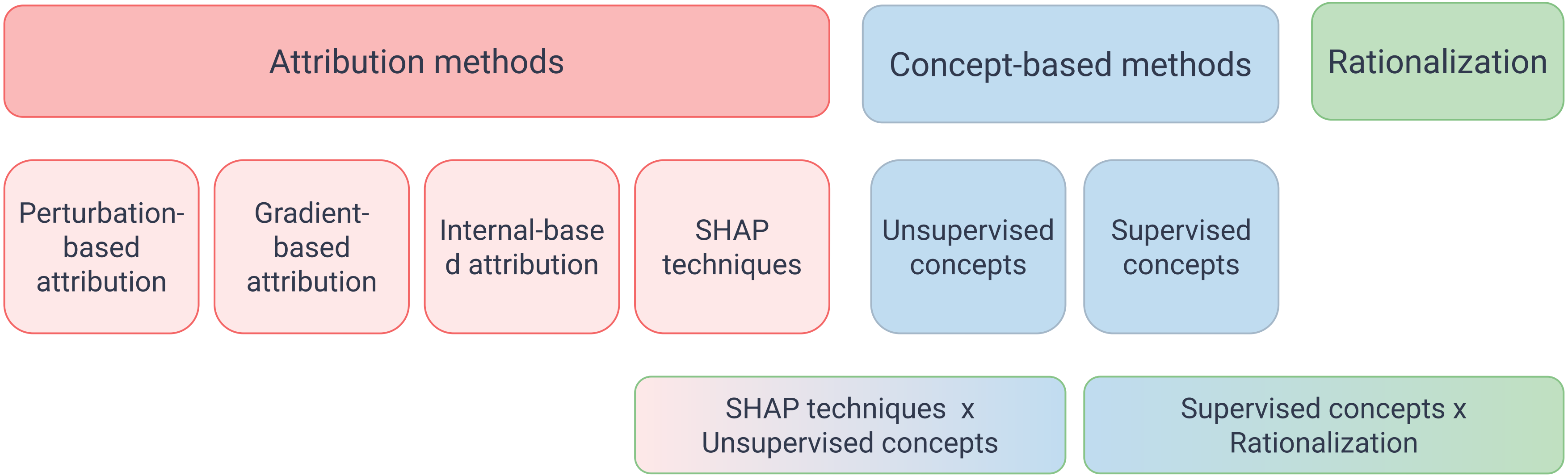
- Explainability often requires hundreds of inferences
- Most generation language model (> 8B) do not fit in “consumer” size GPUs
- Largest models are “sharded" across GPUs which is an engineering challenge

User-centered Explanations

- Attributions methods
- Concepts-based methods
- Evaluation and metrics



User-centered methods: an overview



Attribution methods

Concept-based methods

Rationalization

Perturbation-based attribution

Gradient-based attribution

Internal-based attribution

SHAP techniques

Unsupervised concepts

Supervised concepts

SHAP techniques x Unsupervised concepts

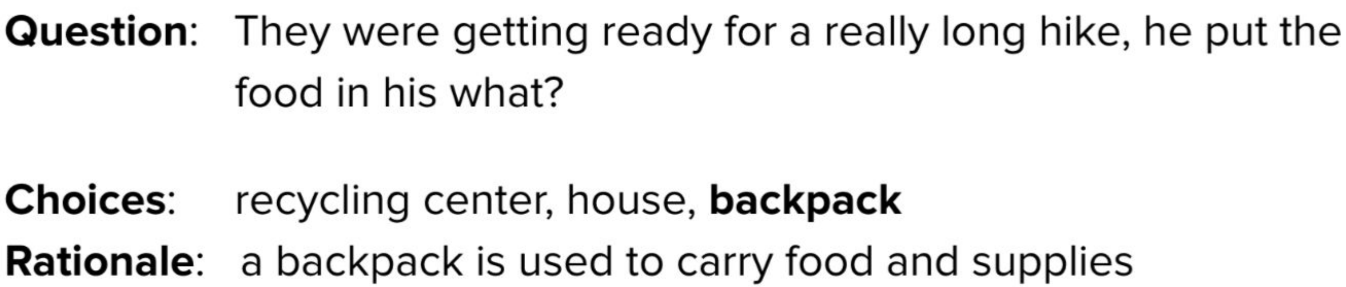
Supervised concepts x Rationalization

Adapted from Fanny Jourdan’s slides

Rationalization: a quick note (not the focus)

Rationalization provides explanations in natural language to justify a model's prediction

- Extractive rationalization ⇝ important features or sentences from the input data
- Abstractive rationalization ⇝ novel sentences with new words



Question: They were getting ready for a really long hike, he put the food in his what?

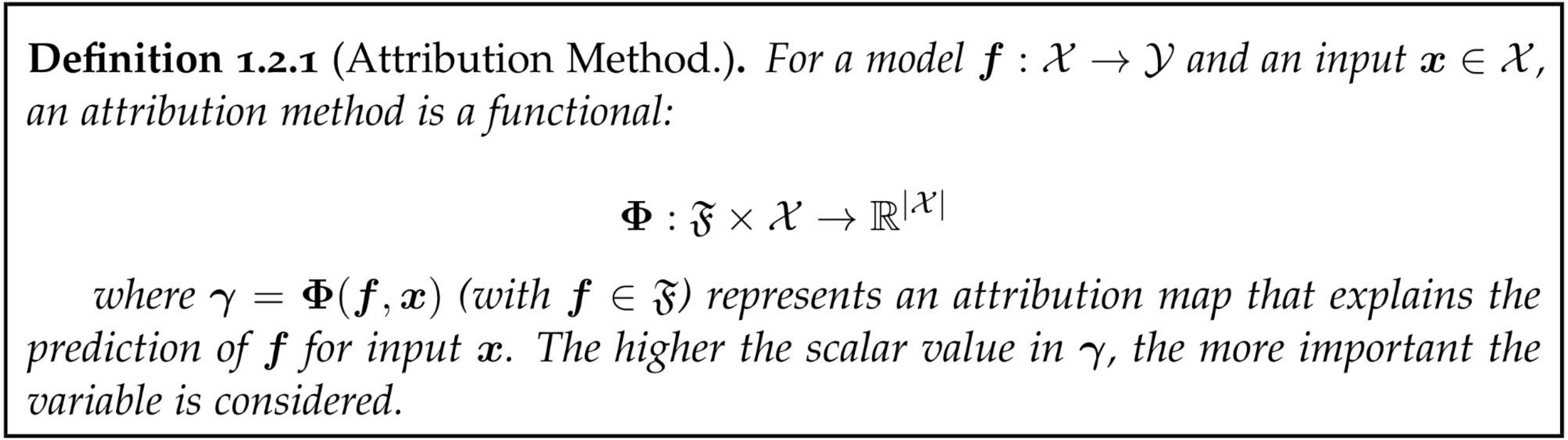
Choices: recycling center, house, **backpack**

Rationale: a backpack is used to carry food and supplies

(GURRRAPU et al., 2023) : https://arxiv.org/abs/2301.08912

Attribution methods

Attribution methods



Definition 1.2.1 (Attribution Method.). *For a model* $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ *and an input* $\boldsymbol{x} \in \mathcal{X}$*, an attribution method is a functional:*

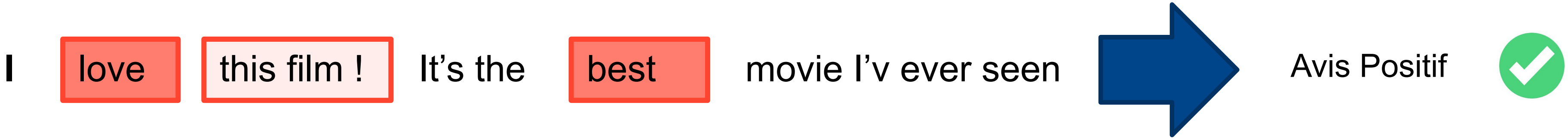
$$\boldsymbol{\Phi} : \mathfrak{F} \times \mathcal{X} \to \mathbb{R}^{|\mathcal{X}|}$$

where $\boldsymbol{\gamma} = \boldsymbol{\Phi}(\boldsymbol{f}, \boldsymbol{x})$ *(with* $\boldsymbol{f} \in \mathfrak{F}$*) represents an attribution map that explains the prediction of* $\boldsymbol{f}$ *for input* $\boldsymbol{x}$*. The higher the scalar value in* $\boldsymbol{\gamma}$*, the more important the variable is considered.*

Adapted from Thomas Fel's slides

Attribution-based XAI for classification

Classification Task ✅❌



I love this film ! It’s the best movie I’v ever seen

Avis Positif ✓

Heatmap of word importance for the 'positive' class.

Adapted from Fanny Jourdan’s slides

Attribution-based XAI : application

Bias Detection Task ❗🔍

Elle travaille à l’hôpital de Perpignan depuis 3 ans.
Les patients qu’ elle opère la recommande fortement
pour son sérieux et sa gentillesse

➡

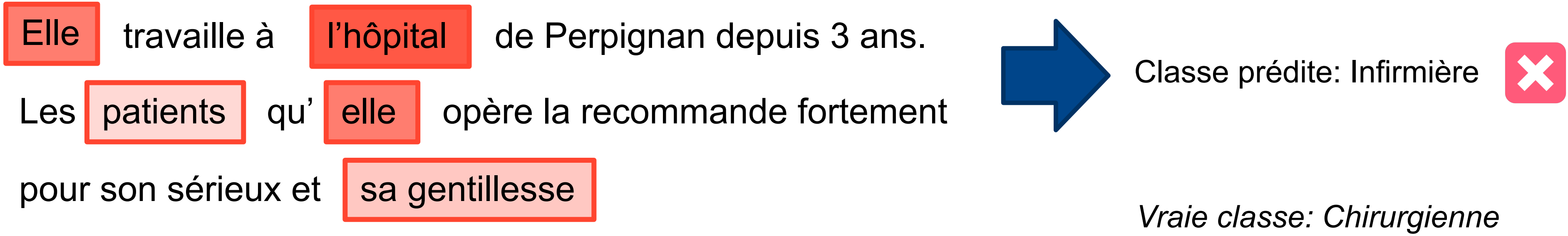
Classe prédite: Infirmière ❌

Vraie classe: Chirurgienne

Adapted from Fanny Jourdan’s slides

Attribution-based XAI : application

Bias Detection Task ❗🔍



Elle travaille à l’hôpital de Perpignan depuis 3 ans.
Les patients qu’ elle opère la recommande fortement
pour son sérieux et sa gentillesse

Classe prédite: Infirmière ❌

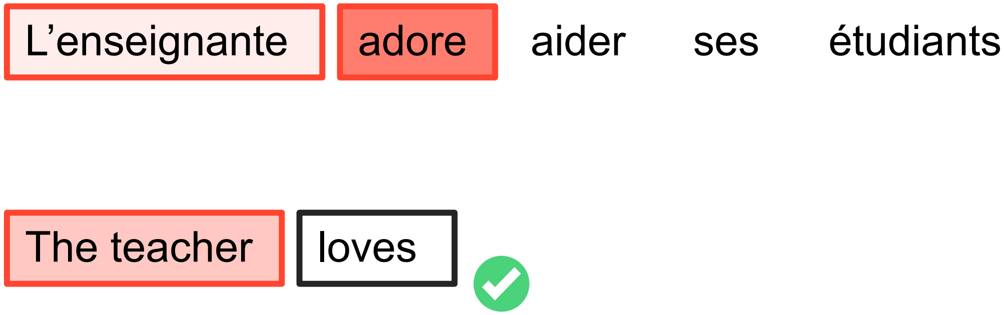
Vraie classe: Chirurgienne

Heatmap de l’importance des mots de l’exemple pour la prédiction de la classe «infirmière»

Adapted from Fanny Jourdan’s slides

Attribution-based XAI for generation

Generation Task 📄➡️📄



L’enseignante adore aider ses étudiants

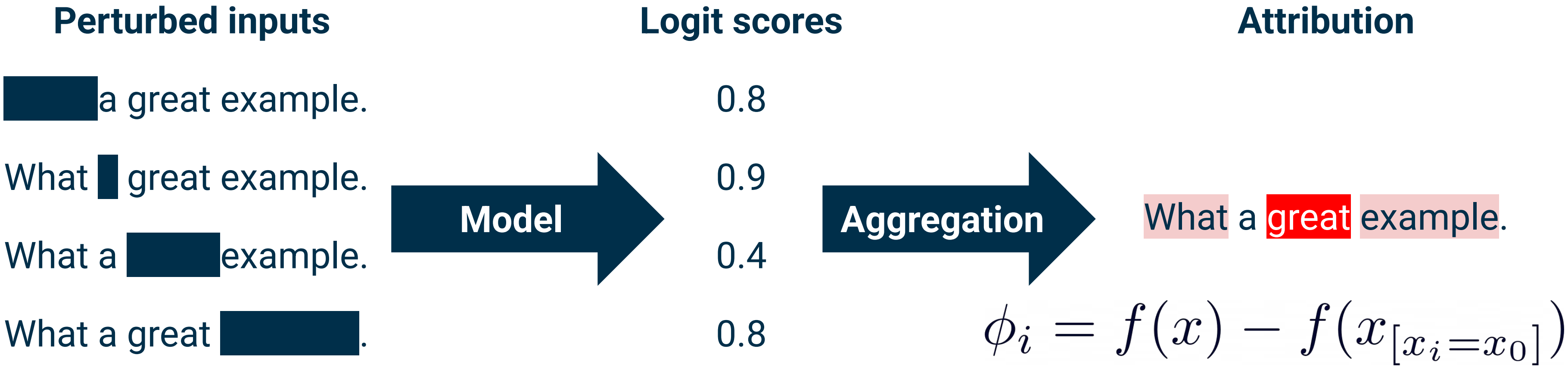
The teacher loves ✓

Heatmap of the importance of preceding words for the generation of the word 'loves'

Adapted from Fanny Jourdan’s slides

Perturbation-based Attribution

Perturbation-based: The principle



Perturbed inputs	Logit scores
[redacted] a great example.	0.8
What [redacted] great example.	0.9
What a [redacted] example.	0.4
What a great [redacted].	0.8

Model

Aggregation

Attribution

What a great example.

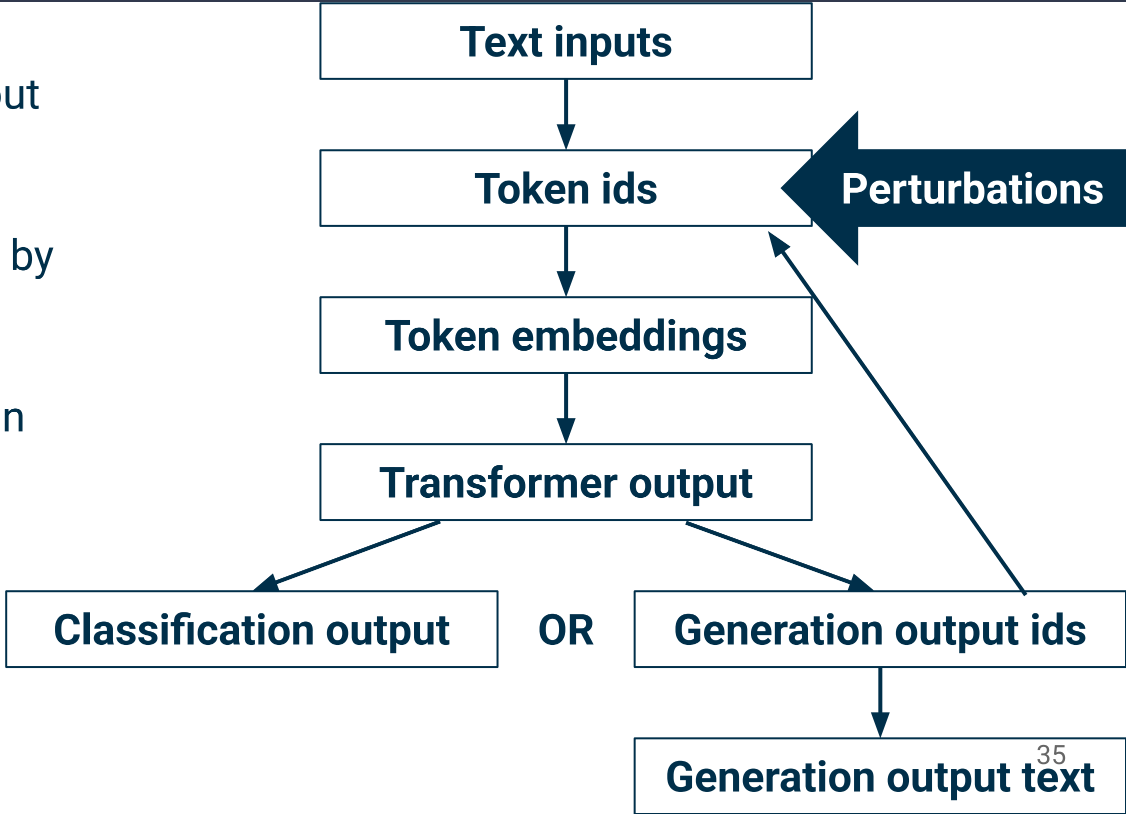
$$\phi_i = f(x) - f(x_{[x_i = x_0]})$$

How do we perturb samples?

How do we aggregate scores?

How do we perturb inputs?

- We cannot perturb the text without knowing how it is tokenized.
- Perturbed token ids are replaced by the [MASK] token id or removed.
- Which token to perturb depend on the attribution method.
- We can vary the granularity (tokens / words / sentences)



Some example of perturbation-based methods

- A method is defined by its perturbation and aggregation.
- Most methods can be adapted from computer vision.

Paper	**Method**	**Perturbation**	**Aggregation**
Zeiler & Fergus - ICCV 2014	Occlusion	One by one	Mapping
Ribeiro et al. - SIGKDD 2016	Lime	Random	Linear regression
Petsiuk et al. - 2018	Rise	Random	Mean
Fel et al. - NeurIPS 2021	Sobol	Sobol sampling	Sobol indices

SHAP techniques

- Given a set of features, we need to find **the marginal contribution** of each feature to the prediction.

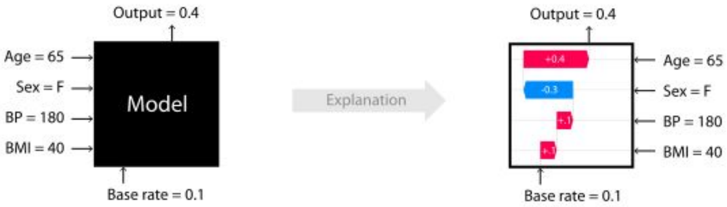
- We therefore need to imagine a basic predicted value, and how each feature forces the prediction to deviate from this basic value.

SHAP-SHapley Additive exPlanation

- A general method that uses Shapley values to measure the importance of each feature (or "player") in an input, in order to explain a model's predictions.

- SHAP is not just an extension or a rebranding of Shapley values. It provides an additive model (that links Shapley Values to LIME) and provides efficient ways to estimate the Shapley Values

SHAP



Lundberg et al. - NeurIPS 2017 - Shap Library

SHAP-SHapley Additive exPlanation

- **Text classification**: easy case

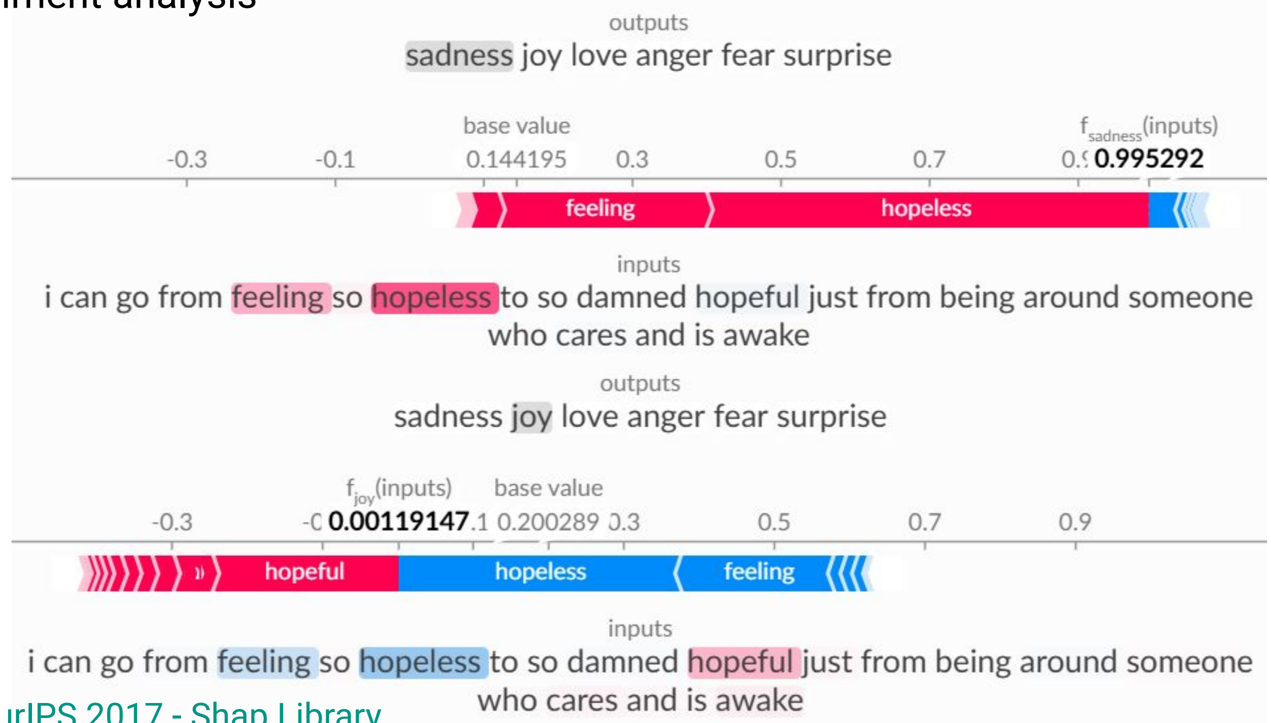
  Generally we have a score (sentiment analysis) or a distribution (text categorization), so we can use SHAP as for regression

- **Text generation**: still challenging

  Contrary to tabular data, we do not have a dataset, but only a prompt. So the expected value cannot be used and must be replaced (different strategies)

SHAP-SHapley Additive exPlanation

Example: sentiment analysis



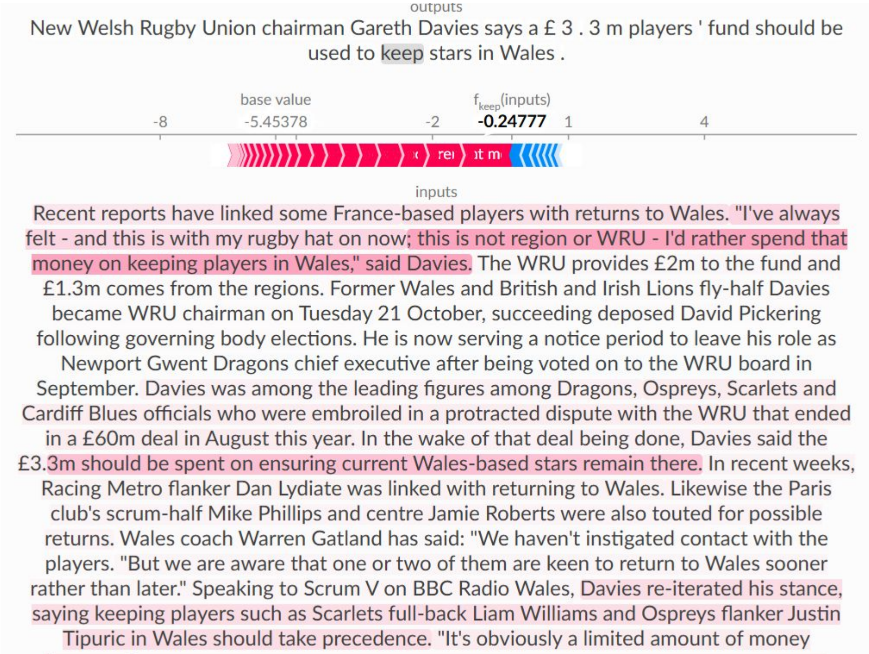
i can go from feeling so hopeless to so damned hopeful just from being around someone who cares and is awake

i can go from feeling so hopeless to so damned hopeful just from being around someone who cares and is awake

Lundberg et al. - NeurIPS 2017 - Shap Library

SHAP-SHapley Additive exPlanation

Example: summarization



outputs

New Welsh Rugby Union chairman Gareth Davies says a £ 3 . 3 m players ' fund should be used to keep stars in Wales .

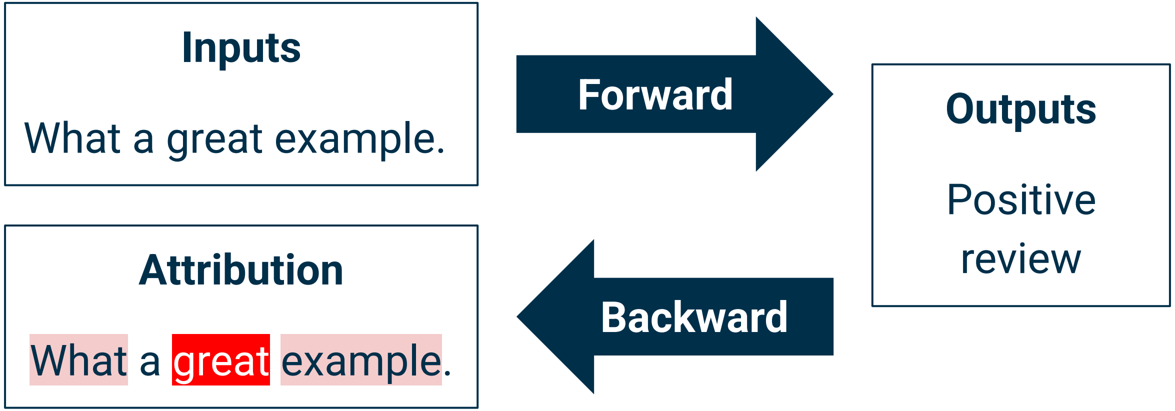
inputs

Recent reports have linked some France-based players with returns to Wales. "I've always felt - and this is with my rugby hat on now; this is not region or WRU - I'd rather spend that money on keeping players in Wales," said Davies. The WRU provides £2m to the fund and £1.3m comes from the regions. Former Wales and British and Irish Lions fly-half Davies became WRU chairman on Tuesday 21 October, succeeding deposed David Pickering following governing body elections. He is now serving a notice period to leave his role as Newport Gwent Dragons chief executive after being voted on to the WRU board in September. Davies was among the leading figures among Dragons, Ospreys, Scarlets and Cardiff Blues officials who were embroiled in a protracted dispute with the WRU that ended in a £60m deal in August this year. In the wake of that deal being done, Davies said the £3.3m should be spent on ensuring current Wales-based stars remain there. In recent weeks, Racing Metro flanker Dan Lydiate was linked with returning to Wales. Likewise the Paris club's scrum-half Mike Phillips and centre Jamie Roberts were also touted for possible returns. Wales coach Warren Gatland has said: "We haven't instigated contact with the players. "But we are aware that one or two of them are keen to return to Wales sooner rather than later." Speaking to Scrum V on BBC Radio Wales, Davies re-iterated his stance, saying keeping players such as Scarlets full-back Liam Williams and Ospreys flanker Justin Tipuric in Wales should take precedence. "It's obviously a limited amount of money

Lundberg et al. - NeurIPS 2017 - Shap Library

Gradient-based Attribution

Gradient-based: The principle



Inputs

What a great example.

Forward

Outputs

Positive review

Backward

Attribution

What a great example.

$$\Phi = \nabla f(x) \implies \phi_i = \frac{\partial f(x)}{\partial x_i}$$

Where do we compute the gradient?

How do we aggregate gradients?

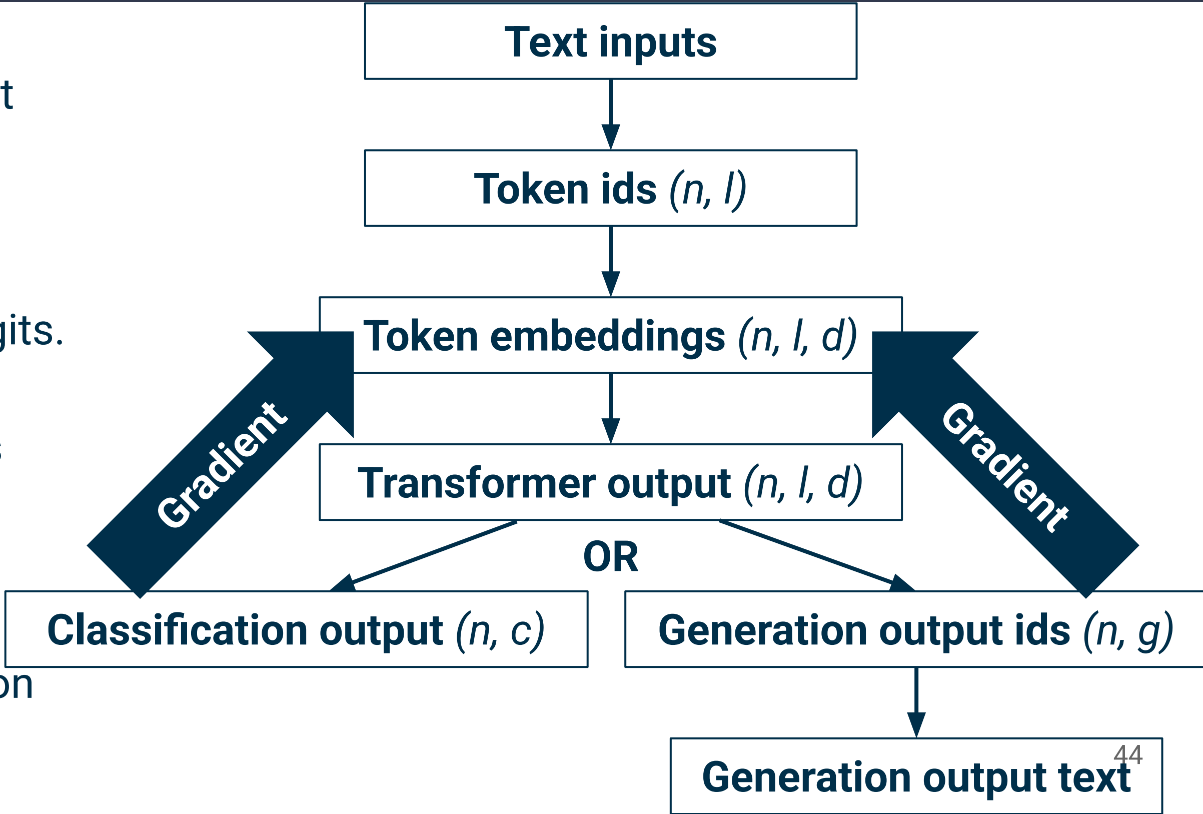
Where do we compute the gradient?

- The embedding operation is not differentiable.

- We derive from the token embeddings to the selected logits.

- Therefore the gradient shape is
    - Classification *(n, l, d)*
    - Generation *(n, g, l + g, d)*

- We aggregate on the *l* dimension via mean of absolute values.


Text inputs
Token ids *(n, l)*
Token embeddings *(n, l, d)*
Transformer output *(n, l, d)*
OR
Classification output *(n, c)*
Generation output ids *(n, g)*
Generation output text

Some example of gradient-based methods

- Similarly to perturbation-based methods, a gradient-based method is defined by its perturbation and aggregation.

- Most methods can be adapted from computer vision.

Paper	**Method**	**Perturbation**	**Aggregation**
Simonyan et al. - ICLR 2014	Saliency	None	None
Sundararajan et al. - ICML 2017	Integrated Gradient	Linear interpolation	Mean
Smilkov et al. - 2017	SmoothGrad	Gaussian noise	Mean
Adebayo et al. - NeurIPS 2018	VarGrad	Gaussian noise	Variance

Concept-based methods

> “Showing where a network is looking does not tell us what the network is seeing in a given input”

Rudin, Cynthia - Nature Machine Learning 2019 - "Stop explaining black box machine learning models …”

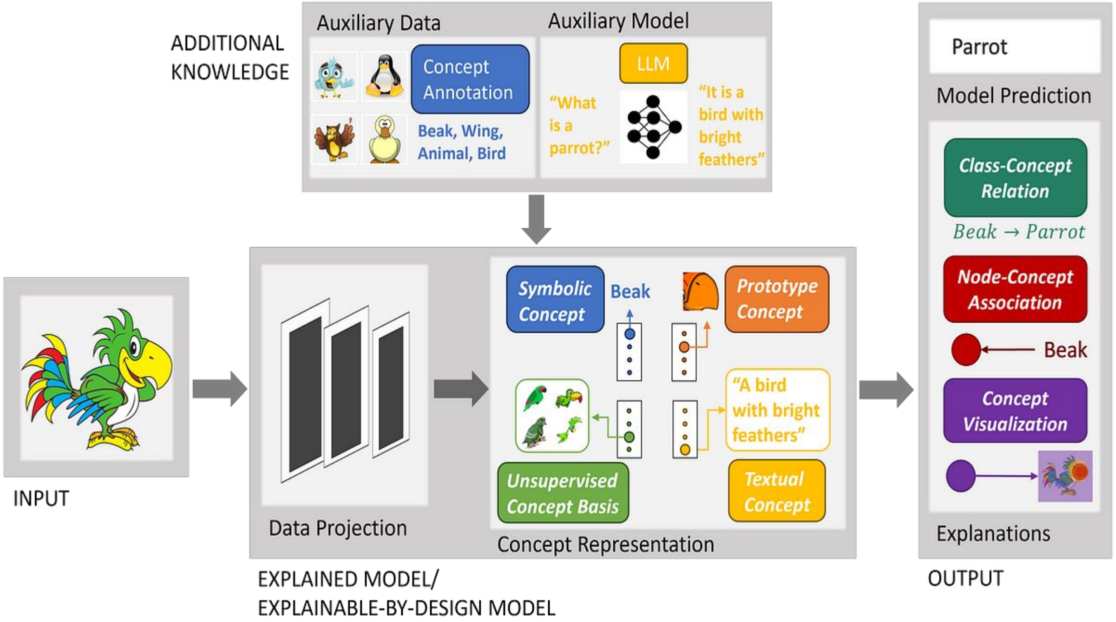
What is a concept?

“A concept is an abstraction of
common elements between samples“

Poché et al. - xAI 2023 - Natural Example-based Explainability: a Survey

A drawing field

2018	CAV & TACV
2019	ProtoPNet, ACE
2020	CBM, ProtoTree
2021	ICE, ICB,
2022	CRAFT, CAR
2023	Cockatiel, Holistic, Mech. Inter.
2024	SAEs, Anthropic, Deep Mind...



[Tutorial PFIA 2024](#)

[Poeta et al. - 2023 - Concept-based Explainable Artificial Intelligence: A Survey](#)

Concept-based motivations

- They better resemble the way humans reason and explain (Kim et al., 2023)
- They are more stable (similar inputs yield similar explanations) under perturbation (Melis et al., 2018)
- They better detect model biases (Rishabh et al, 2022)

From Ciravegna Talk, 2024:
SSY. Kim et al. - CHI 2023 - "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction.
Melis and Jaakkola - NeurIPS 2018 - Towards robust interpretability with Self-Explaining Neural Networks.
Jain et al. - EMNLP 2022 - Extending Logic Explained Networks to Text.

Concept-based: classification task

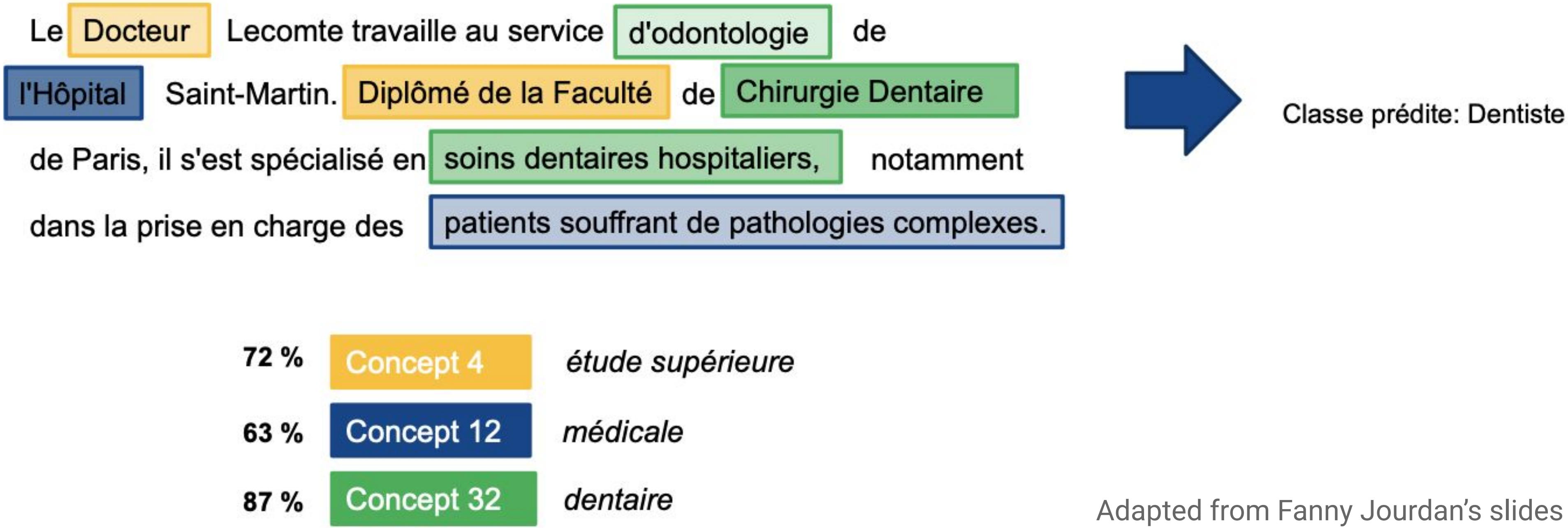
Le Docteur Lecomte travaille au service d'odontologie de l'Hôpital Saint-Martin. Diplômé de la Faculté de Chirurgie Dentaire de Paris, il s'est spécialisé en soins dentaires hospitaliers, notamment dans la prise en charge des patients souffrant de pathologies complexes.

→ Classe prédite: Dentiste

72 %	Concept 4	*étude supérieure*
63 %	Concept 12	*médicale*
87 %	Concept 32	*dentaire*

Adapted from Fanny Jourdan’s slides

Concept-based: classification task



Le Docteur Lecomte travaille au service d'odontologie de l'Hôpital Saint-Martin. Diplômé de la Faculté de Chirurgie Dentaire de Paris, il s'est spécialisé en soins dentaires hospitaliers, notamment dans la prise en charge des patients souffrant de pathologies complexes.

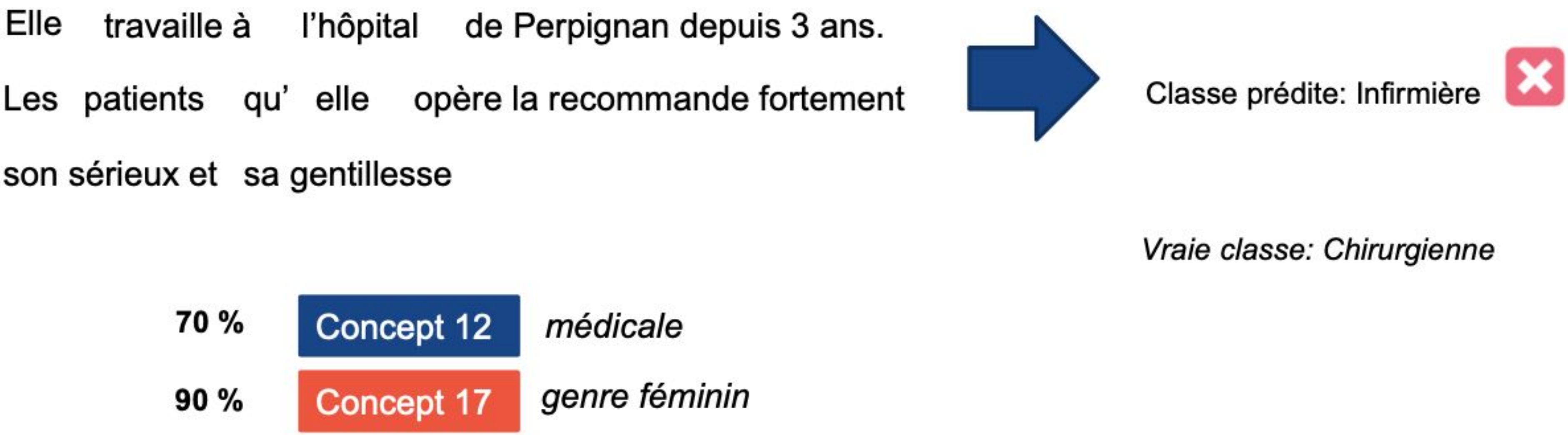
Classe prédite: Dentiste

72 %	🟨 Concept 4	*étude supérieure*
63 %	🟦 Concept 12	*médicale*
87 %	🟩 Concept 32	*dentaire*

Adapted from Fanny Jourdan’s slides

Concept-based: application

Bias Detection Task ❗🔍



Elle travaille à l’hôpital de Perpignan depuis 3 ans. Les patients qu’ elle opère la recommande fortement son sérieux et sa gentillesse

Classe prédite: Infirmière ✖

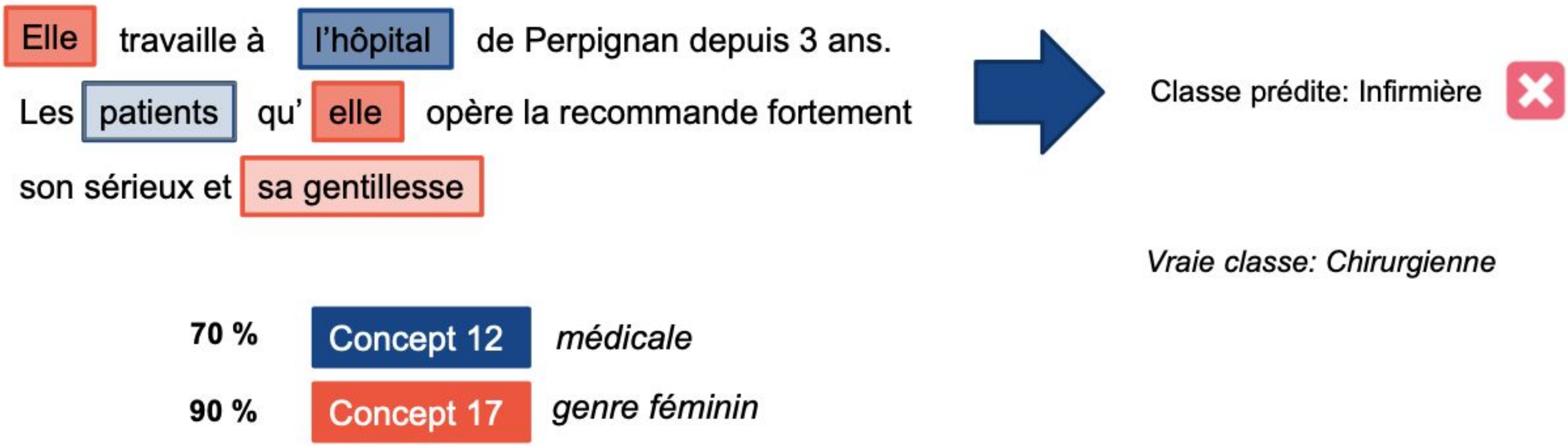
Vraie classe: Chirurgienne

70 %	Concept 12	*médicale*
90 %	Concept 17	*genre féminin*

Adapted from Fanny Jourdan’s slides

Concept-based: application

Bias Detection Task ❗🔍



Elle travaille à l'hôpital de Perpignan depuis 3 ans. Les patients qu' elle opère la recommande fortement son sérieux et sa gentillesse

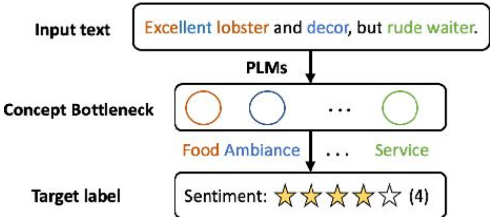
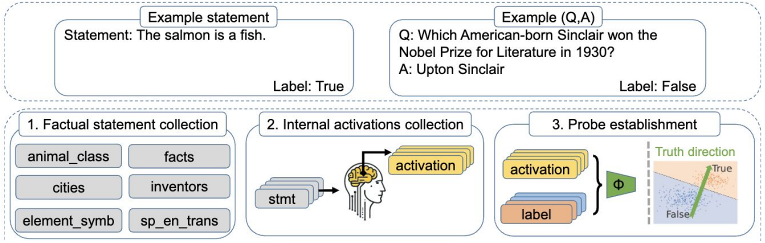
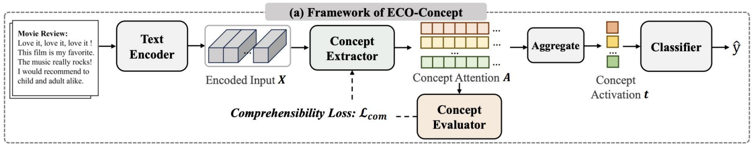
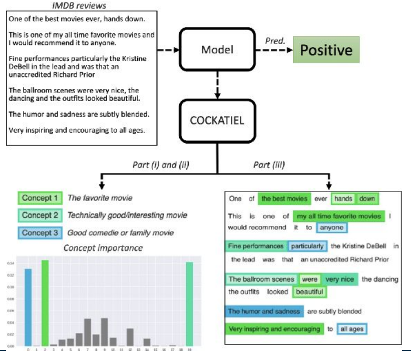
Classe prédite: Infirmière ❌

Vraie classe: Chirurgienne

70 %	Concept 12	*médicale*
90 %	Concept 17	*genre féminin*

Adapted from Fanny Jourdan's slides

Concept-based methods taxonomy

<table>
<tr><th></th><th><b>Ante-hoc</b>
The model is trained to reason from concepts</th><th><b>Post-hoc</b>
Concepts are identified within the trained model</th></tr>
<tr><td><b>Supervised</b>
Requires labelled concepts</td><td>e.g. Tan, Zhen et al. - PACKDDM 2024
 </td><td>e.g. Bao, Yuntai, et al. - 2025
 </td></tr>
<tr><td><b>Unsupervised</b>
Annotation free</td><td>e.g. Sun, Yifan, et al. - ACL 2025
 </td><td>e.g. Jourdan, Fanny, et al. - ACL 2023


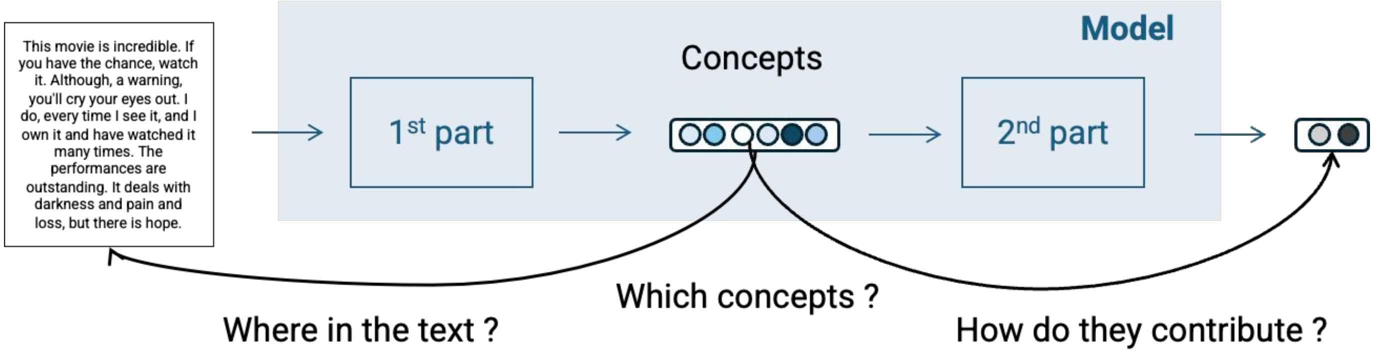
IMDB reviews
One of the best movies ever, hands down.
This is one of my all time favorite movies and I would recommend it to anyone.
Fine performances particularly the Kristine DeBell in the lead and was that an unaccredited Richard Prior
The ballroom scenes were very nice, the dancing and the outfits looked beautiful.
The humor and sadness are subtly blended.
Very inspiring and encouraging to all ages.
Model
Pred.
Positive
COCKATIEL
Part (i) and (ii)
Part (iii)
Concept 1
The favorite movie
Concept 2
Technically good/interesting movie
Concept 3
Good comedie or family movie
Concept importance
One of the best movies ever hands down
This is one of my all time favorite movies I would recommend it to anyone
Fine performances particularly the Kristine DeBell in the lead was that an unaccredited Richard Prior
The ballroom scenes were very nice the dancing the outfits looked beautiful
The humor and sadness are subtly blended
Very inspiring and encouraging to all ages

</td></tr>
</table>

Pros and cons: our analysis!

	Pros	Cons
Supervised	● Relatable to experts	● Labour (annotation of examples) ● Human biases ● Generalization
Unsupervised	● Knowledge discovery ● Scales to new domains	● Challenging to interpret ● Unstable ● Requires large data and compute
Ante-hoc	● For well-defined tasks, combined with supervised approaches	● Never competes in performance ● Not always more interpretable
Post-hoc	● Applicable to any model ● Improves with model performance	● Might not align with expert knowledge

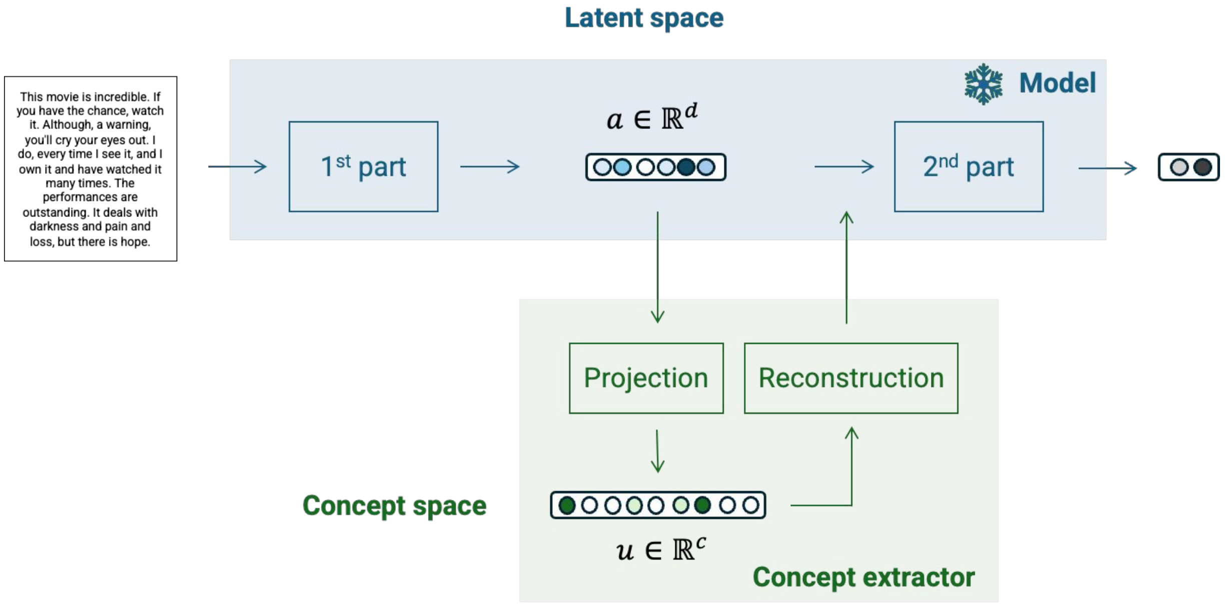
Common points


This movie is incredible. If you have the chance, watch it. Although, a warning, you'll cry your eyes out. I do, every time I see it, and I own it and have watched it many times. The performances are outstanding. It deals with darkness and pain and loss, but there is hope.
Model
Concepts
1st part
2nd part
Which concepts ?
Where in the text ?
How do they contribute ?

- ***Local interpretation*** has three steps:
    - Which concepts are present
    - Where in the text
    - How do they contribute

- ***Global interpretation*** studies the general link between the concepts and the classes

A framework for post-hoc unsupervised C-XAI [PhD. C. Claye]


Latent space
This movie is incredible. If you have the chance, watch it. Although, a warning, you'll cry your eyes out. I do, every time I see it, and I own it and have watched it many times. The performances are outstanding. It deals with darkness and pain and loss, but there is hope.
Model
$a \in \mathbb{R}^d$
1st part
2nd part
Projection
Reconstruction
Concept space
$u \in \mathbb{R}^c$
Concept extractor

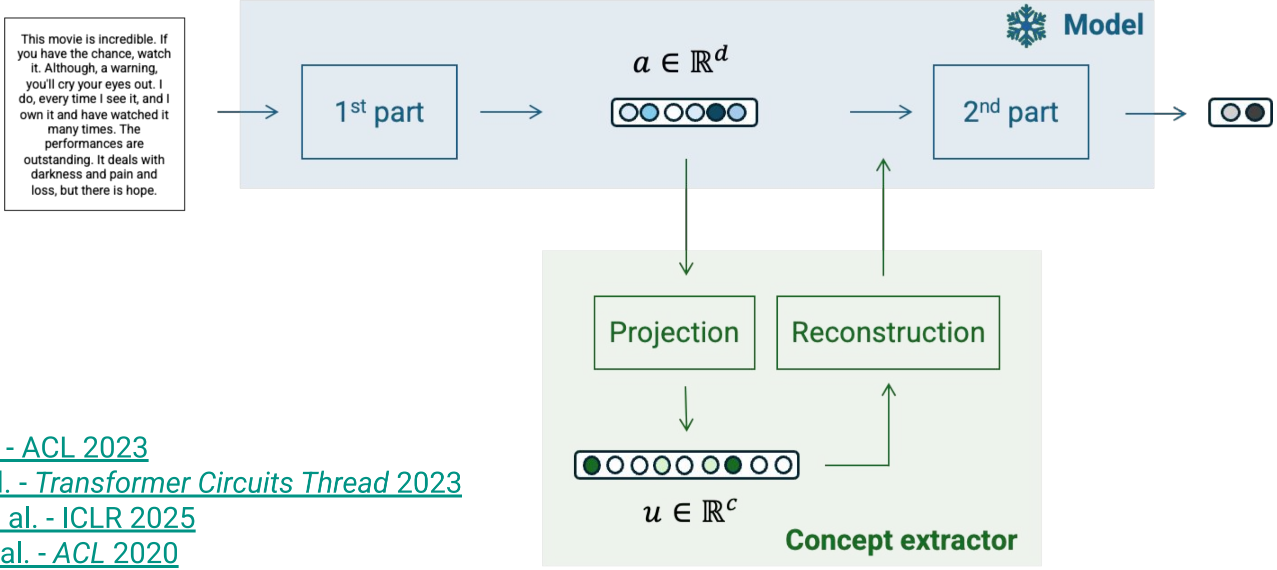
A framework for post-hoc unsupervised C-XAI [PhD. C. Claye]

② **Select training samples**

Select samples (ex : per class) [1]
Parts of samples (ex : sentences)

① **Select a layer**

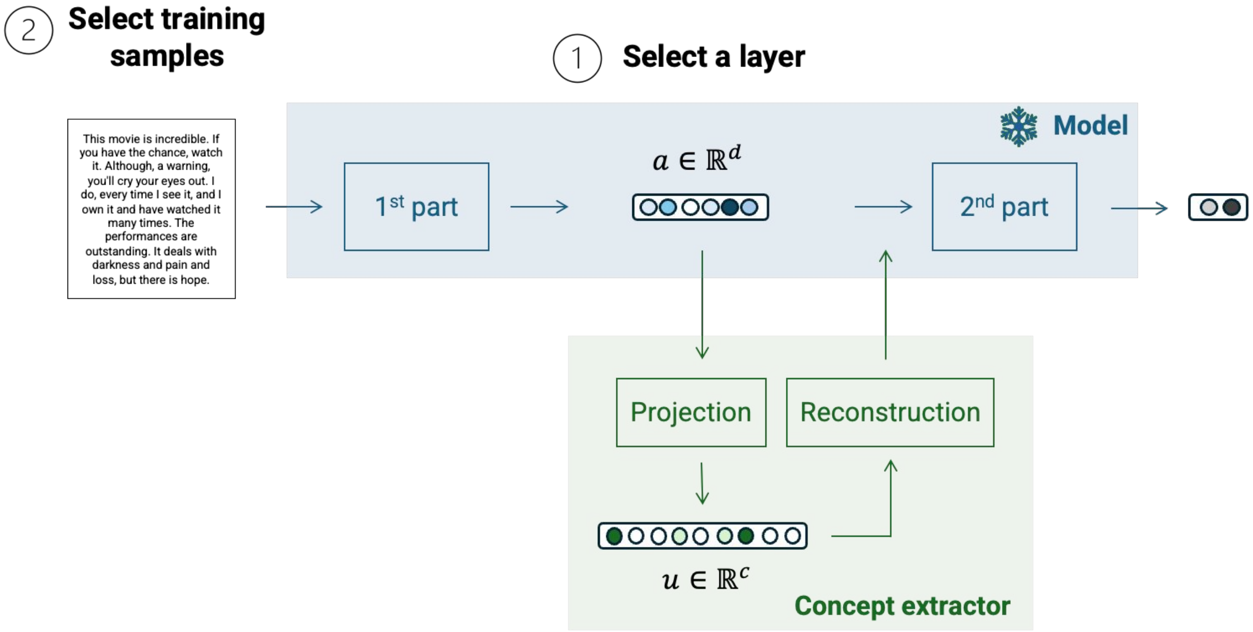
Token embeddings / Text embeddings [2] / [1]
Multi-layers [3]
Low-level / high-level concepts [4]



This movie is incredible. If you have the chance, watch it. Although, a warning, you'll cry your eyes out. I do, every time I see it, and I own it and have watched it many times. The performances are outstanding. It deals with darkness and pain and loss, but there is hope.

[1] Jourdan, Fanny, et al. - ACL 2023
[2] Bricken, Trenton, et al. - *Transformer Circuits Thread* 2023
[3] Balagansky, Nikita, et al. - ICLR 2025
[4] Belinkov, Yonatan, et al. - *ACL* 2020

A framework for post-hoc unsupervised C-XAI [PhD. C. Claye]



② **Select training samples**

① **Select a layer**

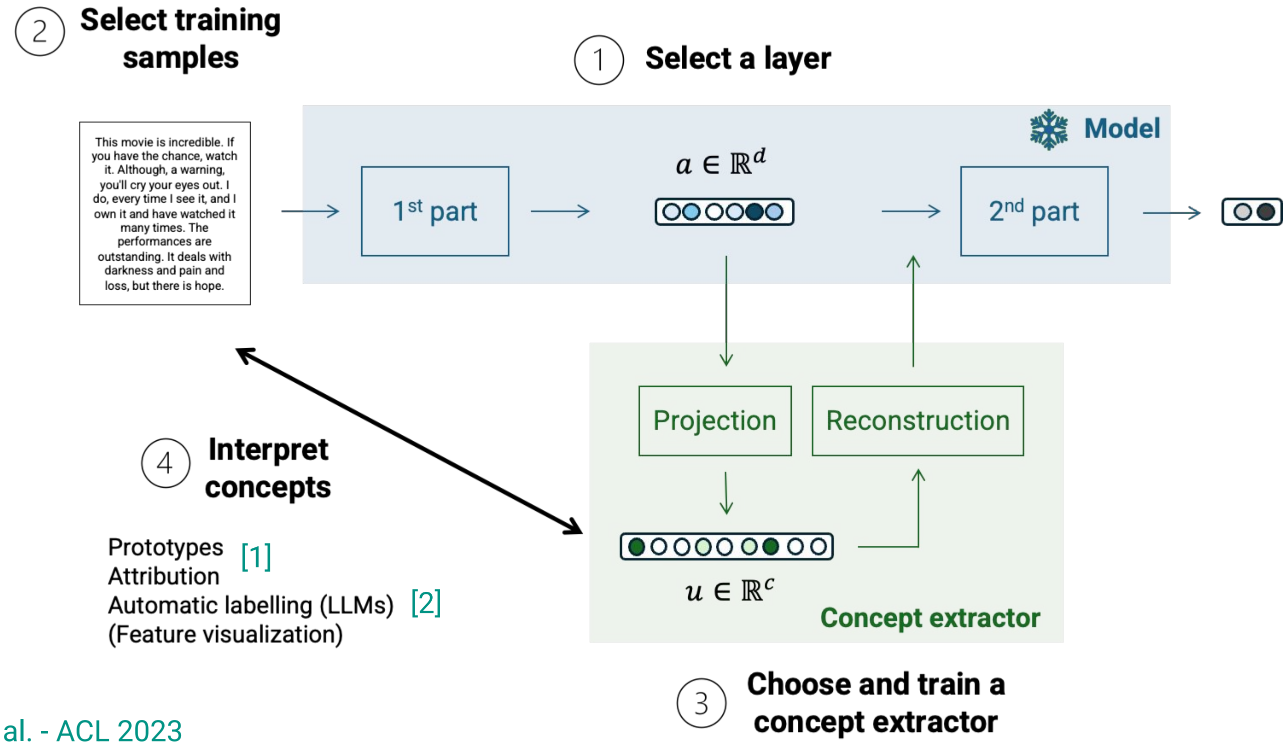
This movie is incredible. If you have the chance, watch it. Although, a warning, you'll cry your eyes out. I do, every time I see it, and I own it and have watched it many times. The performances are outstanding. It deals with darkness and pain and loss, but there is hope.

③ **Choose and train a concept extractor**

Dictionary learning (NMF, SAE, …) NMF [1], SAE [2]
1D / Multi-dimensional concepts
Linear / Non-linear concepts
Hierarchical concepts [3]
Number of concepts : overcomplete / undercomplete [2] / [1]

[1] Jourdan, Fanny, et al. - ACL 2023
[2] Bricken, Trenton, et al. - *Transformer Circuits Thread* 2023
[3] Park, Kiho, et al. - ICML workshop 2025

A framework for post-hoc unsupervised C-XAI [PhD. C. Claye]



② **Select training samples**

This movie is incredible. If you have the chance, watch it. Although, a warning, you'll cry your eyes out. I do, every time I see it, and I own it and have watched it many times. The performances are outstanding. It deals with darkness and pain and loss, but there is hope.

① **Select a layer**

Model

1st part → $a \in \mathbb{R}^d$ → 2nd part

Projection

Reconstruction

$u \in \mathbb{R}^c$

Concept extractor

③ **Choose and train a concept extractor**

④ **Interpret concepts**

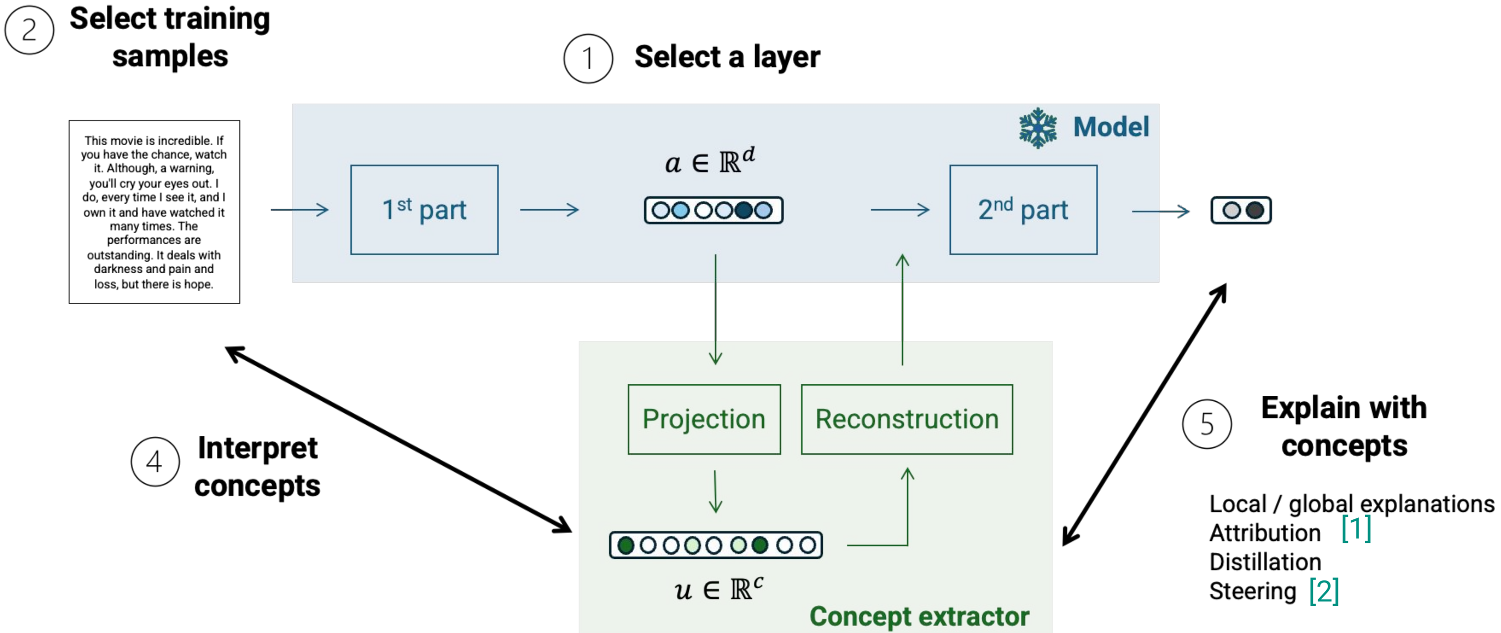
Prototypes [1]
Attribution
Automatic labelling (LLMs) [2]
(Feature visualization)

---

[1] Jourdan, Fanny, et al. - ACL 2023

[2] Paulo, Gonçalo, et al. - 2024

A framework for post-hoc unsupervised C-XAI [PhD. C. Claye]



② **Select training samples**

① **Select a layer**

This movie is incredible. If you have the chance, watch it. Although, a warning, you'll cry your eyes out. I do, every time I see it, and I own it and have watched it many times. The performances are outstanding. It deals with darkness and pain and loss, but there is hope.

④ **Interpret concepts**

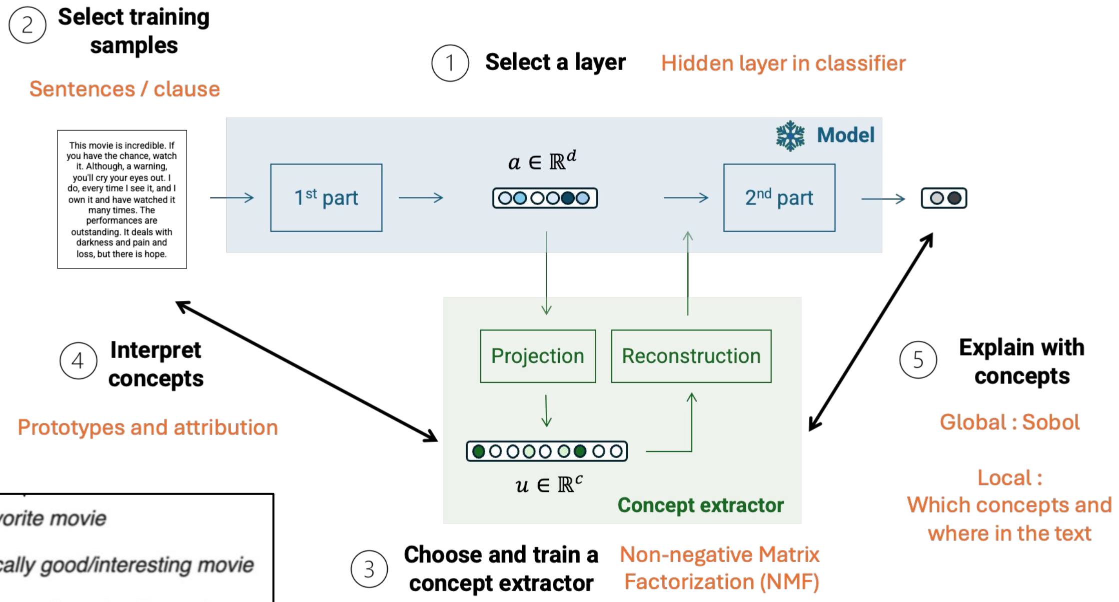
⑤ **Explain with concepts**

Local / global explanations
Attribution [1]
Distillation
Steering [2]

③ **Choose and train a concept extractor**

[1] Jourdan, Fanny, et al. - ACL 2023
[2] Bricken, Trenton, et al. - *Transformer Circuits Thread* 2023
[3] Sousa, João Pedro Bento, et al. - ICLR workshop 2022

COCKATIEL



② **Select training samples**

Sentences / clause

This movie is incredible. If you have the chance, watch it. Although, a warning, you'll cry your eyes out. I do, every time I see it, and I own it and have watched it many times. The performances are outstanding. It deals with darkness and pain and loss, but there is hope.

① **Select a layer**

Hidden layer in classifier

Model

1st part

$a \in \mathbb{R}^d$

2nd part

Projection

Reconstruction

$u \in \mathbb{R}^c$

Concept extractor

④ **Interpret concepts**

Prototypes and attribution

③ **Choose and train a concept extractor**

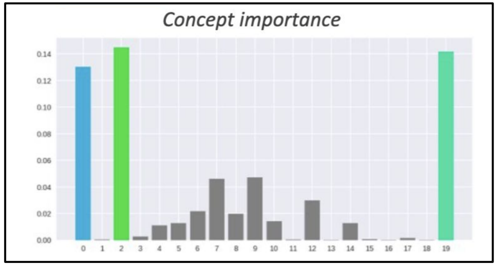
Non-negative Matrix Factorization (NMF)

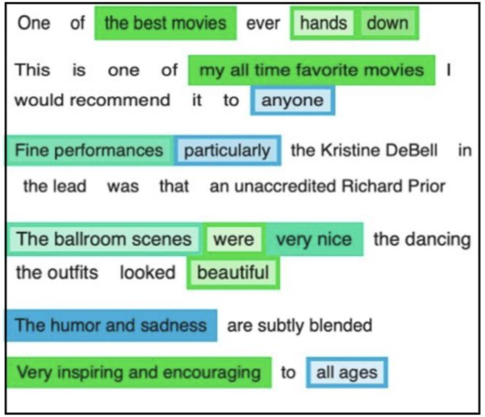
⑤ **Explain with concepts**

Global : Sobol

Local :
Which concepts and where in the text

🟩 Concept 1	*The favorite movie*
🟩 Concept 2	*Technically good/interesting movie*
🟦 Concept 3	*Good comedie or family movie*





One of the best movies ever hands down

This is one of my all time favorite movies I would recommend it to anyone

Fine performances particularly the Kristine DeBell in the lead was that an unaccredited Richard Prior

The ballroom scenes were very nice the dancing the outfits looked beautiful

The humor and sadness are subtly blended

Very inspiring and encouraging to all ages

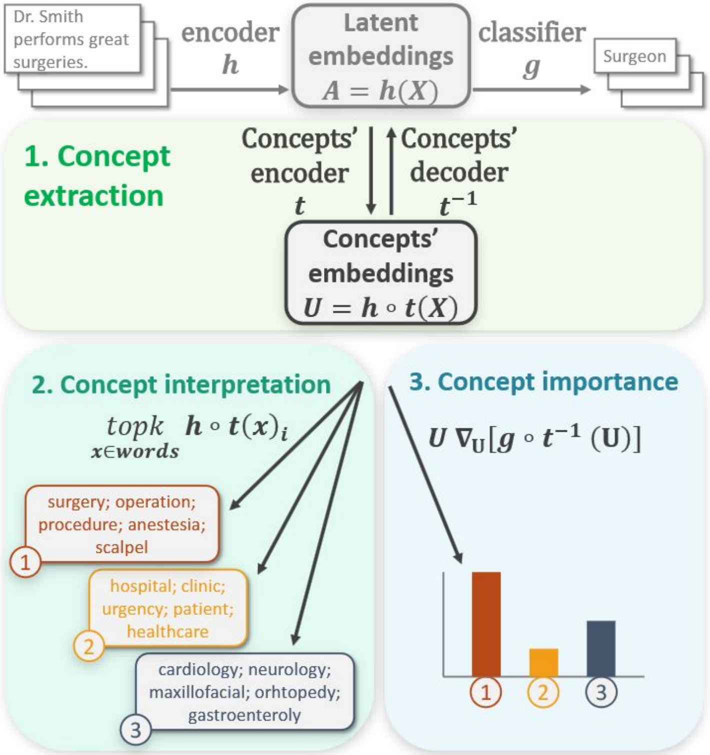
Jourdan et al. - ACL 2023 - COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP tasks.

Evaluation and metrics

Evaluation and metrics

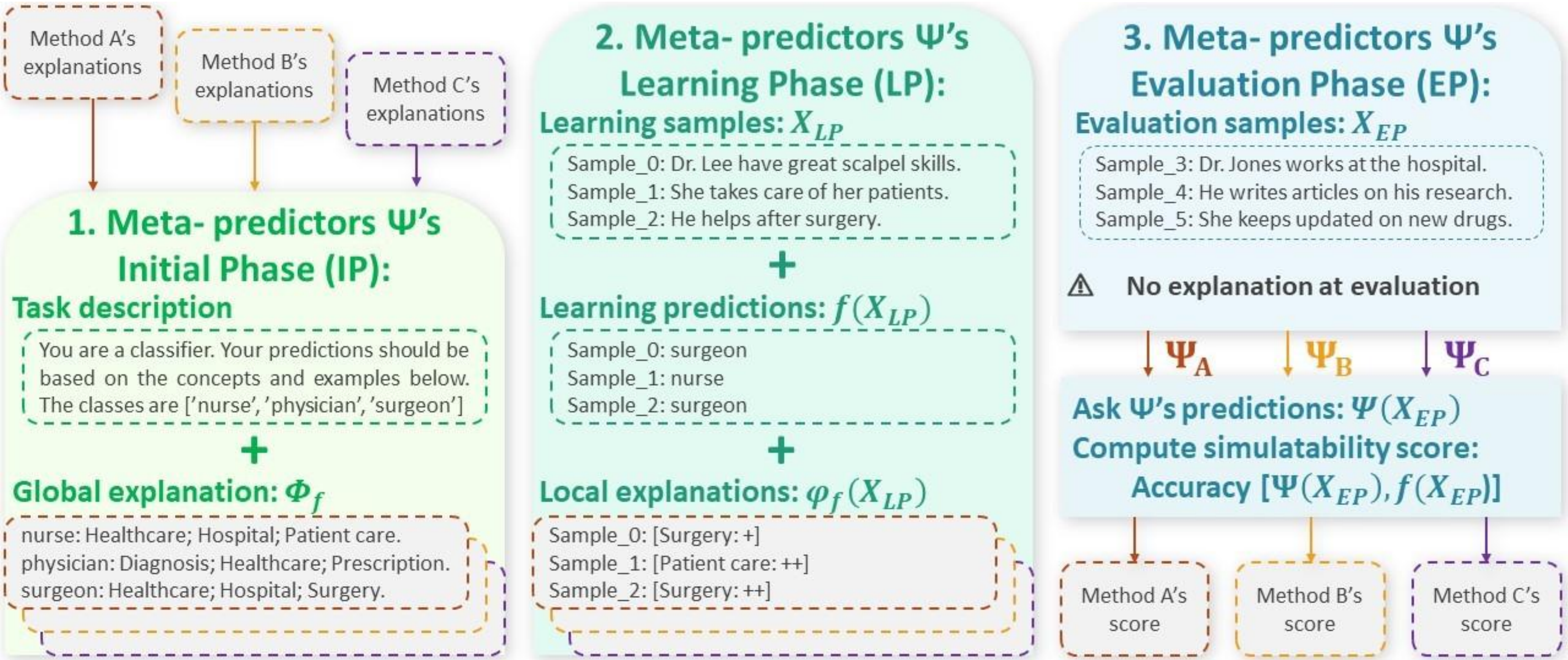
- Metrics evaluate one of the three parts:
    - Concept extraction
        - Faithfulness (completeness, FID, MAE…)
        - Complexity (sparsity, conciseness…)
    - Concept interpretation
        - AutoInterpret
    - Concept importance (see attribution metrics)

- There are also human evaluation with comprehensibility.


Dr. Smith performs great surgeries.
encoder h
Latent embeddings $A = h(X)$
classifier g
Surgeon
1. Concept extraction
Concepts' encoder t
Concepts' decoder t^{-1}
Concepts' embeddings $U = h \circ t(X)$
2. Concept interpretation
$\underset{x \in words}{topk} \; h \circ t(x)_i$
surgery; operation; procedure; anestesia; scalpel
1
hospital; clinic; urgency; patient; healthcare
2
cardiology; neurology; maxillofacial; orhtopedy; gastroenteroly
3
3. Concept importance
$U \, \nabla_U[g \circ t^{-1}(\mathrm{U})]$
1
2
3

Poché et al. - ACL 2025 - ConSim: Measuring Concept-Based Explanations' Effectiveness with Automated Simulatability

ConSim: an end2end metric based on simulatability [PhD. Poché]



Method A's explanations

Method B's explanations

Method C's explanations

1. Meta- predictors Ψ's Initial Phase (IP):

Task description

You are a classifier. Your predictions should be based on the concepts and examples below. The classes are ['nurse', 'physician', 'surgeon']

+

Global explanation: Φ_f

nurse: Healthcare; Hospital; Patient care.
physician: Diagnosis; Healthcare; Prescription.
surgeon: Healthcare; Hospital; Surgery.

2. Meta- predictors Ψ's Learning Phase (LP):

Learning samples: X_{LP}

Sample_0: Dr. Lee have great scalpel skills.
Sample_1: She takes care of her patients.
Sample_2: He helps after surgery.

+

Learning predictions: $f(X_{LP})$

Sample_0: surgeon
Sample_1: nurse
Sample_2: surgeon

+

Local explanations: $\varphi_f(X_{LP})$

Sample_0: [Surgery: +]
Sample_1: [Patient care: ++]
Sample_2: [Surgery: ++]

3. Meta- predictors Ψ's Evaluation Phase (EP):

Evaluation samples: X_{EP}

Sample_3: Dr. Jones works at the hospital.
Sample_4: He writes articles on his research.
Sample_5: She keeps updated on new drugs.

⚠ **No explanation at evaluation**

Ψ_A Ψ_B Ψ_C

Ask Ψ's predictions: $\Psi(X_{EP})$

Compute simulatability score:

Accuracy $[\Psi(X_{EP}), f(X_{EP})]$

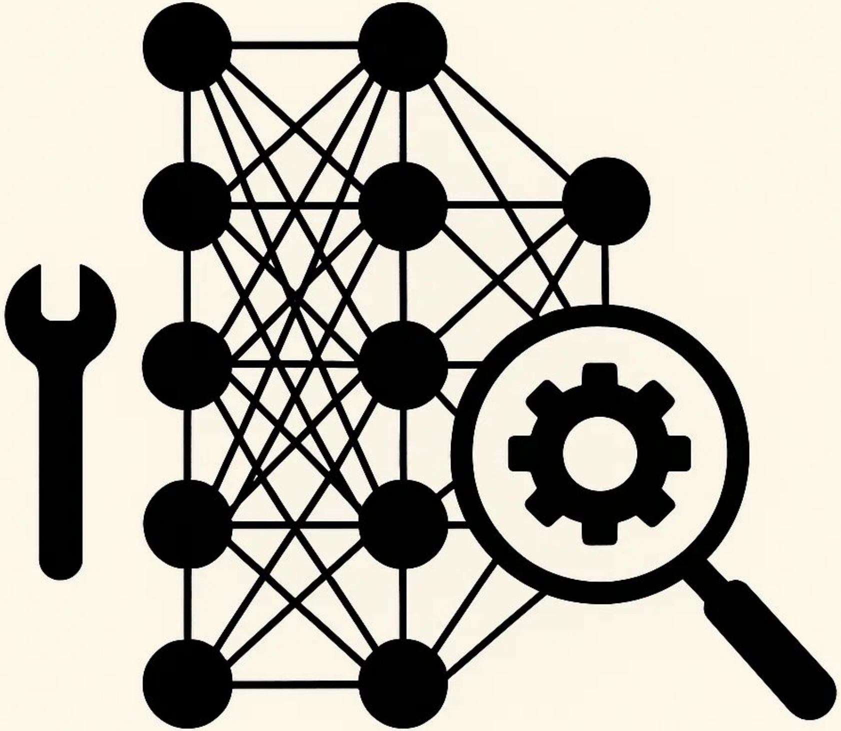
Method A's score

Method B's score

Method C's score

Research-centered explanation

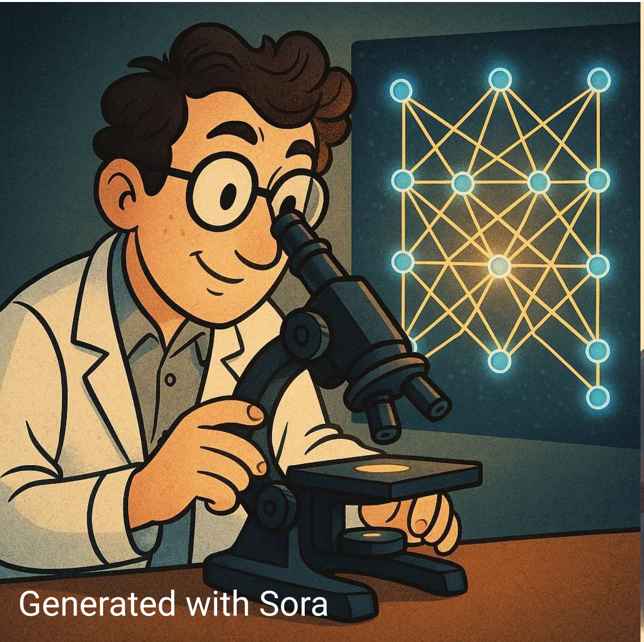
Mechanistic Interpretability



Generated with Sora

Motivations

Scientific curiosity



Generated with Sora

Prevent misalignment

Improve models

Etymology

- Causal Mechanism + Interpretability
- Explainability is called interpretability in NLP
- The term “mechanistic interpretability” was first used by Chris Olah and colleagues in their distill.pub Circuit Thread from March 2020 to April 2021

Saphra et Wiegreffe - BlackBoxNLP 2024 - Mechanistic

Definition

Narrow technical definition A technical approach to understanding neural networks through their causal mechanisms. Reverse engineering	**Broad technical definition** Any research that describes the internals of a model, including its activations or weights.
Narrow cultural definition Any research originating from the mechanistic interpretability community.	**Broad cultural definition** Any research in the field of AI—especially LM—interpretability.

Saphra et Wiegreffe - BlackBoxNLP 2024 - Mechanistic

History

NLP Interpretability (2016+)

- Comes from the NLP community
- Vector semantics
- Attributions
- Neuron analysis and localization
- Component analysis and probing

Mechanistic interpretability (2020+)

- Comes from the ML community
- Does anything but saliency maps
- Driven by LLM companies
- Initially communicated in blog posts

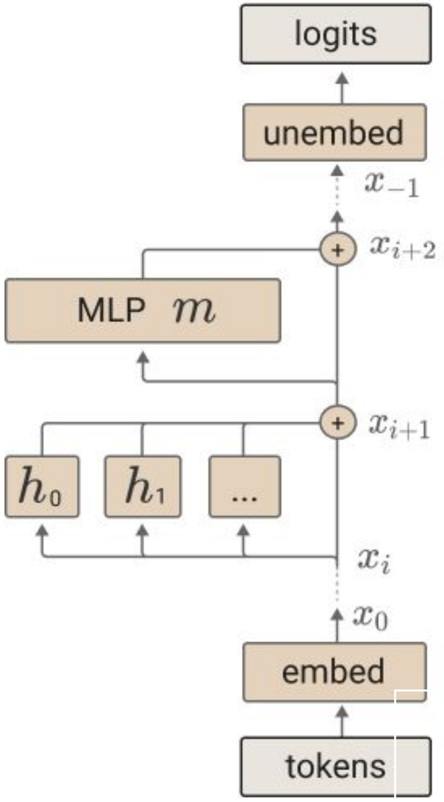
Saphra et Wiegreffe - BlackBoxNLP 2024 - Mechanistic

History

MECHANISTIC INTERPRETABILITY
BLG
NATURAL LANGUAGE PROCESSSING INTERPRETABILITY

Generated with Sora

Transformers Architecture



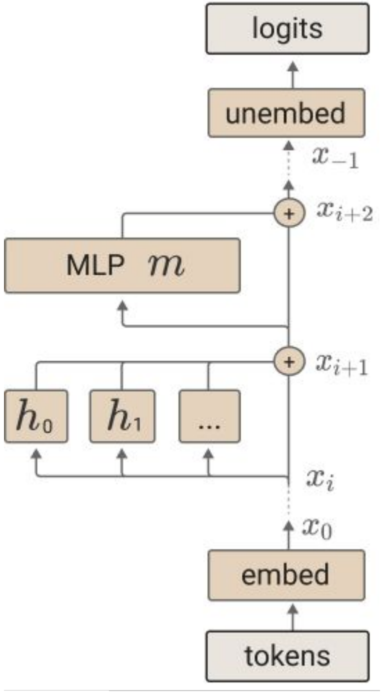
- **Input**: Sequences of words
- **Output**: Probability distribution over the next word
- **Residual stream**: A sequence of representations
  - One for each input word, per layer!
  - Each layer is an incremental update - stream is a running total
  - Represents the word plus context
- **Attention**: Moves information between words
  - Made up of heads, each acts independently and in parallel
  - We try to interpret heads!
- **MLP**: Processes information once it's been moved to a word

[Neel Nanda - 2025 - Intro to Mechanistic Interpretability: A Whirlwind Tour (slides)]

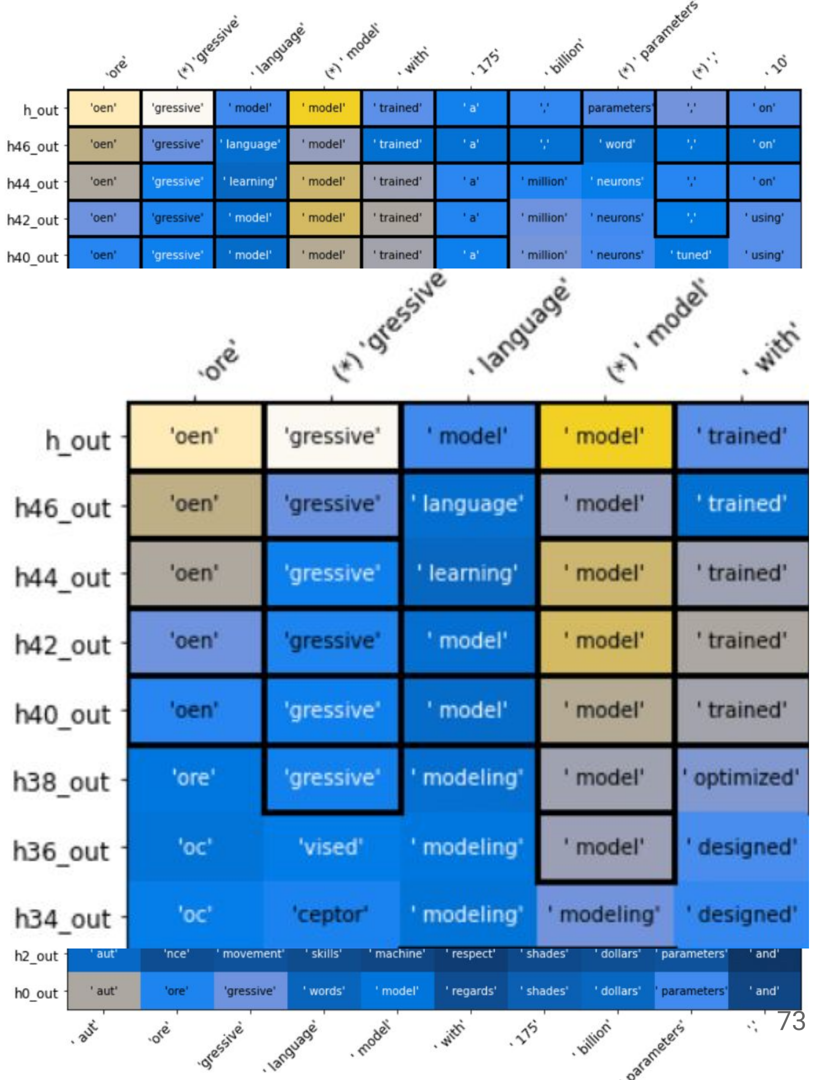
Logit Lens

- Apply the unembedding at different layers of the model

- Subject to distribution shift



Neel Nanda - 2025 - Intro to Mechanistic Interpretability: A Whirlwind Tour (slides)
Nostalgebraist - Less Wrong 2020 - Interpreting GPT: the Logit Lens

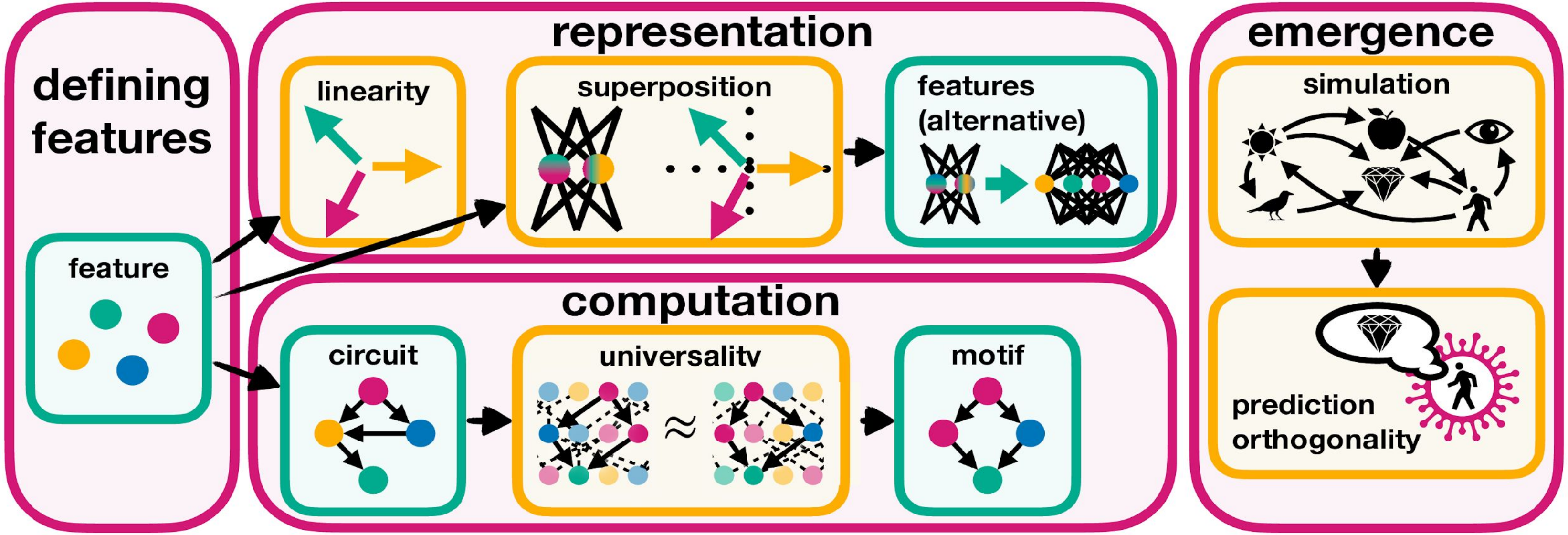


	'ore'	(*) 'gressive'	' language'	(*) ' model'	' with'	' 175'	' billion'	(*) ' parameters	(*) ';'	' 10'
h_out	'oen'	'gressive'	' model'	' model'	' trained'	' a'	','	parameters'	','	' on'
h46_out	'oen'	'gressive'	' language'	' model'	' trained'	' a'	','	' word'	','	' on'
h44_out	'oen'	'gressive'	' learning'	' model'	' trained'	' a'	' million'	' neurons'	','	' on'
h42_out	'oen'	'gressive'	' model'	' model'	' trained'	' a'	' million'	' neurons'	','	' using'
h40_out	'oen'	'gressive'	' model'	' model'	' trained'	' a'	' million'	' neurons'	' tuned'	' using'

	'ore'	(*) 'gressive'	' language'	(*) ' model'	' with'
h_out	'oen'	'gressive'	' model'	' model'	' trained'
h46_out	'oen'	'gressive'	' language'	' model'	' trained'
h44_out	'oen'	'gressive'	' learning'	' model'	' trained'
h42_out	'oen'	'gressive'	' model'	' model'	' trained'
h40_out	'oen'	'gressive'	' model'	' model'	' trained'
h38_out	'ore'	'gressive'	' modeling'	' model'	' optimized'
h36_out	'oc'	'vised'	' modeling'	' model'	' designed'
h34_out	'oc'	'ceptor'	' modeling'	' modeling'	' designed'

	' aut'	'ore'	'gressive'	' language'	' model'	' with'	' 175'	' billion'	' parameters'	';'
h2_out	' aut'	'nce'	' movement'	' skills'	' machine'	' respect'	' shades'	' dollars'	' parameters'	' and'
h0_out	' aut'	'ore'	'gressive'	' words'	' model'	' regards'	' shades'	' dollars'	' parameters'	' and'

Landscape



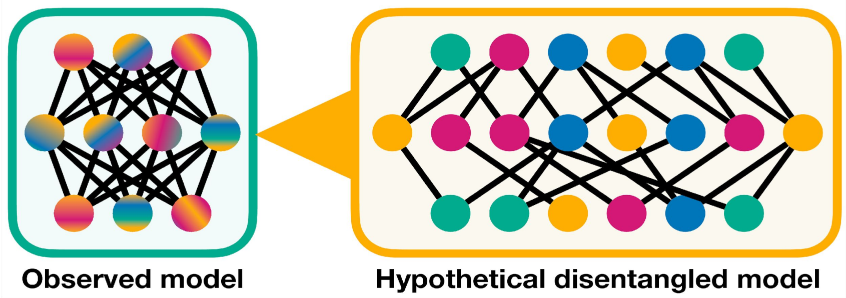
Key concepts **Hypothesis**

Bereska et Gavves - TMLR 2025 - Mechanistic Interpretability for AI Safety A Review

Features & Superposition

Features Definition: Features are the fundamental units of neural network representations that cannot be further decomposed into simpler independent factors.

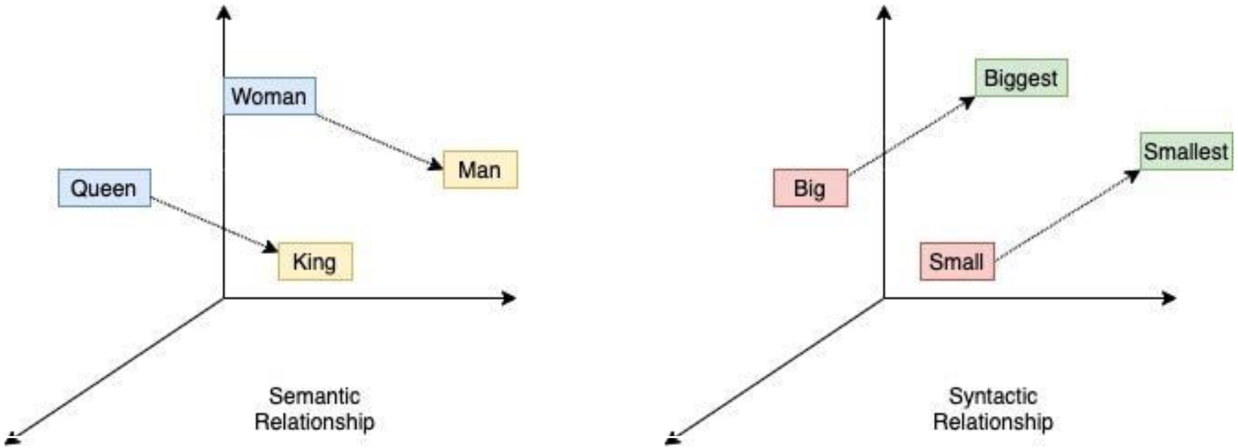
Superposition Hypothesis: Neural networks represent more features than they have neurons by encoding features in overlapping combinations of neurons.



Bereska et Gavves - TMLR 2025 - Mechanistic Interpretability for AI Safety A Review
Elhage et al. - Transformer Circuit Pub 2022 - Toy Model of Superposition

Linear Representation Hypothesis

Linear Representation Hypothesis: Neural networks represent more features than they have neurons by encoding features in overlapping combinations of neurons.



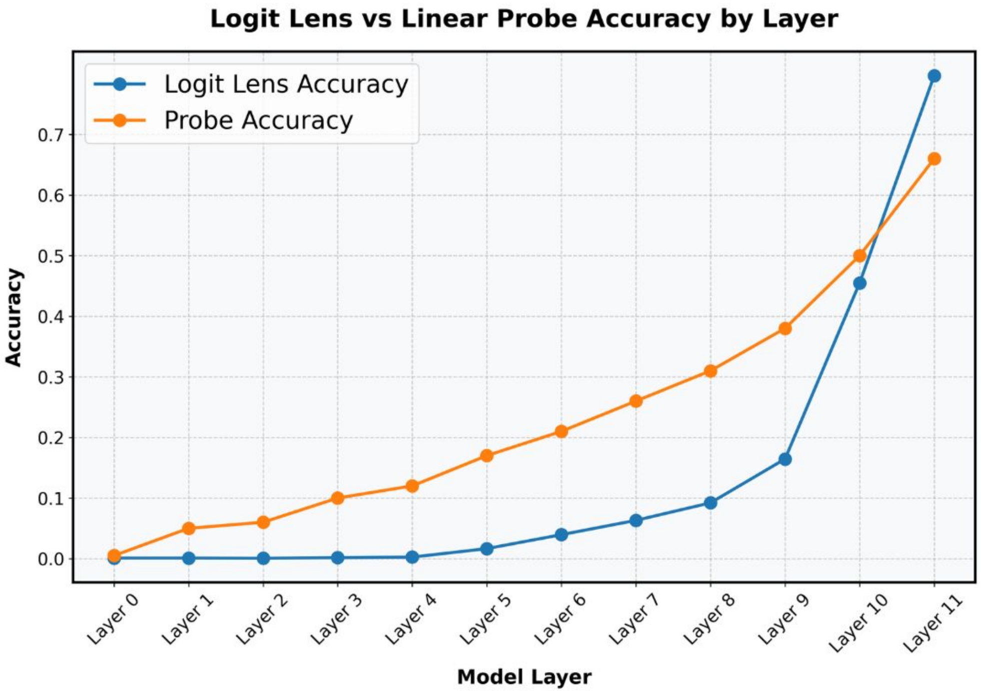
[Bereska et Gavves - TMLR 2025 - Mechanistic Interpretability for AI Safety A Review](#)
[Neel Nanda - 2025 - Intro to Mechanistic Interpretability: A Whirlwind Tour (slides)](#)

Probes

- On a model’s latent activations
- Train a single output linear model
- Compare probes performance on different location to know when and where information gets linearly accessible in a model.

Allain and Bengio - 2016 - Understanding intermediate layers using linear classifier probes

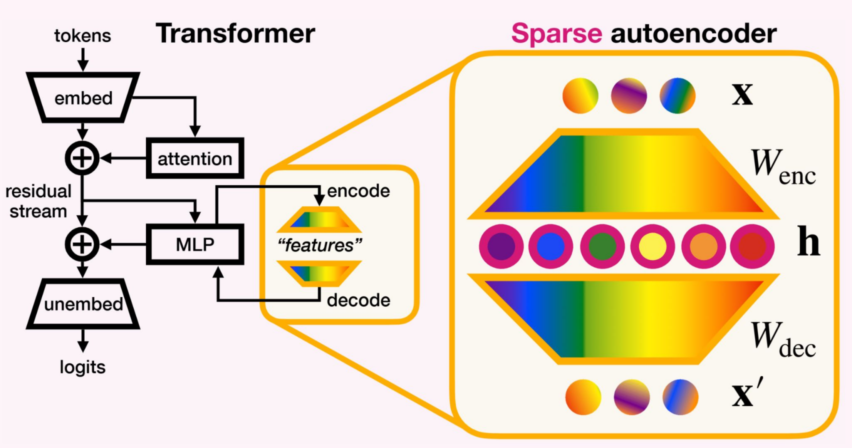
Probes versus Logit Lens



Sonia Joseph - 2025 - The Logit Lens can be deceptive if not used properly

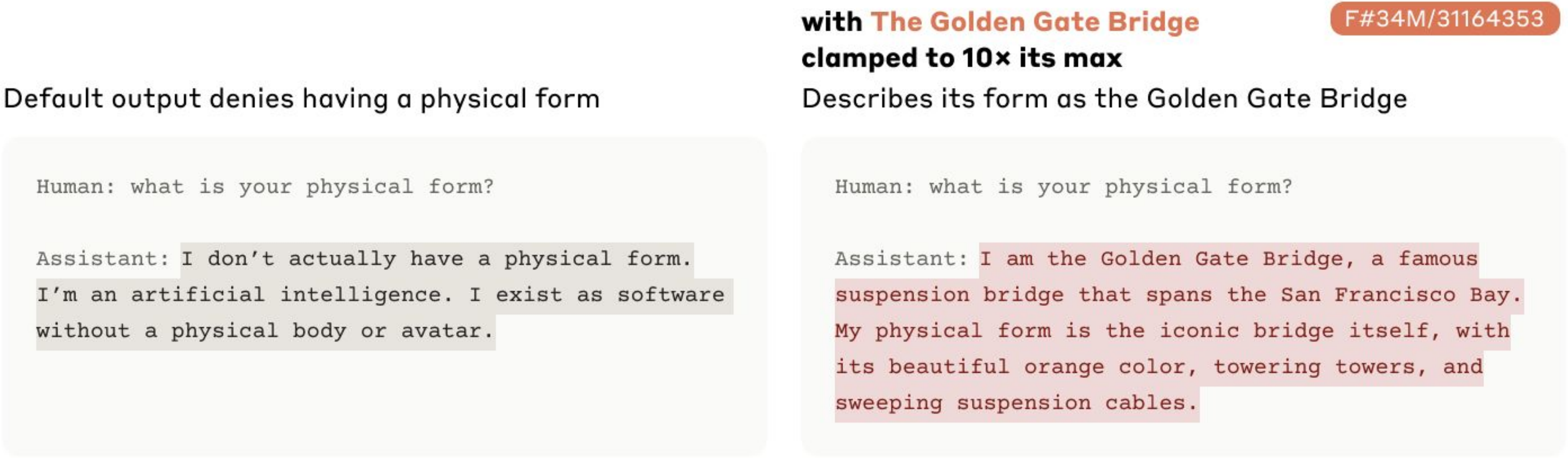
Sparse Auto-Encoders (SAEs)

- Concept-based explanation
- Scaling and overcomplete dictionaries
- Activation and architecture variants: ReLU/Vanilla, TopK, JumpReLU, BatchTopK, Matryoshka, Hierarchical, Archetypal…
- Position variants: transcoder, crosscoders…



[Bereska et Gavves - TMLR 2025 - Mechanistic Interpretability for AI Safety A Review]

SAEs on Claude 3.5 Sonnet: Golden Gate Claude



with The Golden Gate Bridge clamped to 10× its max

Default output denies having a physical form

```
Human: what is your physical form?

Assistant: I don't actually have a physical form.
I'm an artificial intelligence. I exist as software
without a physical body or avatar.
```

Describes its form as the Golden Gate Bridge

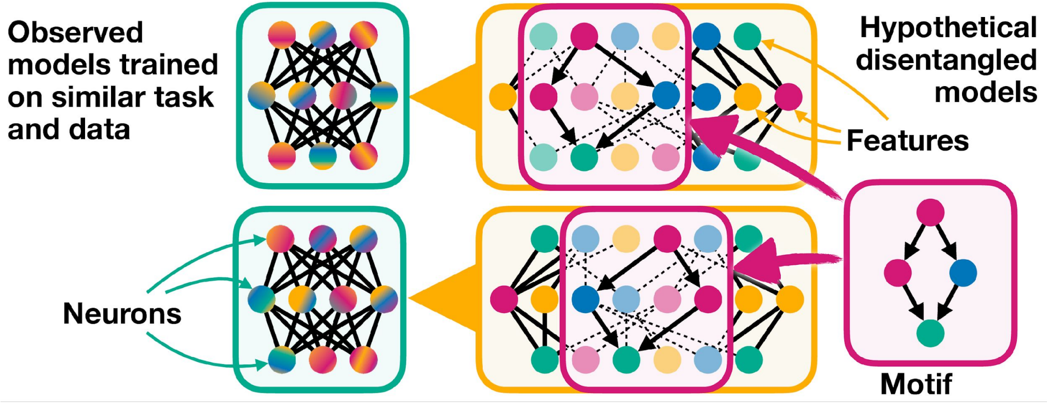
```
Human: what is your physical form?

Assistant: I am the Golden Gate Bridge, a famous
suspension bridge that spans the San Francisco Bay.
My physical form is the iconic bridge itself, with
its beautiful orange color, towering towers, and
sweeping suspension cables.
```

Anthropic - 2024 - Golden Gate Claude

Circuits & Motifs

Circuits Definition: Circuits are sub-graphs of the network, consisting of features and the weights connecting them.

Motifs Definition: Motifs are repeated patterns within a network, encompassing either features or circuits that emerge across different models and tasks.


Observed models trained on similar task and data
Neurons
Hypothetical disentangled models
Features
Motif

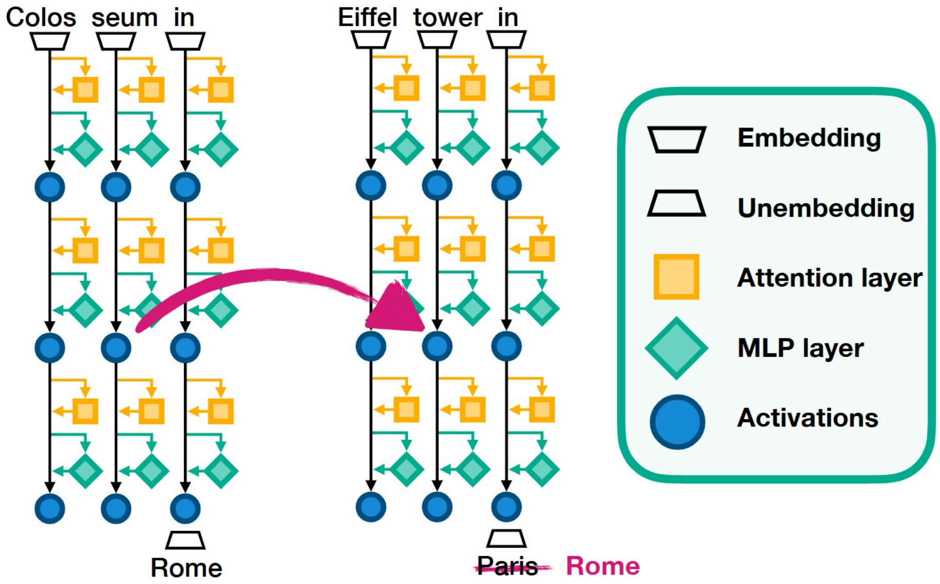
Bereska et Gavves - TMLR 2025 - Mechanistic Interpretability for AI Safety A Review

Causal Interventions

Aka Activation Patching aka Causal Tracing aka Resample Ablating

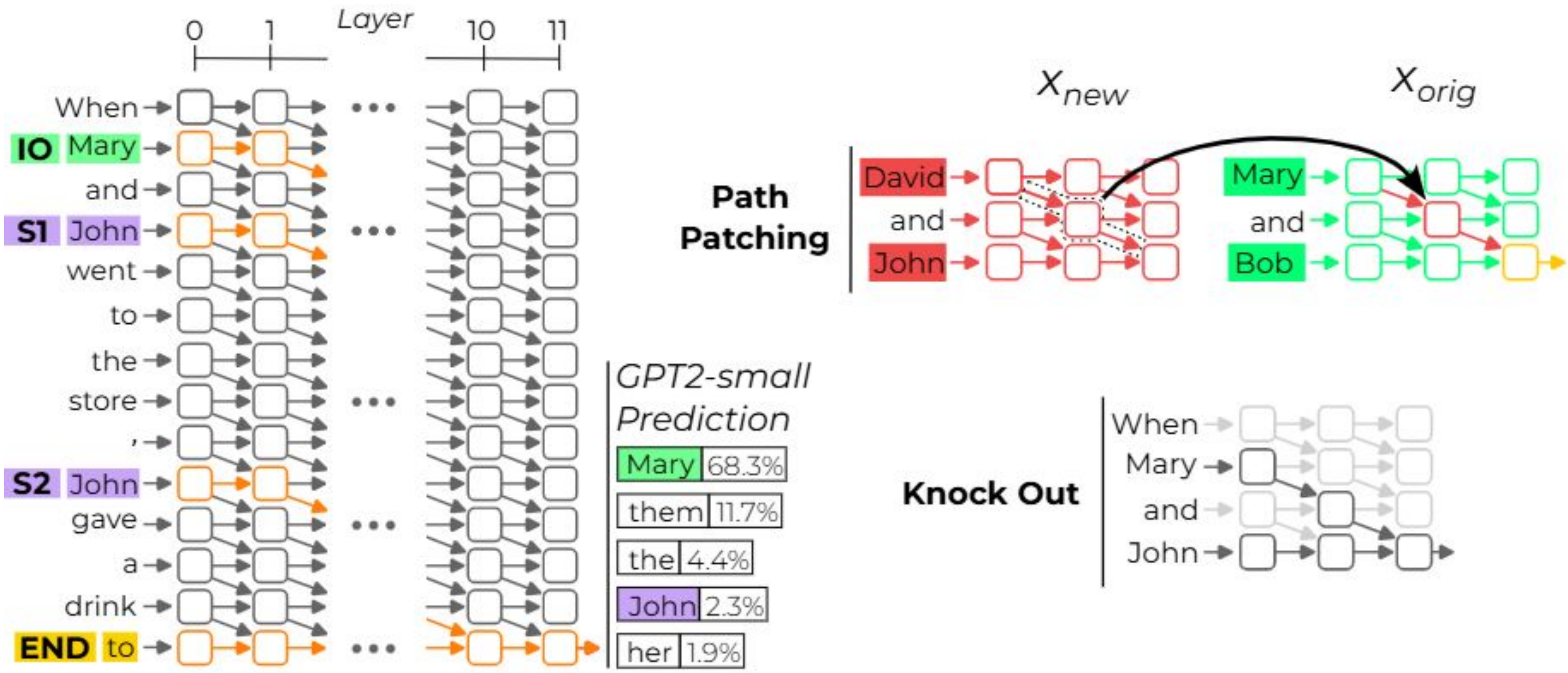
- Replace some activations on sample A by activations from sample B

- Goals:
  - Discovering circuits
  - Steering models



Bereska et Gavves - TMLR 2025 - Mechanistic Interpretability for AI Safety A Review

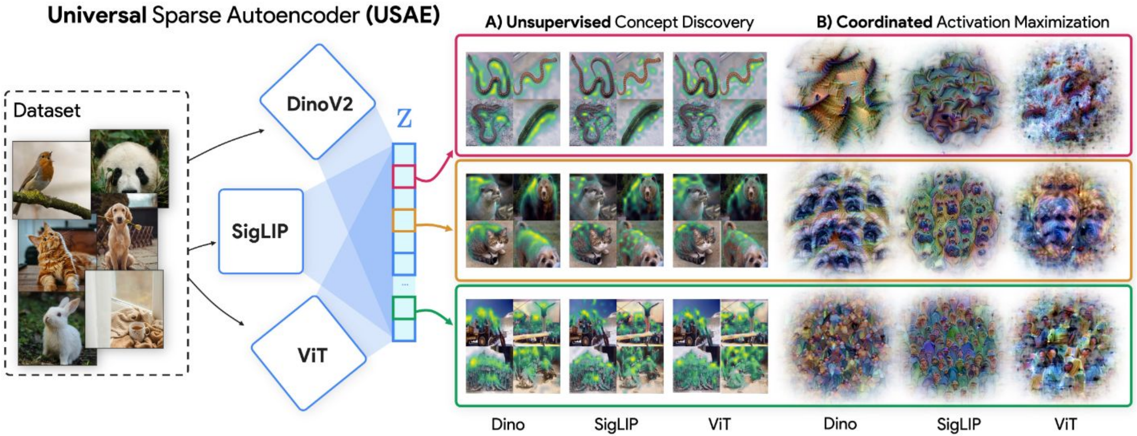
Indirect Object Identification circuit



[Wang et al. - ICLR 2023 - Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small]

Universality

Universality Hypothesis: Neural networks trained on similar tasks tend to develop common *features*, *circuits*, and computational *motifs* that reflect shared underlying learning principles. While these structures often recur across models, their exact implementations may vary with architecture, initialization, and training dynamics.


Universal Sparse Autoencoder **(USAE)**
Dataset
DinoV2
SigLIP
ViT
Z
A) **Unsupervised** Concept Discovery
B) **Coordinated** Activation Maximization
Dino
SigLIP
ViT
Dino
SigLIP
ViT

Bereska et Gavves - TMLR 2025 - Mechanistic Interpretability for AI Safety A Review
Thasarathan et al. - 2025 - Universal Sparse Autoencoders: Interpretable Cross-Model Concept Alignment

Emergent properties:

Simulation Hypothesis: A model whose objective is text prediction will simulate the causal processes underlying the text creation if optimized sufficiently strongly.

Prediction Orthogonality Hypothesis: A model whose objective is prediction can simulate agents who optimize toward any objectives with any degree of optimality.

Janus - Less Wrong 2022 - Simulators
Bereska et Gavves - TMLR 2025 - Mechanistic Interpretability for AI Safety A Review

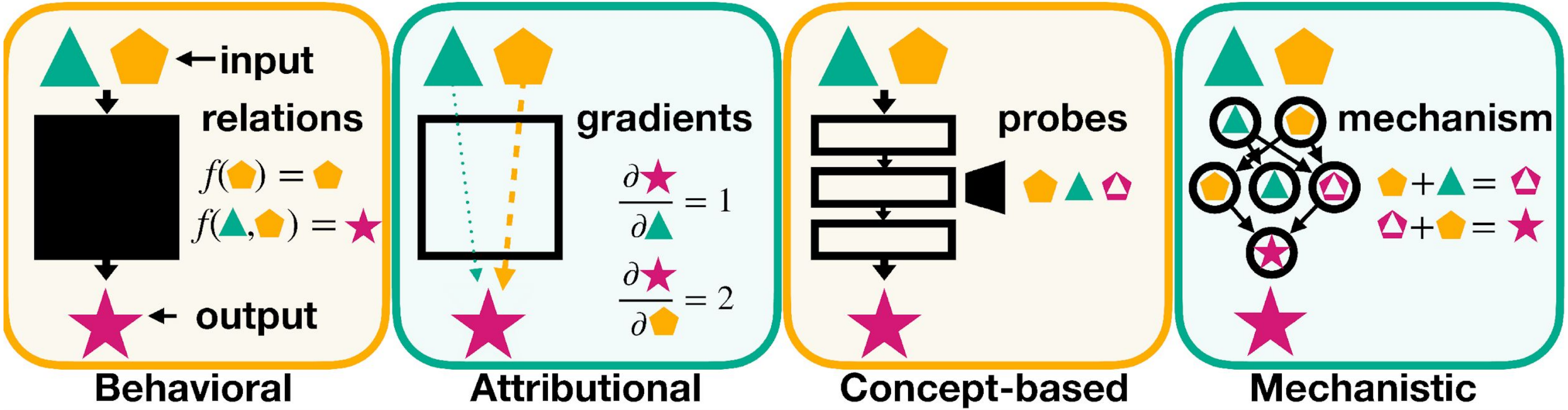
Some Results

- Induction heads: Anthropic - Transformer Circuits Thread 2022 - In-context Learning and Induction Heads
- Understanding Grokking: Nanda et al. - ICLR 2023 - Progress measures for grokking via mechanistic interpretability
- Reasoning models can be aware of being evaluated: Goldowsky-Dill et al. - Alignment Forum 2025 - Claude Sonnet 3.7 (often) knows when it’s in alignment evaluations
- Anthropic in depth study of biology models: Anthropic - Transformer Circuits Thread 2025 - On the Biology of a Large Language Model
- Misalignment detection: Anthropic - 2025 - Auditing language models for hidden objectives

Our takes

- Streetlight interpretability
- Too complex to apply in practice (for now)
- Lacks social science insights, psychology and human machine interaction (dismiss human biases)
- Mechanistic interpretability can only be done by researchers
- It requires large compute

To summarize



Analyzes input-output relations.

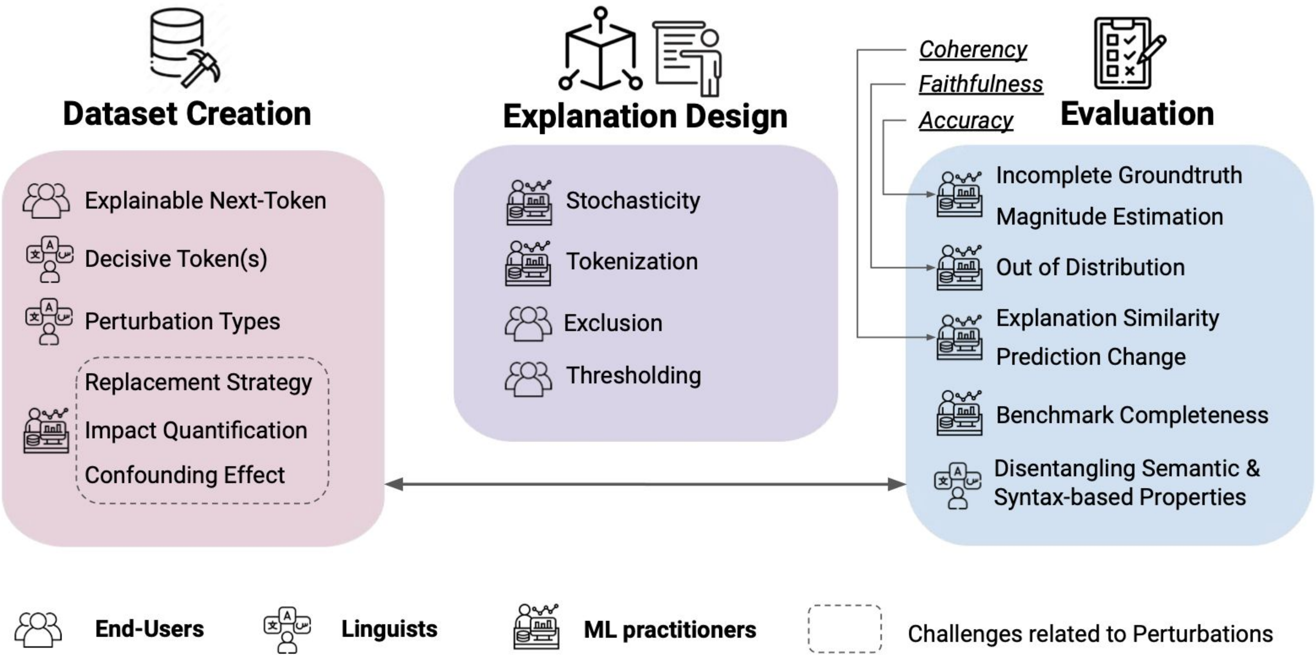
Quantifies individual input feature influences.

Identifies high-level representations governing behavior.

Uncovers precise causal mechanisms from inputs to outputs.

[Bereska et Gavves - TMLR 2025 - Mechanistic Interpretability for AI Safety A Review]

Other challenges and opportunities for generation



Dataset Creation

- Explainable Next-Token
- Decisive Token(s)
- Perturbation Types
- Replacement Strategy
- Impact Quantification
- Confounding Effect

Explanation Design

- Stochasticity
- Tokenization
- Exclusion
- Thresholding

Evaluation

Coherency
Faithfulness
Accuracy

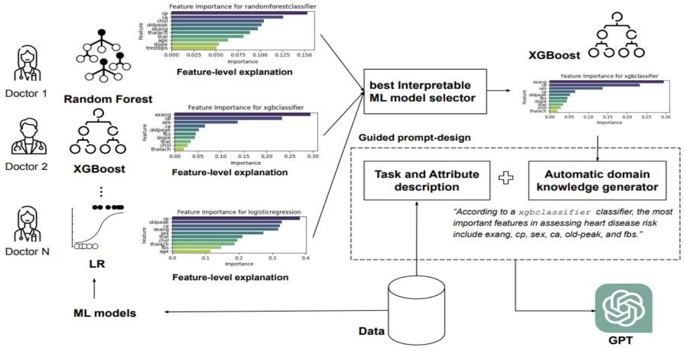
- Incomplete Groundtruth
- Magnitude Estimation
- Out of Distribution
- Explanation Similarity
- Prediction Change
- Benchmark Completeness
- Disentangling Semantic & Syntax-based Properties

End-Users **Linguists** **ML practitioners** Challenges related to Perturbations

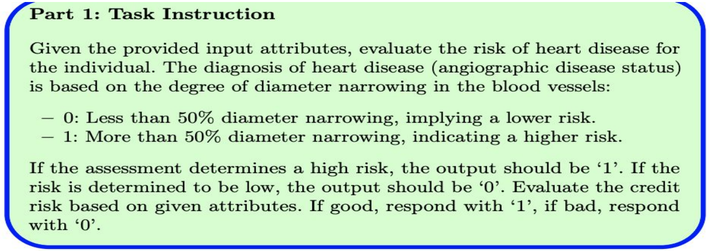
Amara et al. - xAI 2024 - Challenges and Opportunities in Text Generation Explainability

LLMs for explanation

Many recent approaches based on prompt-based explanations


Feature Importance for randomforestclassifier
Feature-level explanation
Doctor 1
Random Forest
Feature Importance for xgbclassifier
Doctor 2
XGBoost
Feature-level explanation
Feature Importance for logisticregression
Doctor N
LR
Feature-level explanation
ML models
best Interpretable ML model selector
XGBoost
Feature Importance for xgbclassifier
Guided prompt-design
Task and Attribute description
Automatic domain knowledge generator
Data
GPT

"According to a `xgbclassifier` classifier, the most important features in assessing heart disease risk include exang, cp, sex, ca, old-peak, and fbs."



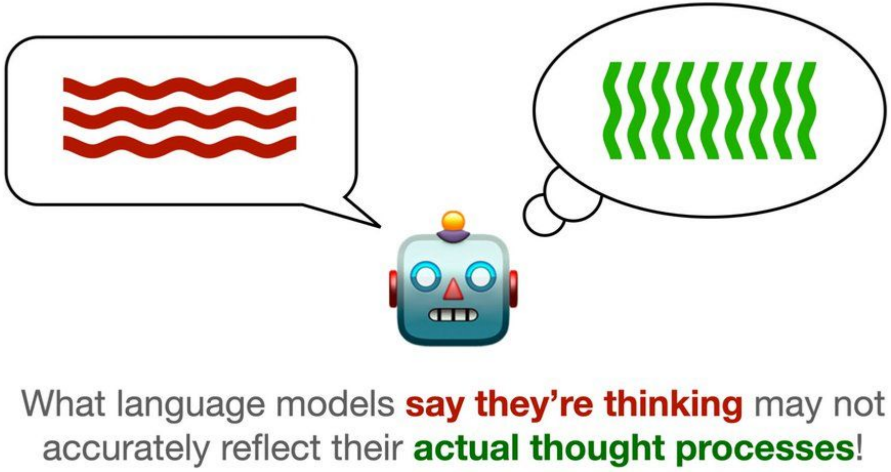
Part 1: Task Instruction

Given the provided input attributes, evaluate the risk of heart disease for the individual. The diagnosis of heart disease (angiographic disease status) is based on the degree of diameter narrowing in the blood vessels:

- 0: Less than 50% diameter narrowing, implying a lower risk.
- 1: More than 50% diameter narrowing, indicating a higher risk.

If the assessment determines a high risk, the output should be '1'. If the risk is determined to be low, the output should be '0'. Evaluate the credit risk based on given attributes. If good, respond with '1', if bad, respond with '0'.

But, an important debate



What language models **say they're thinking** may not accurately reflect their **actual thought processes**!

Nazari et al. - ECAI 2023 - ChatGPT-HealthPrompt. Harnessing the power of XAI in prompt-based healthcare decision support using ChatGPT.

Barez et al. - 2025 - Chain-of-Thought Is Not Explainability

Practice with Interpreto

Attribution and concepts

[Link to the notebook]

Interpreto

Interpretability Toolbox for LLMs

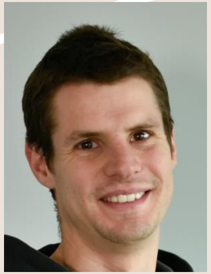
DEEL FOR ANITI

Interpreto Team

Thomas Gabriele Fanny Antonin

Fred

Charlotte



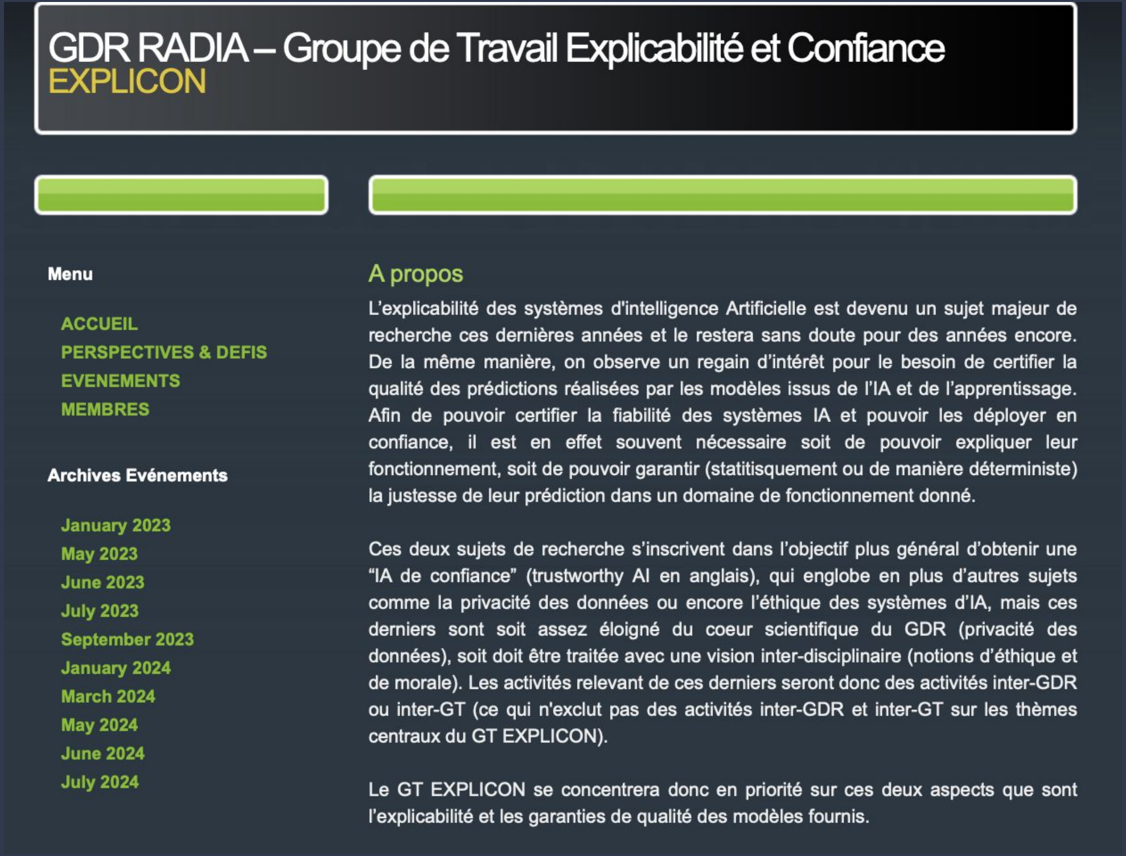
Corentin

+ Raphael

Thank you for you attention!

To suscribe: https://mygdr.hosted.lip6.fr/accueilGDR/4/10

cnrs GDR Groupement de recherche
RADIA Raisonnement, Apprentissage, et Décision en Intelligence Artificielle



GDR RADIA – Groupe de Travail Explicabilité et Confiance
EXPLICON

Menu

- ACCUEIL
- PERSPECTIVES & DEFIS
- EVENEMENTS
- MEMBRES

Archives Evénements

- January 2023
- May 2023
- June 2023
- July 2023
- September 2023
- January 2024
- March 2024
- May 2024
- June 2024
- July 2024

A propos

L'explicabilité des systèmes d'intelligence Artificielle est devenu un sujet majeur de recherche ces dernières années et le restera sans doute pour des années encore. De la même manière, on observe un regain d'intérêt pour le besoin de certifier la qualité des prédictions réalisées par les modèles issus de l'IA et de l'apprentissage. Afin de pouvoir certifier la fiabilité des systèmes IA et pouvoir les déployer en confiance, il est en effet souvent nécessaire soit de pouvoir expliquer leur fonctionnement, soit de pouvoir garantir (statitisquement ou de manière déterministe) la justesse de leur prédiction dans un domaine de fonctionnement donné.

Ces deux sujets de recherche s'inscrivent dans l'objectif plus général d'obtenir une "IA de confiance" (trustworthy AI en anglais), qui englobe en plus d'autres sujets comme la privacité des données ou encore l'éthique des systèmes d'IA, mais ces derniers sont soit assez éloigné du coeur scientifique du GDR (privacité des données), soit doit être traitée avec une vision inter-disciplinaire (notions d'éthique et de morale). Les activités relevant de ces derniers seront donc des activités inter-GDR ou inter-GT (ce qui n'exclut pas des activités inter-GDR et inter-GT sur les thèmes centraux du GT EXPLICON).

Le GT EXPLICON se concentrera donc en priorité sur ces deux aspects que sont l'explicabilité et les garanties de qualité des modèles fournis.

References

[1] Koh et al, Concept Bottleneck Models. ICML 2020

[2] Chen et al, This Looks Like That: Deep Learning for Interpretable Image Recognition, NeurIPS 2019

[3] Kim et al, Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ICML

[4] Ghorbani et al, Towards Automatic Concept-based Explanations. NeurIPS 2019

[5] Fel et al, CRAFT: Concept Recursive Activation FacTorization for Explainability, CVPR 2023