Foundation Models

Jacopo Staiano
jacopo.staiano@unitn.it

Marco Bronzini
marco.bronzini-1@unitn.it

University of Trento
October 15, 2025

Overview

Friday, October 10

Lab 1

- Metrics for Natural Language Generation (NLG)
- Real-world use cases (MT, QA with RAG)

Wednesday, October 15

Lab 2

- Extract latent features from LLM embeddings by training a classification model

November

Lab 3

Lab 4

LLM-based pipeline

1 OBJECTIVE / TASK

LLMs as...

a black-box tool

- Question-answering
- Information extraction
- Text generation
- Machine Translation

a subject of study

- Training
- Interpretability (probing)
- Steering

2 DATA

Open datasets

- GitHub
- Hugging Face
- TensorFlow Datasets

Data collection

- Synthetic data
- Web Scraping
- Humans

Modality

- Text
- Image/video
- Audio

3 METRICS

Classification

- Precision, Recall, F1 score
- Jaccard similarity

Regression

- MAE, MSE, MAPE

Text Generation

- Lexical-based (e.g., ROUGE)
- Semantic-based (e.g., BERTScore)

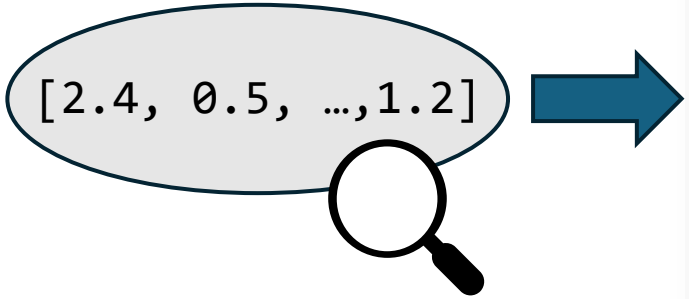
TASK

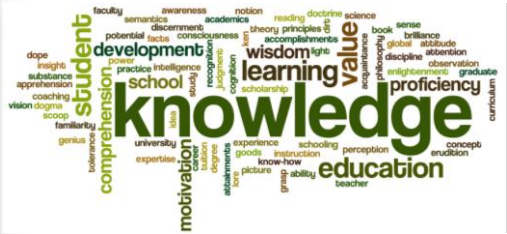
PROBING: Extract latent features from the ***vector space of neural models*** (Information decoding)

Identification of human-interpretable features

a) *Concepts* (e.g., 🐶 , 🐼)
b) *Linguistic features* (e.g., noun, agent)
c) *Abstract properties* (e.g., truthfulness)
d) …

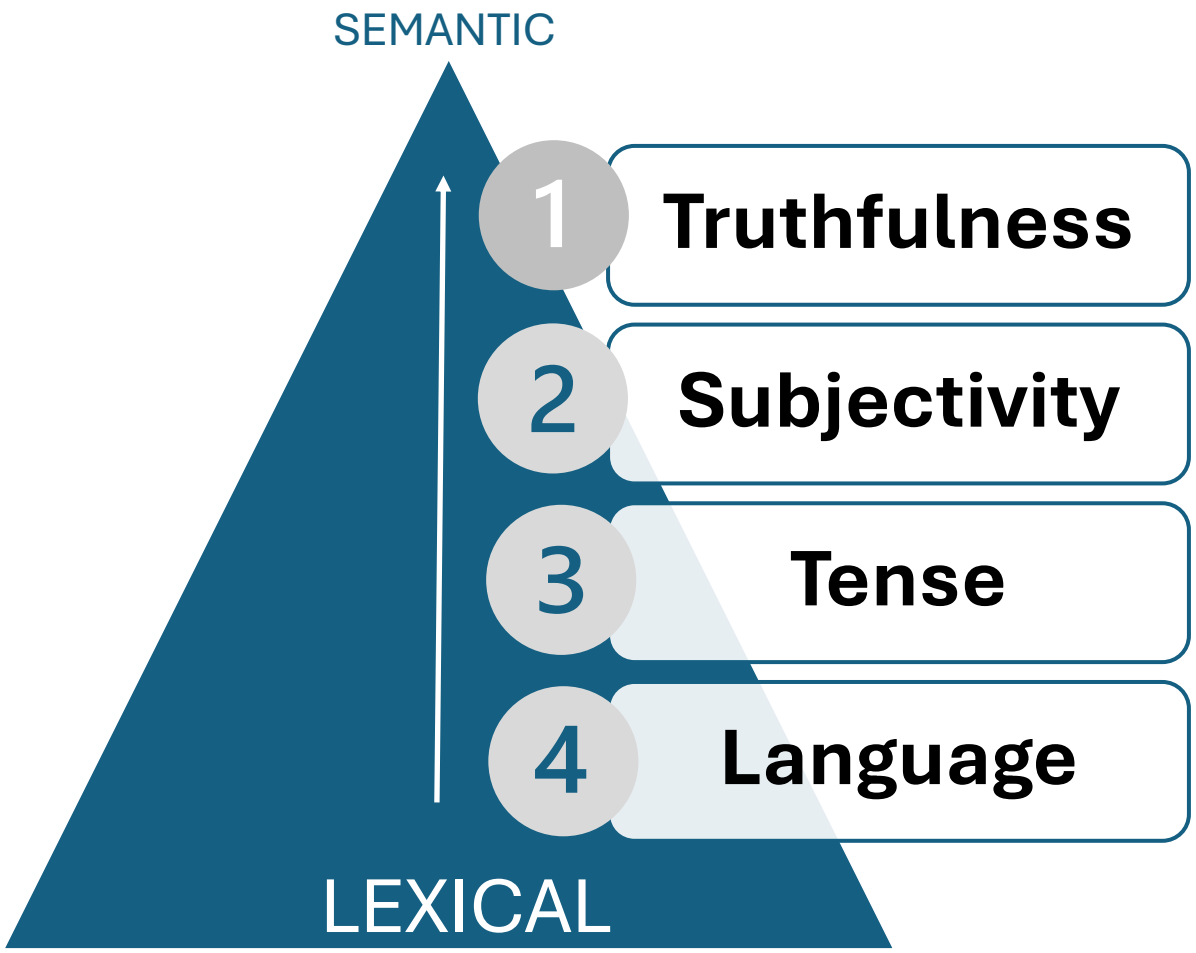
There's *limited understanding* of internal representations of LLMs





INFORMATION DECODING

LATENT FEATURES



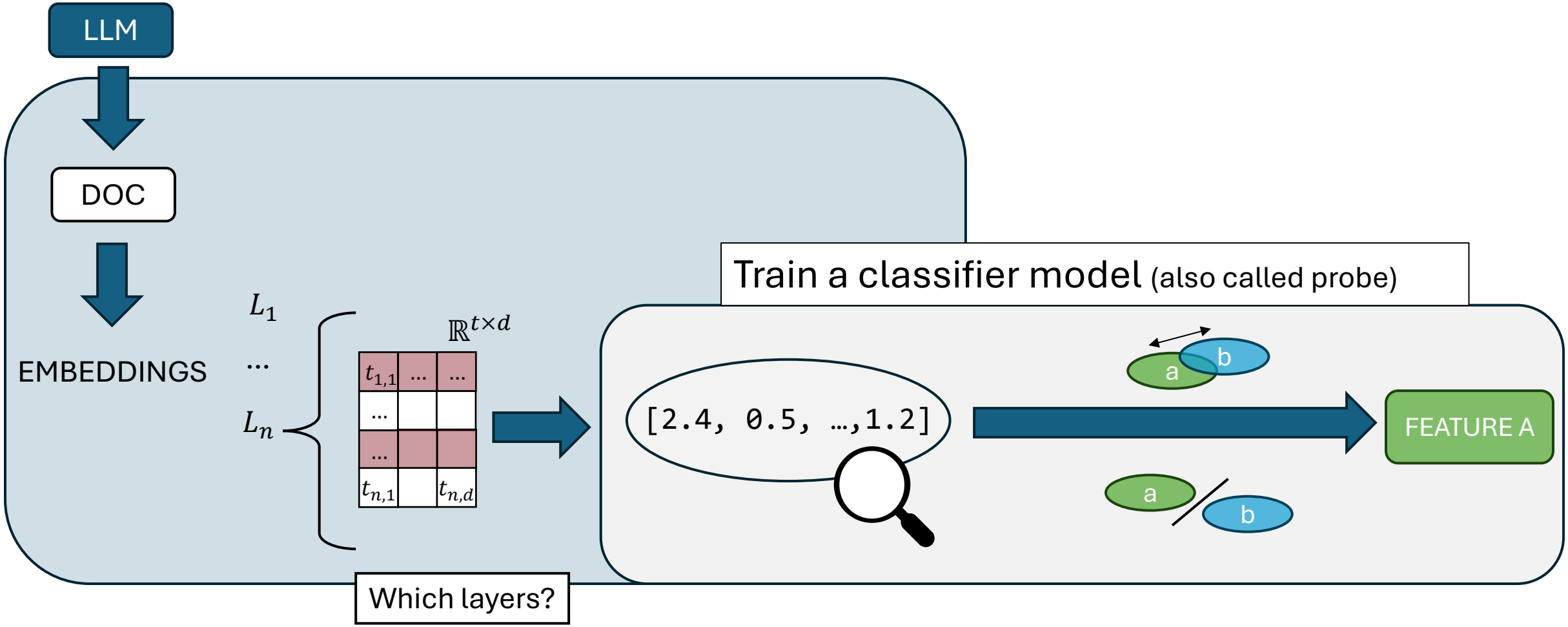
The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets

CLEF 2025

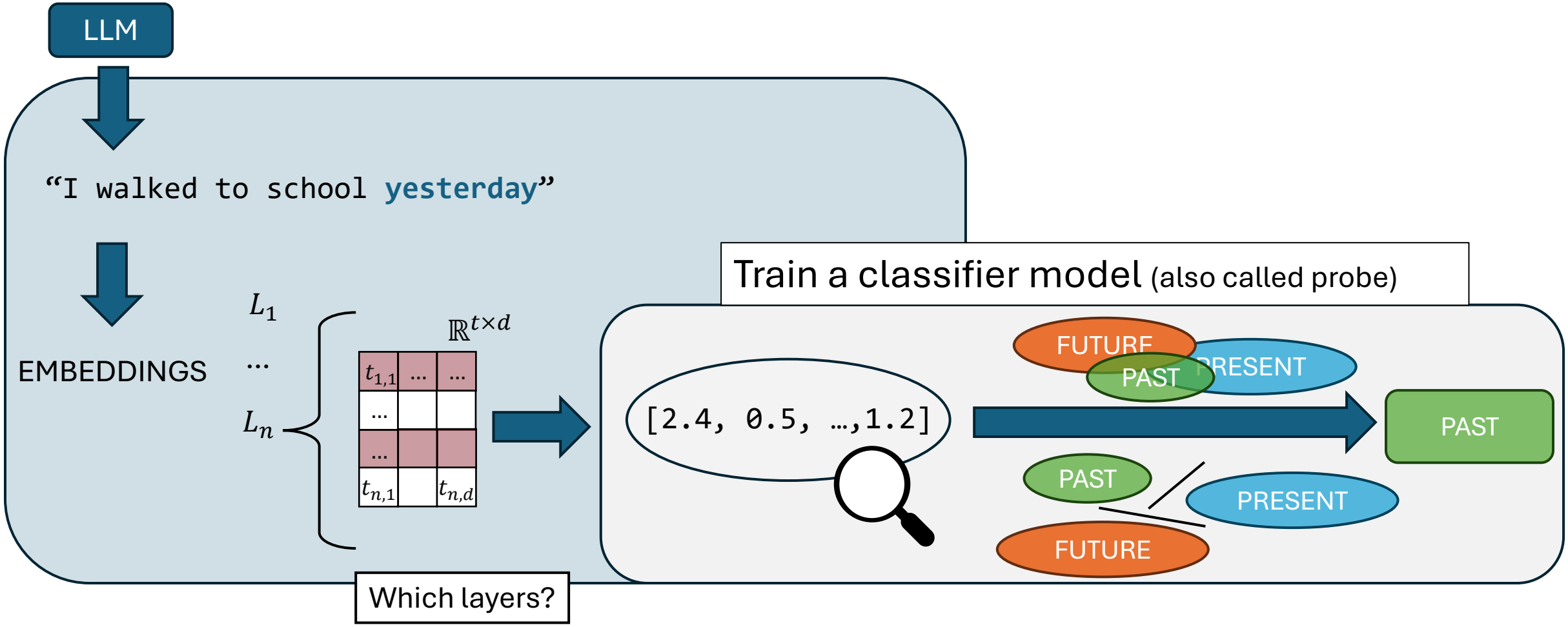
EnglishTense: A large scale English texts dataset categorized into three categories: Past, Present, Future tenses

MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages

PIPELINE


LLM
DOC
EMBEDDINGS
L_1
...
L_n
$\mathbb{R}^{t \times d}$
$t_{1,1}$ ... ...
...
...
$t_{n,1}$ $t_{n,d}$
Which layers?
Train a classifier model (also called probe)
[2.4, 0.5, …,1.2]
a
b
FEATURE A
a
b

PIPELINE


LLM
"I walked to school yesterday"
EMBEDDINGS
L_1
...
L_n
$\mathbb{R}^{t \times d}$
$t_{1,1}$ ... ...
...
...
$t_{n,1}$ $t_{n,d}$
Which layers?
Train a classifier model (also called probe)
[2.4, 0.5, …,1.2]
FUTURE
PAST
PRESENT
PAST
PAST
PRESENT
FUTURE

LOSS FUNCTION

CrossEntropyLoss - PyTorch

CrossEntropyLoss

```
class torch.nn.CrossEntropyLoss(weight=None, size_average=None,
ignore_index=-100, reduce=None, reduction='mean',
label_smoothing=0.0)
```

[source]

This criterion computes the cross entropy loss between input logits and target.

It is useful when training a classification problem with *C* classes. If provided, the optional argument `weight` should be a 1D *Tensor* assigning weight to each of the classes. This is particularly useful when you have an unbalanced training set.

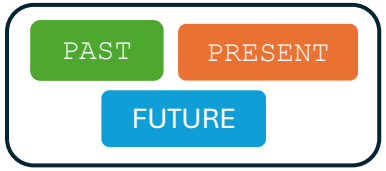
The *input* is expected to contain the unnormalized logits for each class (which do *not* need to be positive or sum to 1, in general). *input* has to be a Tensor of size (C) for unbatched input, $(minibatch, C)$ or $(minibatch, C, d_1, d_2, ..., d_K)$ with $K \geq 1$ for the *K*-dimensional case. The last being useful for higher dimension inputs, such as computing cross entropy loss per-pixel for 2D images.

DATASETS (1): Tense

EnglishTense: A large scale English texts dataset categorized into three categories: Past, Present, Future tenses

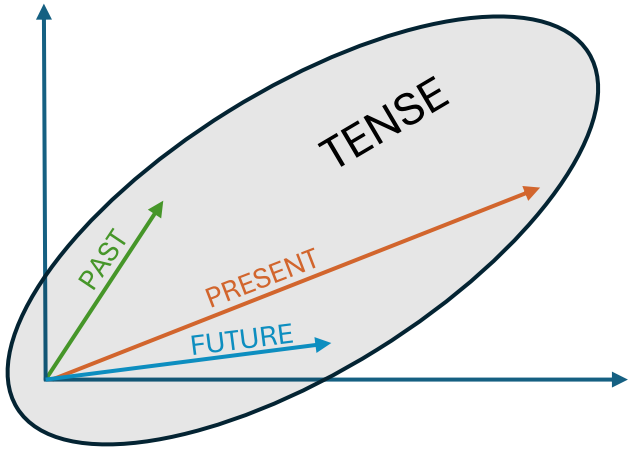
```
CLASSES (3): FUTURE|PRESENT|PAST

                                                    doc    label
0      by 2050 ai architects will have designed selfc...   FUTURE
1      in the future sustainable transportation optio...   FUTURE
2      china has been actively involved in peacekeepi...  PRESENT
3      educational diversity is a hallmark of foreign...  PRESENT
4        the coach substituted an underperforming player     PAST
...                                                  ...      ...
13311                      he will become a good person   FUTURE
13312                     their door opens after eleven  PRESENT
13313                 my mother will cook delicious food   FUTURE
13314                       i am going to win this race   FUTURE
13315               dona will buy a new mobile next month   FUTURE

[13316 rows x 2 columns]
```

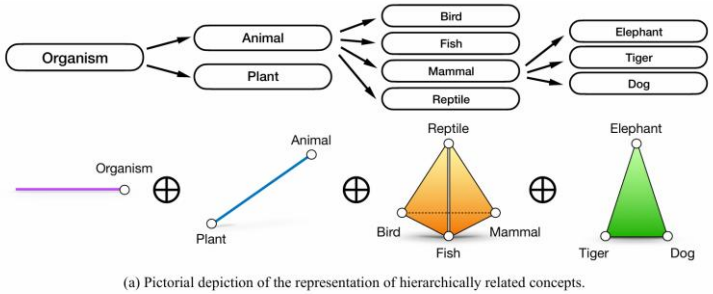
Three classes



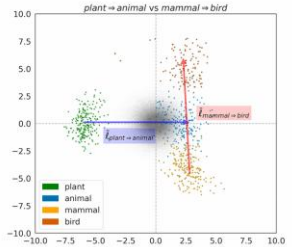
Input features as *directions/subspaces* in the LLM vector space

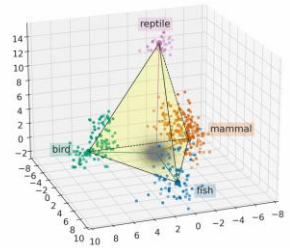
TENSE, LANGUAGE, …





(a) Pictorial depiction of the representation of hierarchically related concepts.





The Geometry of Categorical and Hierarchical Concepts in Large Language Models

DATASETS (2): Language

MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages

1. what is the best italian restaurant around here
2. can you shuffle the light colors in the house
3. send reply to joe

→ en-US

1. comprami un biglietto del treno per modena
2. Inserisci la riunione con paolo domani alle dieci
3. ripeti questo brano tre volte

→ it-IT

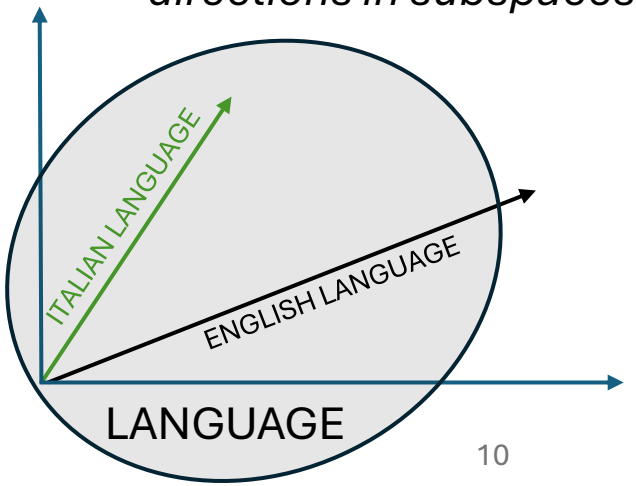
Six languages

it-IT | ja-JP | en-US
es-ES | fr-FR | de-DE

comprami un biglietto del treno per **modena** → [2.4, 0.5, …,1.2]

send reply to **joe** → [9.1, 1.3, …,0.8]

Input features as *directions in subspaces*



DATASETS (3): Subjectivity

CLEF 2025

1. In states with shortages, it's also far more difficult to find teachers for math, science, and special education classes
2. Yet little of substance has been done for Serbia's LGBTQ+ people

→ objective

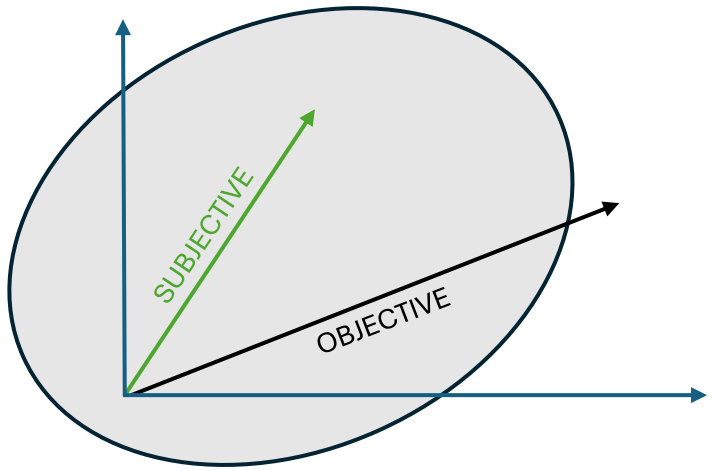
1. If it consolidates power into the hands of the government, we can expect the situation to be pushed hard
2. Gone are the days when they led the world in recession-busting

→ subjective

Two classes

subj | obj

Input features as *directions in subspaces*



From LEXICAL to SEMANTICS

- Not anymore based on the *syntax of the input*
- Based on the *knowledge* and *reasoning* of the language model*

Linguistic analysis → abstract reasoning

* thus, small LLMs might struggle

DATASETS (4): Truthfulness

The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets

1. Apple A5 was created by Google
2. The mother tongue of Thomas Joannes Stieltjes is English

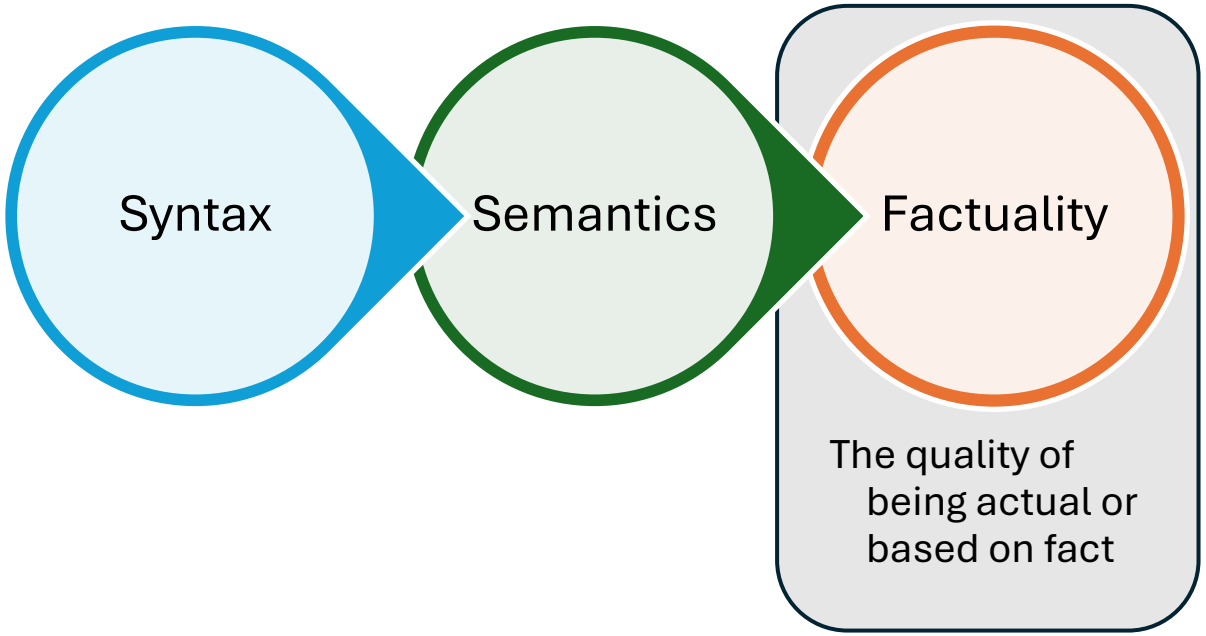
→ FALSE

1. Apple A5 was created by Apple
2. The mother tongue of Thomas Joannes Stieltjes is Dutch

→ TRUE

Two classes

TRUE FALSE



FURTHER STAGE → Claim verification

- A LLM benchmark an input against its internal knowledge

METRICS

We are going to evaluate the performance of our trained prone using **traditional classification metrics**:

True Negative (TN) | True Positive (TP)

False Negative (FN) | False Positive (FP)

Precision

$$precision\ (P) = \frac{TP}{TP+FP}$$

Recall

$$recall\ (R) = \frac{TP}{TP + FN}$$

F1 score

$$F1\ score = \frac{2PR}{P + R}$$

METRICS

We are going to evaluate the performance of our trained prone using **traditional classification metrics**:

Precision

$$precision\ (P) = \frac{TP}{TP+FP}$$

Recall

$$recall\ (R) = \frac{TP}{TP + FN}$$

F1 score

$$F1\ score = \frac{2PR}{P + R}$$

After having compute these metrics for each class, we can compute the average via:

Weighted Average

Average all scores, weighted by the number of true samples in each class

Macro Average

Average all scores equally

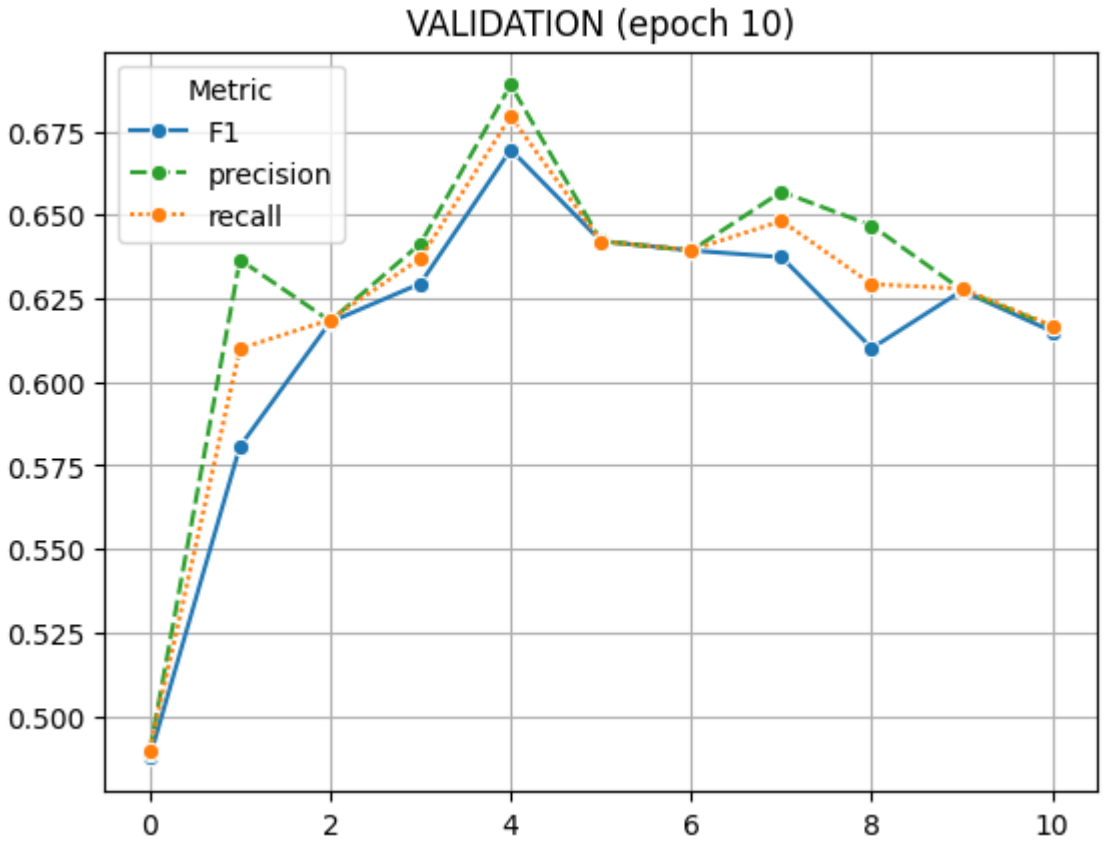
Practical Tutorial

Please, open the following notebook:
https://colab.research.google.com/github/saturnMars/FM_2025/blob/main/Lab2_probing.ipynb

Degree of freedom

1. Latent features being targeted
   - We cannot find what is not encoded
2. Choice of the hidden layer
   - There's not right answer, but the median layer is generally the best one
3. Variability of language models
   - The dimension matters, especially the embedding dimension
4. Number of data
   - More data should lead to increase the generaralization
5. Architecture of the probing model
   - It has a huge impact on findings

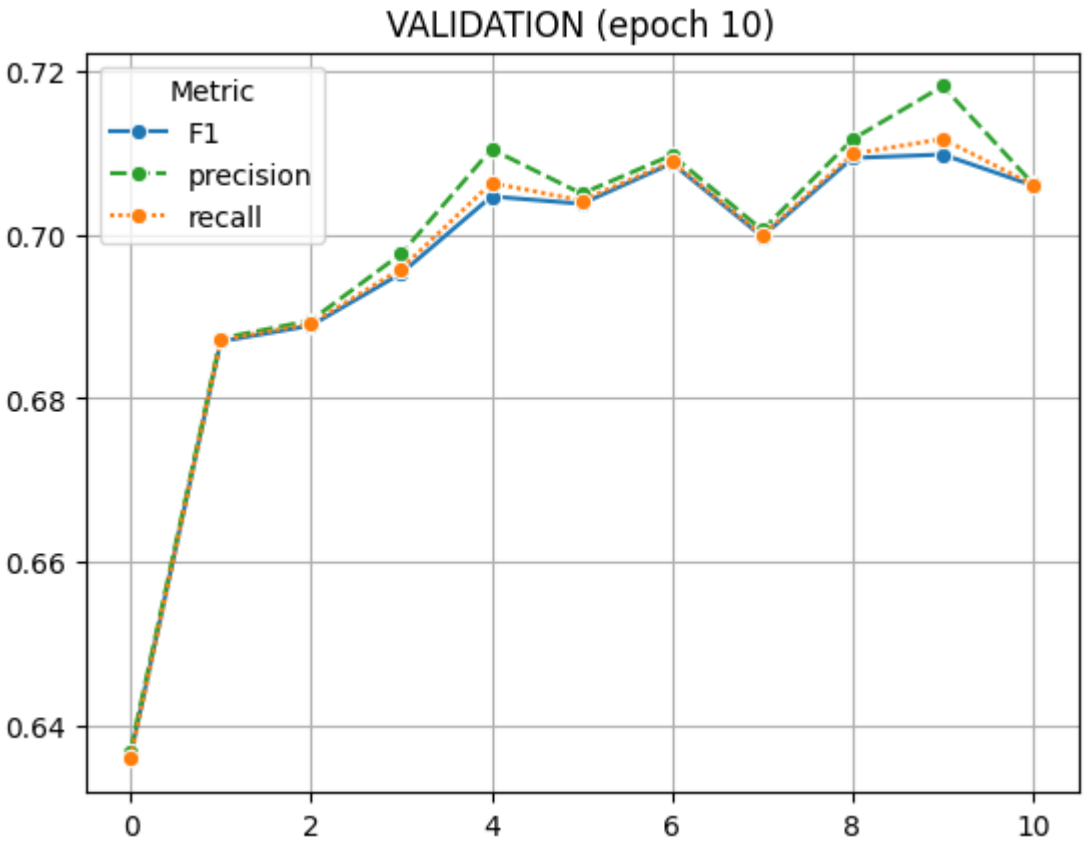
Truthfulness (GPT-2 → 130M, 768 dim)



[TEST] F1 score = 0.65

1) The earth is flat --> False (0.54)
2) Water boils at 100 degrees Celsius --> False (0.94)
3) The sky is green --> False (0.92)
4) The sky is blue --> False (0.78)
5) Renault 8 is produced by Fiat --> False (0.87)

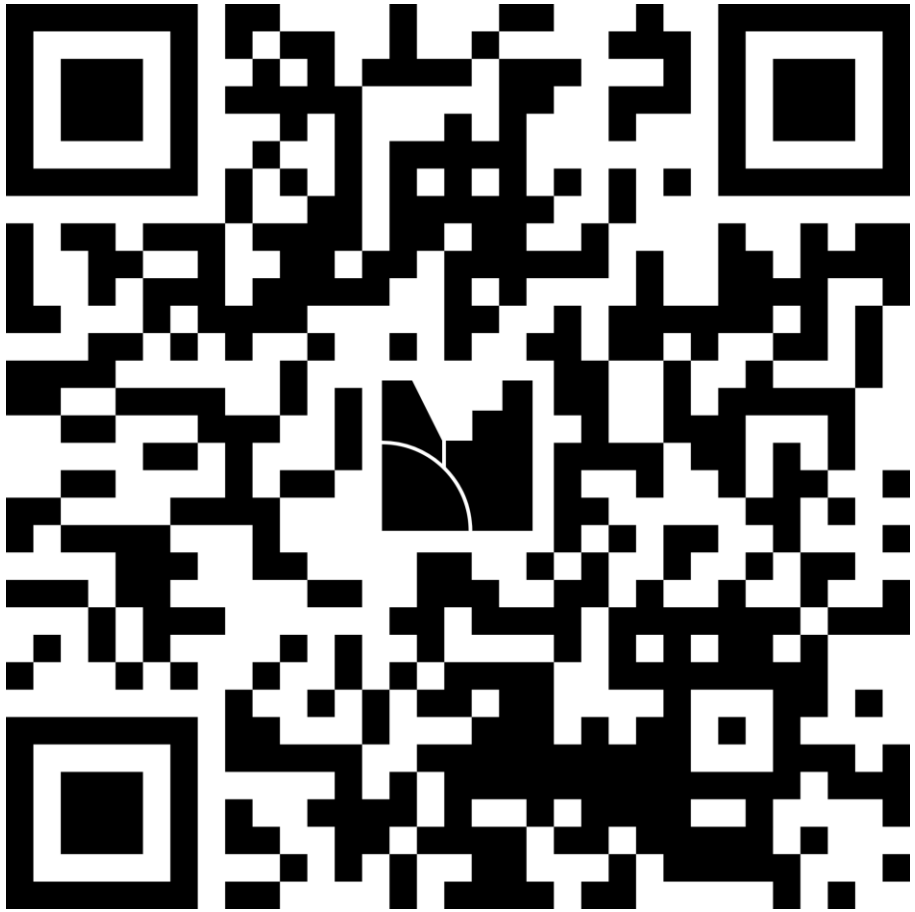
Truthfulness (GPT-2, XL → 1.5B, 1600 dim)



[TEST] F1 score = 0.71

```
1) The earth is flat --> False (0.88)
2) Water boils at 100 degrees Celsius --> True (0.58)
3) The sky is green --> False (0.68)
4) The sky is blue --> True (0.82)
5) Renault 8 is produced by Fiat --> False (0.81)
```

QUIZ



MENTI **(8829 1189)**

https://www.menti.com/al3exyz6vt47