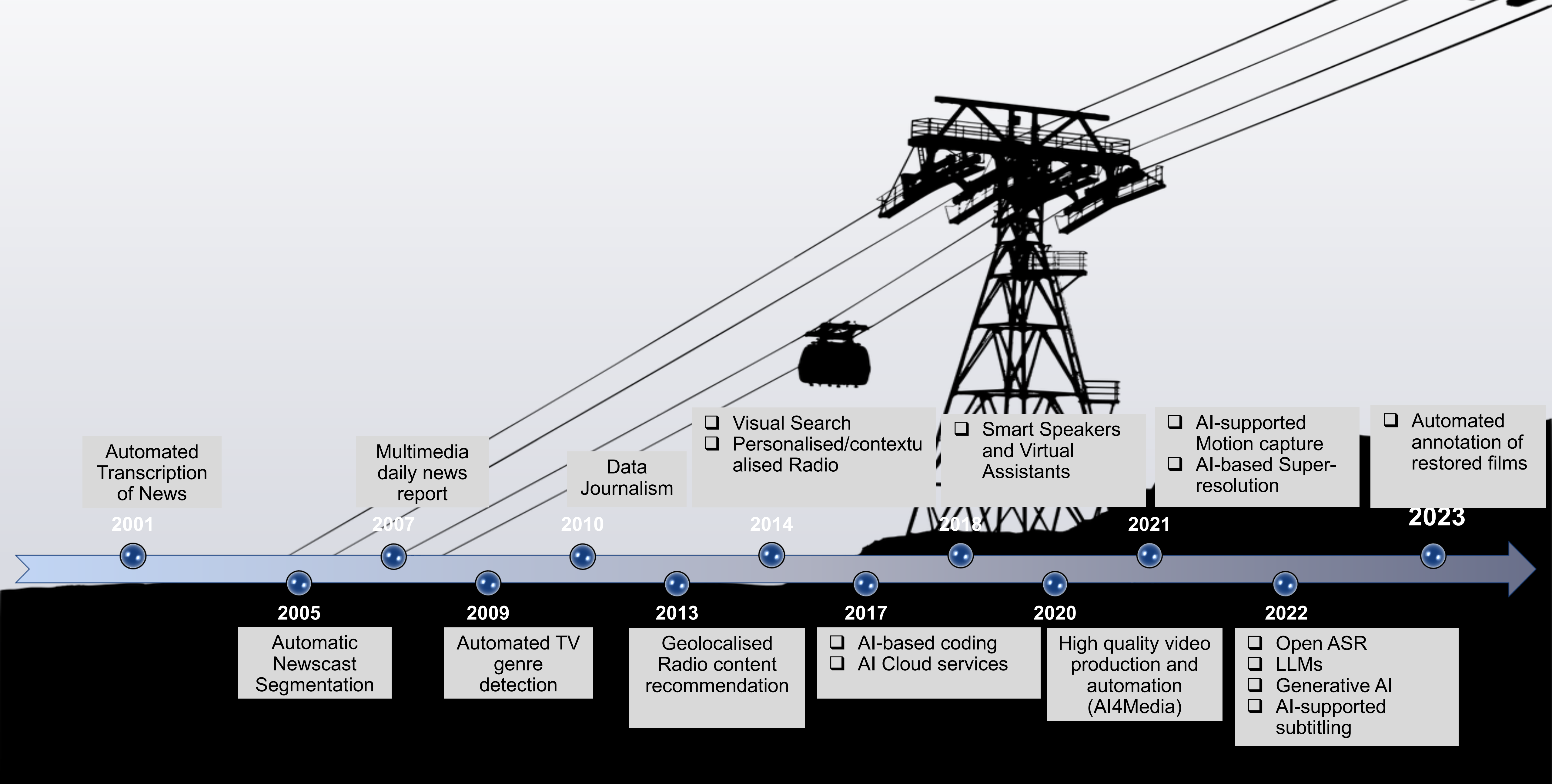
LLM finetuning and benchmark

A few examples

Alberto Messina, Stefano Scotta – CRITS RAI

21/06/2023

RAI R&D agenda on AI



2001
Automated Transcription of News

2005
Automatic Newscast Segmentation

2007
Multimedia daily news report

2009
Automated TV genre detection

2010
Data Journalism

2013
Geolocalised Radio content recommendation

2014
- ❑ Visual Search
- ❑ Personalised/contextualised Radio

2017
- ❑ AI-based coding
- ❑ AI Cloud services

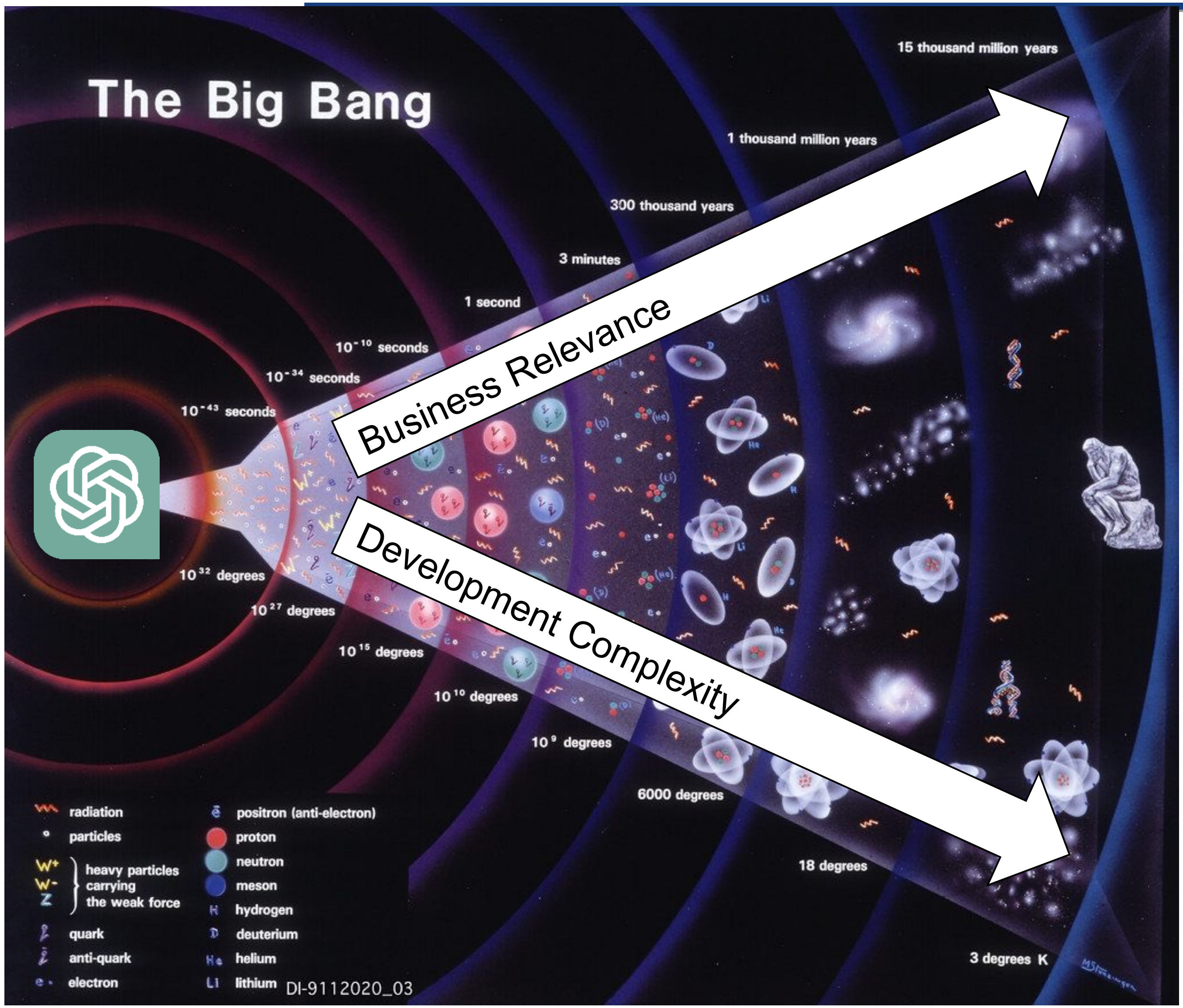
2018
- ❑ Smart Speakers and Virtual Assistants

2020
High quality video production and automation (AI4Media)

2021
- ❑ AI-supported Motion capture
- ❑ AI-based Super-resolution

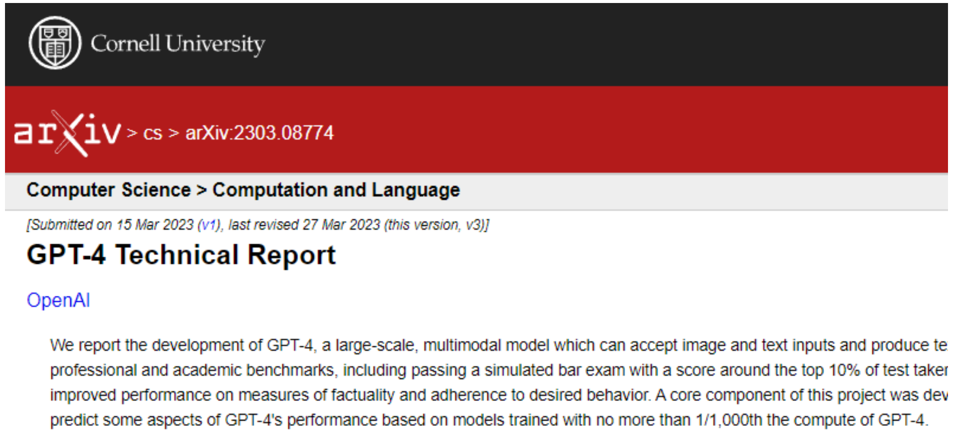
2022
- ❑ Open ASR
- ❑ LLMs
- ❑ Generative AI
- ❑ AI-supported subtitling

2023
- ❑ Automated annotation of restored films



Infinite expansion
or
big crunch?

Is GPT4 powerful? Yes indeed.



Cornell University

arXiv > cs > arXiv:2303.08774

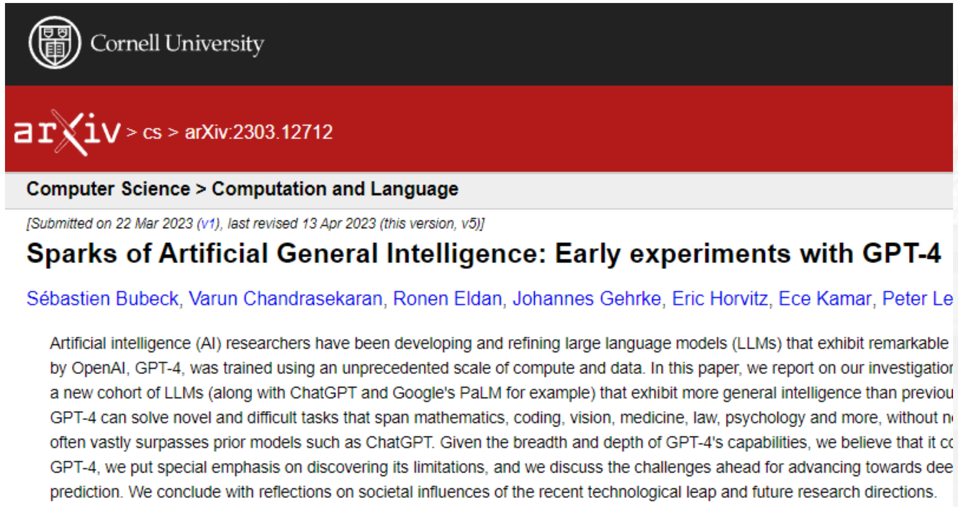
Computer Science > Computation and Language

[Submitted on 15 Mar 2023 (v1), last revised 27 Mar 2023 (this version, v3)]

GPT-4 Technical Report

OpenAI

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce te
professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test taker
improved performance on measures of factuality and adherence to desired behavior. A core component of this project was dev
predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.



Cornell University

arXiv > cs > arXiv:2303.12712

Computer Science > Computation and Language

[Submitted on 22 Mar 2023 (v1), last revised 13 Apr 2023 (this version, v5)]

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Le

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable
by OpenAI, GPT-4, was trained using an unprecedented scale of compute and data. In this paper, we report on our investigatior
a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previou
GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without n
often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it cc
GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards dee
prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

Can GP4 be harmful? Yes Indeed.

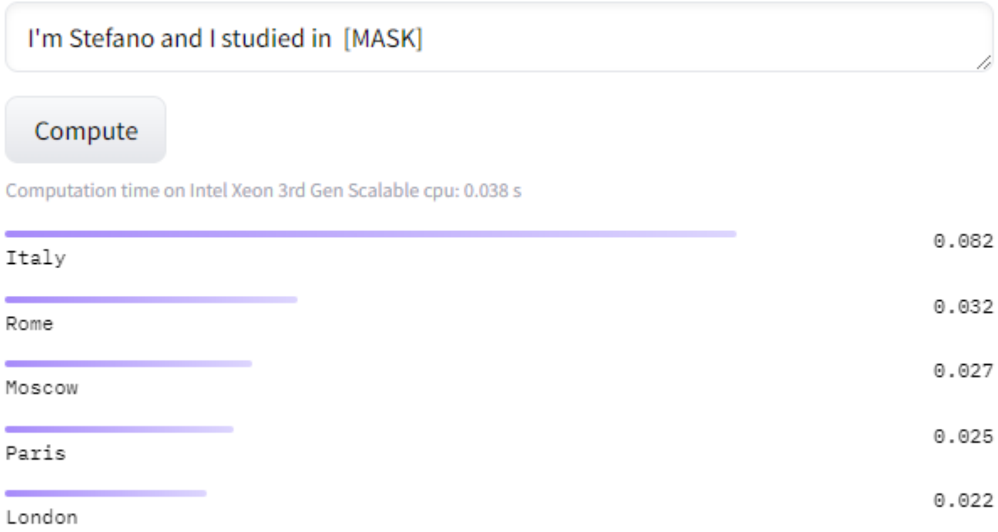
Applications of Large Language Models in Media

- News assistants
- Media annotation
- Online service enhancement
- Disinformation flow analysis
- Social media impact optimisation
- ...

The only limit is your imagination

Large Language Models

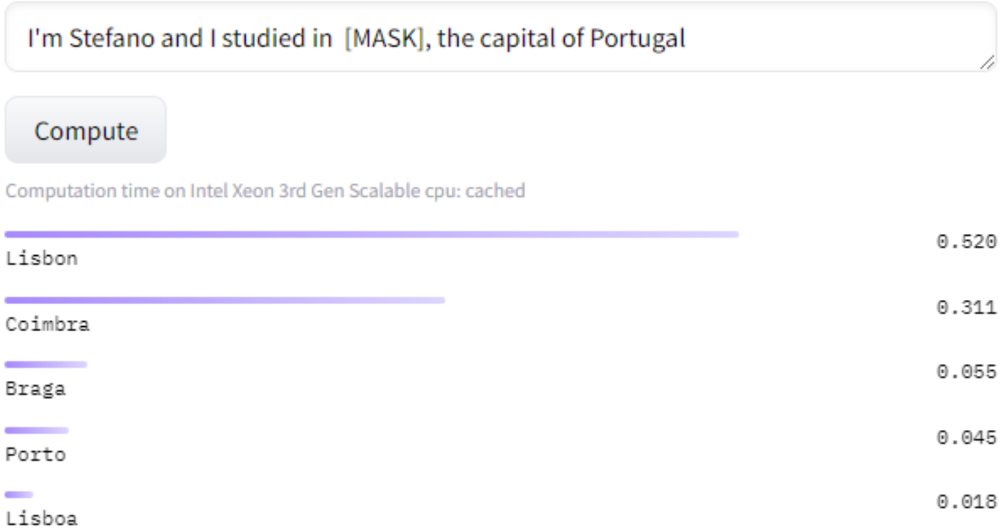
The LLMs are huge models that have absorbed an extremely high amount of information (often from unknown sources) and, above all, a formal “knowledge” of language.



I'm Stefano and I studied in [MASK]

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.038 s

```
Italy      0.082
Rome       0.032
Moscow     0.027
Paris      0.025
London     0.022
```



I'm Stefano and I studied in [MASK], the capital of Portugal

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

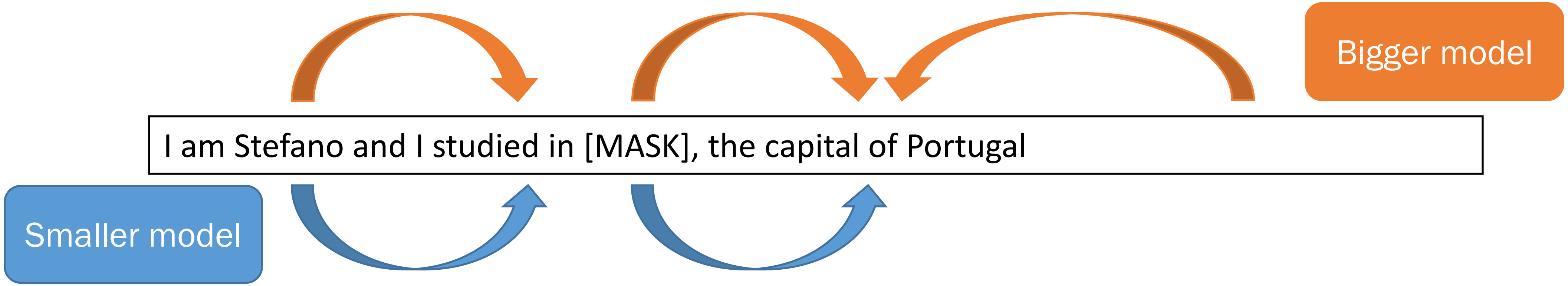
```
Lisbon     0.520
Coimbra    0.311
Braga      0.055
Porto      0.045
Lisboa     0.018
```

Actually, at their “initial stage” they are not capable of performing tasks other than **completing text in a probabilistic way**.

Example based on bert-base-multilingual-cased with **110M** parameters (gpt-3.5 has approx. **175B parameters**)

Large Language Models

In general, the bigger the model the more complex the probability distribution could be, taking in account more context (more parameters = more conditions considered).



Bigger model

I am Stefano and I studied in [MASK], the capital of Portugal

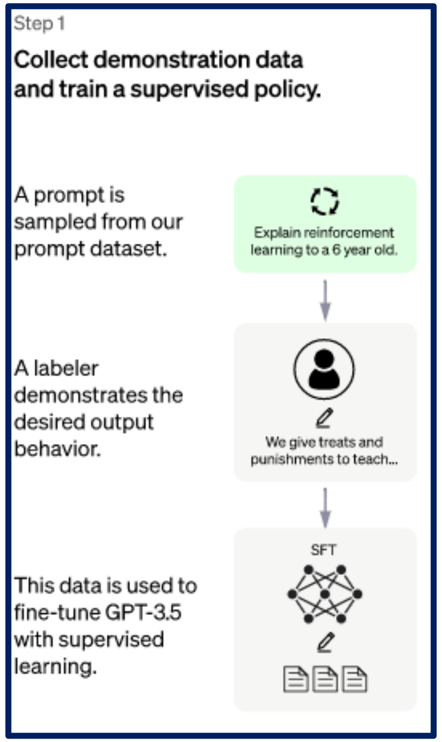
Smaller model

But **that is not all**. In order to achieve great results on human interactions or specific tasks these models have to be optimized (fine-tuned).

Fine Tuning

It is therefore necessary to optimize ("**fine-tuning**") such models to interact with humans in an easy way.

ChatGPT is the result of a fine-tuning process aimed at enabling it to "**answer *any* question or instruction**".



Step 1

Collect demonstration data and train a supervised policy.

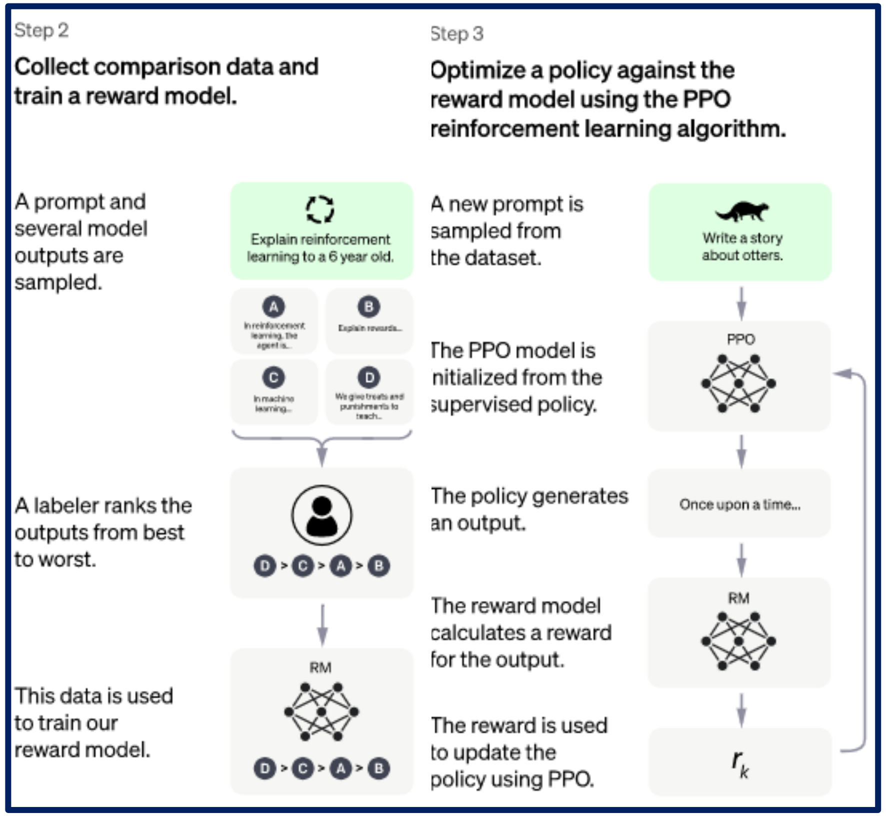
A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT

In a first phase, the model receives a huge number of instruction-response pairs from which it "learns" to answer to each instruction



Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is nitialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

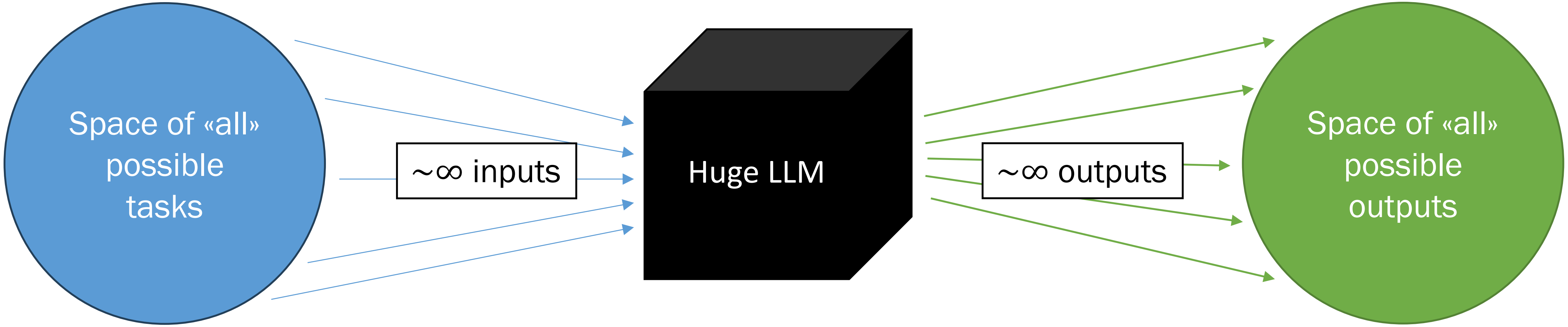
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

Subsequently, the model is further refined (**reinforcement learning**) by having it generate several responses to the same input and evaluating the "best" one. In this way, the model learns to answer "better".

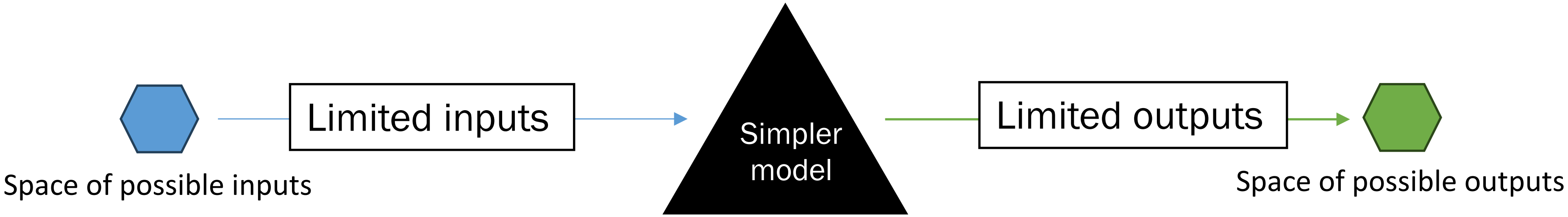
Introducing ChatGPT (openai.com)

Fine Tuning on specific tasks

In order to give a response to **any** instruction models such as ChatGPT needs a lot of resources to be used and trained/fine-tuned locally.

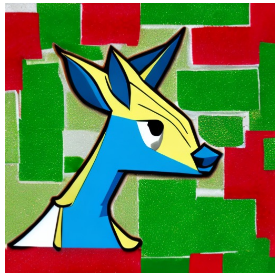

Space of «all» possible tasks
~∞ inputs
Huge LLM
~∞ outputs
Space of «all» possible outputs

Reducing the dimension of the possible inputs, tasks and outputs it is possible to use **simpler/smaller models** and it is possible doing that with much less resources.



Llama – an open source LLM

Stanford Alpaca



Llama

Open source LLM, developed by META available in various “sizes”.

Alpaca

Fine-tuned version of Llama (on 52k instructions/answers generated by a GPT model) to answer instructions and questions.

Camoscio

Similar to Alpaca but the fine tuning was performed on the translation in Italian of the instructions/responses used to obtain Alpaca.

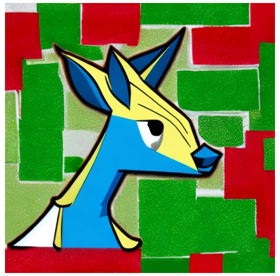
Rai

LLaMA Stanford CRFM Camoscio

Llama and Camoscio logos generated with Stable Diffusion

- Llama: “a llama cartoon drawing for a tech company logo”
- Camoscio: “a chamois cartoon drawing for a tech company logo” plus inpainting: “Green, white, red geometries in the background”

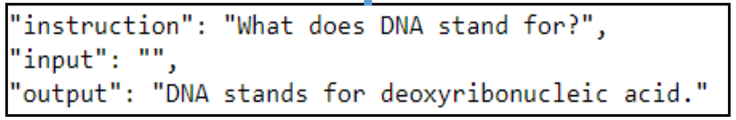
Llama – an open source LLM



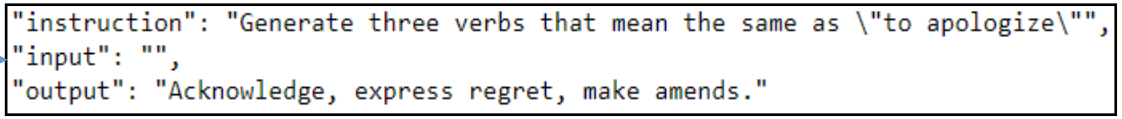
Llama

Alpaca

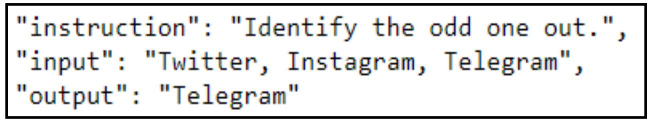
Camoscio



```
"instruction": "What does DNA stand for?",
"input": "",
"output": "DNA stands for deoxyribonucleic acid."
```



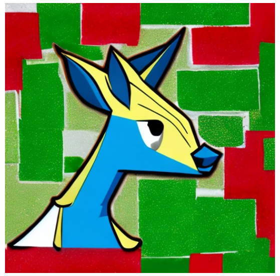
```
"instruction": "Generate three verbs that mean the same as \"to apologize\"",
"input": "",
"output": "Acknowledge, express regret, make amends."
```



```
"instruction": "Identify the odd one out.",
"input": "Twitter, Instagram, Telegram",
"output": "Telegram"
```

Llama – an open source LLM

Stanford Alpaca



Llama

Alpaca

Camoscio

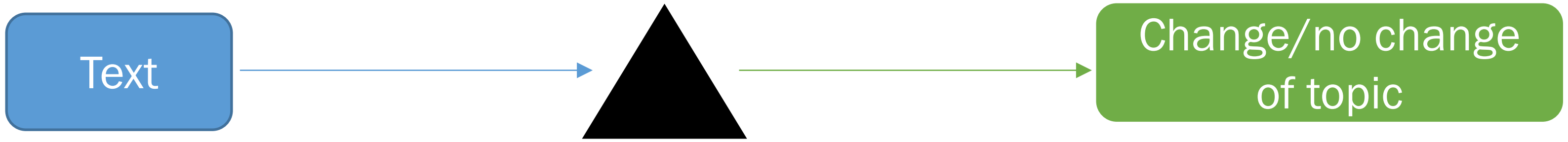
```
"instruction": "\"Categorizza l'oggetto dato in categorie di base.\"",
"input": "'Arancia'",
"output": "\"Frutta, Cibo, Verdura\""
```

```
"instruction": "\"Dimmi un esempio di un rischio che un'azienda potrebbe prendere.\"",
"input": "",
"output": "\"Un'azienda potrebbe correre il rischio di investire in una nuova tecnologia o di entrare in un nuovo mercato.\""
```

```
"instruction": "\"Nomina due sport popolari in Brasile.\"",
"input": "",
"output": "1. Calcio 2. Pallavolo sulla spiaggia"
```

Specific tasks for a F-T LLM

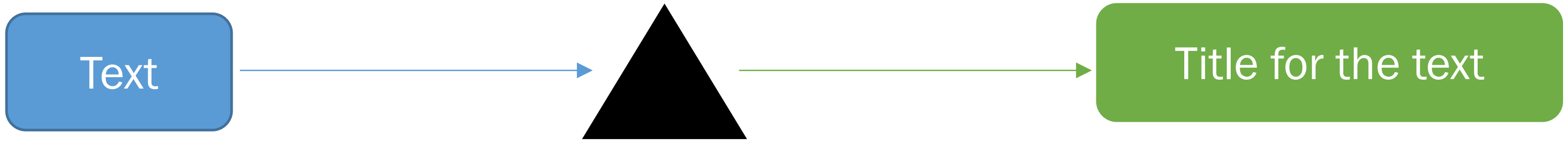
- Detecting a **change of topic** in a text



Text

Change/no change of topic

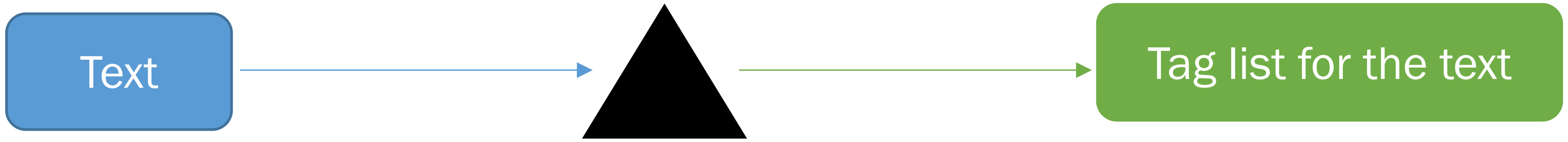
- Proposing a **title** to a news article



Text

Title for the text

- Proposing a list of significant **tags** for a news article given.

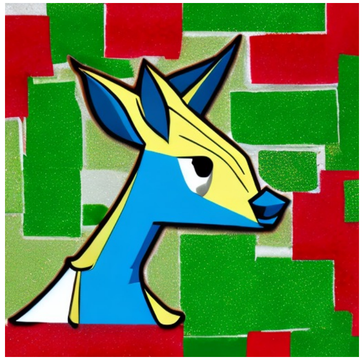


Text

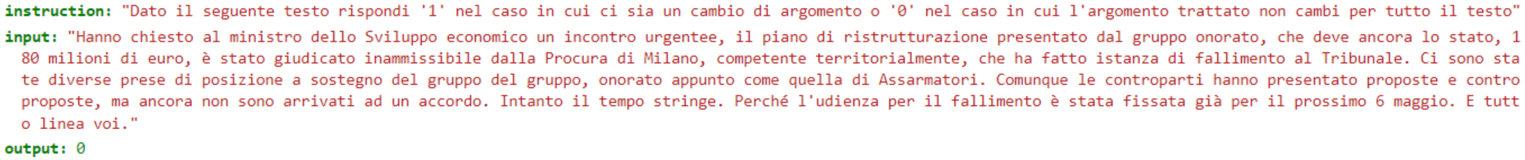
Tag list for the text

Change of Topic detector

Fine-tuned version of Chamois for assigning journalistic titles to news items (texts).



~9000 labelled texts.
Of which ~20% with change of topic



```
instruction: "Dato il seguente testo rispondi '1' nel caso in cui ci sia un cambio di argomento o '0' nel caso in cui l'argomento trattato non cambi per tutto il testo"
input: "Hanno chiesto al ministro dello Sviluppo economico un incontro urgentee, il piano di ristrutturazione presentato dal gruppo onorato, che deve ancora lo stato, 1
  80 milioni di euro, è stato giudicato inammissibile dalla Procura di Milano, competente territorialmente, che ha fatto istanza di fallimento al Tribunale. Ci sono sta
  te diverse prese di posizione a sostegno del gruppo del gruppo, onorato appunto come quella di Assarmatori. Comunque le controparti hanno presentato proposte e contro
  proposte, ma ancora non sono arrivati ad un accordo. Intanto il tempo stringe. Perché l'udienza per il fallimento è stata fissata già per il prossimo 6 maggio. E tutt
  o linea voi."
output: 0
```

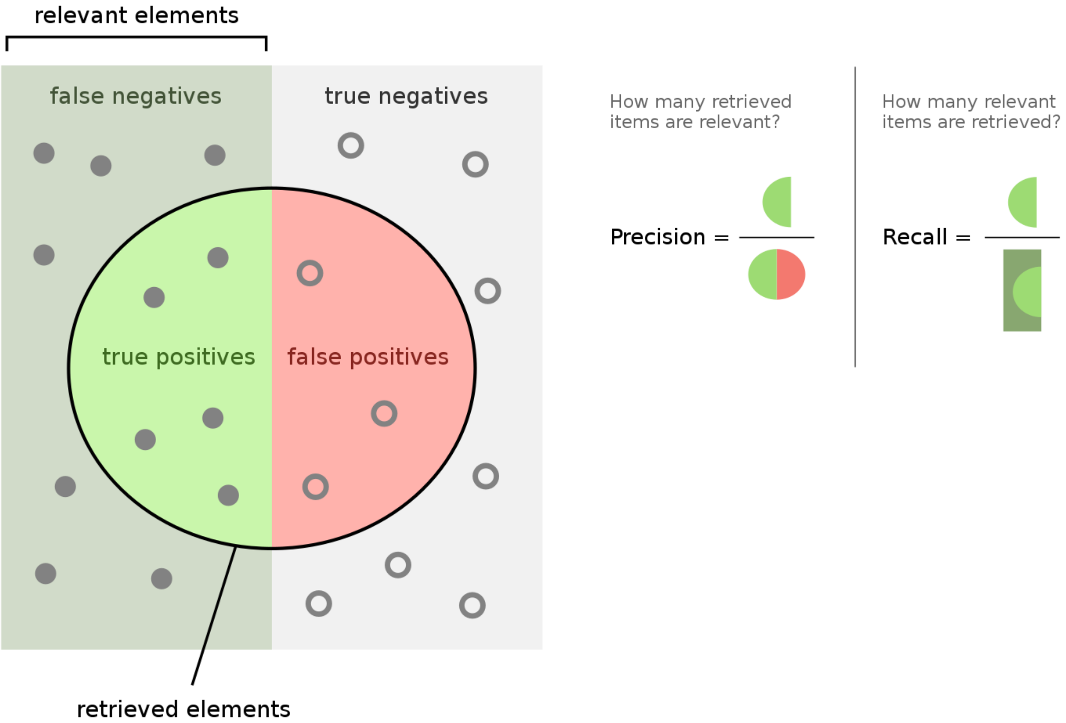
CT detector

Logo Titler generated from Camoscio logo with background inpainting process:
“newspapers first pages with big titles emphasized. Blue and white colors”

Evaluating CT detector

Being the identification of topic changes a **classification task** we used the usual metrics to evaluate the quality of the fine-tuned model on a test set of $\sim$1000 labeled texts.

- **Precision** = $\frac{TP}{TP+FP}$ = **0.79**
- **Recall** = $\frac{TP}{TP+FN}$ = **0.62**
- **Accuracy** = $\frac{TP+TN}{TP+TN+FP+FN}$ = **0.90**
- **F1-Score** = $2 \cdot \frac{1}{\frac{1}{precision}+\frac{1}{recall}}$ = **0.70**



Comparison with GPT-4

We used OpenAI's gpt-4 for the same task and on the same test-set.

Using OpenAI's model via Azure API means deal with the content filter implemented by Azure which "censors" many question/answer. In this case around **4%**.

- **Precision = 0.79**
- Recall = 0.62
- **Accuracy = 0.90**
- **F1-Score = 0.69**

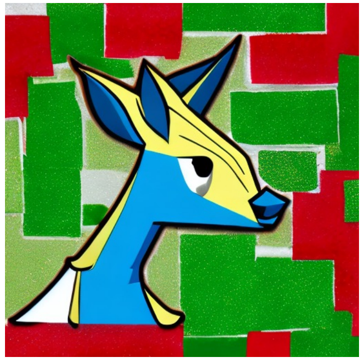
CT detector

- Precision = 0. 31
- **Recall = 0.88**
- Accuracy = 0.61
- F1-Score = 0.46

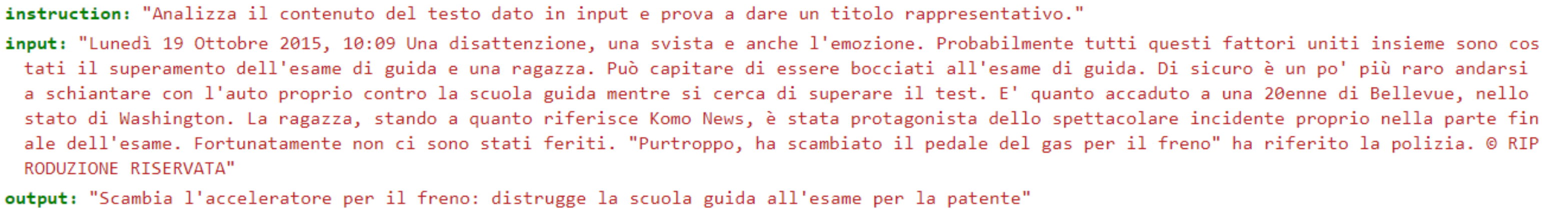
GPT 4

Titler

Fine-tuned version of Camoscio for assigning titles to news articles (texts).



~20000 news article/title couples from Rai News channels



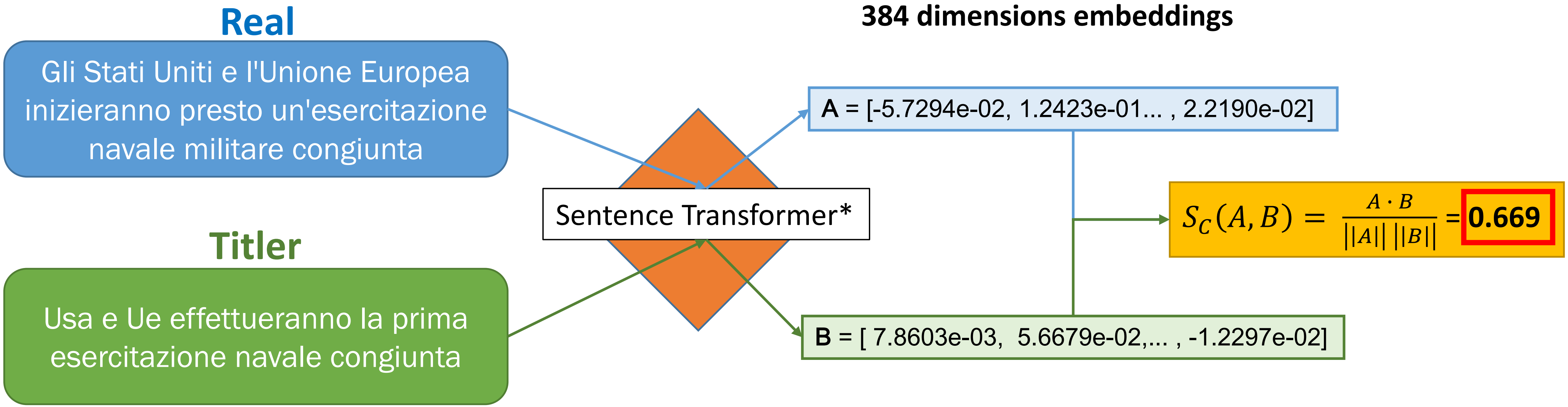
```
instruction: "Analizza il contenuto del testo dato in input e prova a dare un titolo rappresentativo."
input: "Lunedì 19 Ottobre 2015, 10:09 Una disattenzione, una svista e anche l'emozione. Probabilmente tutti questi fattori uniti insieme sono cos
  tati il superamento dell'esame di guida a una ragazza. Può capitare di essere bocciati all'esame di guida. Di sicuro è un po' più raro andarsi
  a schiantare con l'auto proprio contro la scuola guida mentre si cerca di superare il test. E' quanto accaduto a una 20enne di Bellevue, nello
  stato di Washington. La ragazza, stando a quanto riferisce Komo News, è stata protagonista dello spettacolare incidente proprio nella parte fin
  ale dell'esame. Fortunatamente non ci sono stati feriti. "Purtroppo, ha scambiato il pedale del gas per il freno" ha riferito la polizia. © RIP
  RODUZIONE RISERVATA"
output: "Scambia l'acceleratore per il freno: distrugge la scuola guida all'esame per la patente"
```

Titler

Logo Titler generated from Camoscio logo with background inpainting process: “newspapers first pages with big titles emphasized. Blue and white colors”

Evaluating the Titler

The quality of a model that generates titles is much more complicated to «measure». We considered as ground truth the real titles of the articles used to test the model, evaluating the **cosine similarity** between the embeddings of the real title and the one proposed by the model.



Real

Gli Stati Uniti e l'Unione Europea inizieranno presto un'esercitazione navale militare congiunta

Titler

Usa e Ue effettueranno la prima esercitazione navale congiunta

Sentence Transformer*

384 dimensions embeddings

A = [-5.7294e-02, 1.2423e-01... , 2.2190e-02]

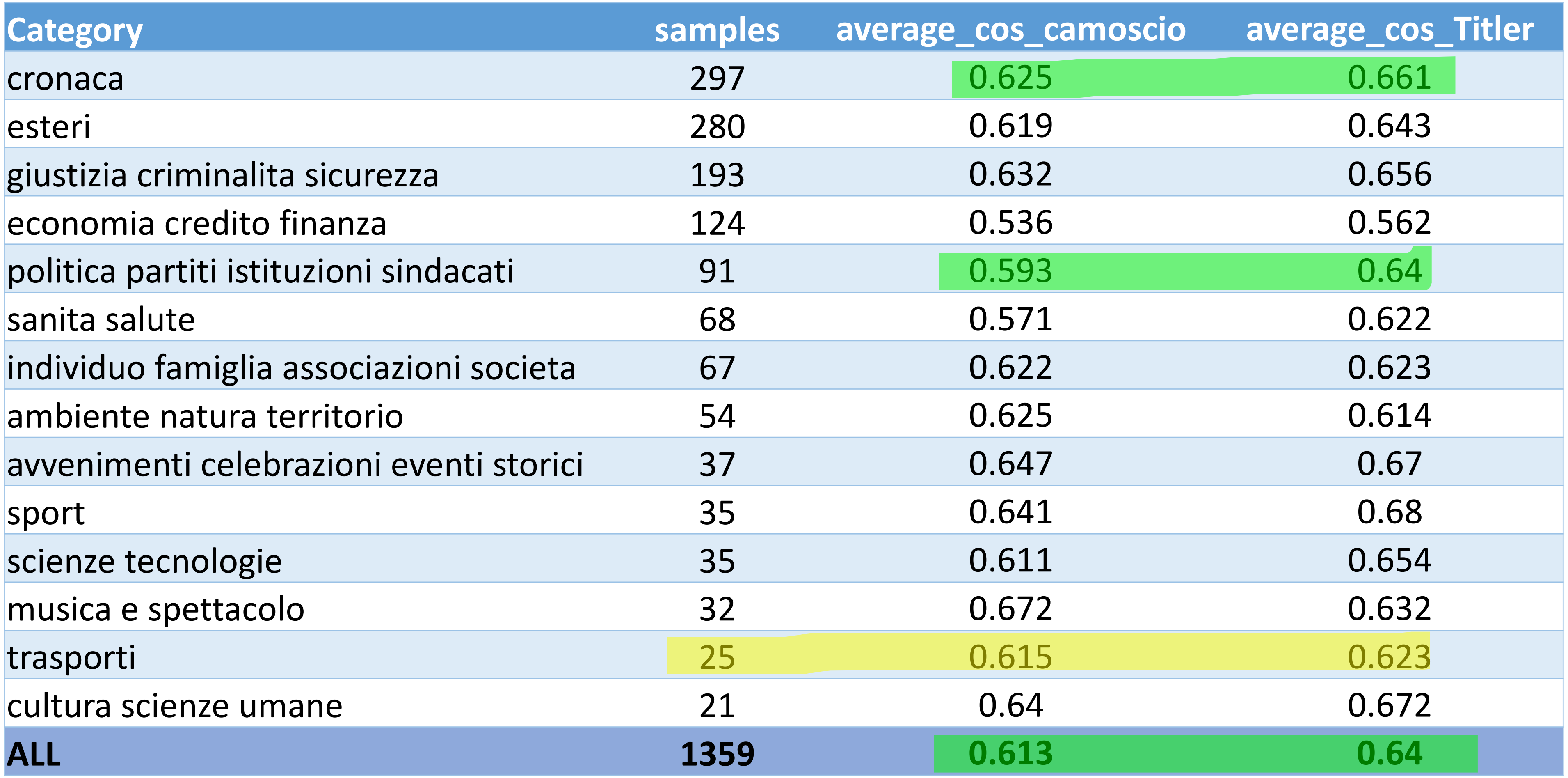
B = [7.8603e-03, 5.6679e-02,... , -1.2297e-02]

$$S_C(A,B) = \frac{A \cdot B}{||A||\,||B||} = \boxed{\mathbf{0.669}}$$

This kind of measure has not an “absolute” value, but it is surely useful to **compare different models**.

sentence-transformers/all-MiniLM-L6-v2 · Hugging Face *

Fine Tuning effectiveness



Category	samples	average_cos_camoscio	average_cos_Titler
cronaca	297	0.625	0.661
esteri	280	0.619	0.643
giustizia criminalita sicurezza	193	0.632	0.656
economia credito finanza	124	0.536	0.562
politica partiti istituzioni sindacati	91	0.593	0.64
sanita salute	68	0.571	0.622
individuo famiglia associazioni societa	67	0.622	0.623
ambiente natura territorio	54	0.625	0.614
avvenimenti celebrazioni eventi storici	37	0.647	0.67
sport	35	0.641	0.68
scienze tecnologie	35	0.611	0.654
musica e spettacolo	32	0.672	0.632
trasporti	25	0.615	0.623
cultura scienze umane	21	0.64	0.672
ALL	**1359**	**0.613**	**0.64**

Original articles and headlines from RaiNews covering the days between 10/03 and 04/05, on categories with at least 20 articles

Comparison with GPT-4

category	samples	average_cos_gpt4	average_cos_Titler
esteri	227	0.652	0.641
cronaca	201	0.652	0.665
giustizia criminalita sicurezza	149	0.661	0.649
economia credito finanza	113	0.567	0.554
politica partiti istituzioni sindacati	79	0.615	0.639
individuo famiglia associazioni societa	62	0.622	0.63
sanita salute	55	0.619	0.618
ambiente natura territorio	46	0.613	0.611
sport	31	0.67	0.668
scienze tecnologie	28	0.689	0.65
musica e spettacolo	28	0.674	0.619
avvenimenti celebrazioni eventi storici	27	0.67	0.653
trasporti	21	0.581	0.621
ALL	**1067**	**0.637**	**0.635**

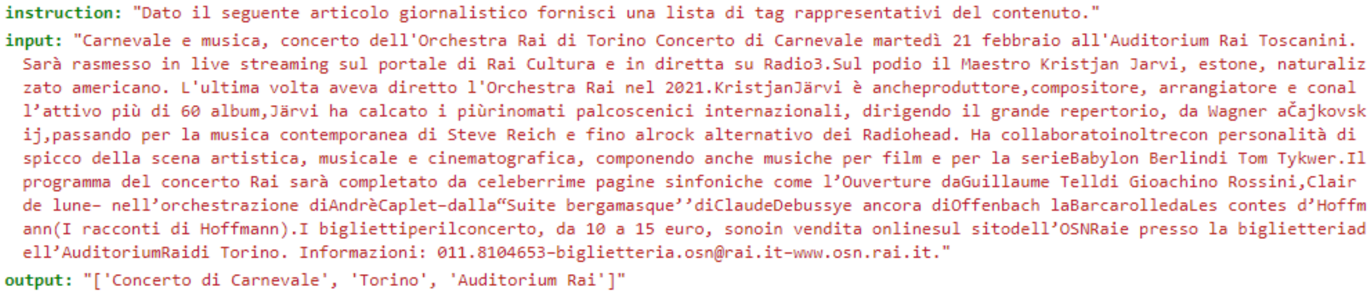
More than 20% of articles "victim" of Azure content filter

Original articles and headlines from RaiNews covering the days between 10/03 and 04/05, on categories with at least 20 articles

Tagger

Fine-tuned version of Camoscio for assigning tags to news items (texts).

~20000 news article/tags couples from Rai News channels



```
instruction: "Dato il seguente articolo giornalistico fornisci una lista di tag rappresentativi del contenuto."
input: "Carnevale e musica, concerto dell'Orchestra Rai di Torino Concerto di Carnevale martedì 21 febbraio all'Auditorium Rai Toscanini.
  Sarà rasmesso in live streaming sul portale di Rai Cultura e in diretta su Radio3.Sul podio il Maestro Kristjan Jarvi, estone, naturaliz
  zato americano. L'ultima volta aveva diretto l'Orchestra Rai nel 2021.KristjanJärvi è ancheproduttore,compositore, arrangiatore e conal
  l’attivo più di 60 album,Järvi ha calcato i piùrinomati palcoscenici internazionali, dirigendo il grande repertorio, da Wagner aČajkovsk
  ij,passando per la musica contemporanea di Steve Reich e fino alrock alternativo dei Radiohead. Ha collaboratoinoltrecon personalità di
  spicco della scena artistica, musicale e cinematografica, componendo anche musiche per film e per la serieBabylon Berlindi Tom Tykwer.Il
  programma del concerto Rai sarà completato da celeberrime pagine sinfoniche come l’Ouverture daGuillaume Telldi Gioachino Rossini,Clair
  de lune- nell’orchestrazione diAndrèCaplet-dalla“Suite bergamasque’’diClaudeDebussye ancora diOffenbach laBarcarolledaLes contes d’Hoffm
  ann(I racconti di Hoffmann).I bigliettiperilconcerto, da 10 a 15 euro, sonoin vendita onlinesul sitodell’OSNRaie presso la biglietteriad
  ell’AuditoriumRaidi Torino. Informazioni: 011.8104653-biglietteria.osn@rai.it-www.osn.rai.it."
output: "['Concerto di Carnevale', 'Torino', 'Auditorium Rai']"
```

Tagger

Logo Tagger generated from Camoscio logo with background inpainting process:
“hashtags list newspaper titles. blue and white colours”

Evaluating the Tagger

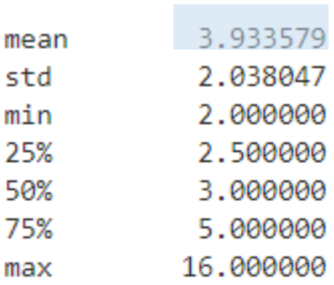
The output of the tagger is a list of words, so it can not be evaluated as the Titler.

We took in account **different measures** to compare the result of our model with the result, for the same task, of gpt-4 and Camoscio (not fine-tuned).

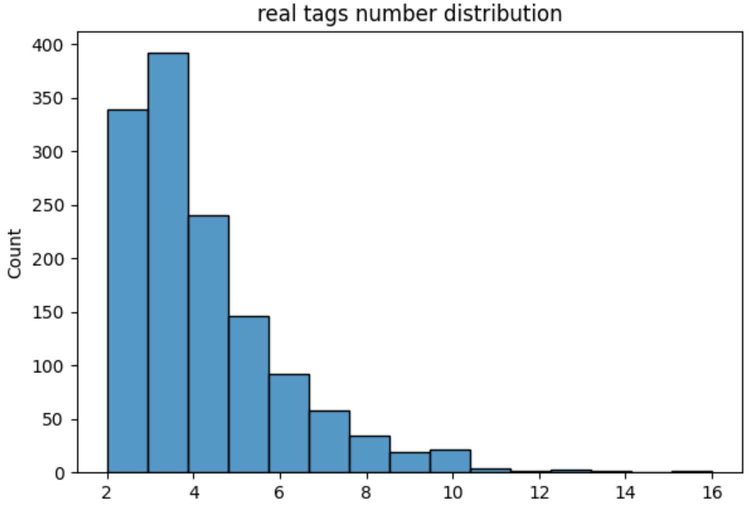
NB: we always considered as «gound truth» the real tags assigned to the articles.

Evaluating the Tagger

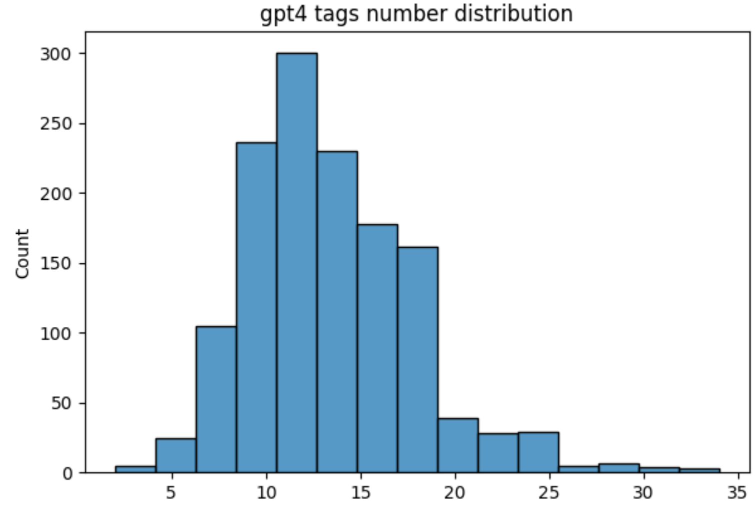
Number of tags distribution



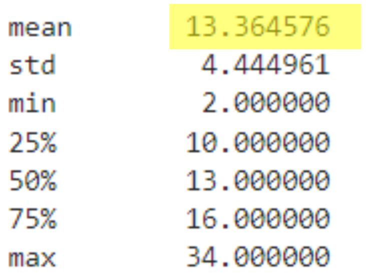
```
mean      3.933579
std       2.038047
min       2.000000
25%       2.500000
50%       3.000000
75%       5.000000
max      16.000000
```



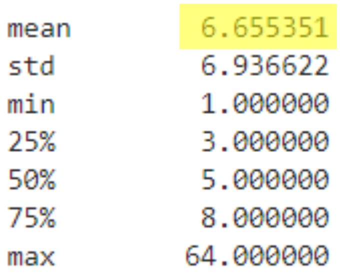
real tags number distribution



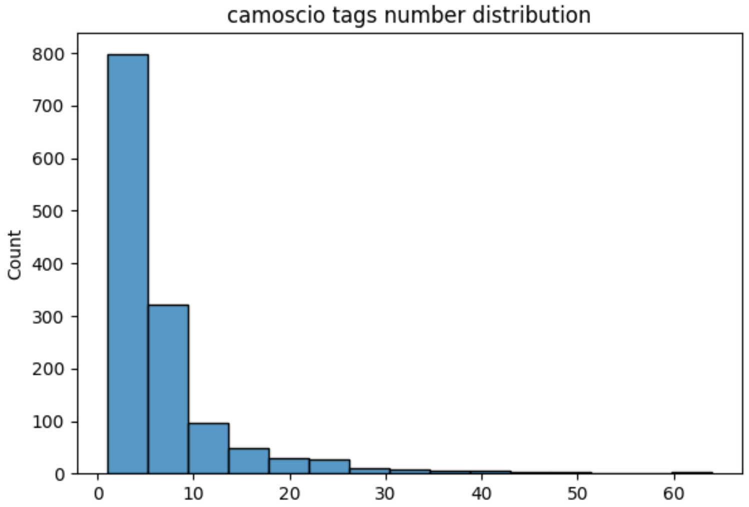
gpt4 tags number distribution



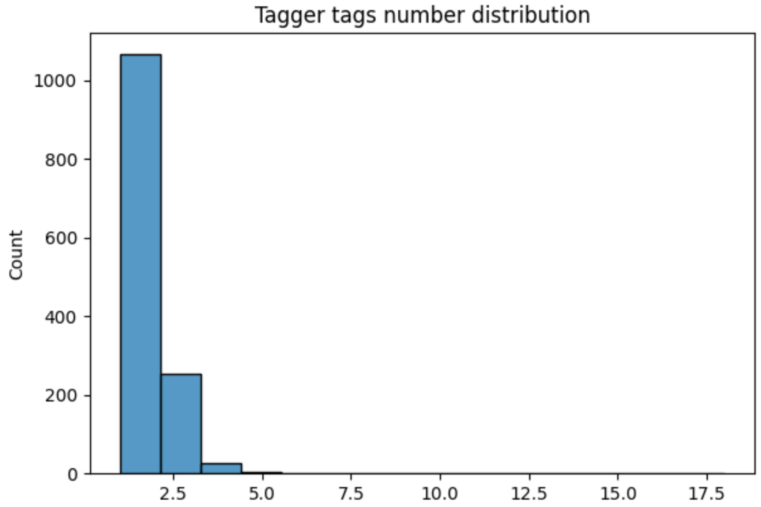
```
mean     13.364576
std       4.444961
min       2.000000
25%      10.000000
50%      13.000000
75%      16.000000
max      34.000000
```



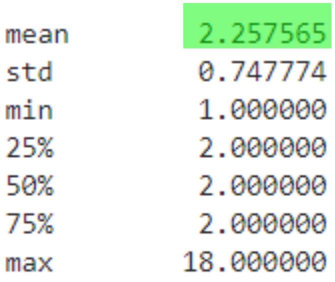
```
mean      6.655351
std       6.936622
min       1.000000
25%       3.000000
50%       5.000000
75%       8.000000
max      64.000000
```



camoscio tags number distribution



Tagger tags number distribution



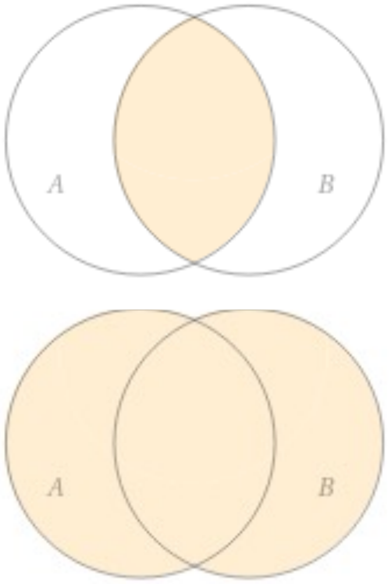
```
mean      2.257565
std       0.747774
min       1.000000
25%       2.000000
50%       2.000000
75%       2.000000
max      18.000000
```

Evaluating the Tagger

Intersection over union

A ≔ set of real tags assigned to an article
B ≔ set of tags assigned by the model to an article

$$IoU_B(article) \coloneqq \frac{|A \cap B|}{|A \cup B|} =$$ fraction of tags assigned by the model equal to the real ones over the total number of tags



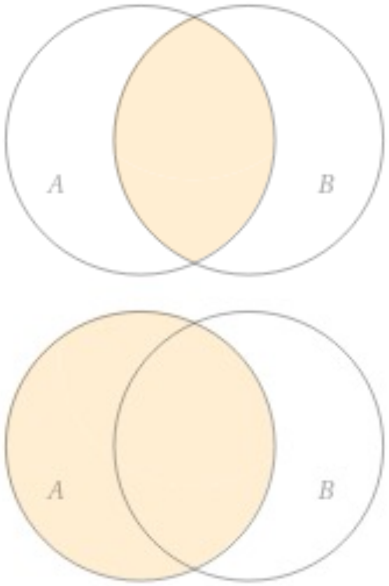
	GPT 4	Camoscio	Tagger
Mean value of IoU for the model considered:	0.094	0.076	**0.270**

Evaluating the Tagger

Intersection over ground truth

A ≔ set of real tags assigned to an article
B ≔ set of tags assigned by the model to an article

$$IoG_B(article) \coloneqq \frac{|A \cap B|}{|A|} =$$ fraction of tags assigned by the model equal to the real ones over the total number of real tags



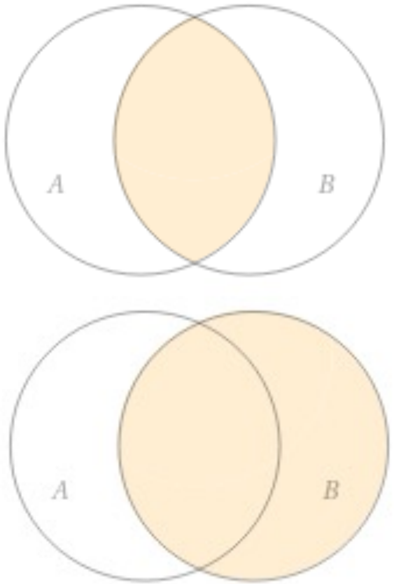
	GPT 4	Camoscio	Tagger
Mean value of IoG for the model considered:	0. 364	0.174	0. 324

Evaluating the Tagger

Intersection over model tags

A ≔ set of real tags assigned to an article
B ≔ set of tags assigned by the model to an article

$$IoM_B(article) := \frac{|A \cap B|}{|B|} =$$ fraction of tags assigned by the model which are real



	GPT 4	Camoscio	Tagger
Mean value of IoM for the model considered:	0.111	0.110	**0.499**

Evaluating the Tagger

Normalized Levenshtein distance between sets of words

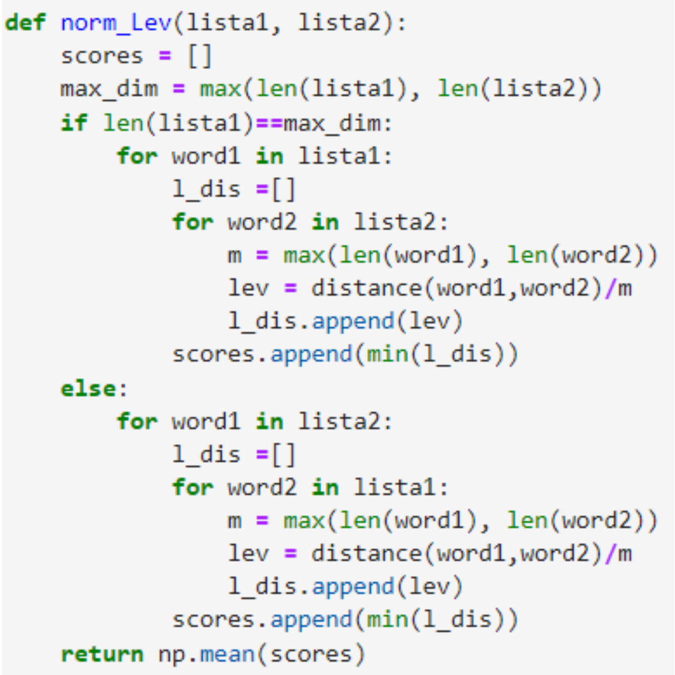
$d_L(word_1, word_2)$ $:=$ number of letter to change to pass from one to the other word

Ex: $d_L(house, home) = 3$

A $:=$ set of real tags assigned to an article

B $:=$ set of tags assigned by the model to an article

$$Lev(A,B) := \begin{cases} \dfrac{1}{|A|} \displaystyle\sum_{w^1 \in A} \min_{w^2 \in B} \left\{ \dfrac{d_L(w^1, w^2)}{\max\{|\mathrm{w}^1|, |\mathrm{w}^2|\}} \right\} & if \quad |A| > |B| \\ \dfrac{1}{|B|} \displaystyle\sum_{w^1 \in B} \min_{w^2 \in A} \left\{ \dfrac{d_L(w^1, w^2)}{\max\{|\mathrm{w}^1|, |\mathrm{w}^2|\}} \right\} & if \quad |B| > |A| \end{cases}$$



```python
def norm_Lev(lista1, lista2):
    scores = []
    max_dim = max(len(lista1), len(lista2))
    if len(lista1)==max_dim:
        for word1 in lista1:
            l_dis =[]
            for word2 in lista2:
                m = max(len(word1), len(word2))
                lev = distance(word1,word2)/m
                l_dis.append(lev)
            scores.append(min(l_dis))
    else:
        for word1 in lista2:
            l_dis =[]
            for word2 in lista1:
                m = max(len(word1), len(word2))
                lev = distance(word1,word2)/m
                l_dis.append(lev)
            scores.append(min(l_dis))
    return np.mean(scores)
```

	GPT 4	Camoscio	Tagger
Mean value of *Lev* for the model considered:	0.579	0.644	0.433

Future works

- Fine tuning other models on the same and other tasks
- Experimenting new techniques of finetuning allowing to fine tune bigger models (qlora)
- Elaborate new techniques to benchmark models on more general tasks
- Combine models to improve results

QLoRA: Efficient Finetuning of Quantized LLMs