A quick overview of Natural Language Processing for Information Extraction and Scientific Surveillance

Xavier Tannier
xavier.tannier@sorbonne-universite.fr

Orphanet, September 17, 2024

Appetizer

1. Allez sur wooclap.com
2. Entrez le code d'événement dans le bandeau supérieur

Code d'événement
UXPXKF

Activer les réponses par SMS

Why Natural Language Processing?

ASSISTANCE HÔPITAUX PUBLIQUE DE PARIS
HÔTEL - DIEU

SCINTIGRAPHIE AU GALLIUM 67

Centre Hospitalier Régional Universitaire
7, rue Pasteur
67001 STRASBOURG Cedex

COMPTE RENDU OPERATOIRE
En date du : 15/06/2010

Cholécystectomie - Cholangiographie - Cœlioscopie

Top sujets

Sujets de la catégorie Diabète

Le Parc Astérix fête ses 30 ans, venez passer un bon moment en famille

Relation Extraction from News Articles (RENA): A Tool for Epidemic Surveillance

ALIGNING LARGE LANGUAGE MODELS FOR CLINICAL TASKS

Supun Manathunga, University of Peradeniya, NeuraSense Research
Isuru Hettigoda, NeuraSense Research

NLP tasks

9 hémocultures positives le 26/6/15 à staphylocoque aureus méticilline sensible. BACTERIAL RESISTANCE

Patient sans signe clinique évident de traumatisme crânien
[C0018674] (Craniocerebral Trauma)

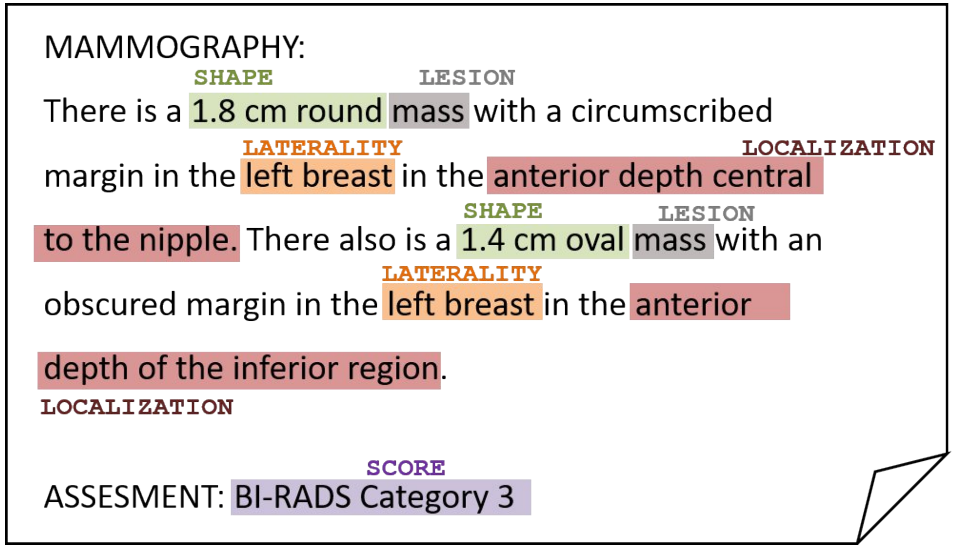
VASCULAR DISEASE
Patient atteint d'ulcères artériels des membres inférieurs, suivis à St Joseph (Dr Wyliana) avec greffe cutanée en octobre 2015

Patiente avec antécédent de chirurgie bariatrique
[C1456587] (Bariatric Surgery)

PROCEDURE MEDICAL DEVICE ANATOMY
La contraception par les dispositifs intra utérins

localization
Des nodules suspects sont observés dans le sein droit.
localization
L'examen est noté ACR 2 à gauche et à droite.
localization

NLP tasks



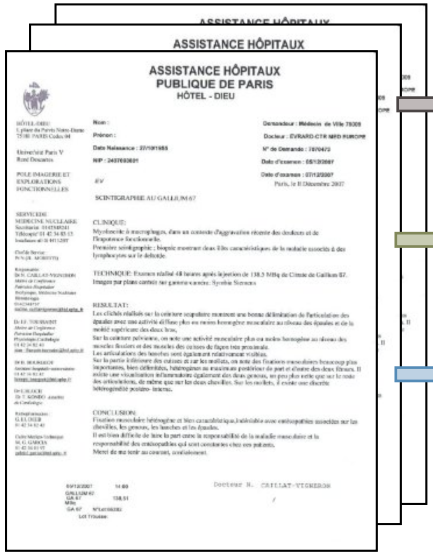
SCORE	LATERALITY	LESION	LATERALITY	LOCALIZATION	SHAPE
BI-RADS 3	LEFT	mass	LEFT	QSI	1.8
BI-RADS 1	RIGHT	mass	LEFT	QI	1.4

NLP tasks


ASSISTANCE HÔPITAUX
PUBLIQUE DE PARIS
HÔTEL - DIEU

Reason for admission = low-kinetic fracture?

Family history?


ASSISTANCE HÔPITAUX
PUBLIQUE DE PARIS
HÔTEL - DIEU

Diabetic patient?

Smoker?

Able to make his/her own decision?

Objectives

Pseudonymising patient records

Structuring data

Indexing & Retrieving

Selecting similar patients

Selecting patients matching criteria

Methods

1. **Rule-based systems**
2. Supervised learning systems
3. Generative, large language models
4. Retrieval-Augmented Generation

Rule-based systems

1. Terminological approach

VASCULAR DISEASE

- *intermittent claudication*
- *arteriopathy of the limbs*
- *arteritis of limbs*
- *arterial ulcer*
- *venous ulcer*
- *arterial insufficiency*
- *gangrene*
- *acute limb ischemia*
- *acute peripheral ischemia*
- *aortic aneurysm*

(Approximate) term matching in the documents

VASCULAR DISEASE

Patient with arterial ulcers of the lower limbs, followed at St Joseph with skin graft in October 2015

Rule-based systems

1. Terminological approach
2. Additional rules

BACTERIA RESISTANCE

- *<BACTERIA> + .* sensitive*
- *<BACTERIA> + .* resistant*
- *<BACTERIA> + sensitive to methicillin*
- …

9 blood cultures positive on 6/26/15 for methicillin-sensitive Staphylococcus aureus.

BACTERIA RESISTANCE

Rule-based systems

1. Terminological approach
2. Additional rules
3. Trickier rules

Emmanuelle Kempf et al., 2023

secondary, progressive, fractured or suspected osteocondensation	*ost[eé]ocondensa. {1,20} (suspect|secondaire|[ée]volutive)|(l[ée]sion|anomalie|image). {1,20}os. {1,30} (suspect|secondaire|[ée]volutive)|os. {1,30} (l[ée]sion|anomalie|image). {1,20} (suspect|secondaire|[ée]volutive)|(l[ée]sion|anomalie|image). {1,20}L[I,Y]tique|(l[ée]sion|anomalie|image). {1,20}condensant. {1,20} (suspect|secondaire|[ée]volutive)|fracture. {1,30} (suspect|secondaire|[ée]volutive)|((l[I[ée]sion|anomalie|image|nodule). {1,80} (secondaire))|((l[I[ée]sion|anomalie|image|nodule)s.{1,40}suspec?ts?).*
post-operative anatomopathological tumor stage (pTNM)	*([ycpP]{1,2}\s? (T([01234x]|is)[abcdx]?) [,\s] {0,2} [ycp] {0,2}\s? (N[xo01234\+][abcdx]?) *\s? (M[o01]? [\+x]?)?)|((T([01234x]|is)[abcdx]?) [,\s] {0,2} [ycp] {0,2}\s? (N[xo01234\+][abcdx]?) \s? (M[o01]? [\+x]?)?)*

Rule-based systems: Pros & Cons

	Rules
General performance	(yellow)
Ease of implementation	(green)
Need for human expertise	(yellow)
Explainability	(green)
Material resources	(green)
Energy consumption	(green)
Ease of maintenance	(yellow)
Generalization to a different problem/context	(red)

Methods

1. Rule-based systems
2. **Supervised learning systems**
3. Generative, large language models
4. Retrieval-Augmented Generation

(What is supervised learning?)

Input

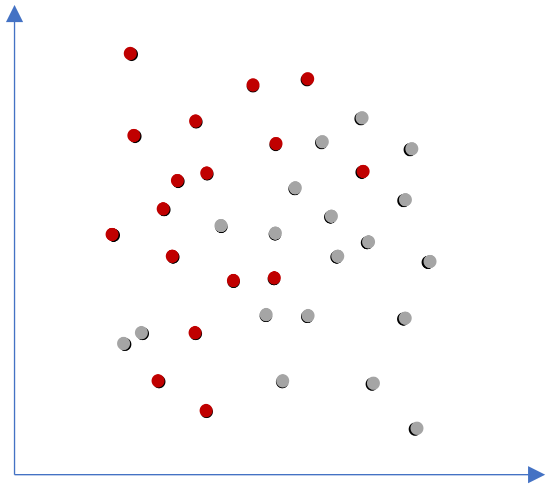
Output

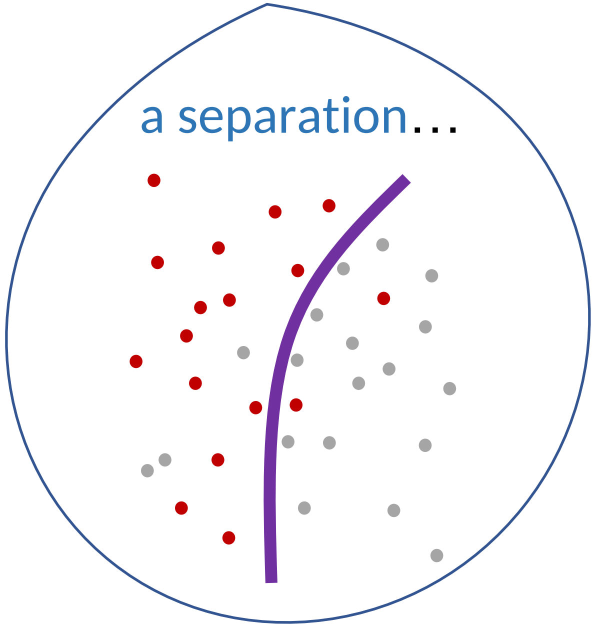
Objects represented by variables (*features*)

Classes associated with each instance (*annotated/labeled data*)

Training (“learning”)

A *model* “trained” on labeled objects, capable of distinguishing the classes



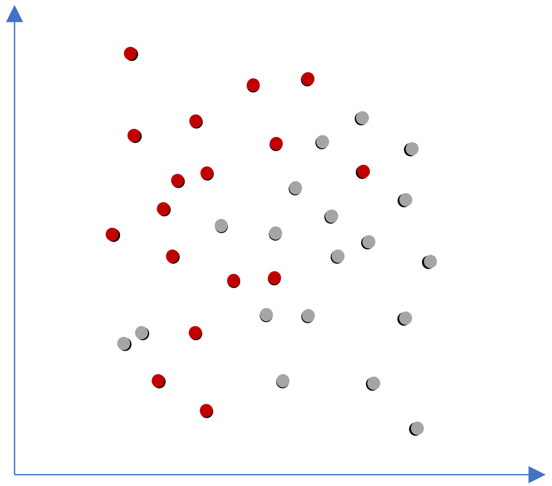


(What is supervised learning?)

Input

Objects represented by variables (*features*)

Classes associated with each instance (*annotated/labeled data*)

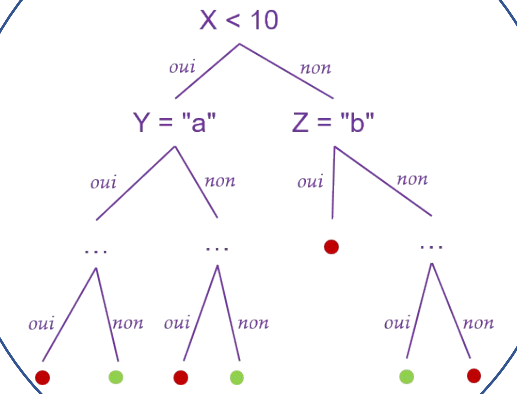


Training
(“learning”)

Output

A *model* “trained” on labeled objects, capable of distinguishing the classes

...a decision tree...



eg : J48

(What is supervised learning?)

Input

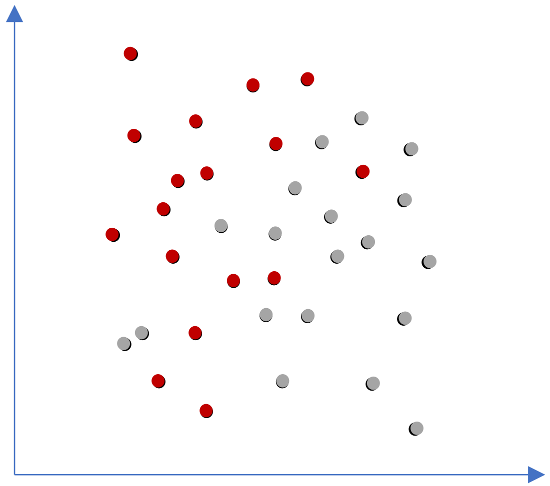
Objects represented by variables (*features*)

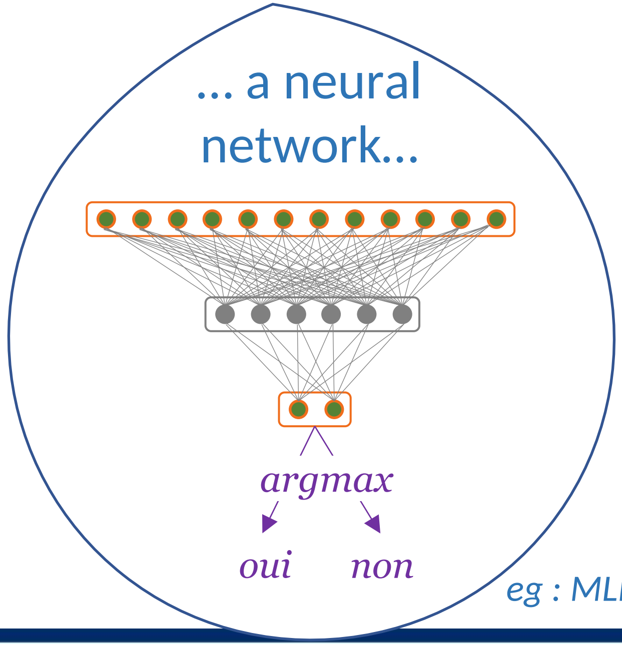
Classes associated with each instance (*annotated/labeled data*)

Training
("learning")

Output

A *model* "trained" on labeled objects, capable of distinguishing the classes





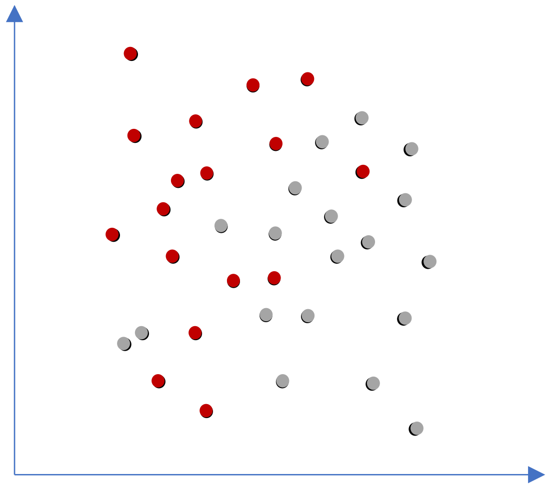
eg : MLP

(What is supervised learning?)

Input

Objects represented by variables (*features*)

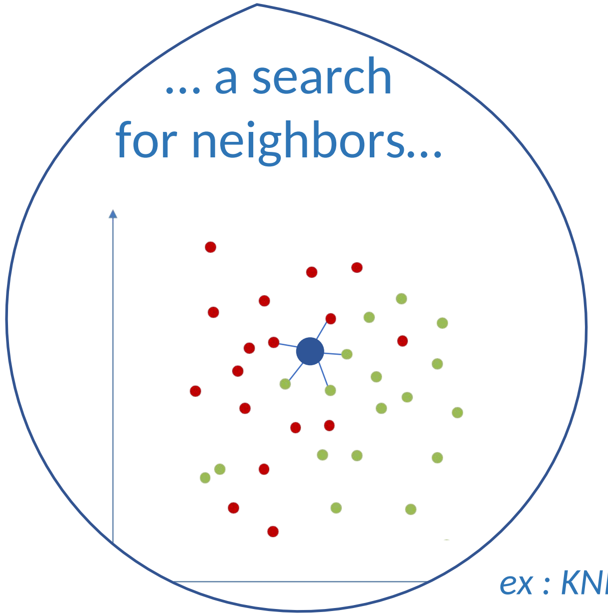
Classes associated with each instance (*annotated/labeled data*)



Training
(“learning”)

Output

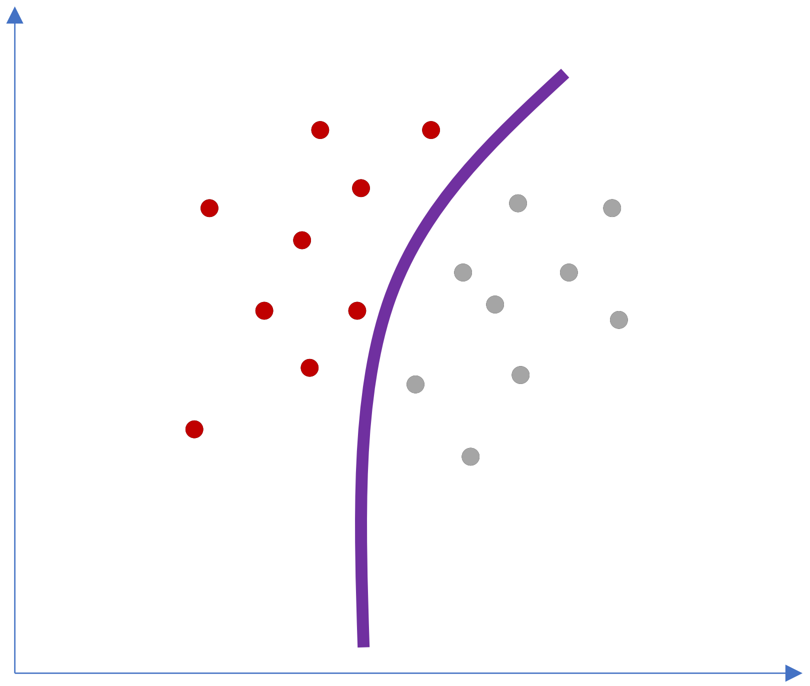
A *model* “trained” on labeled objects, capable of distinguishing the classes



ex : KNN

(What is supervised learning?) Inference

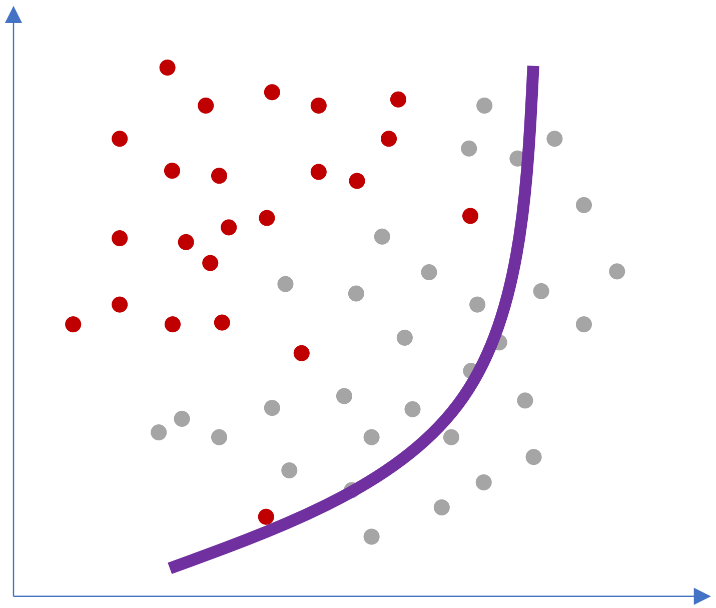
- The model is then applied to new, “unlabeled” data.



(What is supervised learning?) Training

- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.

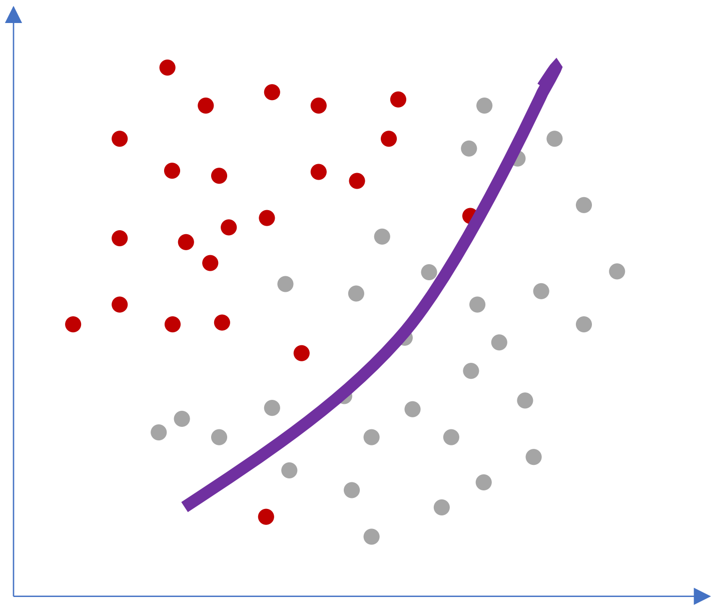
- Often an iterative process



(What is supervised learning?) Training

- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.

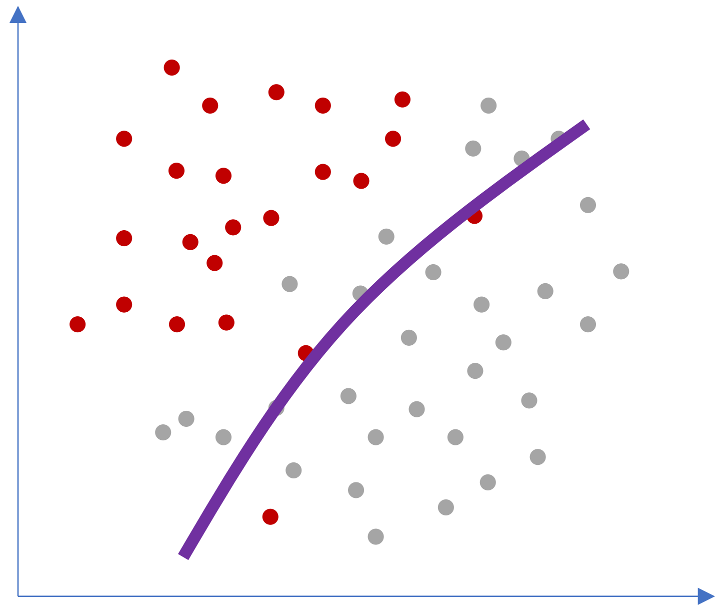
- Often an iterative process



(What is supervised learning?) Training

- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.

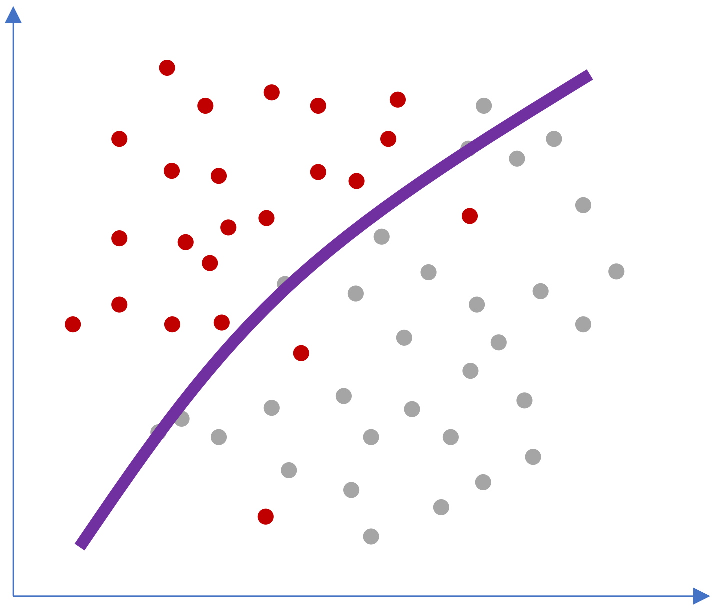
- Often an iterative process



(What is supervised learning?) Training

- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.

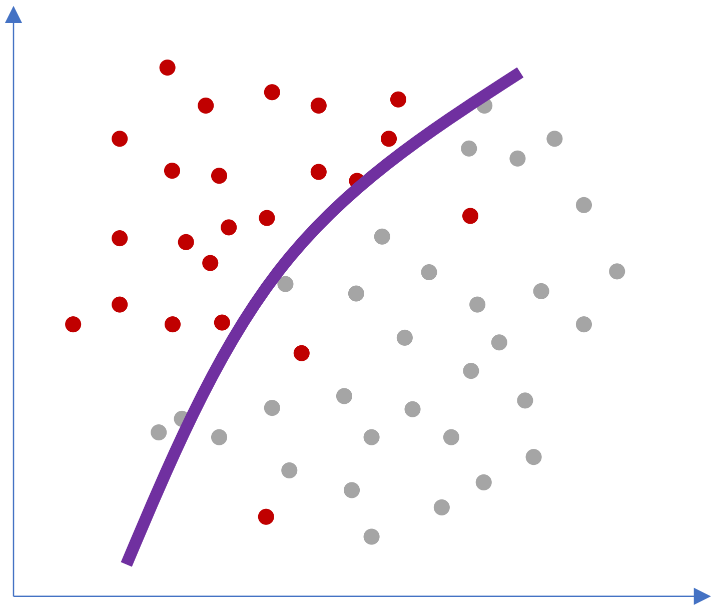
- Often an iterative process



(What is supervised learning?) Training

- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.

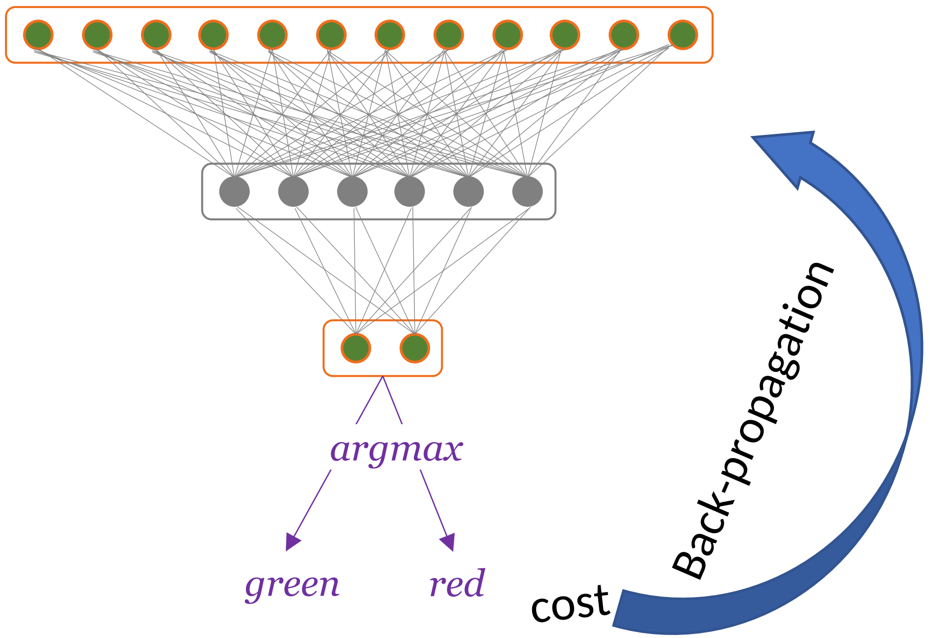
- Often an iterative process



(What is supervised learning?) Training

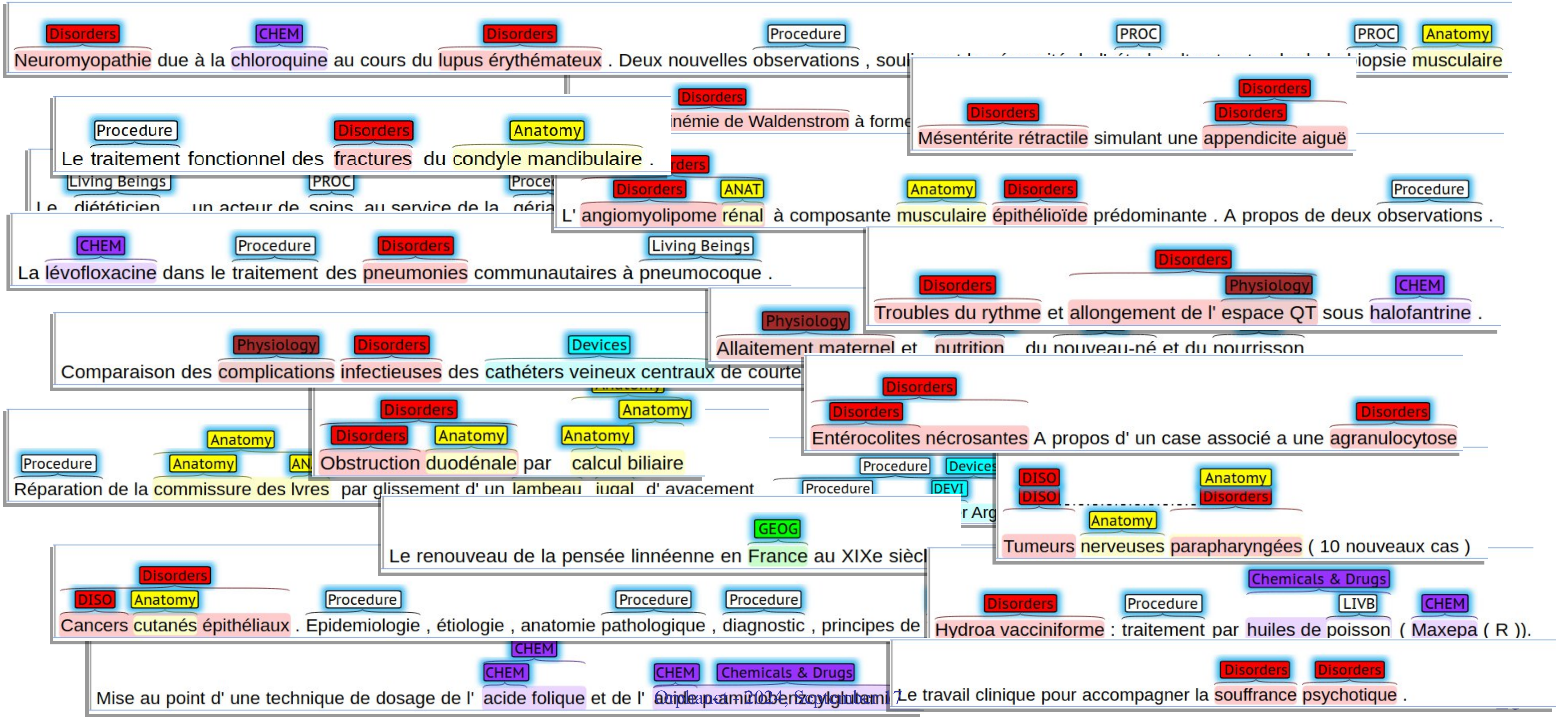
- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.

- Often an iterative process



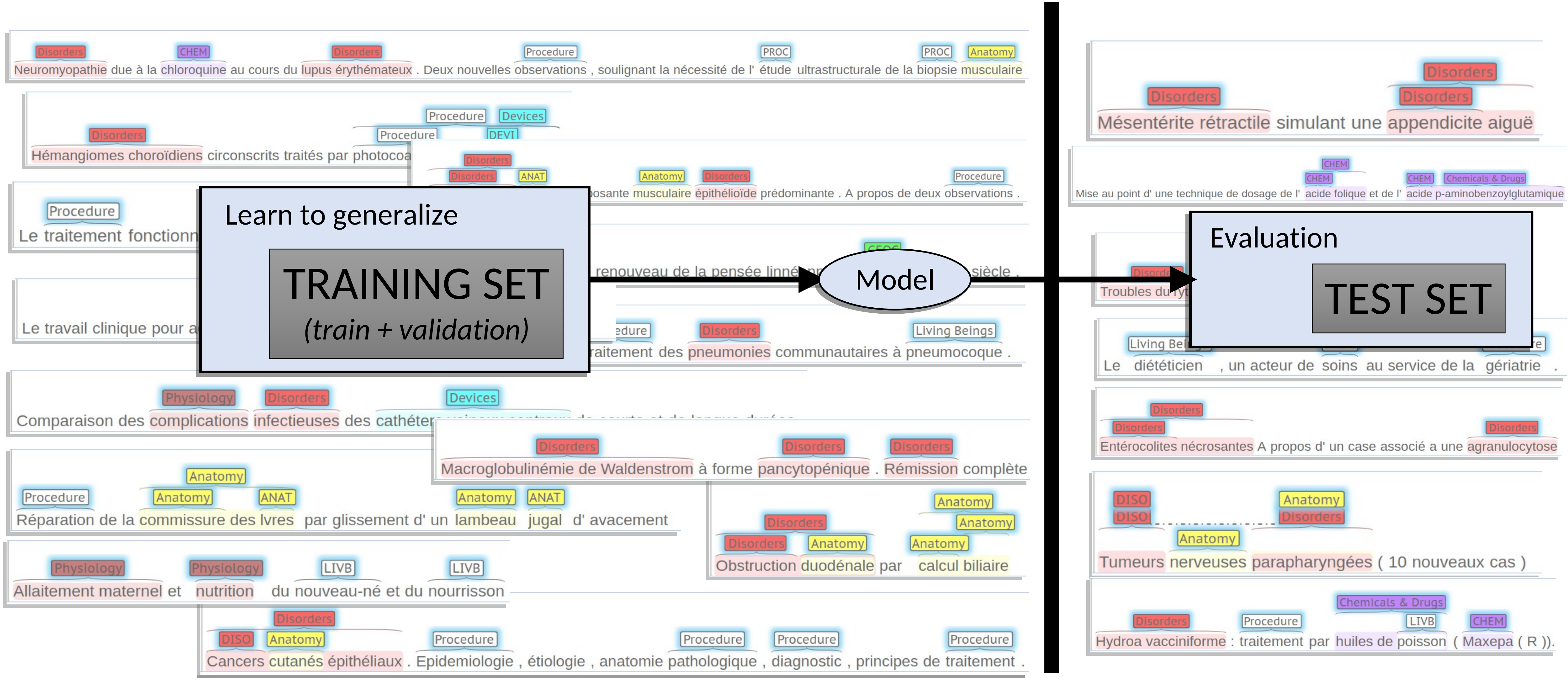
Supervised learning systems

Névéol, A., Grouin, C., Leixa, J., Rosset, S., & Zweigenbaum, P. (2014).
The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization.

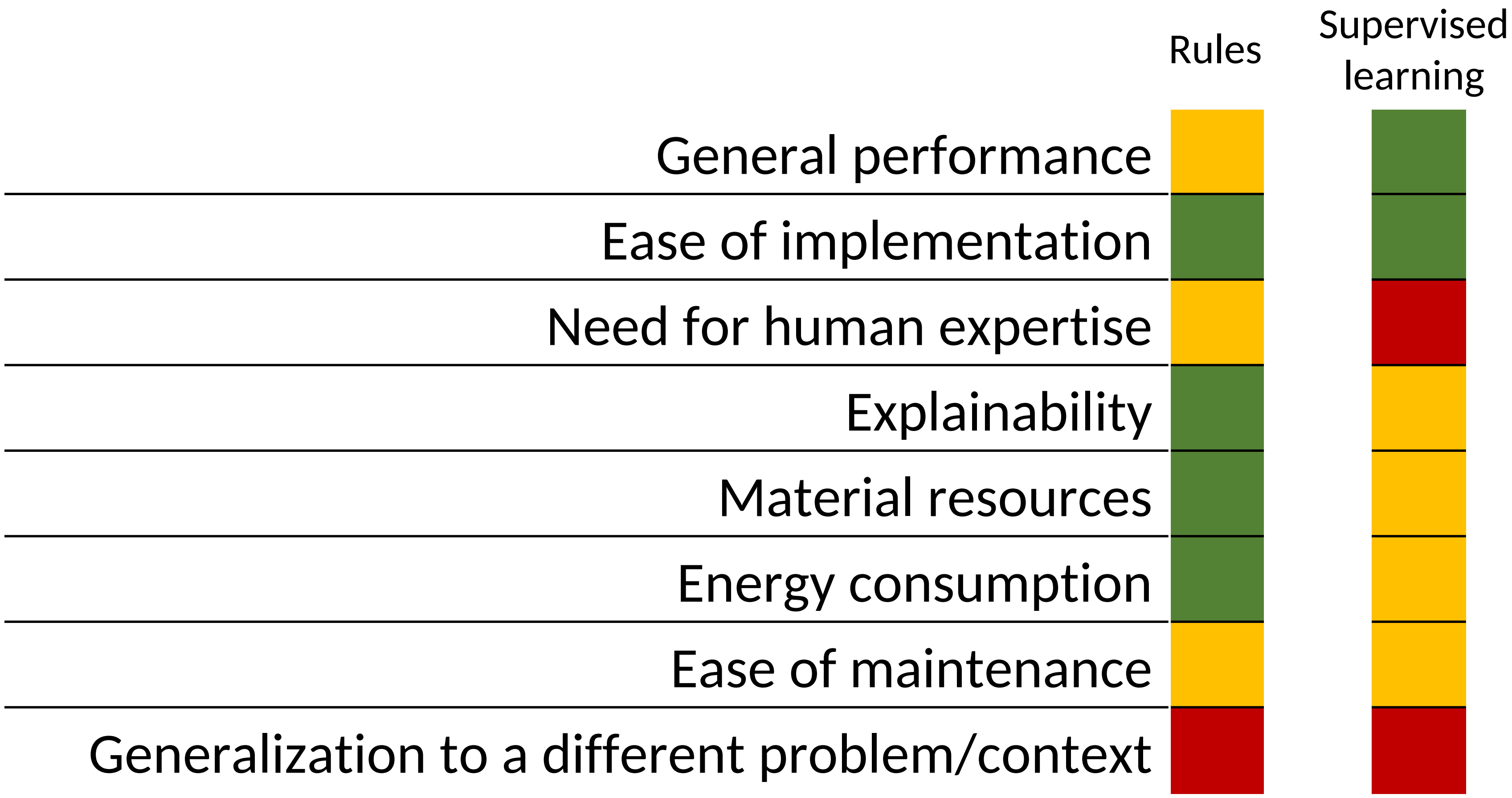


Supervised learning systems

Névéol, A., Grouin, C., Leixa, J., Rosset, S., & Zweigenbaum, P. (2014).
The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization.



Supervised learning systems: Pros & Cons



Methods

1. Rule-based systems
2. Supervised learning systems
3. **Generative, large language models**
4. Retrieval-Augmented Generation

Large Language Models

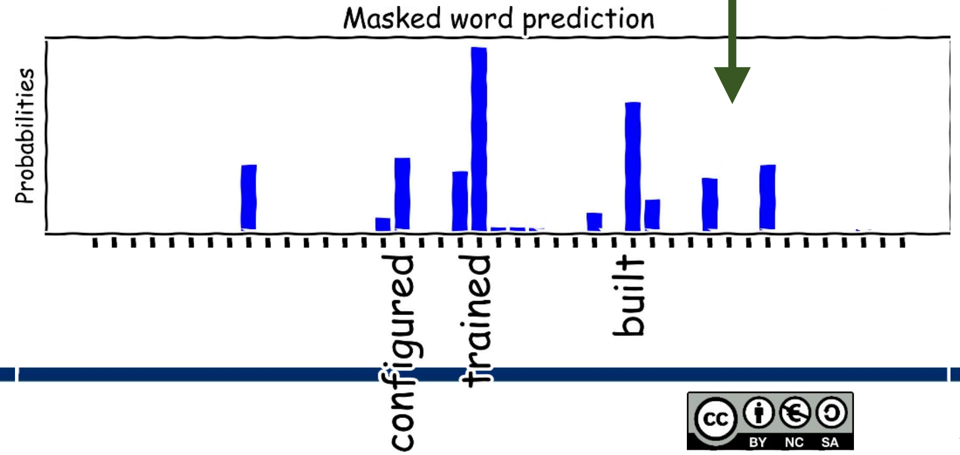
- Generative large language models (LLMs) are trained to produce human-like text.
- They simulate human understanding by predicting and generating text based on the input they receive.
- They can be finetuned for different tasks, e.g rewriting a prompt in another styles.

Auto-regressive language models (*decoder-style, e.g. GPT*)

```
User: What is a large language model?
Assistant: A large language model is a type of artificial intelligence (AI)
model designed to understand and generate human-like language. These models
are
```

GPT-like language model

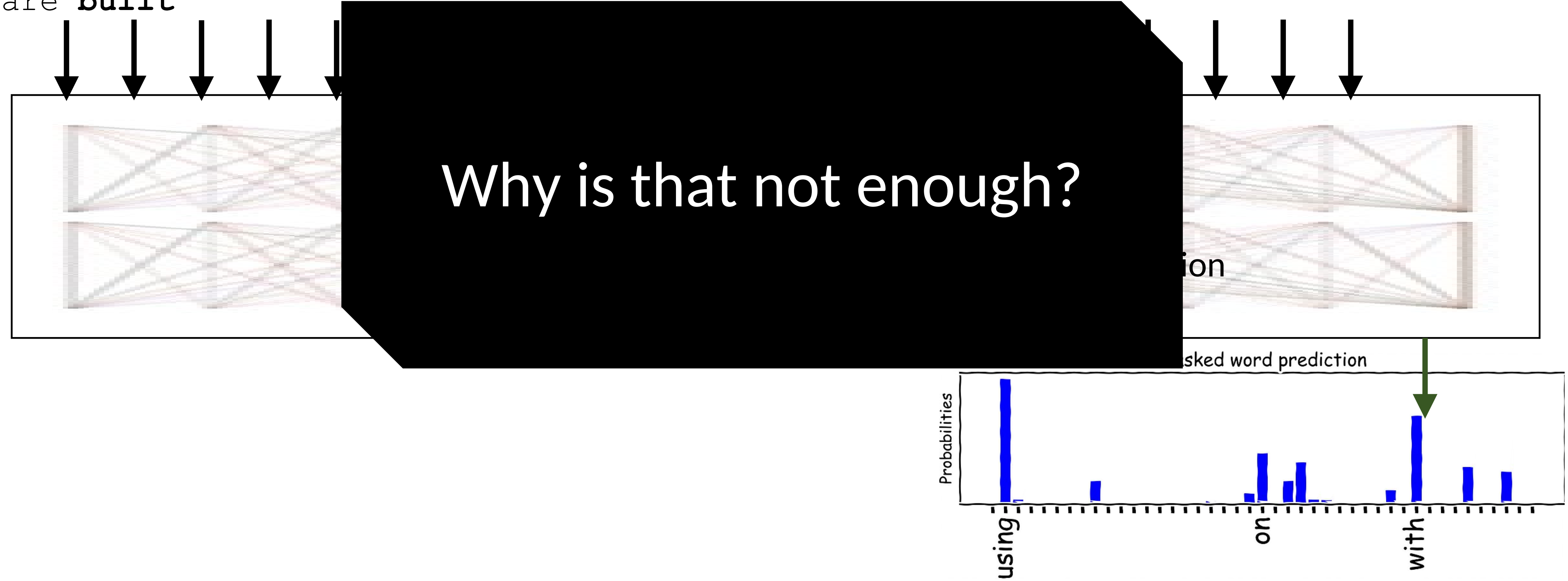
transformers-based model with multiple layers of attention



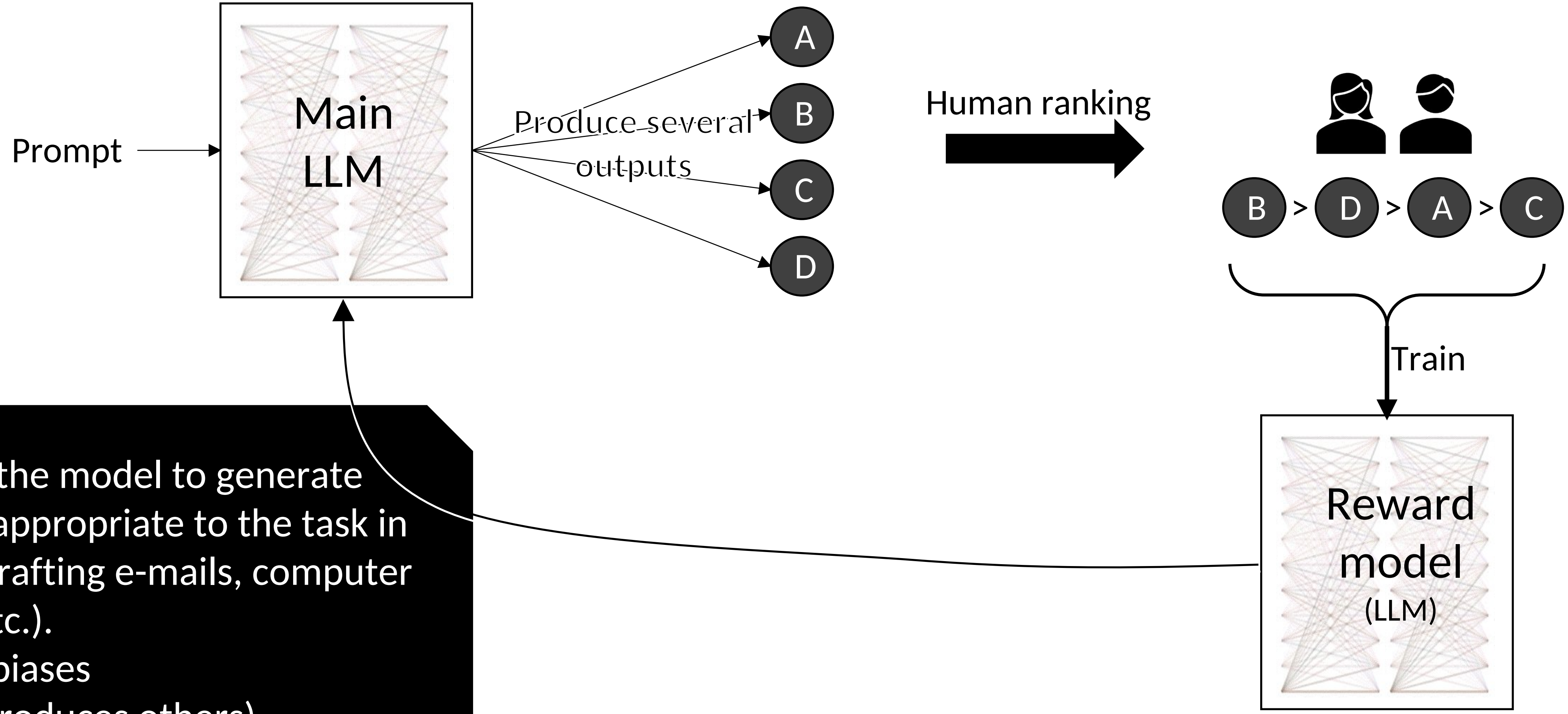
Next tokens are randomly selected from the predicted distribution, introducing variability in the generated output

Auto-regressive language models (*decoder-style, e.g. GPT*)

```
User: What is a large language model?
Assistant: A large language model is a type of artificial intelligence (AI)
model designed to understand and generate human-like language. These models
are built
```



Reinforcement Learning with Human Feedback



- Adapts the model to generate results appropriate to the task in hand (drafting e-mails, computer code, etc.).
- Avoids biases
  (and introduces others)

Why are LLMs so good?

- LLMs are good at producing relevant, well-written and convincing information, thanks to:
  - Representation learning
  - Contextual understanding and attention mechanisms
    (capture long-range dependencies and relationships in the data)
  - Scalability and massive amount of training data
    (wide range of linguistic nuances and topics)
  - Massive finetuning with human feedback
    (produce the right kind of results for the task prompted by the user)

- LLMs are NOT good at:
  - Factuality / knowledge
  - High precision
  - Humor, creativity, originality

How can we use LLMs?

- Direct prompt: *"How do genetic mutations in the ALMS1 gene contribute to the pathophysiology of Alström Syndrome?"*
  ► only knowledge from the pretraining

- Prompt with persona: *"**You're an assistant specialized in research on rare diseases**. How do genetic mutations in the ALMS1 gene contribute to the pathophysiology of Alström Syndrome?"*
  ► allows to guide the answer and its style

- Prompt with document: *"**Given this document**, how do genetic mutations in the ALMS1 gene contribute to the pathophysiology of Alström Syndrome?"*
  ► knowledge from the pretraining + supporting document

How can we use LLMs?

- Few-shot prompting: *“*

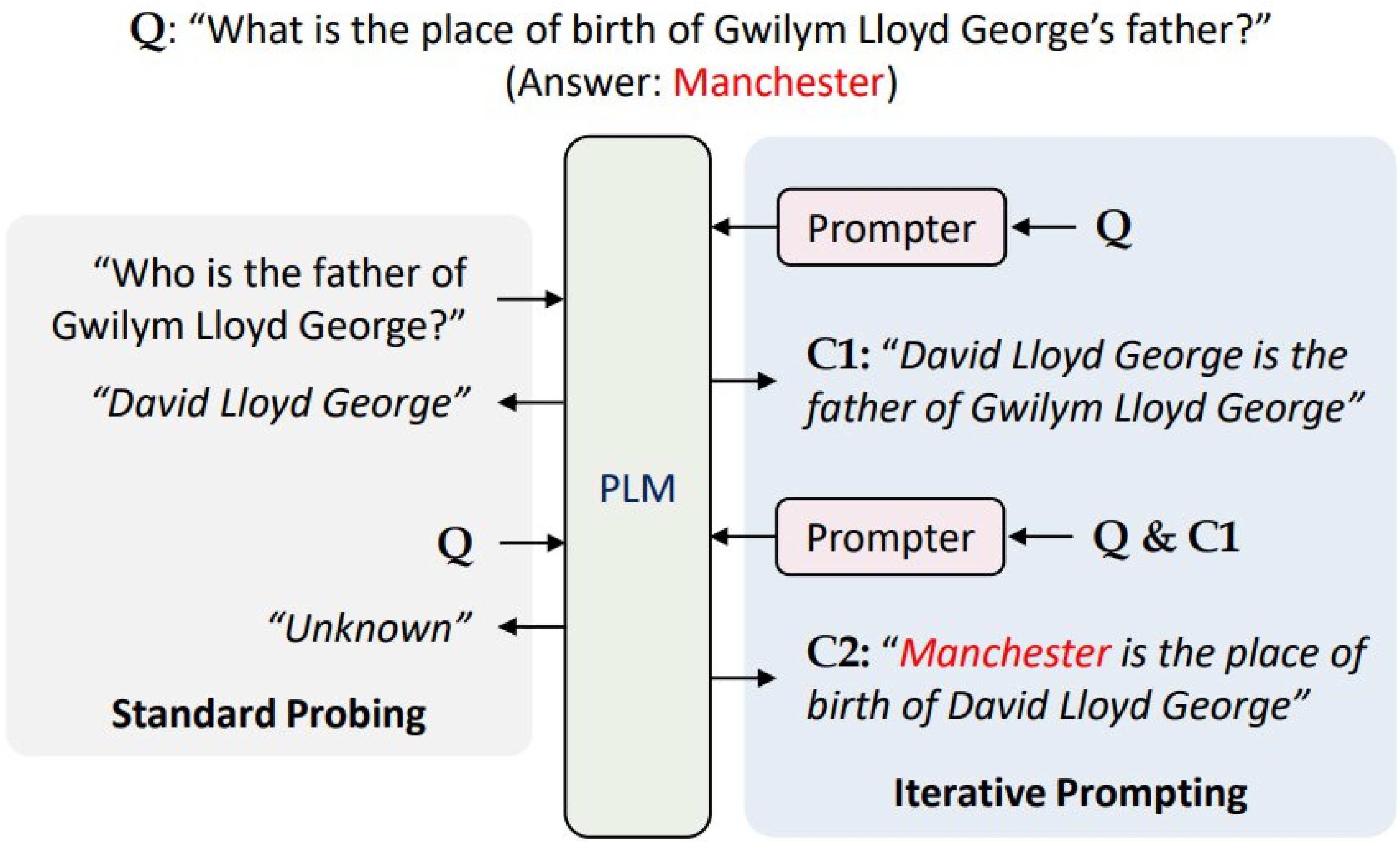
  Classify the following rare diseases into their appropriate categories: Genetic Disorder or Neurodegenerative Disorder

  Input: "Alström Syndrome"
  Output: "Genetic Disorder"
  Input: "Batten Disease"
  Output: "Neurodegenerative Disorder"
  Input: "Huntington's Disease"
  Output: "Neurodegenerative Disorder"
  Input: "Marfan Syndrome"
  Output: "Genetic Disorder"
  Input: "Parkinson's Disease"
  Output: "Neurodegenerative Disorder"
  Input: "Cystic Fibrosis"
  Output:

”

How can we use LLMs?

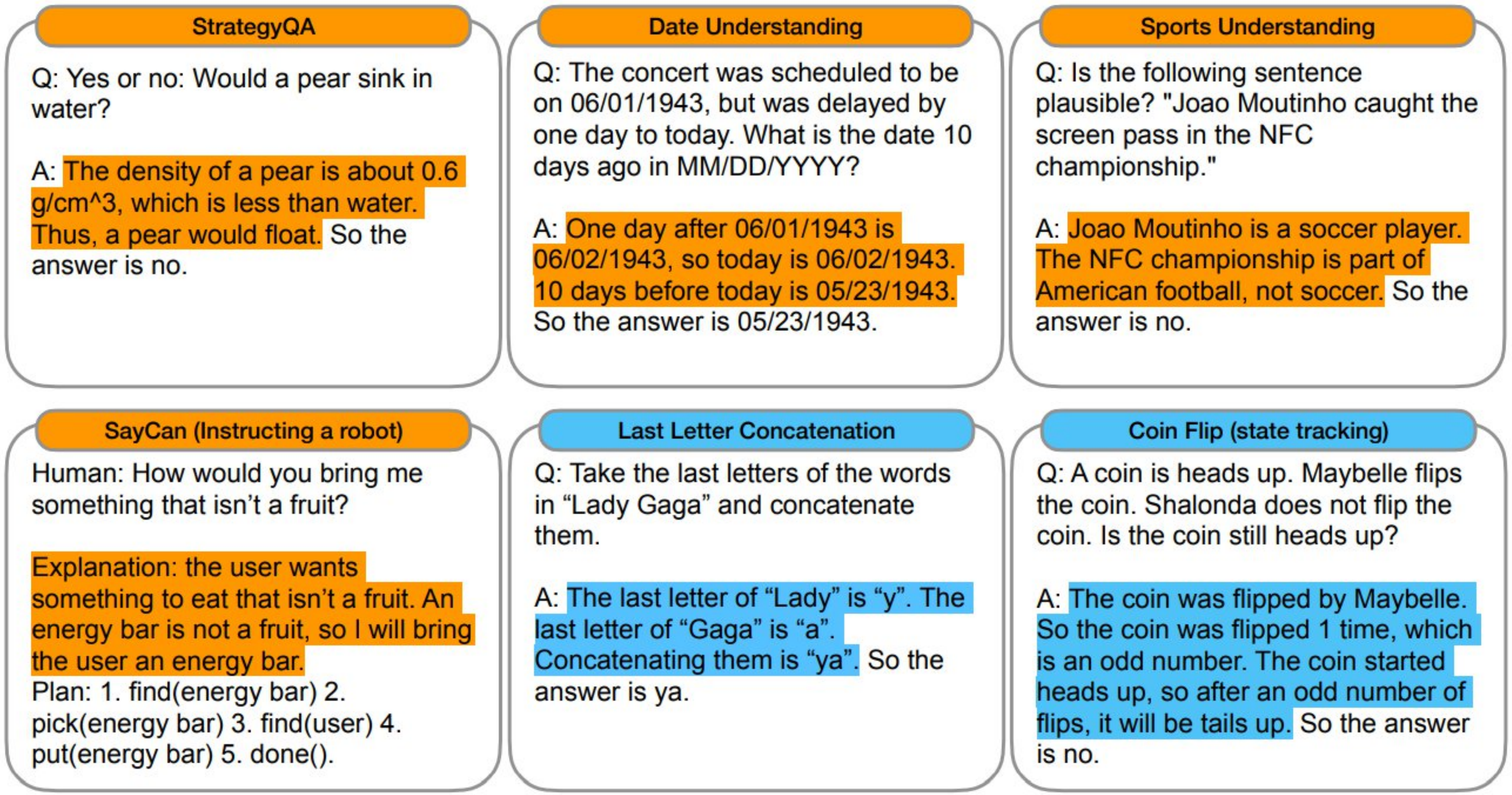
- Prompt chaining:



Wang, B., Deng, X., & Sun, H. (2022). Iteratively Prompt Pre-trained Language Models for Chain of Thought. *Conference on Empirical Methods in Natural Language Processing*.

How can we use LLMs?

- Chain of thoughts:

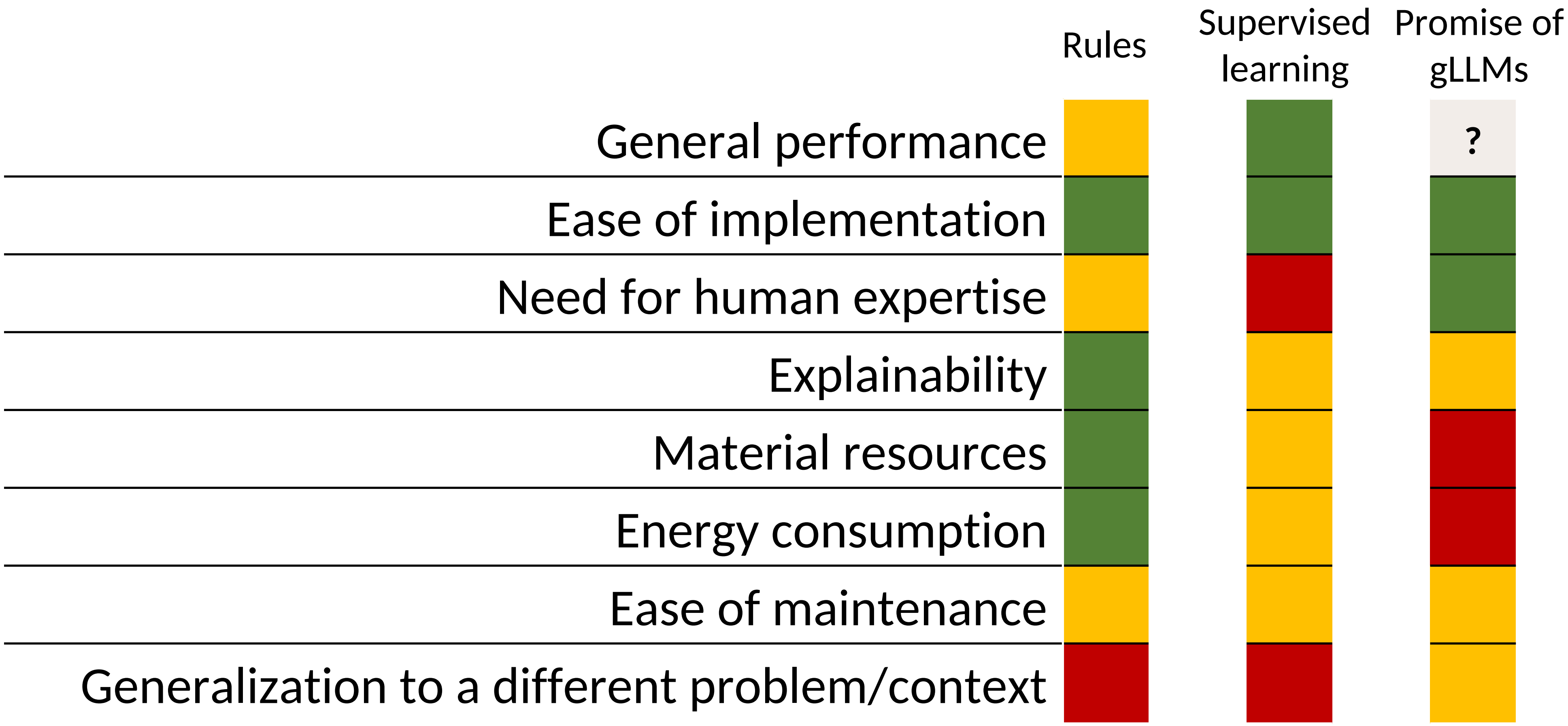


Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *Conference on Neural Information Processing Systems*.

How can we use LLMs?

- Self-Consistency
- Generated Knowledge Prompting
- Tree of Thoughts (ToT)
- Automatic Prompt Engineer (APE)
- Active-Prompt
- Directional Stimulus Prompting
- Program-Aided Language Models
- … (see for example https://www.promptingguide.ai/)

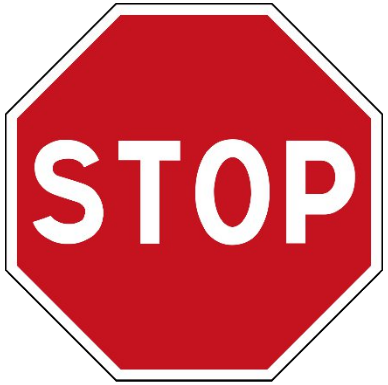
Generative LLMs: Pros & Cons



A thorough evaluation of these models is always necessary!

Whatever the method (rules, ML, LLMs…), build a test set for the evaluation

Do not prompt public or API-based LLMs with sensitive data!

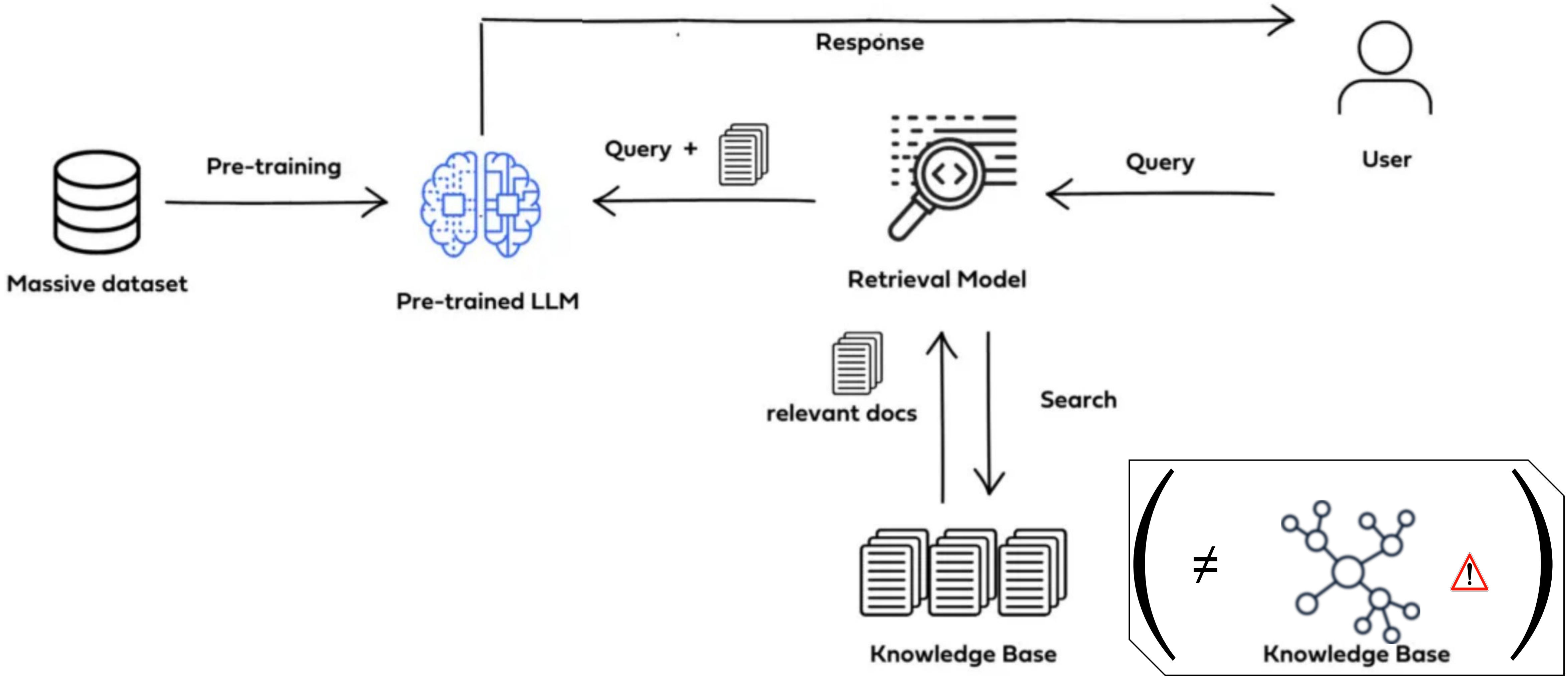


Methods

1. Rule-based systems
2. Supervised learning systems
3. Generative, large language models
4. **Retrieval-Augmented Generation**

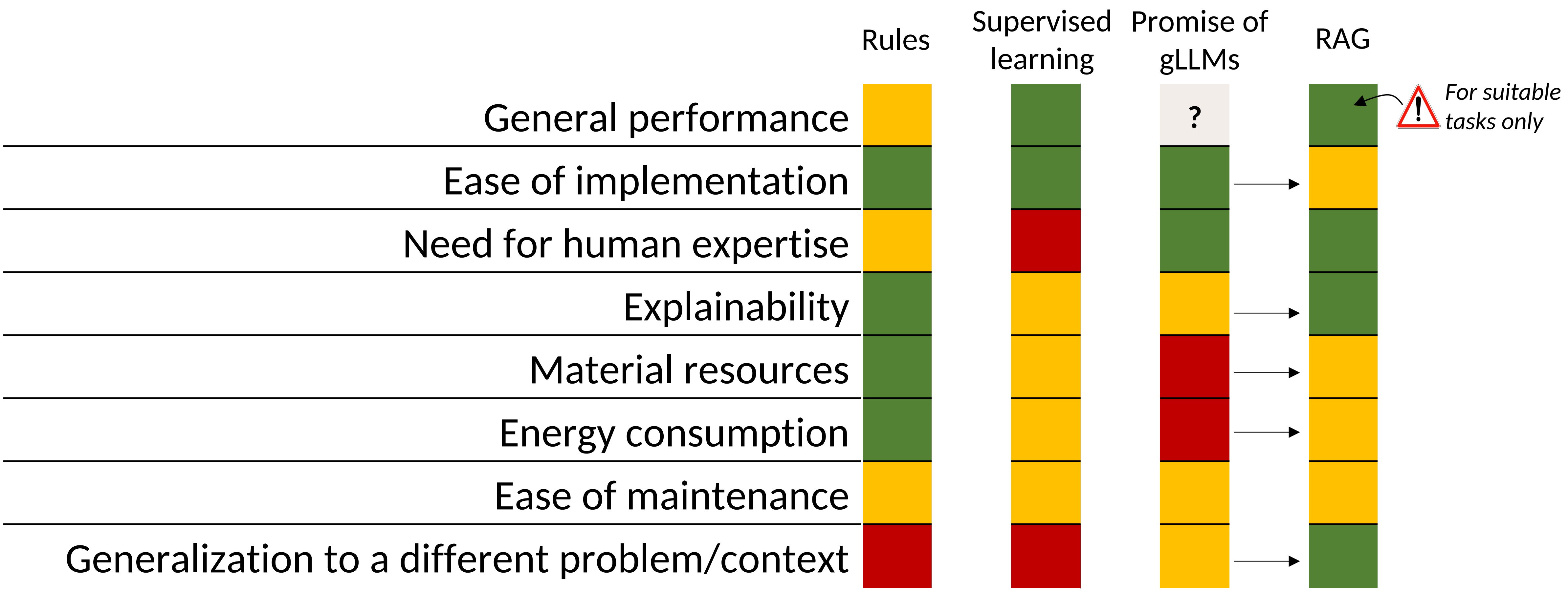
The new search engine

Retrieval-Augmented Generation

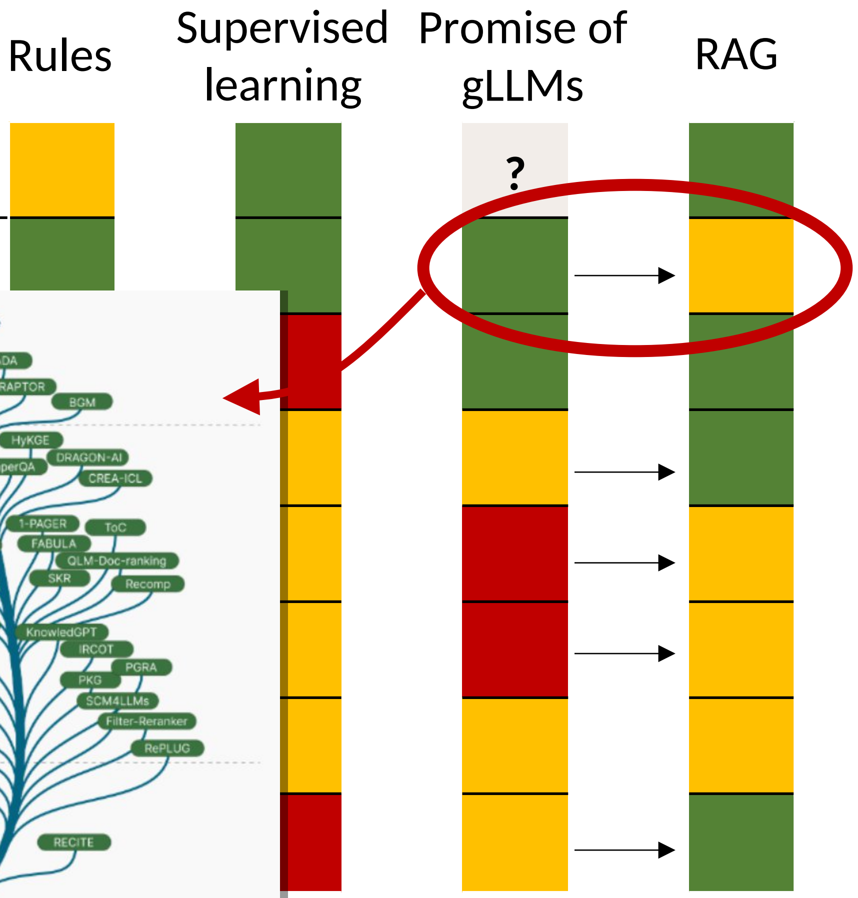


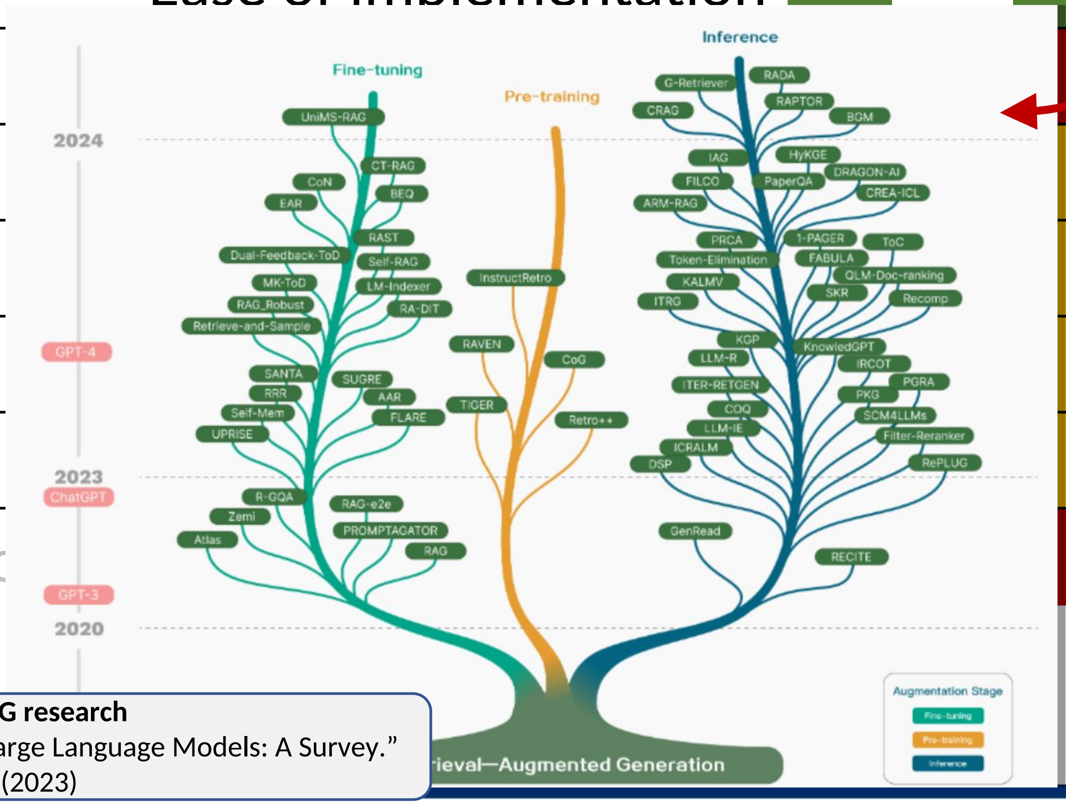
https://medium.com/@krtarunsingh/introduction-to-retrieval-augmented-generation-rag-and-its-transformative-role-in-ai-c07e35da7f01

RAG: Pros & Cons



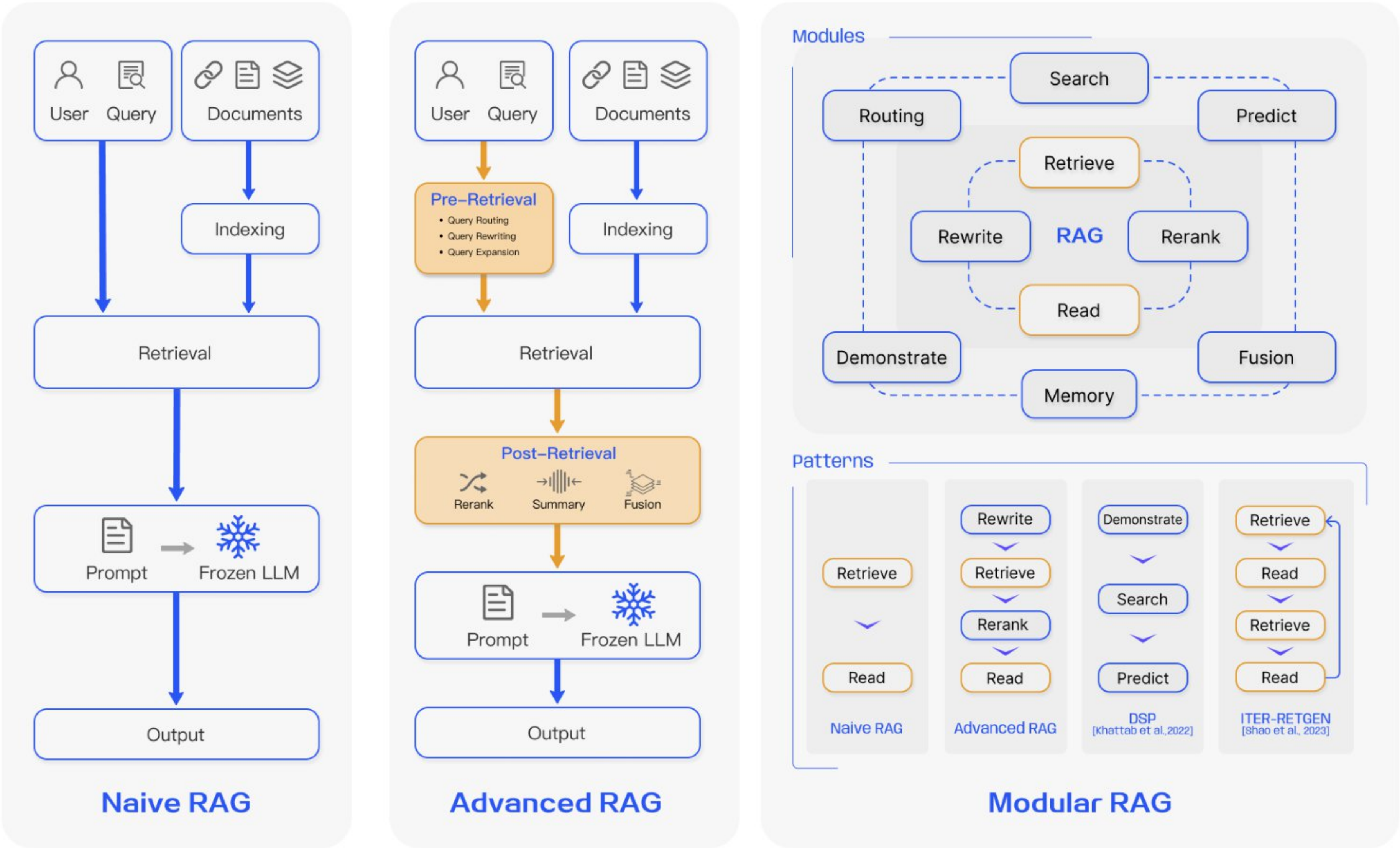
RAG: Pros & Cons





The technology tree of RAG research
Gao et al. “Retrieval-Augmented Generation for Large Language Models: A Survey.” *ArXiv* abs/2312.10997 (2023)

RAG



Gao et al. "Retrieval-Augmented Generation for Large Language Models: A Survey."
ArXiv abs/2312.10997 (2023)

NLP for scientific surveillance

Positioning NLP and LLMs in the process of curation

Data Collection
- Prioritization
- Summarization

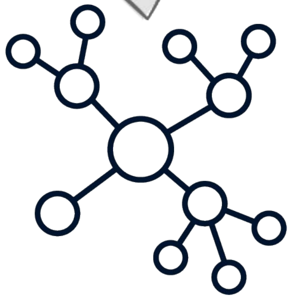
Data Annotation
- Entities
- Relations
- Classification

Data Normalization
- Link to knowledge bases
- Standard descriptions

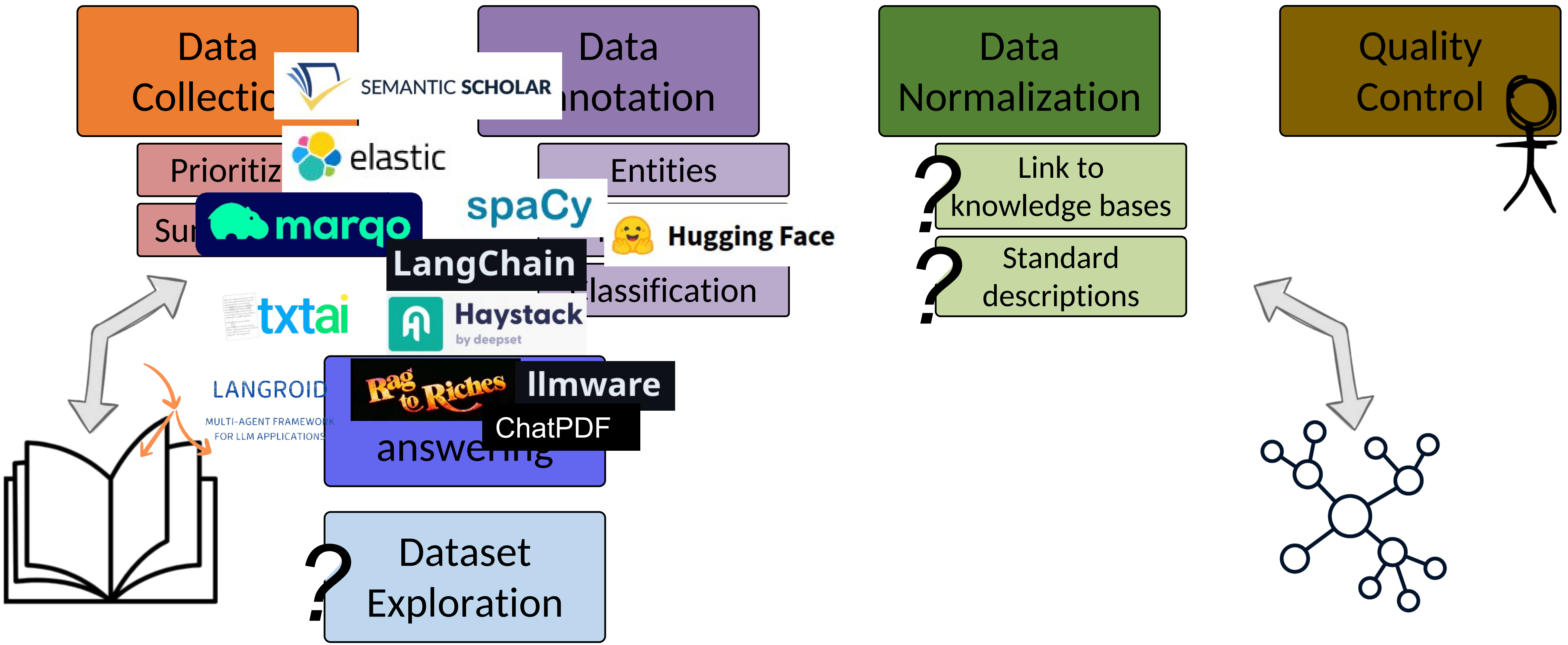
Quality Control

Question answering

Dataset Exploration

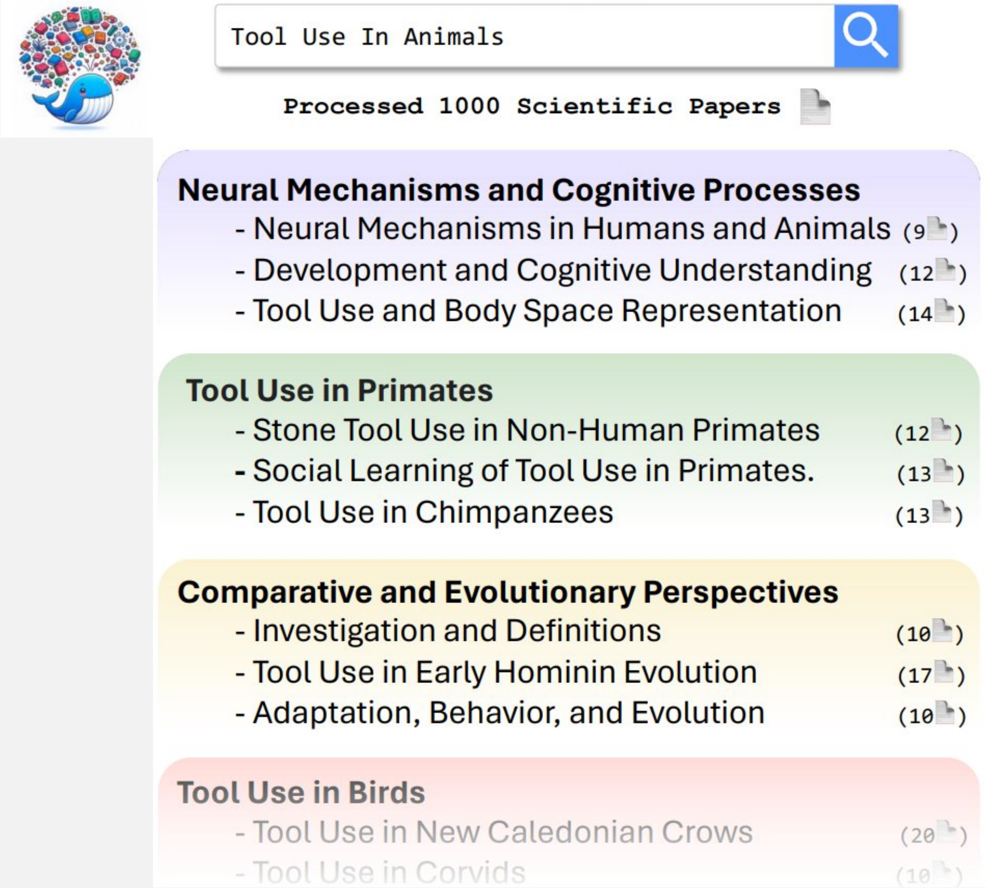


Positioning NLP and LLMs in the process of curation



Positioning NLP and LLMs in the process of curation

Knowledge Navigator: LLM-guided Browsing Framework for Exploratory Search in Scientific Literature, U. Katz et al. 2024



Dataset Exploration

Positioning NLP and LLMs in the process of curation

LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT, LP Meyer et al, 2023

Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction, S Carta et al, 2023

Let's Chat to Find the APIs: Connecting Human, LLM and Knowledge Graph through AI Chain, Q Huang et al, 2023

Knowledge Graph Prompting for Multi-Document Question Answering, Y Wang et al, 2023

Enhancing Knowledge Graph Construction Using Large Language Models, M Trajanoska et al, 2023

Exploring Large Language Models for Knowledge Graph Completion, L. Yao, 2023

Knowledge Graph Large Language Model (KG-LLM) for Link Prediction, D. Shu, 2024

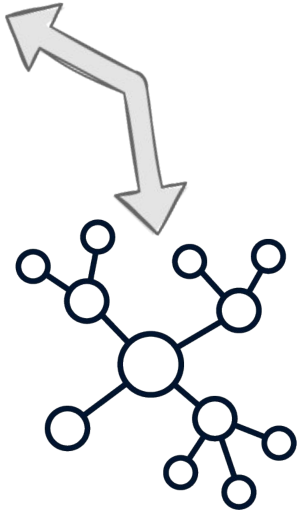
…

Data Normalization

Link to knowledge bases

Standard descriptions

"Large Language Models struggle to learn long-tail knowledge"
(N. Kandpal et al, 2023)



Knowledge Prompting: How Knowledge Engineers Use Large Language Models

ELISAVET KOUTSIANA*, King's College London, United Kingdom

JOHANNA WALKER*, King's College London, United Kingdom

MICHELLE NWACHUKWU, King's College London, United Kingdom

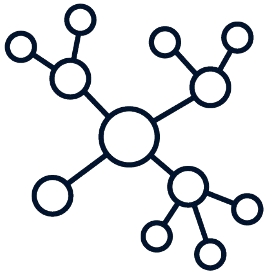
ALBERT MEROÑO-PEÑUELA, King's College London, United Kingdom

ELENA SIMPERL*, King's College London, United Kingdom

Despite many advances in knowledge engineering (KE), challenges remain in areas such as engineering knowledge graphs (KGs) at scale, keeping up with evolving domain knowledge, multilingualism, and multimodality. Recently, KE has used LLMs to support semi-automatic tasks, but the most effective use of LLMs to support knowledge engineers across the KE activites is still in its infancy. To explore the vision of LLM copilots for KE and change existing KE practices, we conducted a multimethod study during a KE hackathon. We investigated participants' views on the use of LLMs, the challenges they face, the skills they may need to integrate LLMs into their practices, and how they use LLMs responsibly. We found participants felt LLMs could contribute to improving efficiency when engineering KGs, but presented increased challenges around the already complex issues of evaluating the KE tasks. We discovered prompting to be a useful but undervalued skill for knowledge engineers working with LLMs, and note that natural language processing skills may become more relevant across more roles in KG construction. Integrating LLMs into KE tasks needs to be mindful of potential risks and harms related to responsible AI. Given the limited ethical training, most knowledge engineers receive solutions such as our suggested 'KG cards' based on data cards could be a useful guide for KG construction. Our findings can support designers of KE AI copilots, KE researchers, and practitioners using advanced AI to develop trustworthy applications, propose new methodologies for KE and operate new technologies responsibly.

https://www.arxiv.org/abs/2408.08878

How to link (indirectly) formal knowledge and LLMs

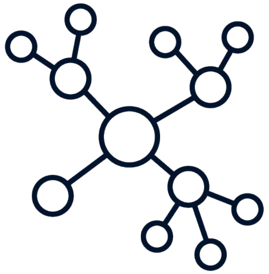


Dialogue with LLM with constraints and "multi-hop" reasoning

- More intelligent
- Expert in your specific domain
- More intuition

- Reads faster
- Better memory

How to link (indirectly) formal knowledge and LLMs



Dialogue with LLM with constraints and “multi-hop” reasoning

Mix LLMs with more traditional NLP approaches
(concept similarity, knowledge embeddings)

Mix LLM/NLP techniques with smart interfaces

Fine-tune LLMs with specific knowledge

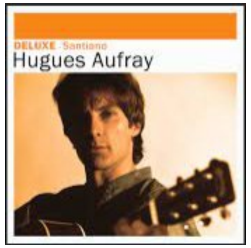
… ?

Thanks!

Questions ?

Backup slides...

Words



D' y penser j' avais le cœur gros

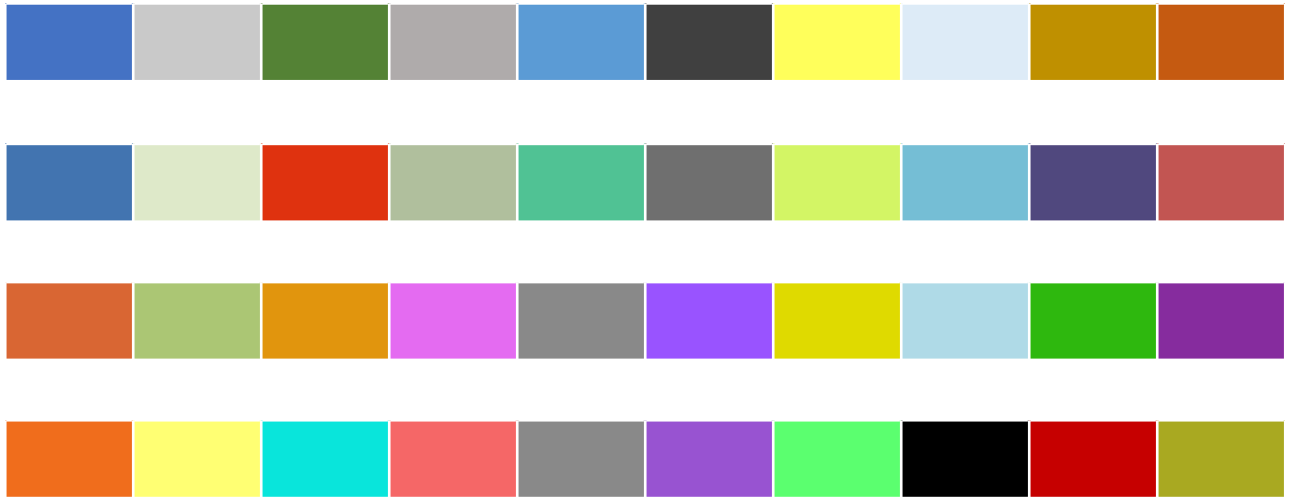
En doublant les feux de Saint-Malo

"Bag of words"

a	0	…	…	…	…	gouvernées	0	pensée	0
à	0	coéternels	0	fanages	0	gouvernement	0	pensées	0
abaissa	0	coéternité	0	fanai	0	gouvernementabilité	0	pensement	0
abaissable	0	coéternités	0	fanaient	0	gouvernemental	0	pensements	0
abaissables	0	**cœur**	**1**	…	…	gouvernementale	0	pensent	0
abaissai	0	cœurs	0	fétuques	0	gouvernementalement	0	**penser**	**1**
…	…	coexista	0	fétus	0	gouvernementales	0	…	…
avachît	0	coexistai	0	**feu**	**1**	gouvernementalisme	0	pertuisaniers	0
avachîtes	0	…	…	feudataire	0	gouvernementalismes	0	pertuiser	0
avaient	0	doublâmes	0	feudataires	0	gouvernementaliste	0	perturba	0
avais	**1**	**doublant**	**1**	…	…	gouvernementaux	0	perturbai	0
avait	0	doublard	0	groomer	0	gouvernementiste	0	perturbaient	0
aval	0	doublards	0	grooms	0	gouvernementistes	0	…	0
avala	0	…	…	**gros**	**1**	gouvernementomane	0	phonocinèse	0
avalable	0	écardons	0	gros-bec	0	gouvernements	0	phonocinétique	0
…	…	écarlate	0	groschen	0	gouvernent	0	phonocontrôle	0
bains	0	écarlates	0	…		gouverner	0	phonocapteur	0
baïonette	0	écarquilla	0	gouttons	0	…	…	…	…

Dense representation (*embeddings*)

- Word embeddings = vector representation of tokens
- Tokens with some degree of similarity 🎭 close to each other in space



Dense representation (*embeddings*)

- **Intuition 1**. Each word of a language is associated with a composition of hidden factors (often unintelligible)

  e.g : *cat = 10 (animal) + 5 (soft) – 10 (loyal)*
  dog = 10 (animal) + 3 (soft) + 10 (loyal)

- **Intuition 2**. Distributional hypothesis

  « You shall know a word by the company it keeps » (Firth, 1957)

  Two words close in vector space = two words that often share similar contexts

  e.g : the … scratches ; … is a felid

  $occurrence(cat) \sim occurrence(tiger)$

  $w_{cat} \cdot w_{context} \sim w_{tiger} \cdot w_{context}$

  $w_{cat} \sim w_{tiger}$

Dogs and cats and tigers

The *cat* is sleeping on the couch

The *dog* is sleeping on the couch

The *cat* is running in the garden

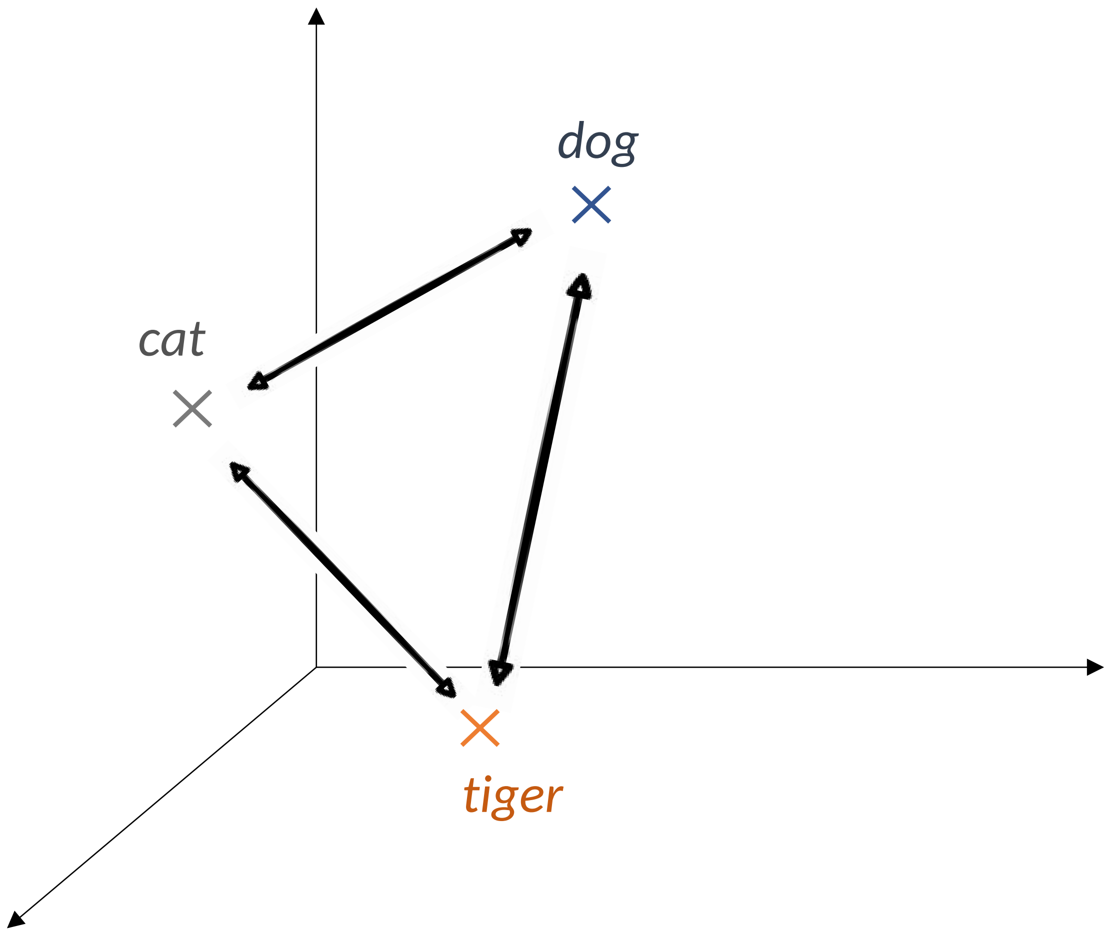
The *dog* is walking in the garden

The *tiger* is a big felid

The *cat* is a small felid

The *tiger* gnaws on an antelope bone

The *dog* gnaws on a chicken bone



Dense representation (*embeddings*)

- Word embeddings = vector representation of tokens
- Tokens close to each other in space 🎭 some degree of similarity between them

e.g. *bike* vs *bicycle*

Dense representation (*embeddings*)

- Word embeddings = vector representation of tokens
- Tokens close to each other in space 🎭 some degree of similarity between them

e.g. *tiger* vs *cat* vs *dog*

Language models

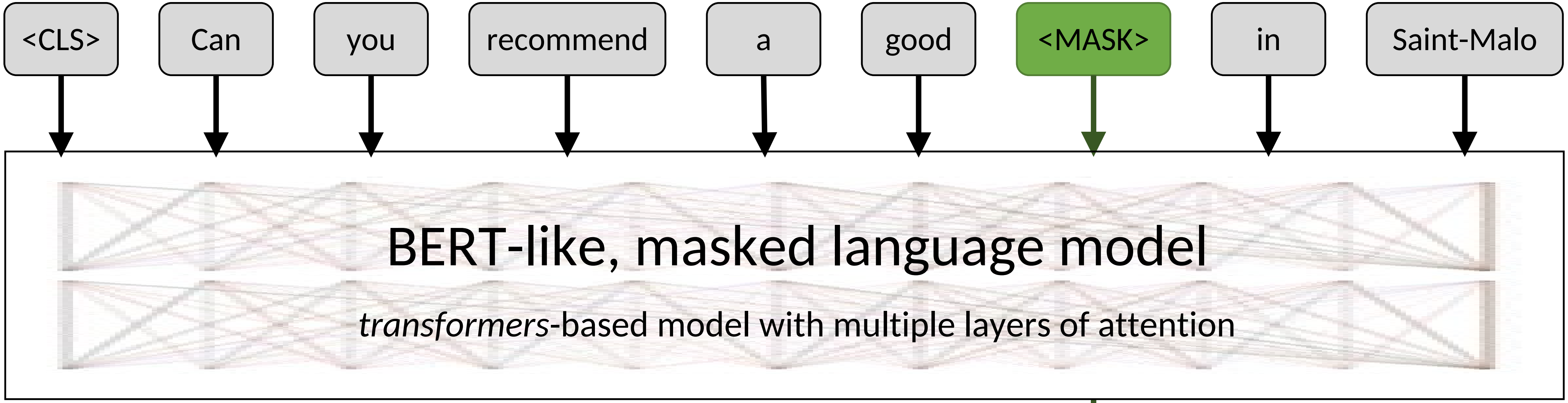
These models belong to the class of *language models*.

A language model is a model without manual supervision, with two different main training settings:

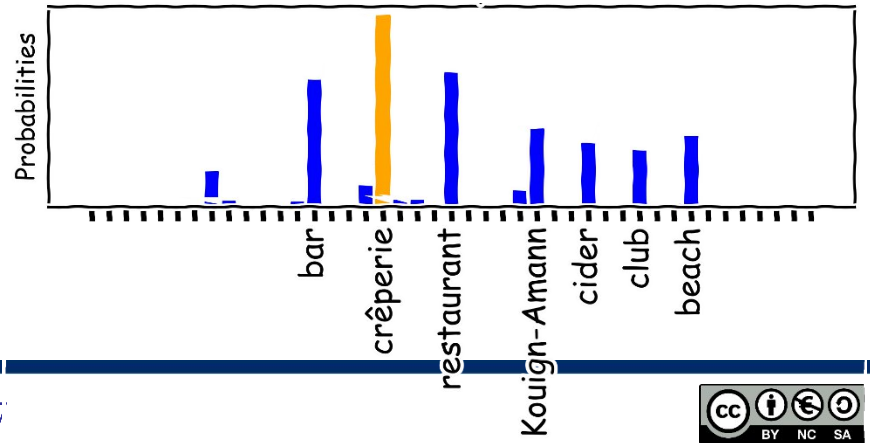
- Predict masked tokens within a sequence (*masked language models*, MLM, e.g. BERT).
- Predict next token (*autoregressive models*, *e.g*. GPT)

These models are at the heart of the vast majority of current NLP systems. They are known as LLMs (*Large Language Models*).

Masked language models (*encoder-style*, *e.g. BERT*)



word prediction



Masked language models (*encoder-style, e.g. BERT*)

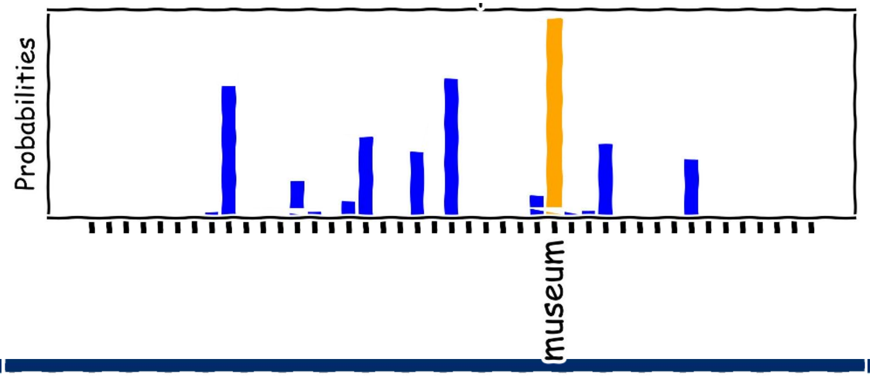
<CLS> | Can | you | recommend | a | good | <MASK> | in | Amsterdam

BERT-like, masked language model

transformers-based model with multiple layers of attention

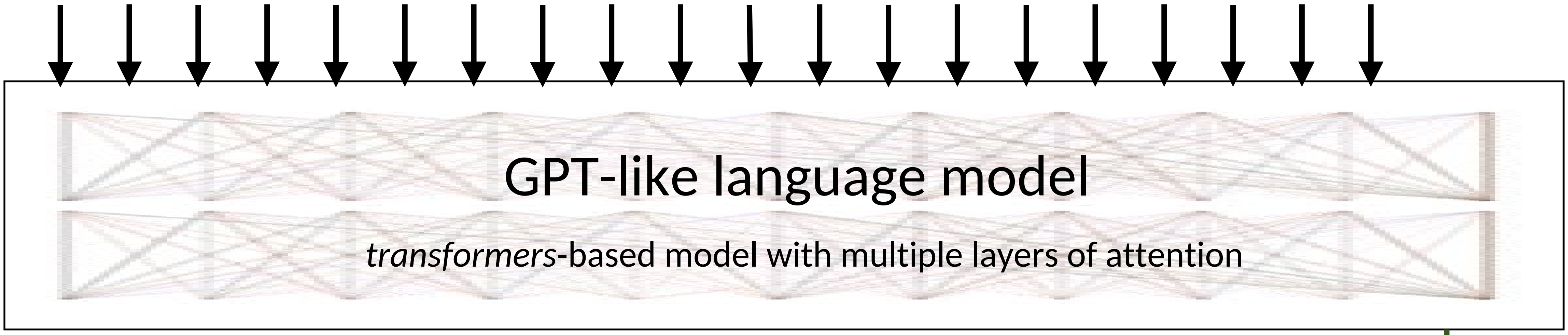
The attention mechanism makes it possible to consider the context for the prediction.

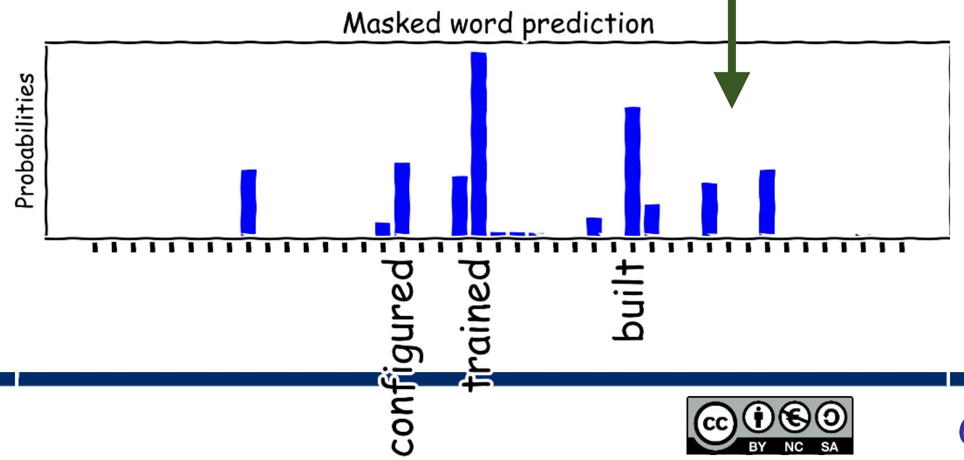
word prediction



Auto-regressive language models (*decoder-style, e.g. GPT*)

```
User: What is a large language model?
Assistant: A large language model is a type of artificial intelligence (AI)
model designed to understand and generate human-like language. These models
are
```

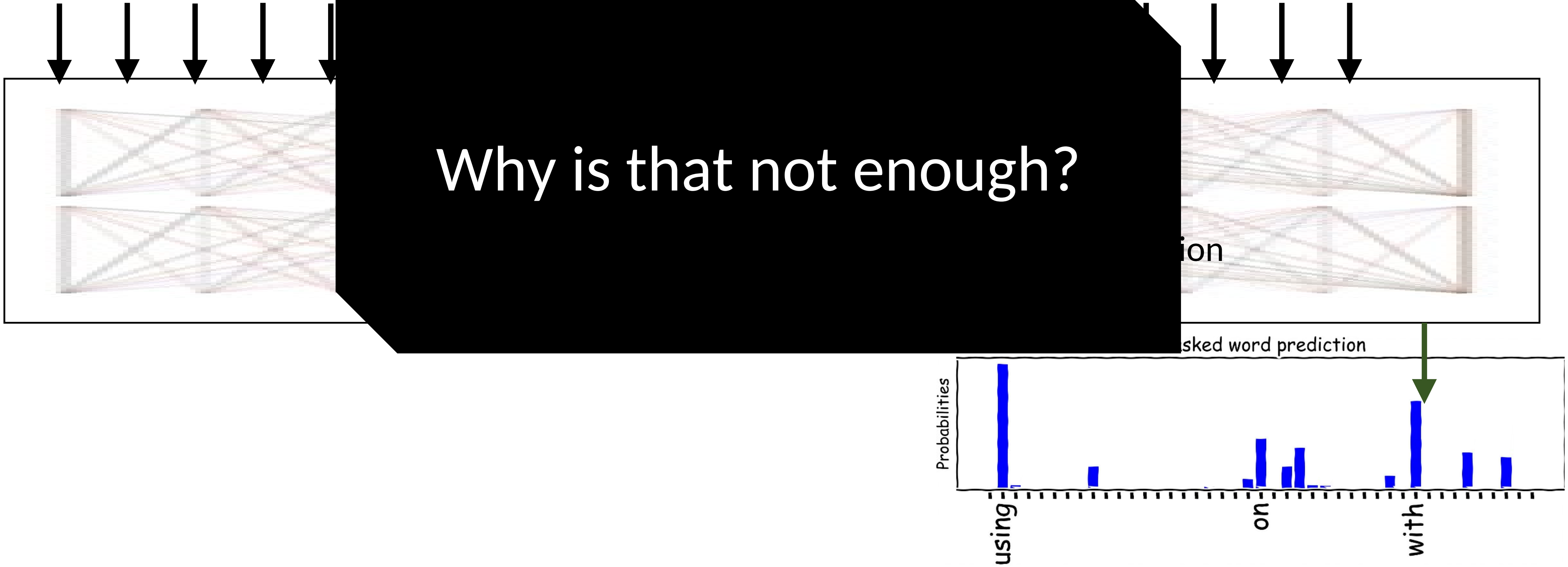




Next tokens are randomly selected from the predicted distribution, introducing variability in the generated output

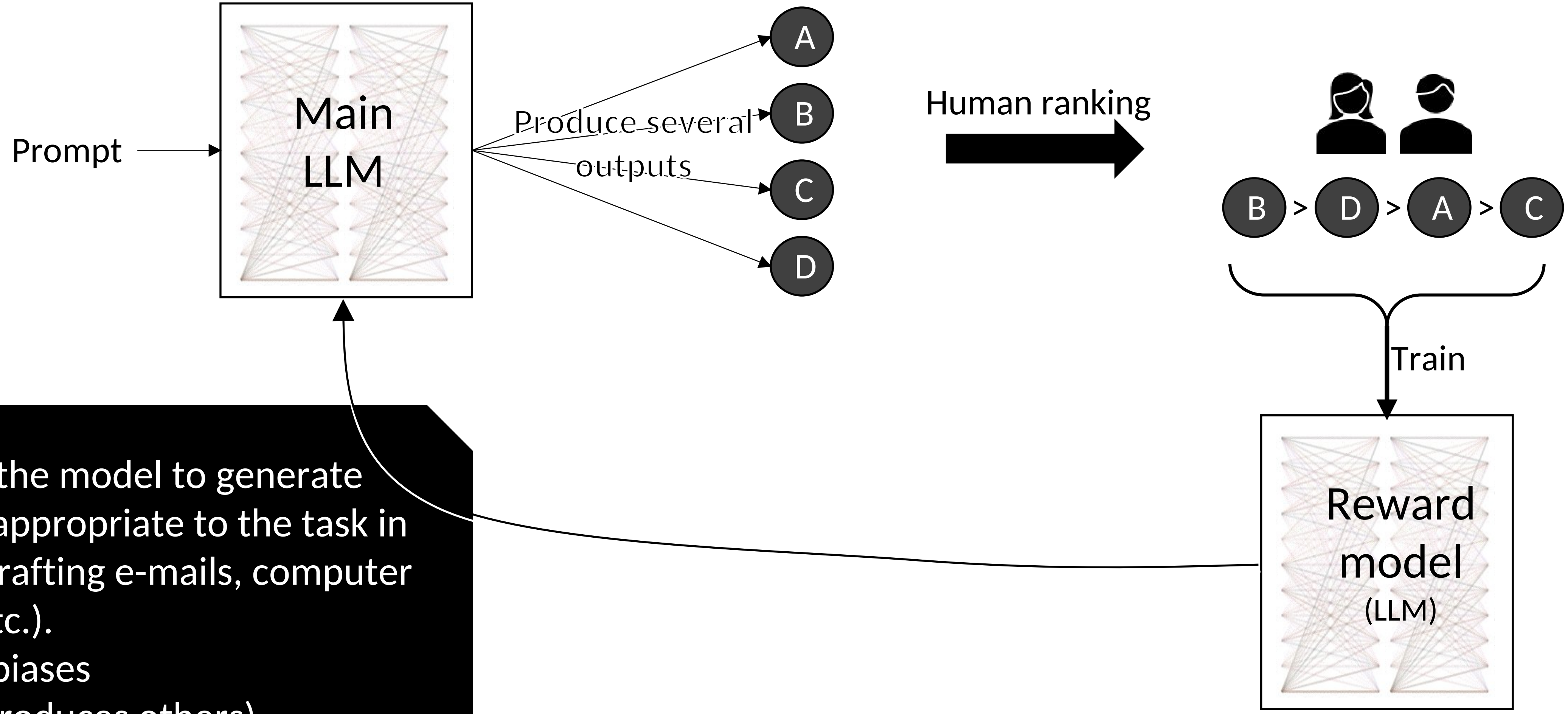
Auto-regressive language models (*decoder-style*, *e.g. GPT*)

User: What is a large language model?
Assistant: A large language model is a type of artificial intelligence (AI) model designed to understand and generate human-like language. These models are **built**



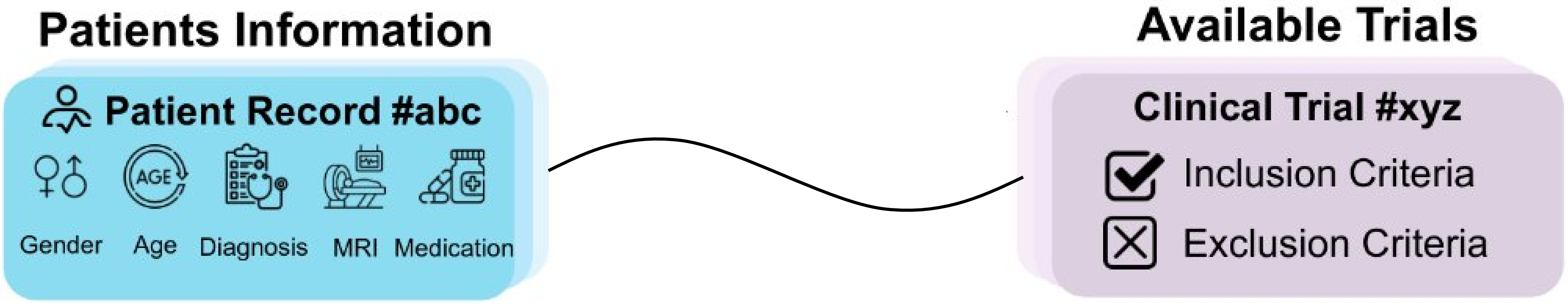
Why is that not enough?

Reinforcement Learning with Human Feedback



- Adapts the model to generate results appropriate to the task in hand (drafting e-mails, computer code, etc.).
- Avoids biases
(and introduces others)

LLMs for patient selection



Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching.
Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, Xia Hu. AMIA Annual Symposium, March 2023

Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Language Models

D. Hamer, P. Schoor, +1 author Daniel Kapitan • Published in arXiv.org 14 April 2023 • Computer Science, Medicine

Transforming clinical trials: the emerging roles of large language models

Jong-Lyul Ghim, Sangzin Ahn • Published in Translational and Clinical... 1 September 2023 • Medicine, Computer Science, Lingu

Distilling Large Language Models for Matching Patients to Clinical Trials

Mauro Nievas, Aditya Basu, Yanshan Wang, Hrituraj Singh less • Published in arXiv.org 15 December 2023 • Computer Science, Medicine

Matching Patients to Clinical Trials with Large Language Models

Qiao Jin, Zifeng Wang, +2 authors Zhiyong Lu • Published in arXiv.org 27 July 2023 • Computer Science, Medicine

Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching.

Jiayi Yuan, Ruixiang Tang, +1 author Xia Hu • Published in AMIA ... Annual Symposium... 24 March 2023 • Computer Science, Medicine • AMIA ... Annual Symposium proceedings. AMIA Symposium

Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology

Cliff Wong, Sheng Zhang, +8 authors Hoifung Poon • Published in Machine Learning in Health... 4 August 2023 • ine, Computer Science

AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models

Surabhi Datta, Kyeryoung Lee, +9 authors Xiaoyan Wang • Published in J. Am. Medical Informatics... 11 November 2023 • Computer Science, Medicine • Journal of the American Medical Informatics Association : JAMIA

LLM for Patient-Trial Matching: Privacy-Aware Data Augmentation Towards Better Performance and Generalizability

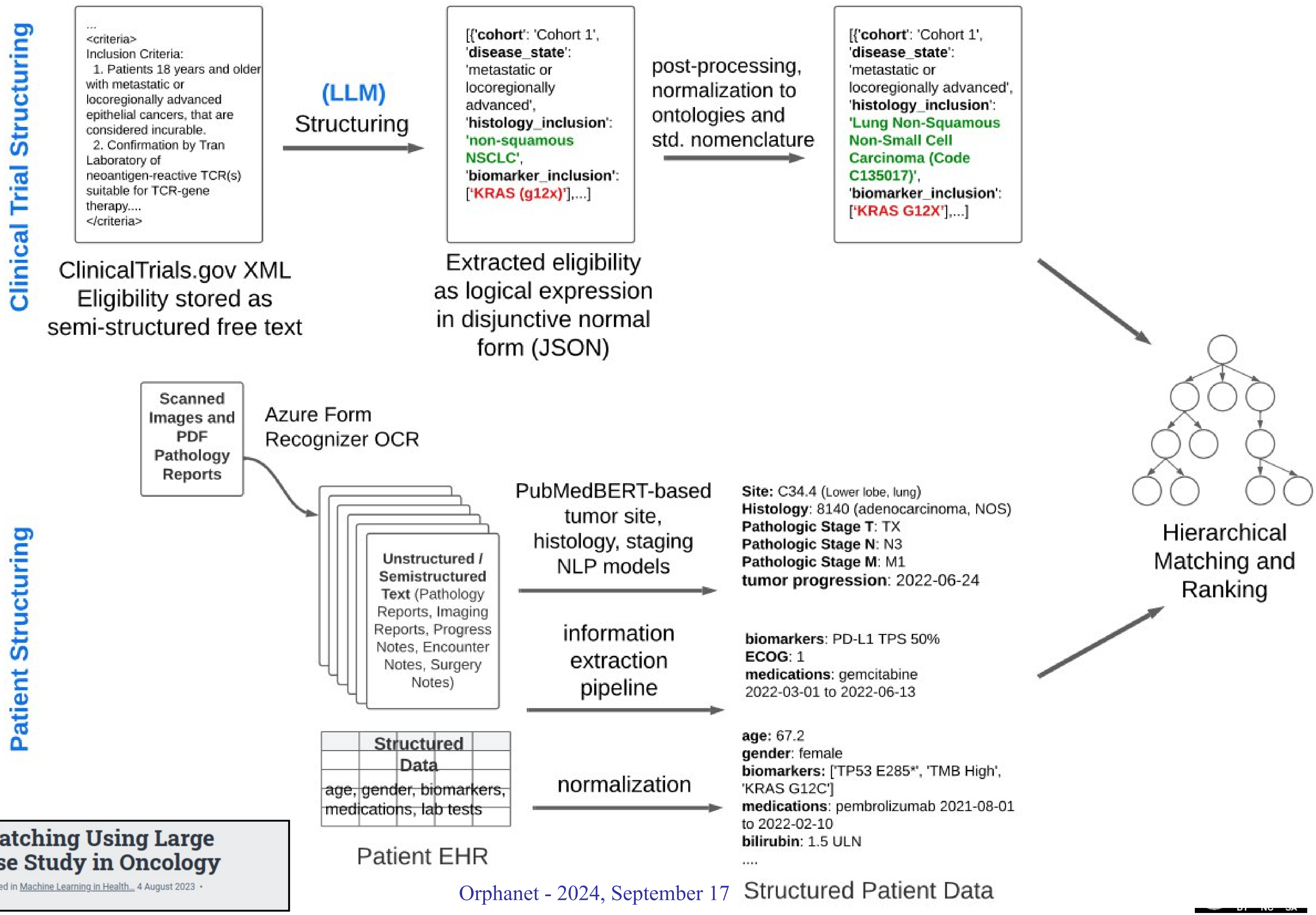
Yuan, Ruixiang Tang, Xiaoqian Jiang, Xia Hu less • Published in arXiv.org 2023 • Computer Science, Medicine

Zero-Shot Clinical Trial Patient Matching with LLMs

Michael Wornow, A. Lozano, +3 authors Nigam H. Shah • Published in arXiv.org 5 February 2024 • Computer Science, Medicine

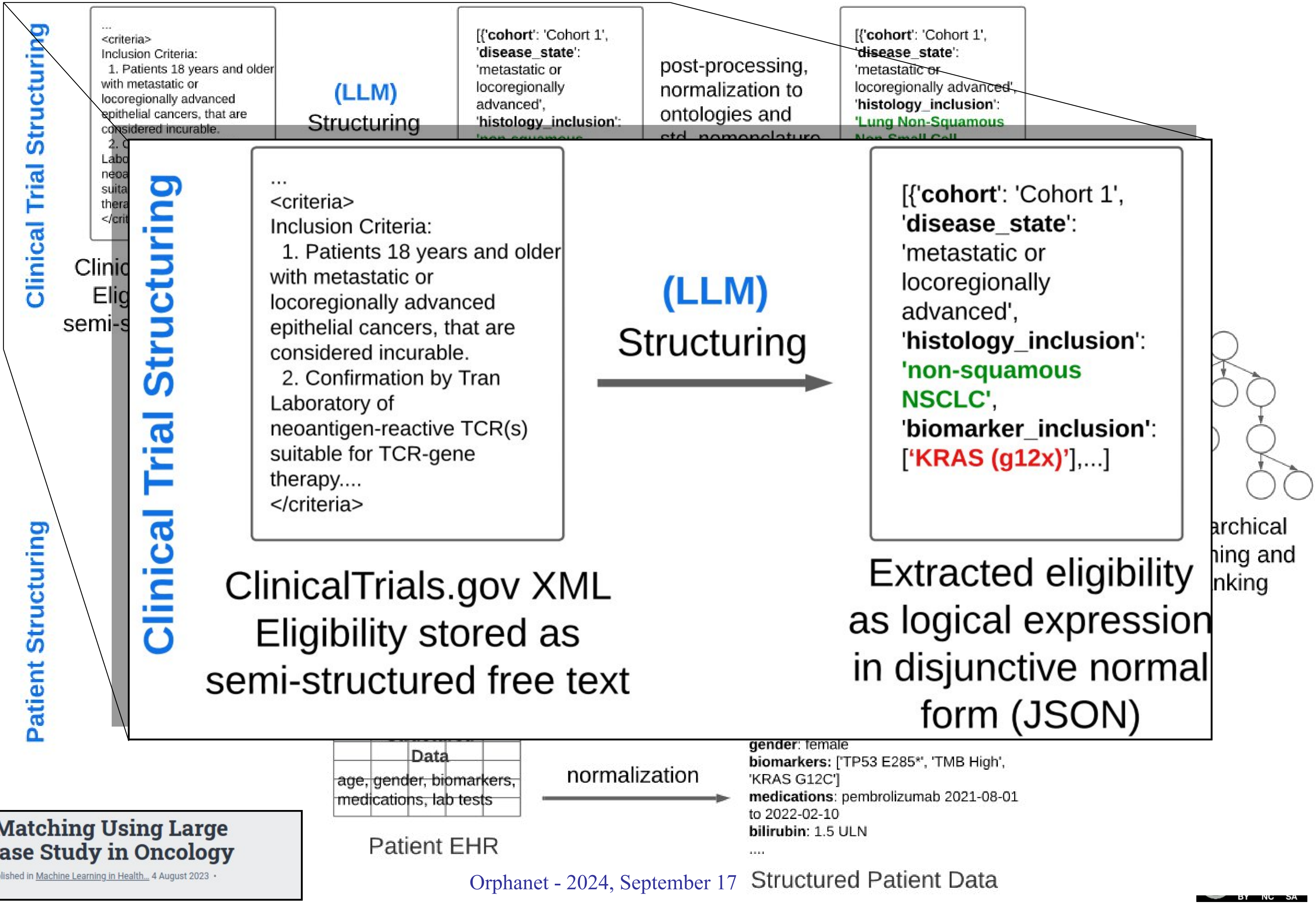
LLMs for patient selection : different possible strategies

- Convert free-text eligibility criteria to formal and structured elements



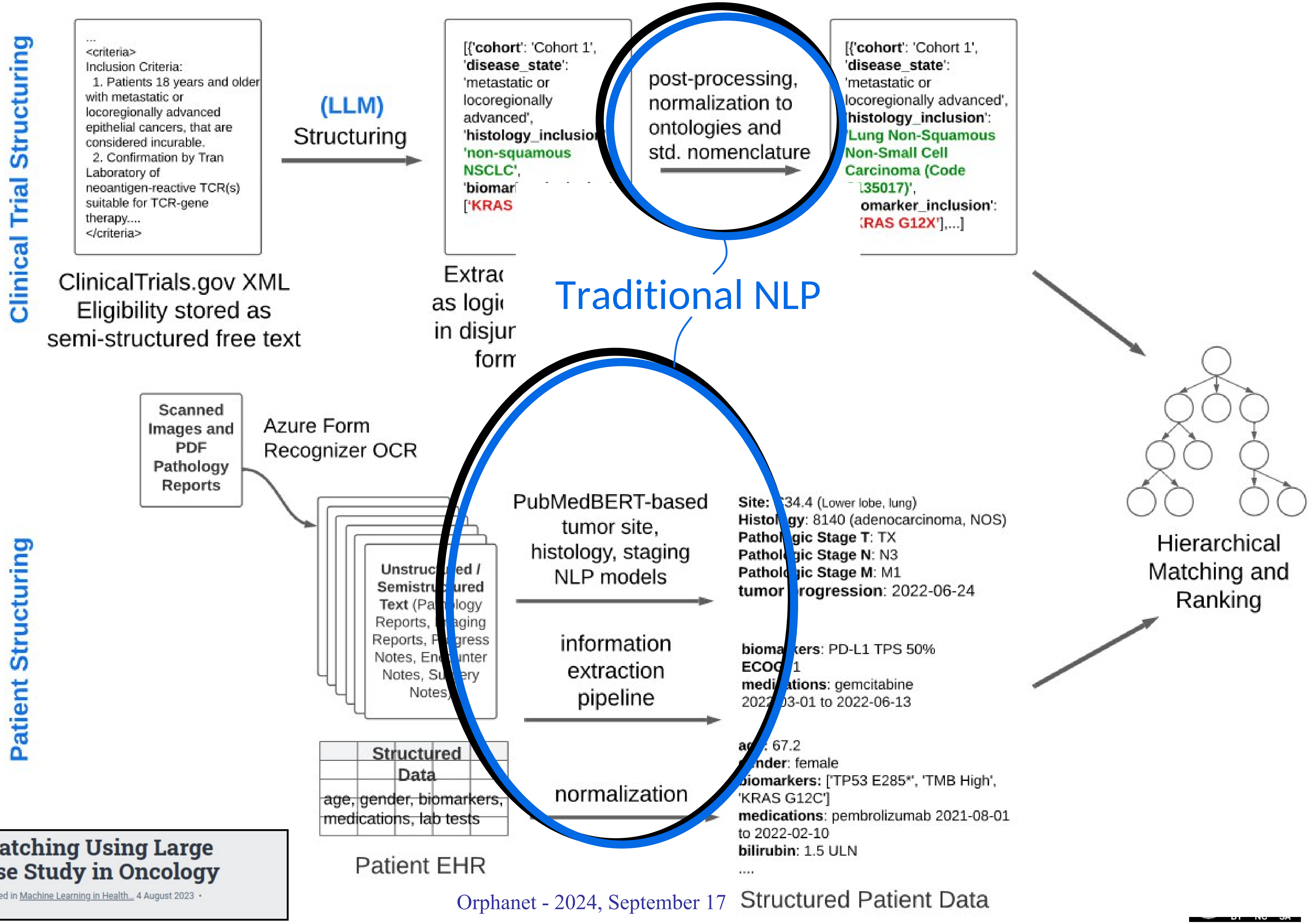
Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology

Cliff Wong, Sheng Zhang, +8 authors Hoifung Poon • Published in Machine Learning in Health... 4 August 2023 • Medicine, Computer Science



Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology

Cliff Wong, Sheng Zhang, +8 authors Hoifung Poon • Published in Machine Learning in Health... 4 August 2023 • Medicine, Computer Science

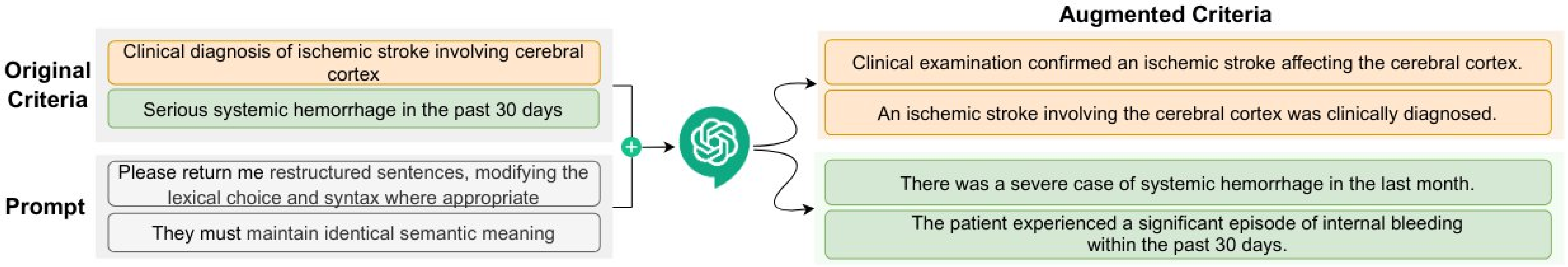


Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology

Cliff Wong, Sheng Zhang, +8 authors Hoifung Poon • Published in Machine Learning in Health... 4 August 2023 • Medicine, Computer Science

LLMs for patient selection : different possible strategies

- Convert free-text eligibility criteria to formal and structured elements
- Convert eligibility criteria to EHR-like sentences

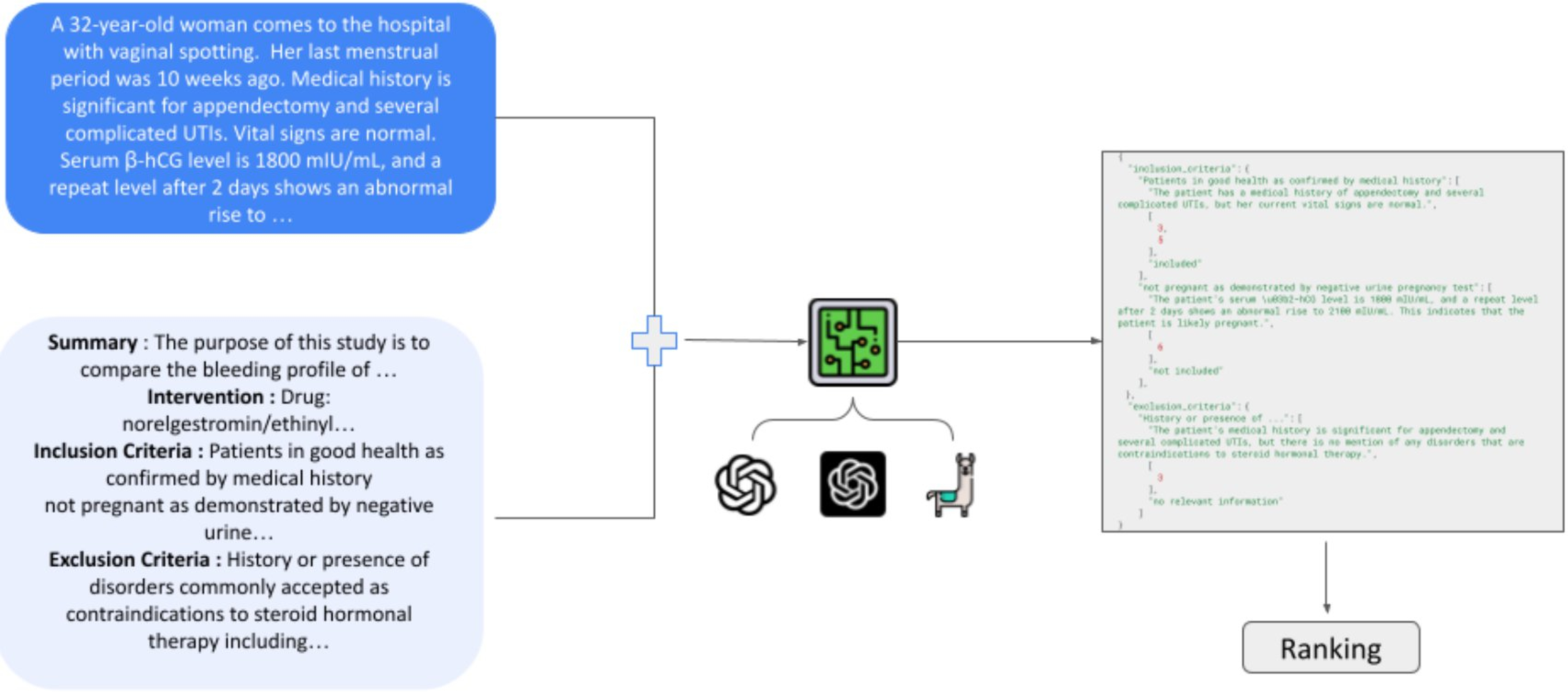


Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching.

Jiayi Yuan, Ruixiang Tang, +1 author Xia Hu • Published in AMIA ... Annual Symposium... 24 March 2023 • Computer Science, Medicine • AMIA ... Annual Symposium proceedings. AMIA Symposium

LLMs for patient selection : different possible strategies

- Convert free-text eligibility criteria to formal and structured elements
- Convert eligibility criteria to EHR-like sentences
- Ask the LLM to compare patients' EHRs and eligibility critera, all at once



Distilling Large Language Models for Matching Patients to Clinical Trials

Mauro Nievas, Aditya Basu, Yanshan Wang, Hrituraj Singh less · Published in arXiv.org 15 December 2023 ·
Computer Science, Medicine

Conclusion (1/2)

“ Clinical trial matching is a key process in health delivery and discovery. In practice, it is plagued by overwhelming unstructured data and unscalable manual processing.

[…]

Initial findings are promising: out of box, cutting-edge LLMs, such as GPT-4, can already **structure elaborate eligibility criteria of clinical trials and extract complex matching logic (e.g., nested AND/OR/NOT)**. **While still far from perfect, LLMs substantially outperform prior strong baselines and may serve as a preliminary solution to help triage patient-trial candidates with humans in the loop.** Our study also reveals a few significant growth areas for applying LLMs to end-to-end clinical trial matching, such as context limitation and accuracy, especially in **structuring patient information from longitudinal medical records**. ”

Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology

Cliff Wong, Sheng Zhang, +8 authors Hoifung Poon • Published in Machine Learning in Health... 4 August 2023 •
Medicine, Computer Science

Conclusion (2/2)

“

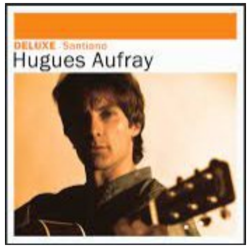
The integration of LLMs into clinical practice is not without its challenges, especially from legal and quality assurance perspectives.

LLMs are **susceptible to generating misleading or incorrect information**, a phenomenon known as “hallucination”, and **ensuring quality control may prove to be demanding**. The complexity and flexibility of LLMs correspondingly make validation regarding accuracy, safety, and clinical efficacy particularly challenging. The inherent opacity of AI models further exacerbates the difficulty of their application in critical, real-world scenarios. Techniques such as chain-of-thought prompting may guide the language model to reveal the reasoning process behind its outputs, thereby increasing the models' explainability. ”

Transforming clinical trials: the emerging roles of large language models

Jong-Lyul Ghim, Sangzin Ahn • Published in Translational and Clinical... 1 September 2023 • Medicine, Computer Science, Linguistics

Words



D' y penser j' avais le cœur gros

En doublant les feux de Saint-Malo

Lebensversicherungsgesellschaftsangestellter (Agglutinative language)

(employee of a life insurance company)

學而不思則罔，思而不學則殆

東京、マドリード、イスタンブール（トルコ）が争う２０２０年夏季五輪の開催地は７日（日本時間８日）、ブエノスアイレスでの国際オリンピック委員会（ＩＯＣ）総会で、ＩＯＣ委員約１００人の投票で決まる。

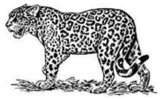
"Bag-of-words" representation

Human-readable	
Explicit semantics	
Dimensionality	
Semantic similarity	
Polysemy	
Spelling variants	
Multi-word expressions	
Unknown words	

Language = symbols

Vocabulary = all the words

bike ≠ bicycle ≠ mountain bike,
dog ≠ cat, Paris ≠ London

 = 

Tuebingen ≠ Tübingen,
tomorrow ≠ tommorrow

Dense representation (*embeddings*)

Two words close in vector space = two words that often share similar contexts

~~Two words close in vector space = two words with a close meaning~~

$w_{bike} \sim w_{bicycle}$ $\quad$ $w_{cat} \sim w_{tiger}$ $\quad$ $w_{bad} \sim w_{good}$

$w_{person} \sim w_{persons}$ $\quad$ $w_{Paris} \sim w_{London}$ $\quad$ $w_{small} \sim w_{tall}$

Dense representation

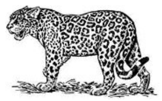
Human-readable		
Explicit semantics		
Dimensionality		Vector = dozens or hundreds of dimensions
Semantic similarity		By construction, but implicit
Polysemy		
Spelling variants		
Multi-word expression		Arithmetic operations on word vectors
Unknown words	?	Depending on the method, see later

Contextual representation

Human-readable	
Explicit semantics	
Dimensionality	
Semantic similarity	
Polysemy	
Spelling variants	
Multi-word expression	
Unknown words	

Vector = dozens or hundreds of dimensions

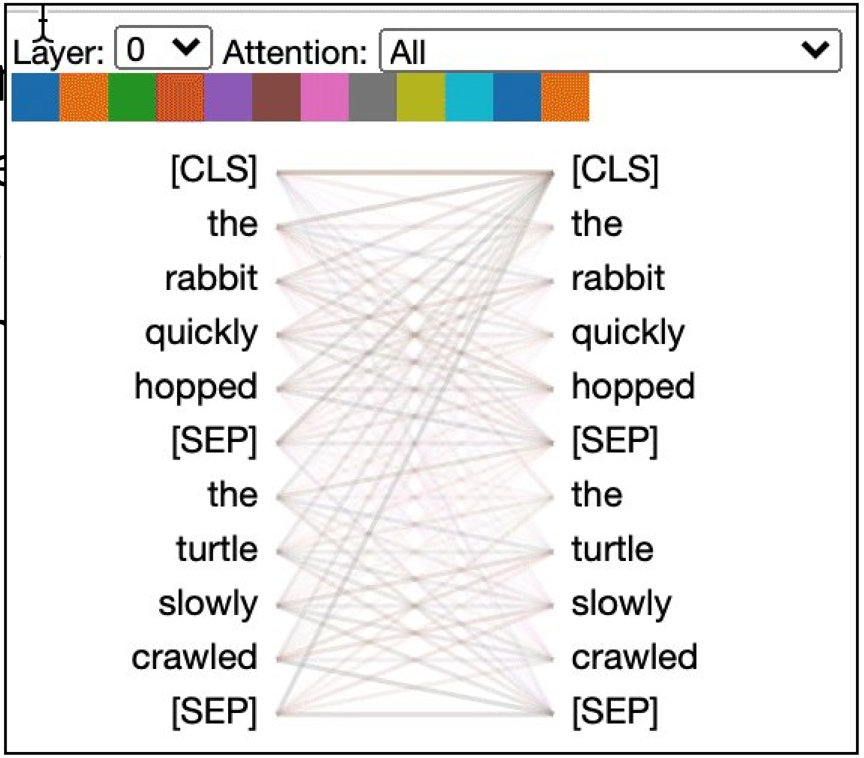
By construction, but implicit

Contextual dense representation

- Static representation: one token = one vector
  - We handle an « embedding matrix » (*N* x *d*)
  - The token vector is the same for each of its occurrences in the corpus

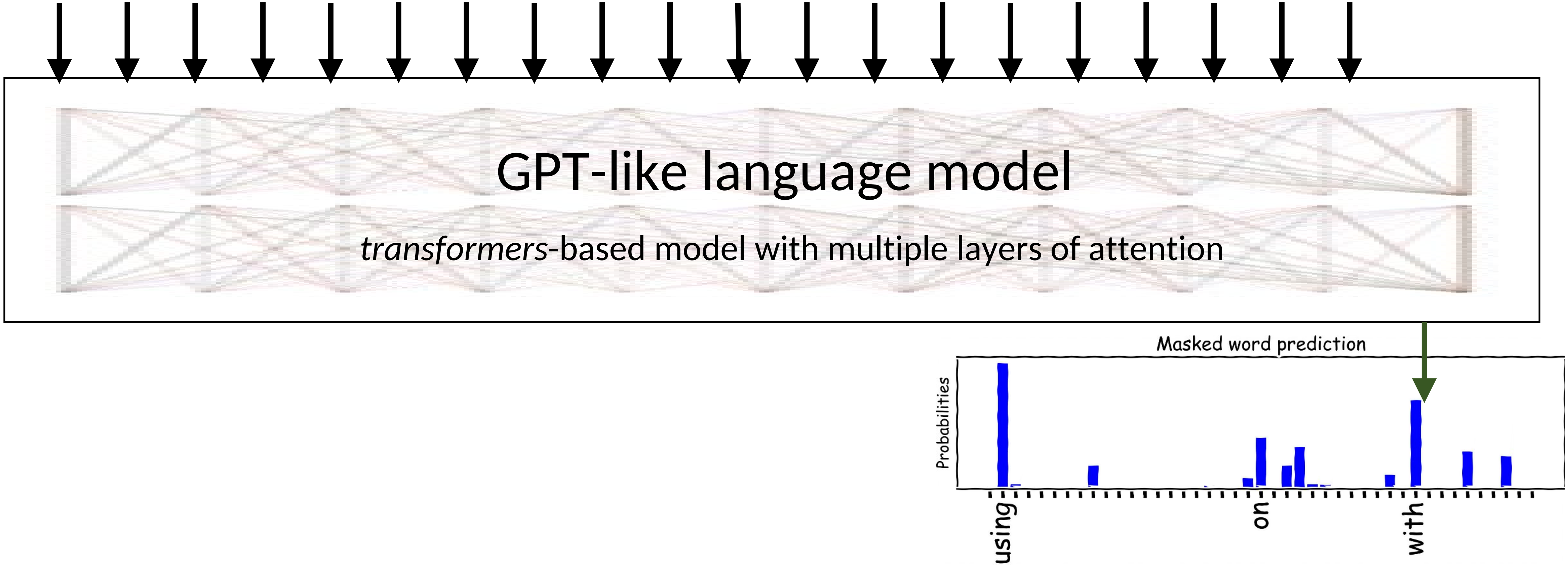
- Contextual representation: vector calculation in co
  - The calculation of the representation is integrated in the
  - e.g: ELMo, ULMFit, transformer-based models such as B
  - The preceding (and following) words affect the represen (usually through an attention mechanism...)



https://github.com/jessevig/bertviz/

Auto-regressive language models (*decoder-style*, *e.g. GPT*)

```
User: What is a large language model?
Assistant: A large language model is a type of artificial intelligence (AI)
model designed to understand and generate human-like language. These models
are built
```



By the way, not all of this is true. LLMs are not predicting words

```
c ' est un fame ##ux trois - mats , fin com ##me un o ##ise ##au
( hiss ##ez ha ##ut , sant ##iano )
di ##x - hui ##t n ##œ ##ud ##s , qu ##at ##re cents ton ##nea ##ux
je sui ##s fi ##er d ' y et ##re mate ##lot

tie ##ns bon la vague , et tie ##ns bon le vent
( hiss ##ez ha ##ut , sant ##iano )
si die ##u ve ##ut , to ##uj ##ours dr ##oit dev ##ant
( no ##us iron ##s ju ##s ##qu ' a san francisco )

je par ##s pour de long ##s moi ##s en lai ##ssa ##nt margot
( hiss ##ez ha ##ut , sant ##iano )
d ' y pens ##er , j ' ava ##is le c ##œ ##ur gr ##os
( en do ##ub ##lan ##t les fe ##ux de saint mal ##o )
```

WordPieces
- A vocabulary of predefined size, composed of ngrams of characters
- Vocabulary chosen to maximize the frequency of ngrams
- Possibility of a multilingual tokenizer

Multimodal dense representation

Another advantage of dense representations:

they can be mixed with representations of something else

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever
Zero-Shot Text-to-Image Generation

Multimodal dense representation

Another advantage of dense representations:

they can be mixed with representations of something else

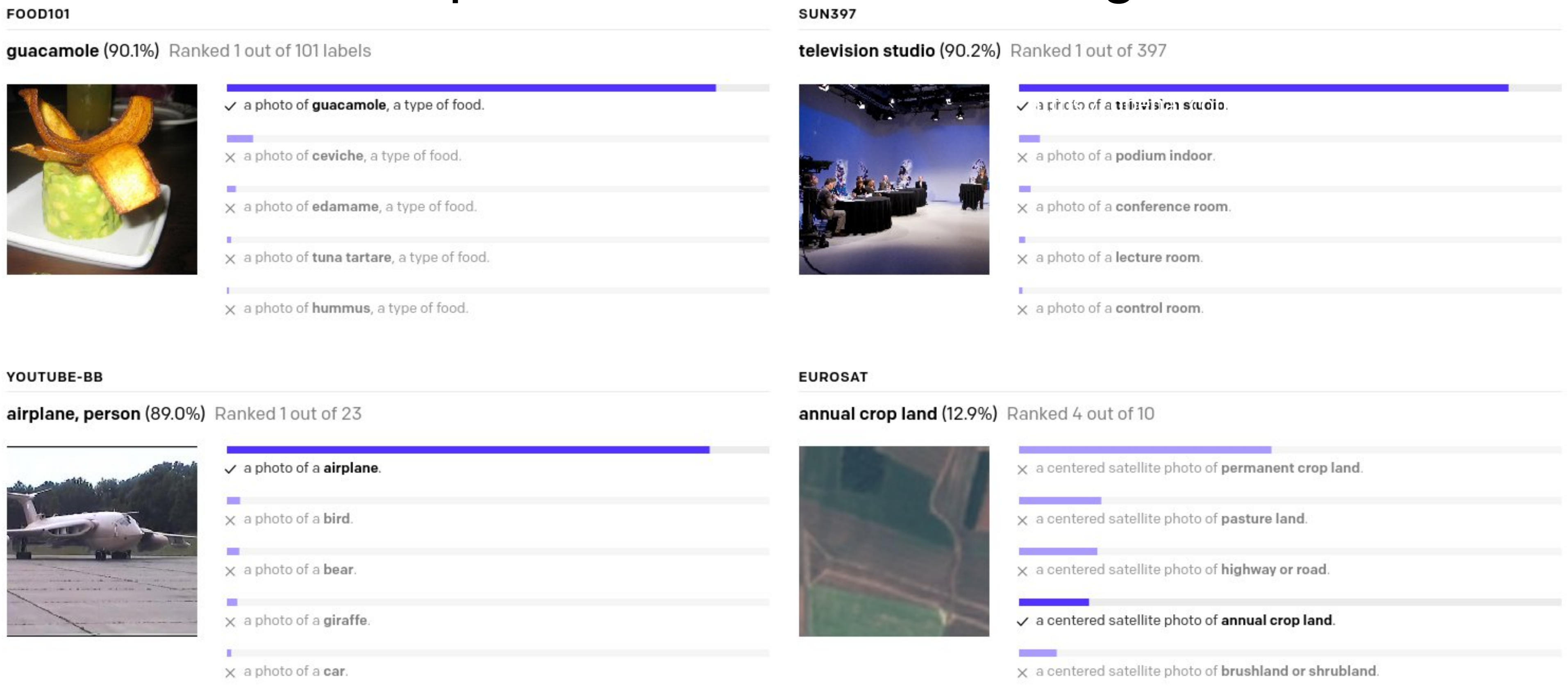
DALL·E 2

https://openai.com/dall-e-2/

Multimodal dense representation

Another advantage of dense representations:

they can be mixed with representations of something else



https://openai.com/blog/clip/

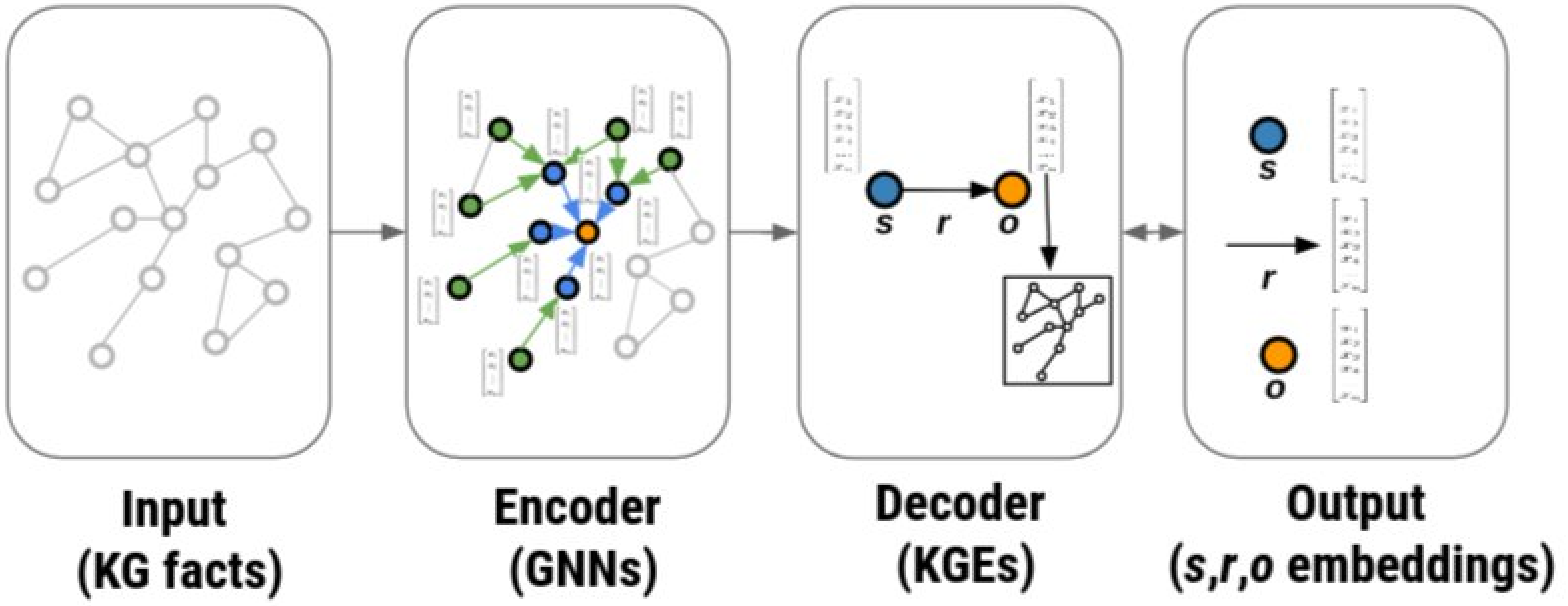
Alec Radford et al.
Learning Transferable Visual Models From Natural Language Supervision
Feb. 2021

Multimodal dense representation

Another advantage of dense representations:
they can be mixed with representations of something else

Knowledge graph embeddings

Luca Costabello, Sumit Pai, Nicholas McCarthy, Adrianna Janik
Knowledge Graph Embeddings Tutorial: From Theory to Practice
ECAI 2020



(Source:
Giuseppe Futia)