



#### A quick overview of Natural Language

### **Processing for Information Extraction**

## and Scientific Surveillance

#### Xavier Tannier xavier.tannier@sorbonne-universite.fr

Orphanet, September 17, 2024



Attribution – NonCommercial - ShareAlike CC BY-NC-SA

# Appetizer





Allez sur wooclap.com

Entrez le code d'événement dans le bandeau supérieur

Activer les réponses par SMS

Code d'événement

### Why Natural Language Processing?



l, place du Parvis Notre-Dame Prénon 75181 PARIS Cedex 04

Date Naissance : 27/10/1955

SCINTIGRAPHIE AU GALLIUM 67

Myofasciite à macrophages, dans un contexte d'aggravat

Première scintigraphie ; biopsie montrant deux ilôts carac

TECHNIQUE: Examen réalisé 48 heures après injection

Images par plans centrés sur gamma-caméra: Symbia Sie

épaules avec une activité diffuse plus ou moins homogèn

Sur la ceinture pelvienne, on note une activité musculair

muscles fessiers et des muscles des cuisses de façon très

Les articulations des hanches sont également relativeme

Sur la partie inférieure des cuisses et sur les mollets, on n

importantes, bien délimitées, hétérogènes au maximum pe

existe une visualisation inflammatoire également des deu

des articulations, de même que sur les deux chevilles. Su

Fixation musculaire hétérogène et bien caractéristique, in

Il est bien difficile de faire la part entre la responsabilité

responsabilité des entésopathies qui sont constantes chez

Docteu

chevilles, les genoux, les hanches et les épaules.

Merci de me tenir au courant, cordialement.

NIP : 240709380

CLINIOUE:

l'impotence fonctionnelle

lymphocytes sur le deltoïde

moitié supérieure des deux bras

hétérogénéité postéro- interne

CONCLUSION:

EV

Univet'sité Paris V René Descartes

POLE IMAGERIE ET **EXPLORATIONS FONCTIONNELLES** 

SERVICEDE MEDECINE NUCLEAIRE Secrétariat: 0142348241 Télécopie' 01 42 34 83 13 nstallation nO IS 441 L2B

Chef de Service Pr N (IL MORETTI) Dr N. CAILLAT-VIGNERON

Majtre de Conférenc alricien Hospitalier Biohysique, Médecine Nucléain Hématologie 0142348737

RESULTAT: adine, caillatvigneron/athtd aphp, fr Les clichés réalisés sur la ceinture scapulaire montrent un

Dr J.F. TOUSSAINT Maitre de Conférence Patricien Hospitalier Physiologie, Cardiologie 01 42 34 82 40 jean - françois toussaint/a/htd.aphp.fr Dr B. BOURGEOT

ssistant hospitalo-s 01 42 34 82 42 benopit bourgeot@htd aphp fr

Dr E.BLOCH Dr T. KONDO Attachés de Cardiologie

G. EL DEEB 1 42 34 82 42 Cadre Médico-Techniqu M. G. GARCIA

42 34 85 97 gabriel.garcia@htd.aphp.fr

> 05/12/2007 14-00 GALLIUM 67 138.51 GA 67 N°Lot:66282 Lot Trousse

#### **ASSISTANCE HÔPITAUX** PUBLIQUE DE PARIS HÔTEL - DIEU

Demandeur : Médecin de Ville 75009 Docteur : EVRARD-CTR MED EUROPE

#### Nº de Demande : 7070473

Date d'examen : 05/12/2007

Centre Hospitalier Régional Universitaire 7, rue Pasteur 67001 STRASBOURG Cedex Tél. 03 88 12 34 78

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

Né le 15/05/195

E., Edouard Opérateurs : Didier D..., Fabien F...

Assistant : Jeanne J.. Anesthésiste : Gérard G. Anesthésie : anesthésie général

Cholécystectomie - Cholangiographie - Cœlioscopie

#### Indication

- Patient de cinquante-cinq ans, adressé par le Dr A... pour la prise en charge d'une lithéas vésiculaire symptomatique connue de longue date.
- Il a comme antécédent une annendicectomie Son poids est de 105 kg pour 1.70 m. Il n'a pas de traitement en cours
- Il est donneur de sang de manière régulière : a priori il n'y a pas d'anémie. Le patient signale des douleurs abdominales diffuses qui remontent à 2005. Depuis un an, il décrit plus particulièrement des douleurs d'apparition rapide, avec platea
- disparition dans les deux heures cui suivent, au niveau de l'hypochondre droit. La plainte principale concerne ces douleurs dont l'intensité a été progressivement crois-

ce jour. Echographie des voies biliaires (01/06/2010) : calcification hyperéchogène mobile, décli 2 x 2 cm avec cône d'ombre postérieur

Indication de cholécystectomie par cœlioscopie plutôt que laparotomie, vu l'absen complications. Information modalités, bénéfices, risques.

Création du pneumopéritoine par méthode « ouverte ». Mise en place de quatre trocarts (optique à l'ombilie, épigastrique, latéral droit, opérateur à de la ligne médiane).

Corlioscopie montrant une vésicule tendue et volumineuse, aux parois ordémaciées reco d'importantes adhérences de l'épiploon, de libération difficile.

[E... Edouard, CRO 15/06/2010, 1/2]

Sujets de la catégorie Diabète		Auteur du sujet	Nombre de Rép. Lus		Date du dernier message
Le Parc Astérix fête ses 30 ans, ven Jouez et tentez de remporter 4 entre			omen	t en fa	mille
A lire avant de poster : La modération sur Doctissimo		Jacoline	0	626	30/08/2016 à 12:36 Jacoline
<ul> <li>Diabètes, insuline, activités</li> <li>physiques : comment gérer après 50</li> <li>ans ?</li> </ul>		Doctissimo	37	14 690	13/11/2015 à 20:03 Cyril
B Dopics utiles.		James	0	49 5	
😂 "le jeu à la con" bis repetita	1629	carangevine	81 405	837 9	
on poste sa glycémie du matin!	1214	Cyril	60 677	886 0	ALIGNI
abeloop un système qui a du plomb dans le	5	Cyril	203	2 99	
CEST ?	13	Crodiledundee	615	3 27	
angement de traitement car changement diagnostic = hypo et glycémie basse		Cris29	2	25	
nimes 640g		Mimichoco	3	43	
WS PERMIS DE CONDUIRE	2	ledom35	78	838	
ur se dire bonsoir, faire un coucou ou tres nouvelles	108	lol17dc	5 395	30 5	
abète et perte de poids :???:		lem50zg	41	21 2	Large capacit
ssible DT1 ?		toune54	5	73	natural challer with fa
yennes lecteurs de glycémie	99	petitsoleil33	4 941	91 5	LLMs includi has sig
potages: nos difficultés autour du ibète		Crodiledundee	12	165	questio A preli 70,639
psoils Et Resottes Rous Prévenis Le					10.05 /

Top suiets

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

Jaeff Hong\*2, Duong Dung\*1, Danielle Hutchinson1, Zubair Akhtar1, Rosalie Chen1, Rebecca Dawson<sup>1</sup>, Aditya Joshi<sup>+2</sup>, Samsung Lim<sup>13</sup>, C Raina MacIntyre<sup>1</sup> and Deepti Gurdasani<sup>145</sup>,

<sup>1</sup>Kirby Institute, University of New South Wales, Sydney, Australia <sup>2</sup>School of Computer Science & Engineering, University of New South Wales, Sydney, Australia 3School of Civil and Environmental Engineering, University of New South Wales, Sydney, Australia William Harvey Research Institute, Oueen Mary University of London, London, UK 5 School of Medicine, University of Western Australia, WA, Australia jaeff.hong@gmail.com, {duong.dung, danielle.hutchinson, zubair.akhtar, rosalie.chen, rebecca.dawson, aditya.joshi, s.lim, r.macintyre, d.gurdasani @unsw.edu.au

#### Abstract

Isuru Hettigoda

NeuraSense Research

Relation Extraction from News Articles (RENA) is a browser-based tool designed to extract key entities and their semantic relationships in English language news articles related to infectious diseases. Constructed using the React News Articles (RENA), uses decoder-only foundation models (also known as 'large language models', i.e., LLMs) to extract semantic relations in infectious disease-related news articles in the English language1 at the document level. When an epidemiologist enters a news article, RENA ex-

racts a list of entities and relations present in the article. This streamlines an automated process for epidemiologists nd researchers who need a method to acquire a large conregate of structured relations from their selected article. By aking use of RENA, they can be aided in their research without having to read a large set of news articles or delve to manual data curation.

Whilst many previous relation extraction (RE) tasks have cused on the sentence level, it is evident that many relaions exist between different sentences, presenting another hallenge in extracting relations at the document level (Xu, Chen, and Zhao 2021). We assume a simplistic definition f a relation: a relation connects exactly two entities. In he context of epidemic intelligence, the sentence 'A patient ied due to COVID-19 today' results in the relations {death umber: '1', relation: 'death of', infectious disease: COVID-9'}. {death number: '1', relation: 'occurred on', event date: oday'} and {infectious disease: COVID-19, relation: 'ocurred on', event date; 'today' }. In this specific case, if two elations are identified, the third can be inferred.

RENA can be used by epidemiologists, public health ofcials and teachers or students of public health to extract formation from news articles of interest. While the utility RENA is for infectious disease-related epidemic intellience, it can potentially be used for news articles across doains and application areas, including journalists/news pubshers who want to verify and investigate information across nultiple sources.

#### Architecture

gure 1 shows the architecture of RENA. In the model aining phase, we generated a set of 300 synthetic arcles, annotated with relevant entities and relationships

Orphanet - 2024, Septembe

#### ABSTRACT

ALIGNING LARGE LANGUAGE MODELS FOR CLINICAL TASKS

Supun Manathunga

University of Peradeniya

NeuraSense Research

Large Language Models (LLMs) have demonstrated remarkable adaptability, showcasing their capacity to excel in tasks for which they were not explicitly trained. However, despite their impressive natural language processing (NLP) capabilities, effective alignment of LLMs remains a crucial challenge when deploying them for specific clinical applications. The ability to generate responses with factually accurate content and to engage in non-trivial reasoning steps are crucial for the LLMs to be eligible for applications in clinical medicine. Employing a combination of techniques including instruction-tuning and in-prompt strategies like few-shot and chain-of-thought prompting has significantly enhanced the performance of LLMs. Our proposed alignment strategy for medical question-answering, known as 'expand-guess-refine', offers a parameter and data-efficient solution. A preliminary analysis of this method demonstrated outstanding performance, achieving a score of 70.63% on a subset of questions sourced from the USMLE dataset

s · Clinical Applications · Alignment Strategy · Medical Ouestion-Answering

ence (AI) research was mainly focusing on specific tasks like mastering the game of e advancement of the deep learning techniques, particularly the transformer models s interact with AI models, especially in the realm of Natural Language Processing cture has laid the groundwork for Large Language Models (LLMs), which exhibit tasks they weren't explicitly trained for, a phenomenon observed as these models [4]. The development of these expansive LLMs may be bringing us closer to the neral Intelligence [5], [6].

ext corpora containing medical knowledge and this knowledge becomes ingrained izing on the task-agnostic nature of LLMs, they find utility across a spectrum of m information retrieval and summarization to decision-making and diagnostics [4] of clinical medicine, it is imperative for these models to aptly grasp the nuances of and engage in reasoned analysis with a certain level of discernment. Mechanisms cination, guarding against harmful content, and ensuring the model's alignment with

P capabilities exhibited by LLMs, they need to be aligned before deploying for

#### NLP tasks

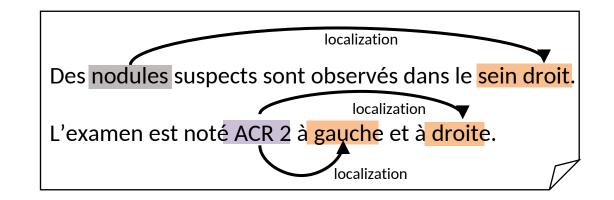
9 hémocultures positives le 26/6/15 à staphylocoque aureus méticilline sensible. BACTERIAL RESISTANCE

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

PROCEDUREMEDICAL DEVICEANATOMYLa contraception par les dispositifs intra utérins

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

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







#### NLP tasks

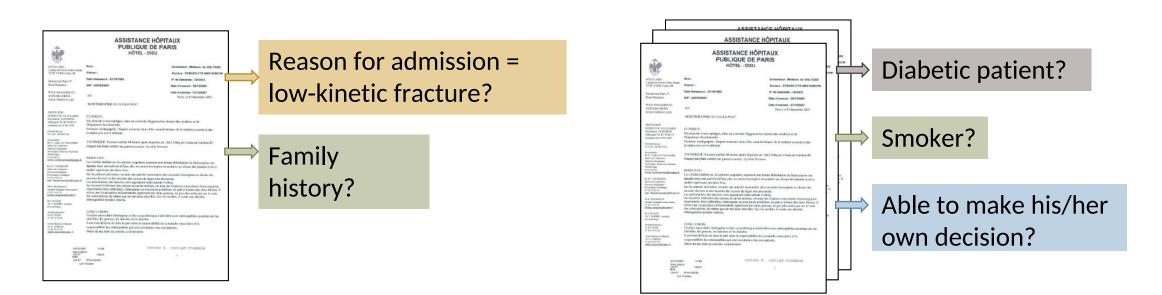
MAMMOGRAPHY: SHAPE LESION There is a 1.8 cm round mass with a circumscribed LATERALITY LOCALIZATION margin in the left breast in the anterior depth central SHAPE LESION to the nipple. There also is a 1.4 cm oval mass with an LATERALITY obscured margin in the left breast in the anterior depth of the inferior region. LOCALIZATION SCORE ASSESMENT: BI-RADS Category 3

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









#### **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 DISEASEintermittent claudication arteriopathy of the limbs arterial ulcer venous ulcer arterial insufficiency gangrene acute limb ischemia(Approximate) term matching in the documentsVASCULAR DI Patient with arterial ulcers of lower limbs, followed at St J with skin graft in October 20	f the oseph
--	----------------





#### **Rule-based systems**

- 1. Terminological approach
- 2. Additional rules







10

#### **Rule-based systems**

1. Terminological approach

#### 2. Additional rules

#### 3. Trickier rules

secondary, progressive, fractured or suspected osteocondensation	ost[eé]ocondensa. {1,20} (suspect   secondaire   [ée]volutive)   (l[ée]sion   anoma{1,20}os.{1,30} (suspect   secondaire   [ée]volutive)   os. {1,30} (l[ée]sion   anomalie   imag(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[l[ée]sion   anomalie   image   nodule). {1,8(suspect   secondaire   [ée]volutive)   ((l[l[ée]sion   anomalie   image   nodule). {1,8(secondaire))   ((l[l[ée]sion   anomalie   image   nodule)s. {1,40}suspec?ts?).	e). {1,20}
post-operative anatomopathological tumor stage (pTNM)	([ycpP]{1,2}\s? (T([01234x] is)[abcdx]?) [,\s] {0,2} [ycp] {0,2}\s? (N[xo01234\+ (M[o01]? [\+x]?)?) ((T([01234x] is)[abcdx]?) [,\s] {0,2} [ycp] {0,2}\s? (N[xo012 \s? (M[o01]? [\+x]?)?)	
	Orphanet - 2024, September 17	<b>©000</b> BY NC SA 11

#### Rule-based systems: Pros & Cons

Rules

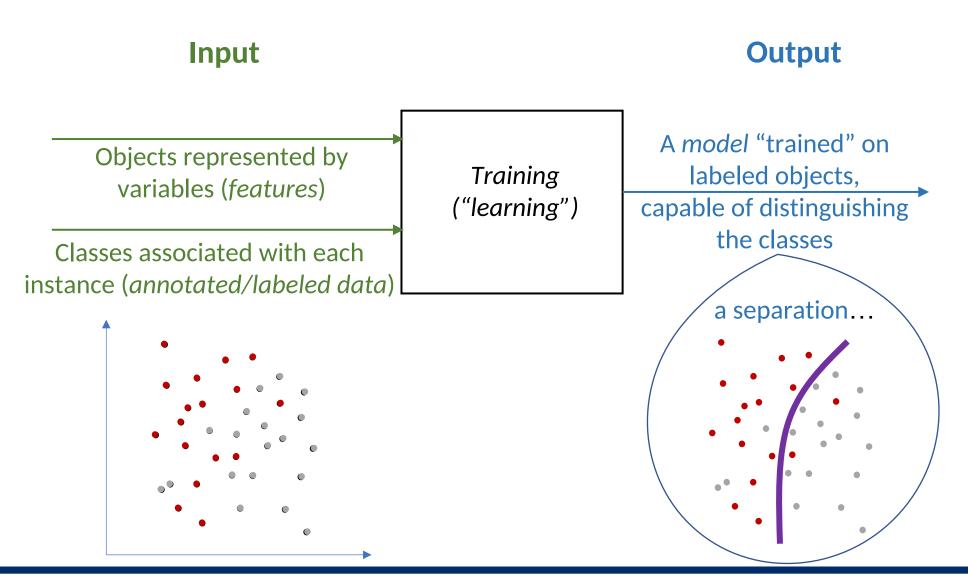
General performance	
Ease of implementation	
Need for human expertise	
Explainability	
Material resources	
Energy consumption	
Ease of maintenance	
Generalization to a different problem/context	





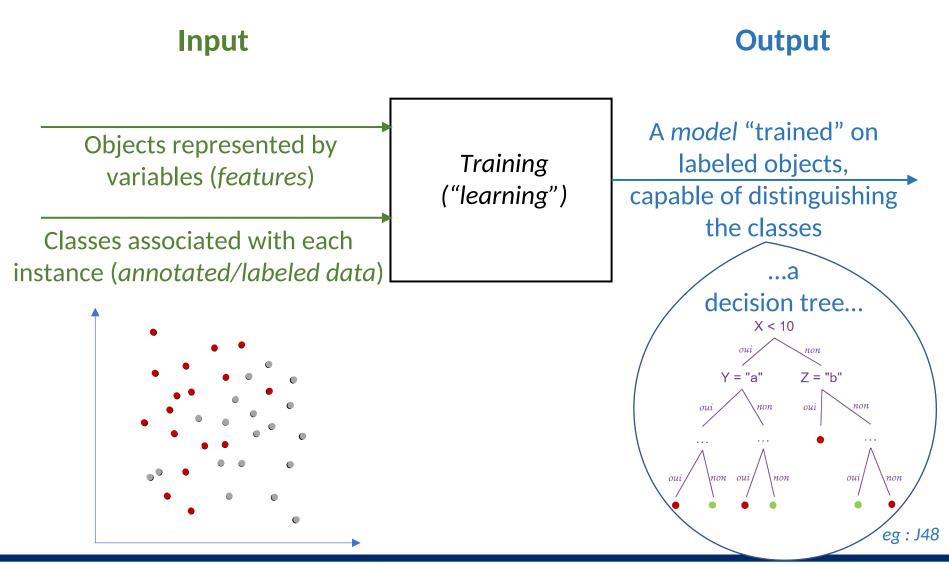
# Methods

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





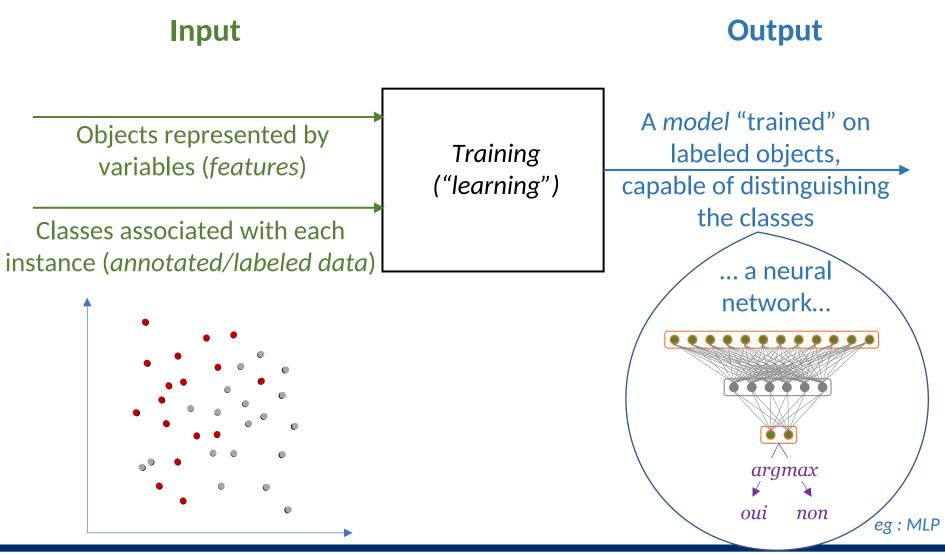






Orphanet - 2024, September 17

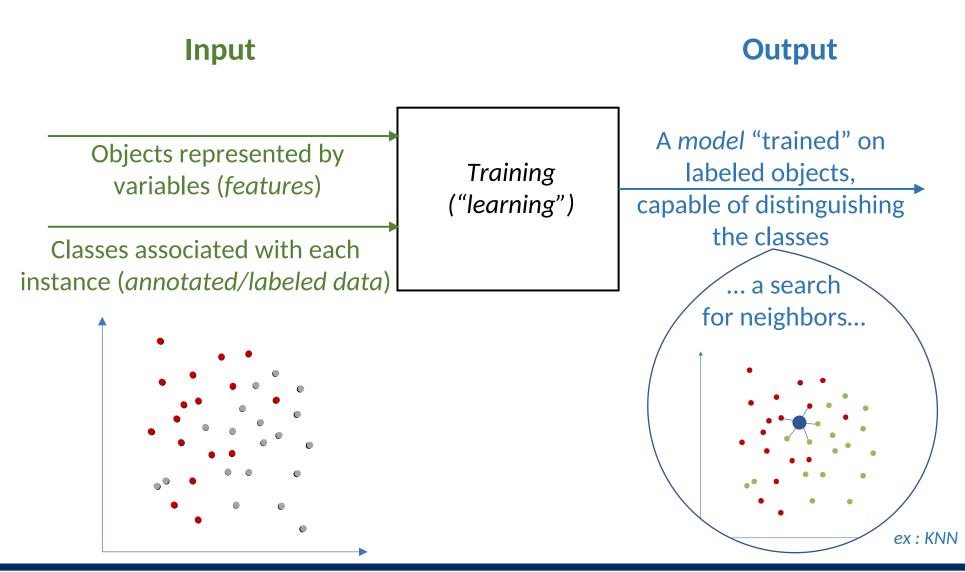






Orphanet - 2024, September 17







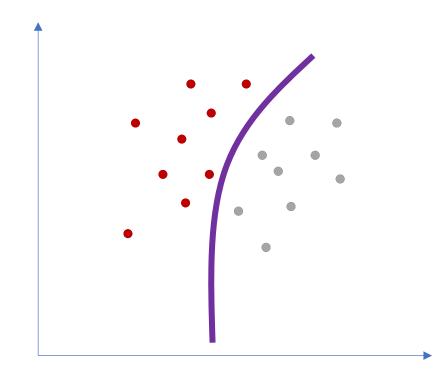


Orphanet - 2024, September 17

**30** NC SA 17

### (What is supervised learning?) Inference

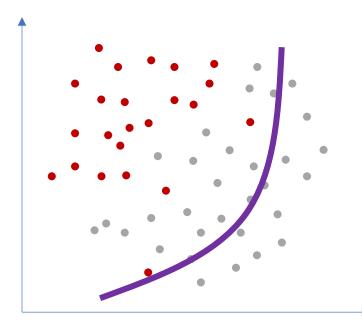
• The model is then applied to new, "unlabeled" data.







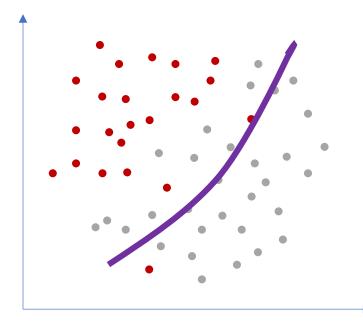
- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.
- Often an iterative process







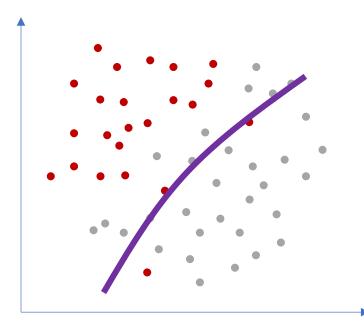
- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.
- Often an iterative process







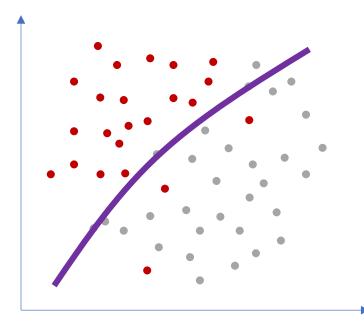
- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.
- Often an iterative process







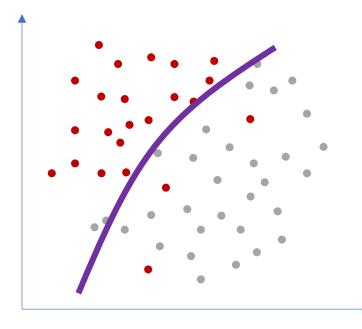
- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.
- Often an iterative process







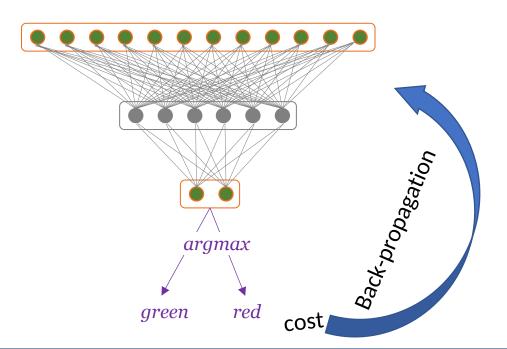
- Most often, learning (*training*) consists of minimizing the error committed by the system (*cost function* or *objective function*) by refining its parameters.
- Often an iterative process







- 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

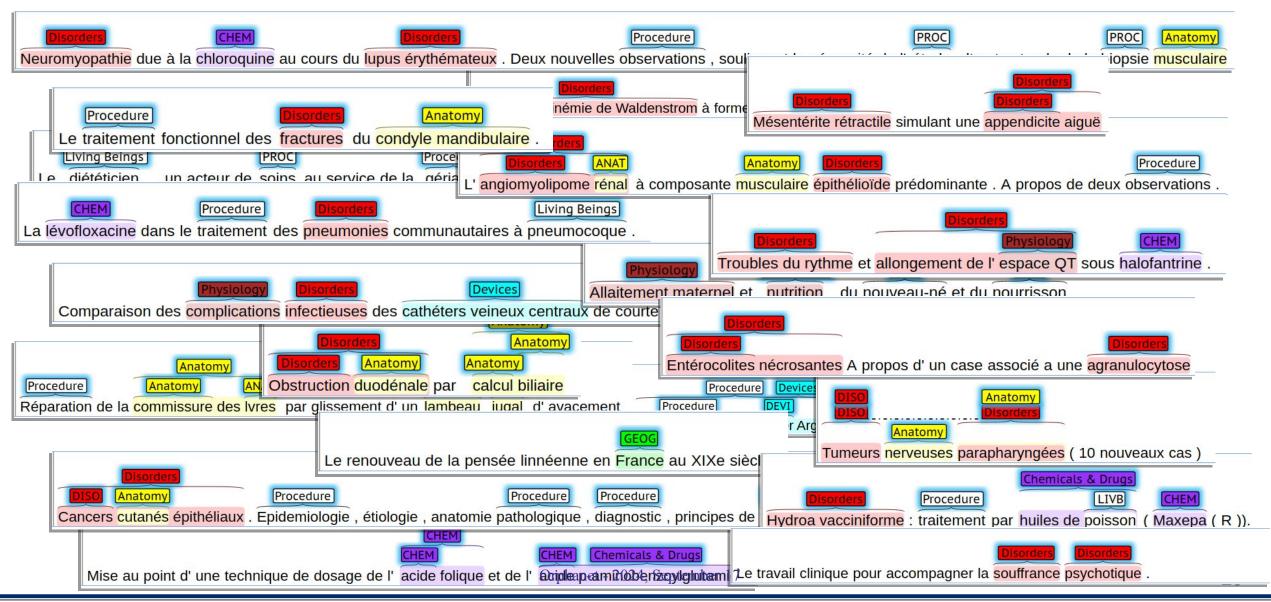






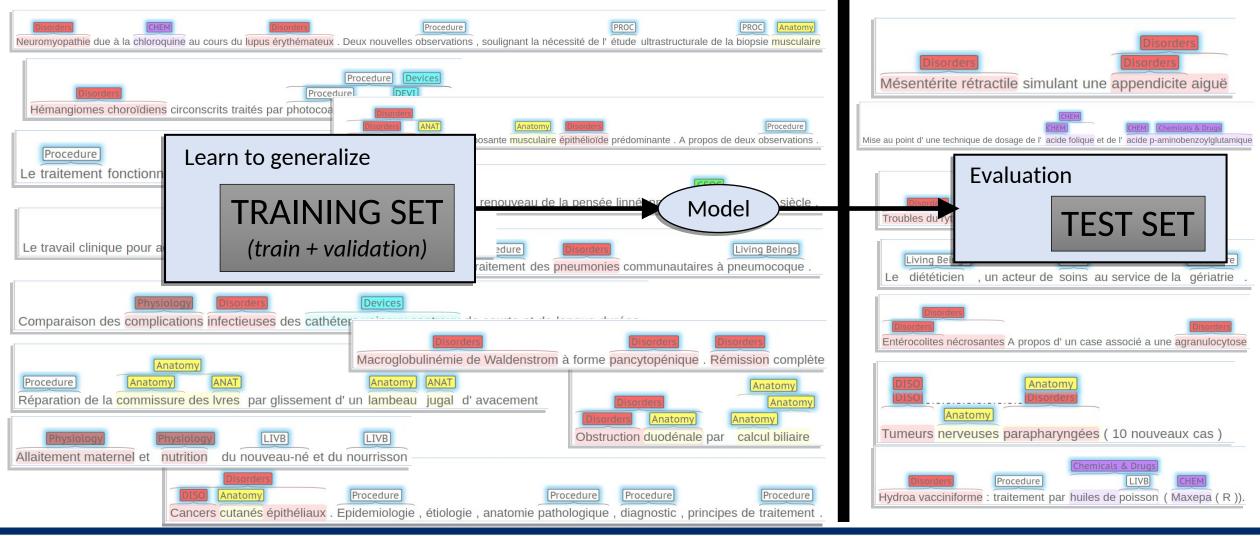
#### Supervised learning systems

#### 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.



Q

SORBONNE UNIVERSITÉ

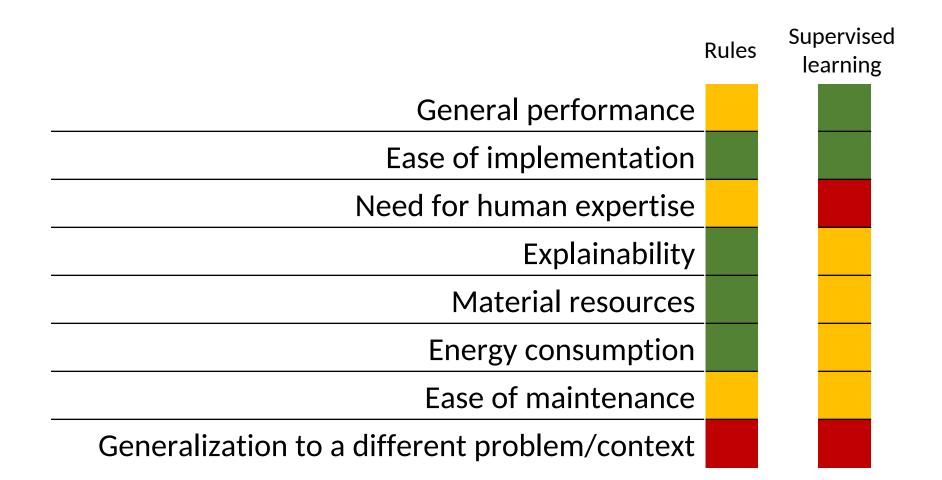
LIMICS 🝕

Orphanet - 2024, September 17



26

#### Supervised learning systems: Pros & Cons





27



# 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.





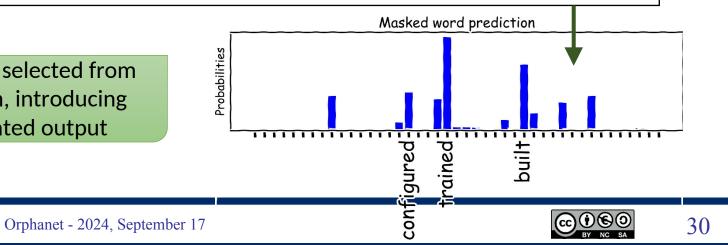
### Auto-regressive language models (decoder-style, e.g. GPT)

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

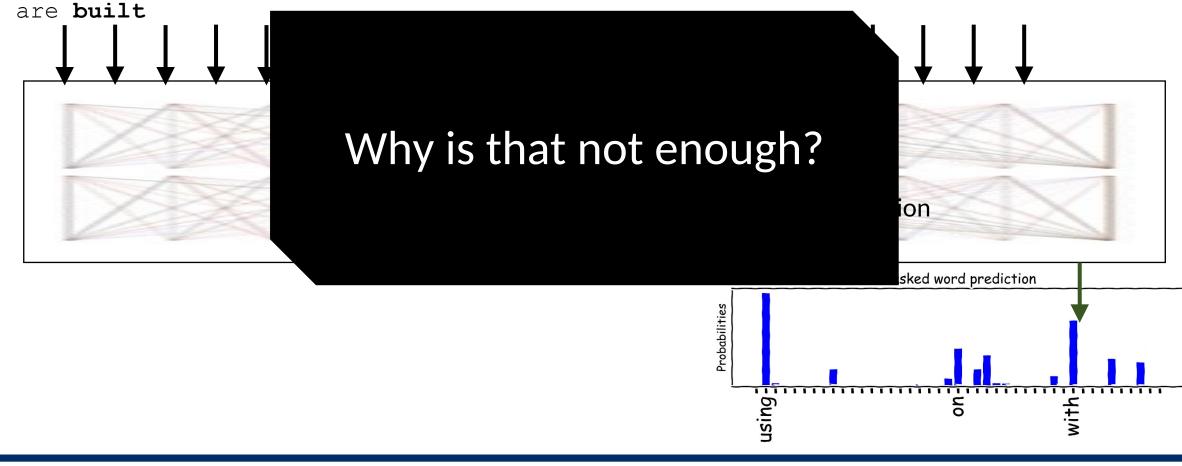
Next tokens are randomly selected from the predicted distribution, introducing variability in the generated output





### 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

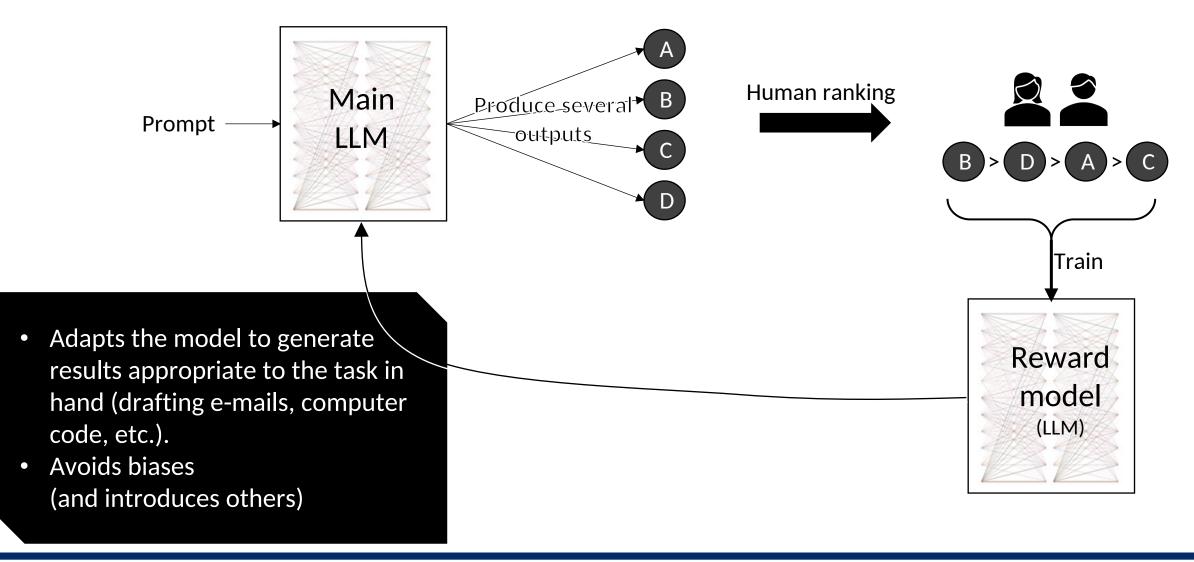


Orphanet - 2024, September 17

SORBONNE UNIVERSITÉ 



### **Reinforcement Learning with Human Feedback**







### 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:

(Answer: Manchester) Prompter - Q "Who is the father of Gwilym Lloyd George?" C1: "David Lloyd George is the father of Gwilym Lloyd George" "David Lloyd George" PLM - Q & C1 Prompter "Unknown" C2: "Manchester is the place of birth of David Lloyd George" **Standard Probing Iterative Prompting** 

Q: "What is the place of birth of Gwilym Lloyd George's father?"

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:

hts:	StrategyQA	Date Understanding	Sports Understanding
	Q: Yes or no: Would a pear sink in water? A: The density of a pear is about 0.6	Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?	Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."
	g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.	A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.	A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.
	SayCan (Instructing a robot)	Last Letter Concatenation	Coin Flip (state tracking)
	Human: How would you bring me something that isn't a fruit?	Q: Take the last letters of the words in "Lady Gaga" and concatenate them.	Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?
	Explanation: the user wants		
	something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring	A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a".	A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which
	the user an energy bar.	Concatenating them is "ya". So the	is an odd number. The coin started
	Plan: 1. find(energy bar) 2.	answer is ya.	heads up, so after an odd number of
	pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().		flips, it will be tails up. So the answer is no.

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*.





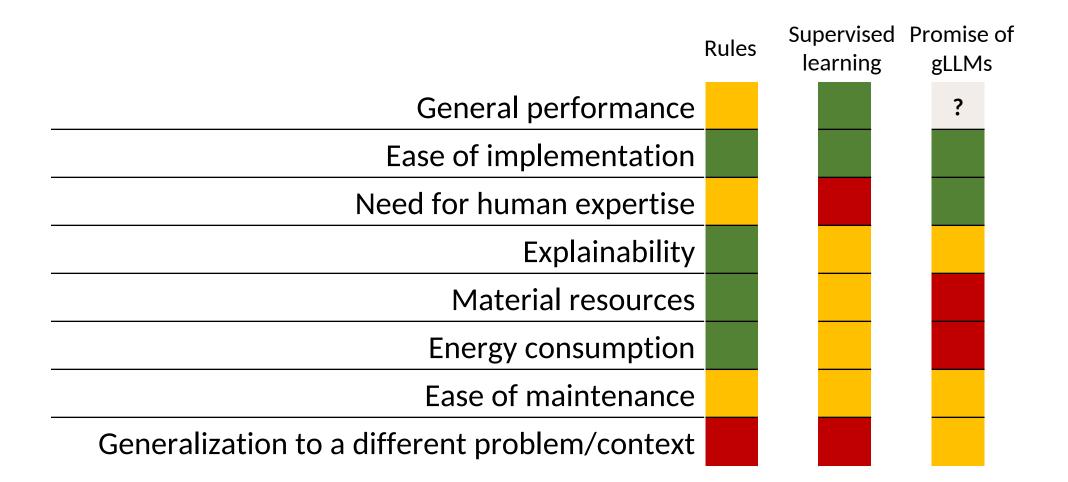
#### How can we use LLMs?

- 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 <u>https://www.promptingguide.ai/</u>)





#### Generative LLMs: Pros & Cons









Q

SORBONNE UNIVERSITÉ



# 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!





S

SORBONNE UNIVERSITÉ



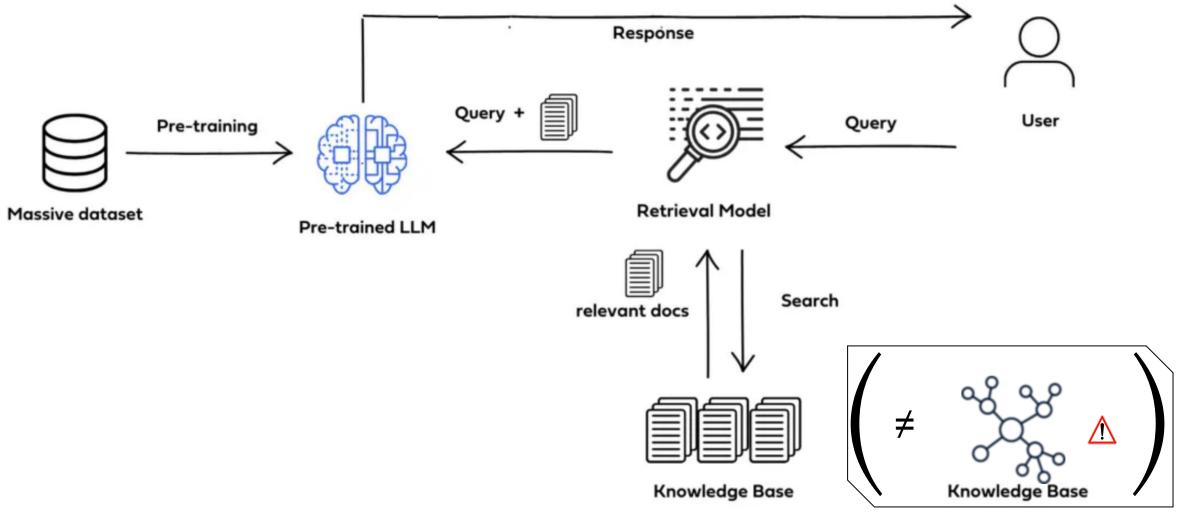


## Methods

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

The new search engine

#### **Retrieval-Augmented Generation**



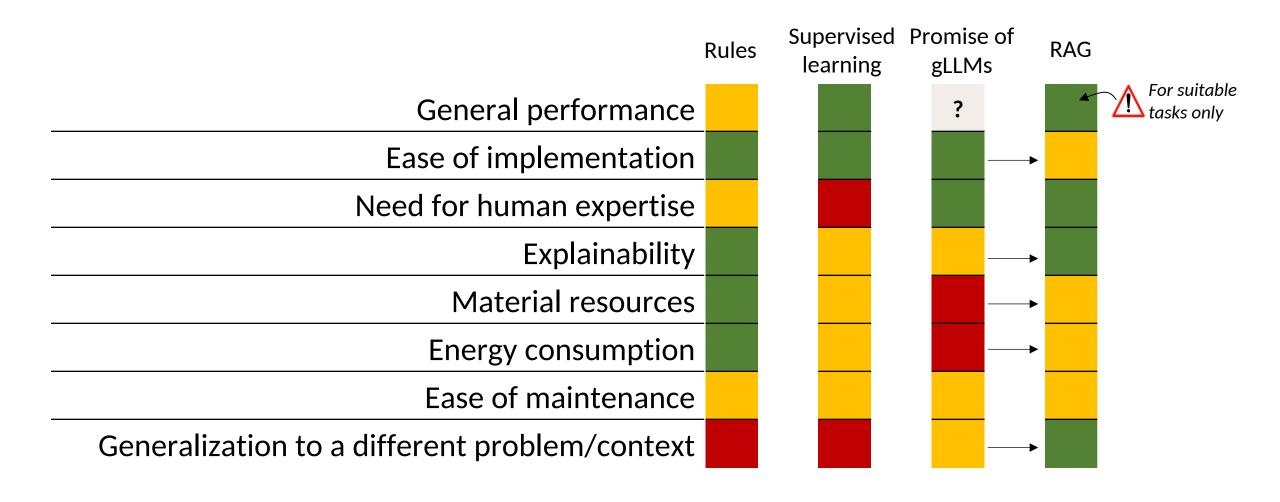
https://medium.com/@krtarunsingh/introduction-to-retrieval-augmented-generation-rag-and-its-transformative-role-in-ai-c07e35da7f01



Orphanet - 2024, September 17

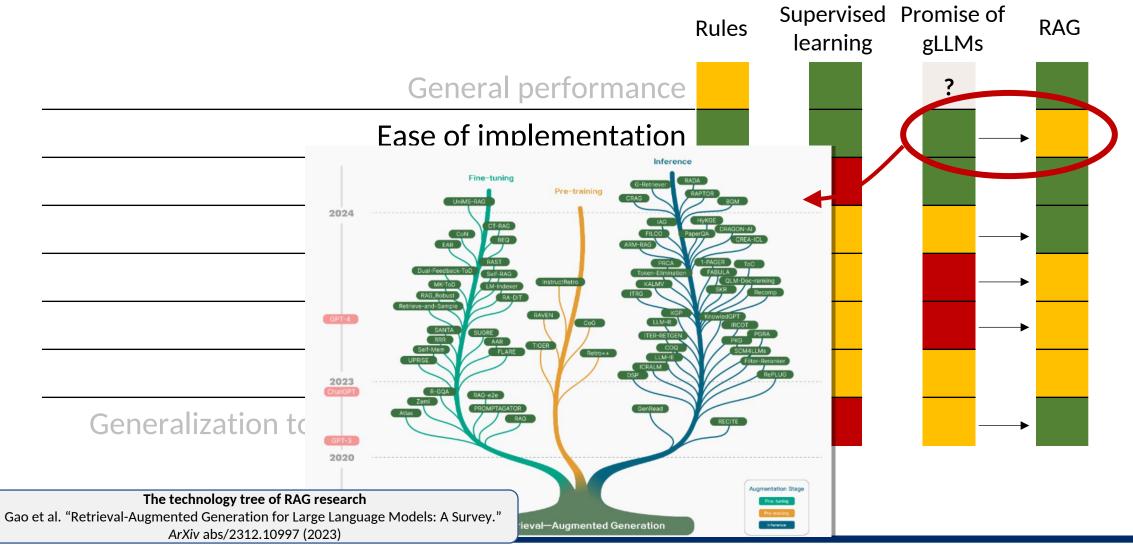


#### RAG: Pros & Cons





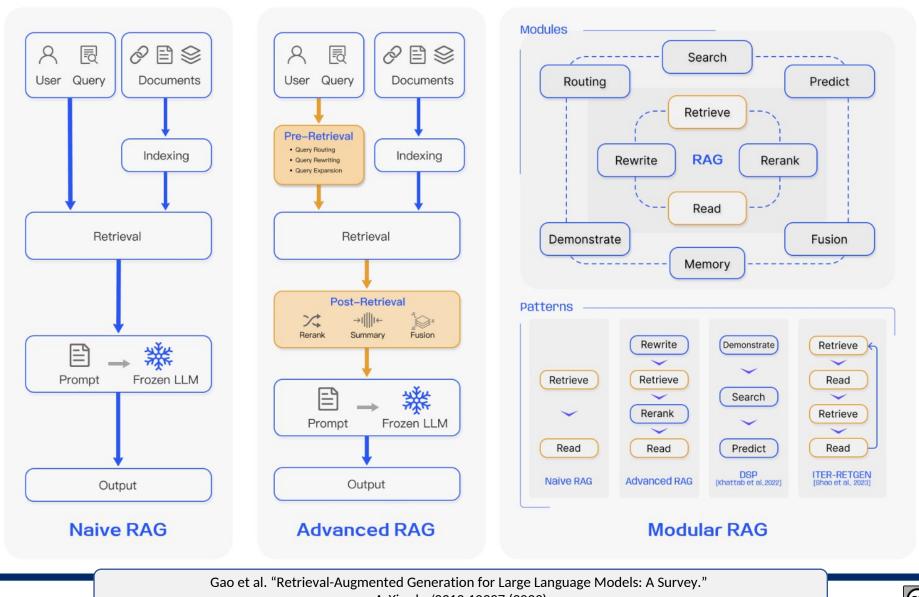
#### RAG: Pros & Cons



Orphanet - 2024, September 17



RAG

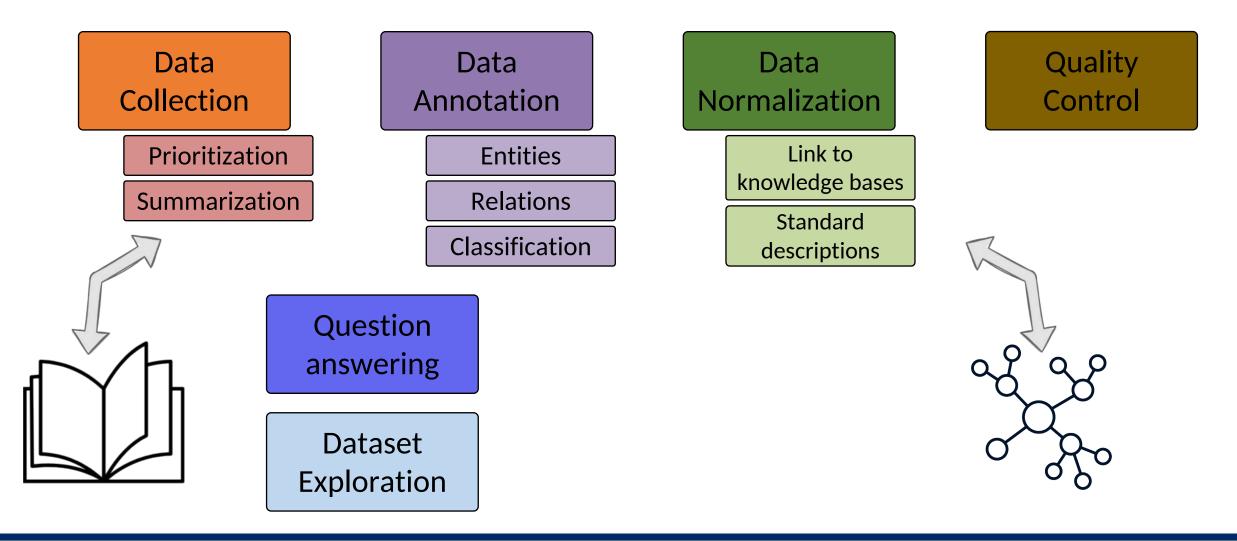




ArXiv abs/2312.10997 (2023)



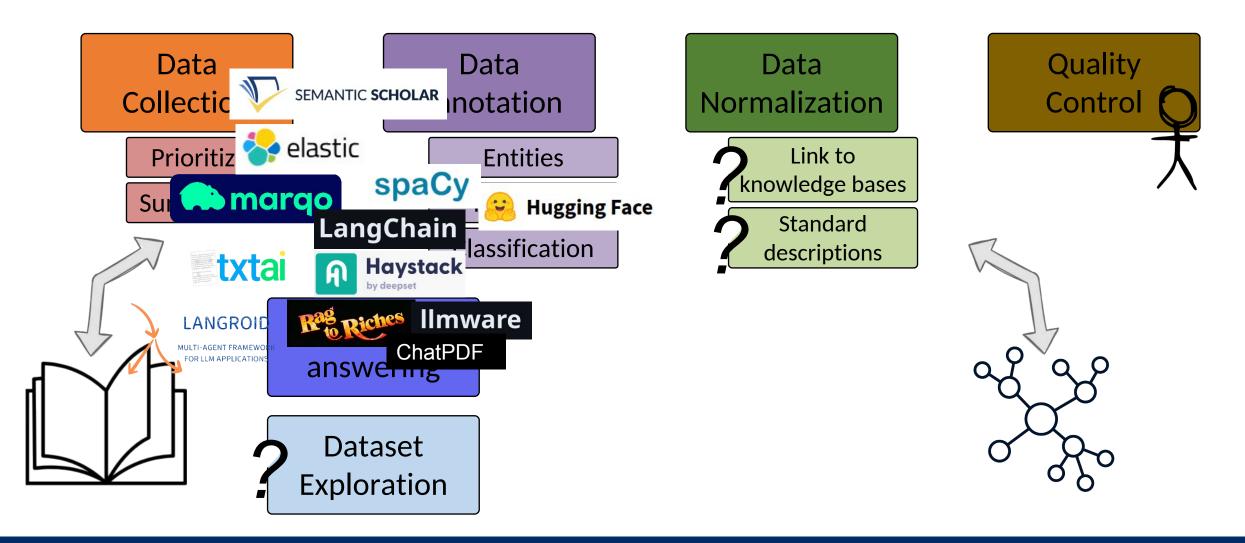
## NLP for scientific surveillance





Orphanet - 2024, September 17







Orphanet - 2024, September 17



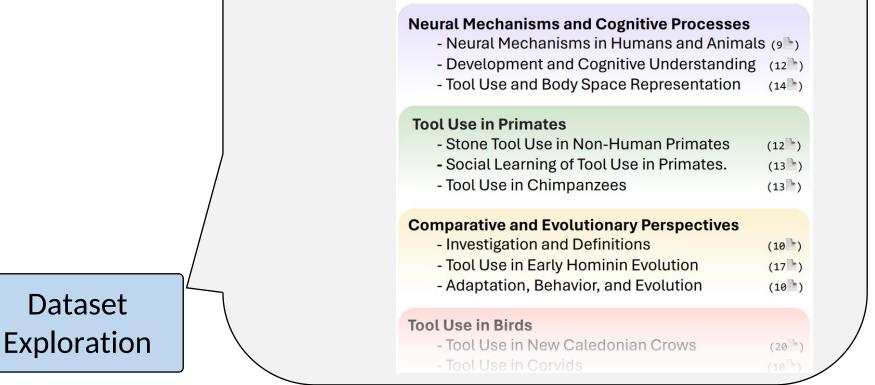




Tool Use In Animals

Q

Processed 1000 Scientific Papers







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

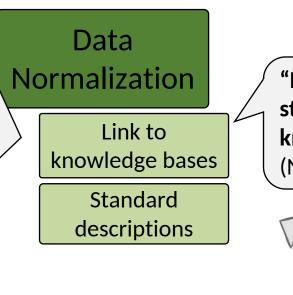
Enhancing Knowledge Graph Construction Using Large Language Models, M Trajanoska et al, 2023

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

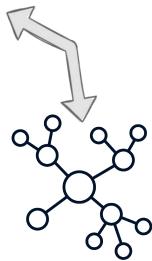
Knowledge Graph Large Language Model (KG-LLM) for Link Prediction, D. Shu, 2024

. . .

SORBONNE UNIVERSITÉ



"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<sup>\*</sup>, King's College London, United Kingdom JOHANNA WALKER<sup>\*</sup>, 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<sup>\*</sup>, 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.

#### How to link (indirectly) formal knowledge and LLMs

Dialogue with LLM with constraints and "multi-hop" reasoning  $\begin{array}{c}
 & & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\
 & & \\$ 





#### 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



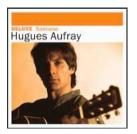


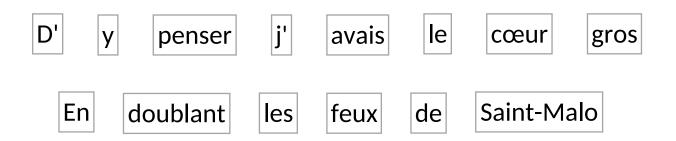
# Thanks!

Questions?

Backup slides...

Words









### "Bag of words"

1 1

а	0	
à	0	coéternels
abaissa	0	coéternité
abaissable	0	coéternités
abaissables	0	cœur
abaissai	0	cœurs
		coexista
avachît	0	coexistai
avachîtes	0	
avaient	0	doublâmes
avais	1	doublant
avait	0	doublard
aval	0	doublards
avala	0	
	0 0	 écardons
	-	
avala avalable  bains	0	écardons
avalable 	0	écardons écarlate

•••	•••
0	fanages
0	fanai
0	fanaient
1	
0	fétuques
0	fétus
0	feu
	feudataire
0	feudataires
1	
0	groomer
0	 grooms
	 gros
0	gros-bec
0	 groschen
0	
0	gouttons
	Orphanet -

			gouvernées
anages	0		gouvernement
anai	0		gouvernementabilité
anaient	0		gouvernemental
	•••		gouvernementale
étuques	0		gouvernementalement
étus	0		gouvernementales
eu	1		gouvernementalisme
eudataire	0		gouvernementalismes
eudataires	0		gouvernementaliste
	•••		gouvernementaux
roomer	0		gouvernementiste
rooms	0		gouvernementistes
ros	1		gouvernementomane
ros-bec	0		gouvernements
roschen	0		gouvernent
			gouverner
outtons	0		
Orphanet - 2024, Sep	tembe	r 17	

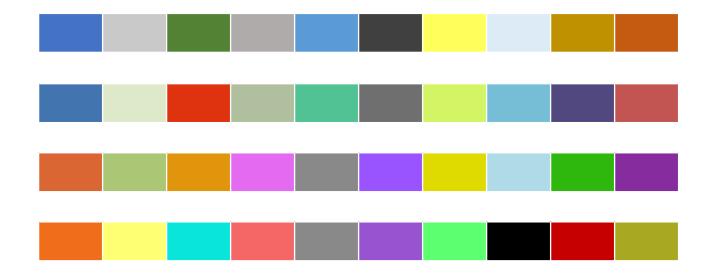
0	pensée
0	pensées
0	pensement
0	pensements
0	pensent
0	penser
0	
0	pertuisaniers
0	pertuiser
0	perturba
0	perturbai
0	perturbaient
0	
0	phonocinèse
0	phonocinétique
0	phonocontrôle
0	phonocapteur
••	

0	
0	
0	
0	
0	
0	
•••	
57	

•••

#### 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) ~ occurrence (tiger)

 $w_{cat}.~w_{context} \sim w_{tiger}$  .  $w_{context}$ 

 $W_{cat} \sim W_{tiopr}$ 

(© Perceval Wajsbürt)

#### Dogs and cats and tigers

The *cat* is sleeping on the couch

The *dog* is sleeping on the couch

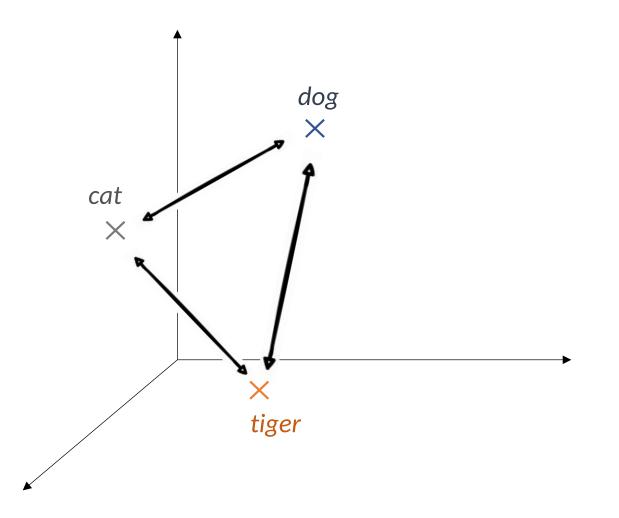
The cat is running in the garden

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

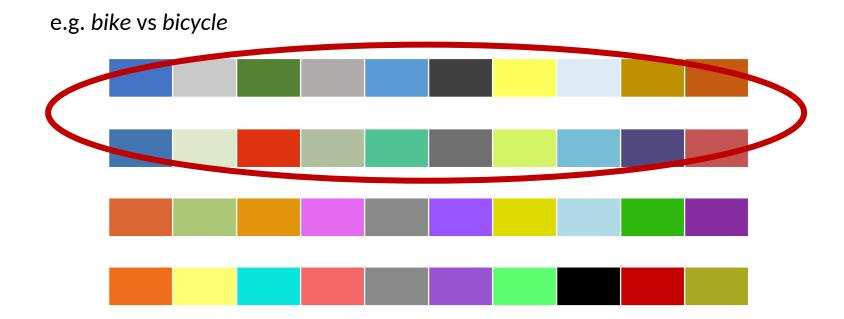






#### Dense representation (embeddings)

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

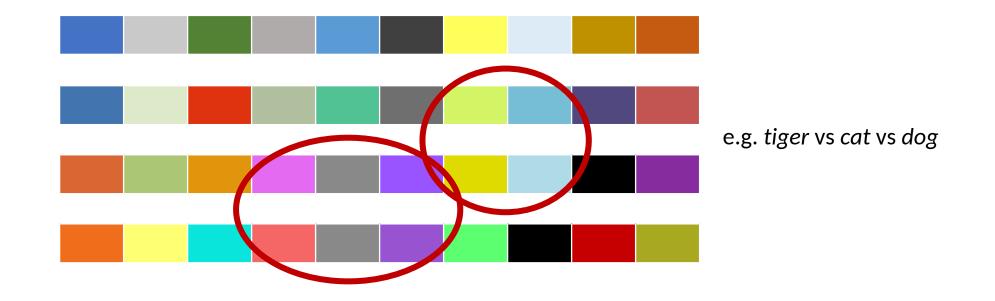






#### Dense representation (embeddings)

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









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

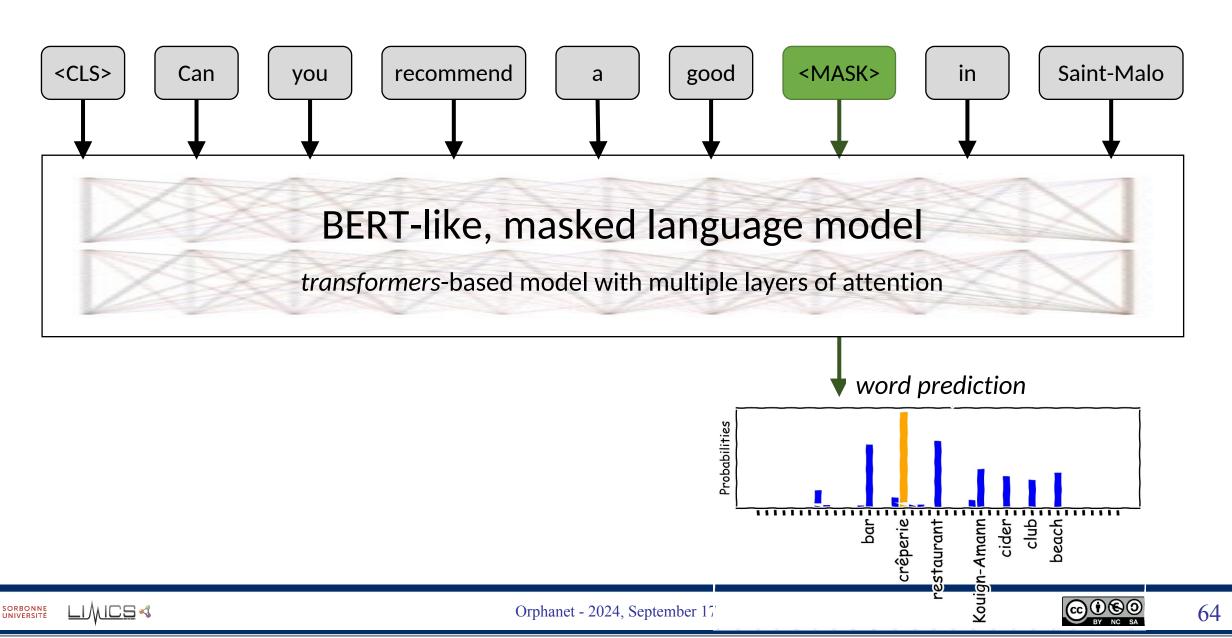
A language model is a model without manual supervision, with two different main training settings:

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

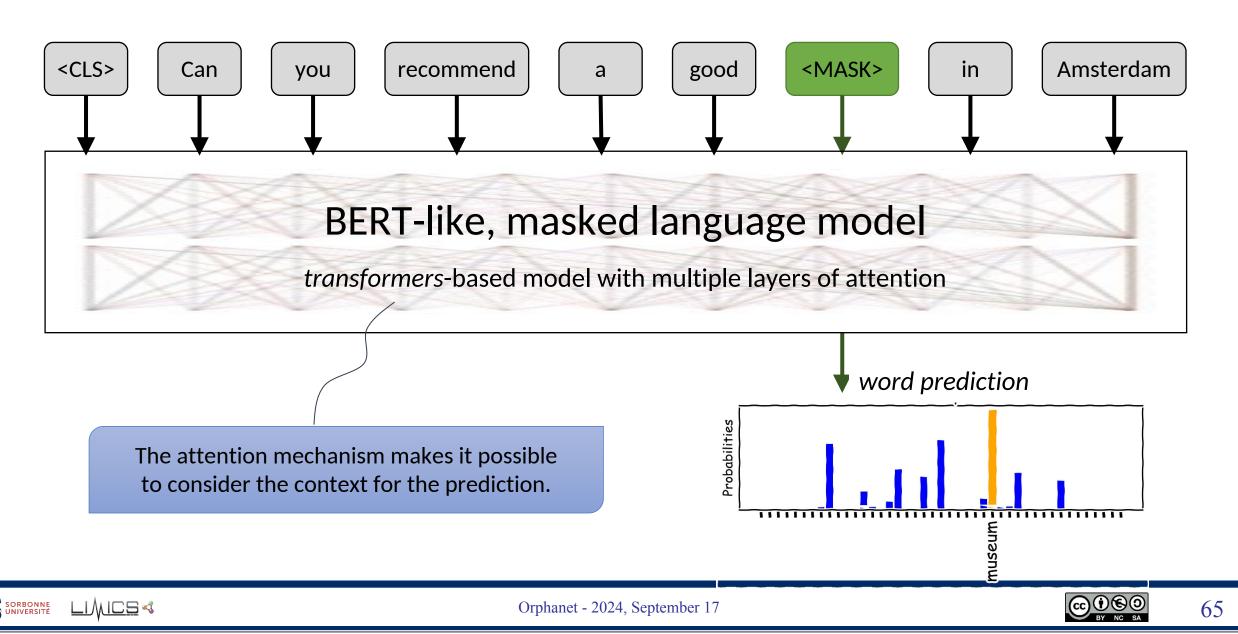
These models are at the heart of the vast majority of current NLP systems. They are known as LLMs (*Large Language Models*).

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

Q



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



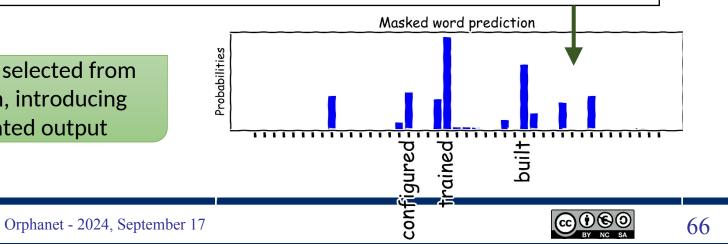
#### Auto-regressive language models (decoder-style, e.g. GPT)

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

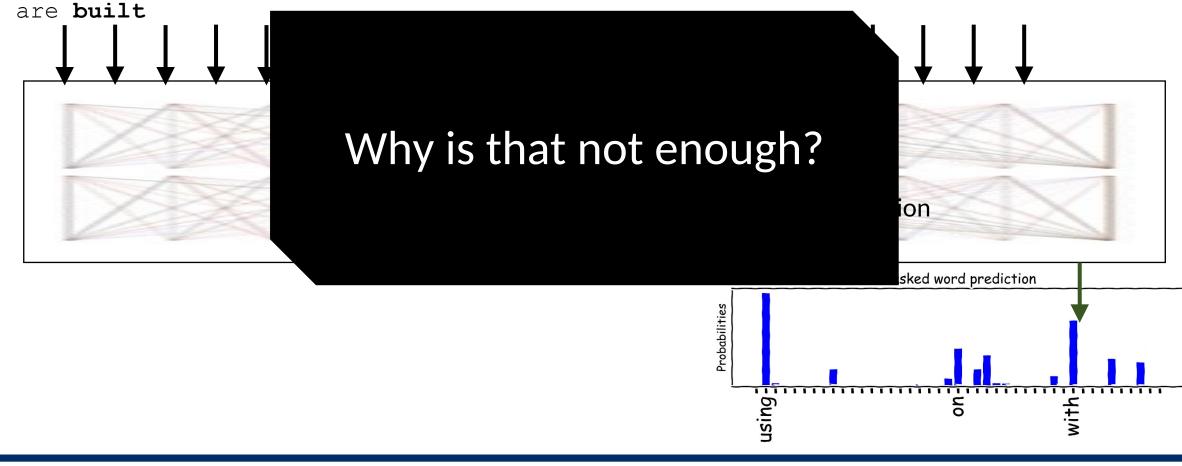
Next tokens are randomly selected from the predicted distribution, introducing variability in the generated output





#### 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





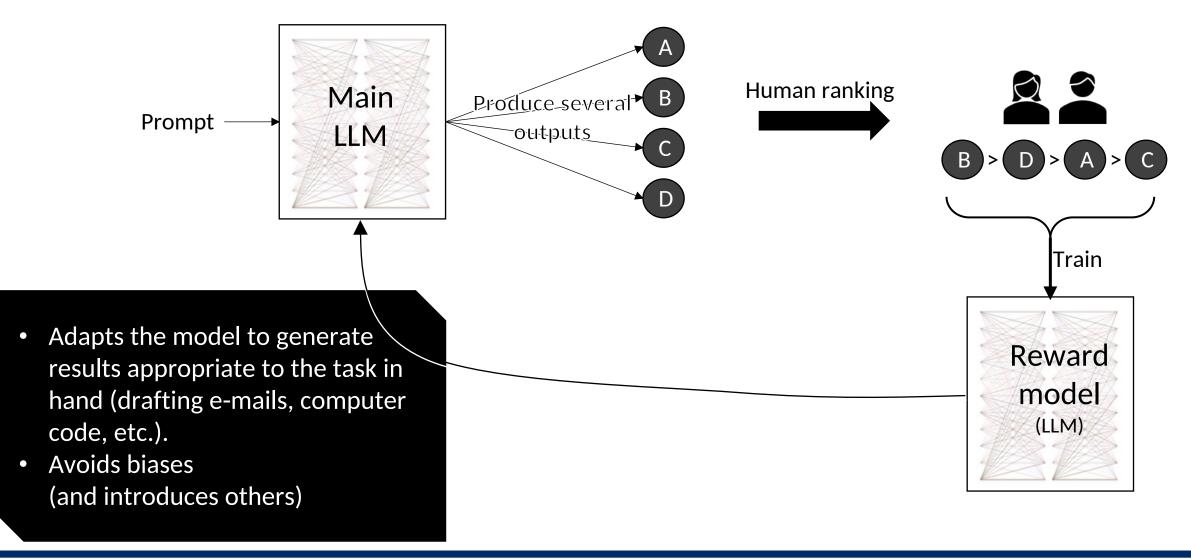
Orphanet - 2024, September 17



#### **Reinforcement Learning with Human Feedback**

Ç

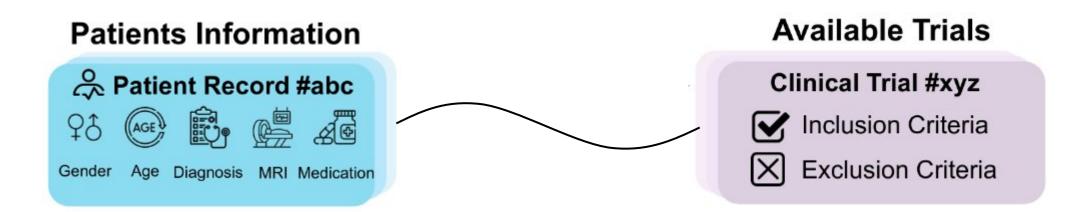
SORBONNE UNIVERSITÉ







#### 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





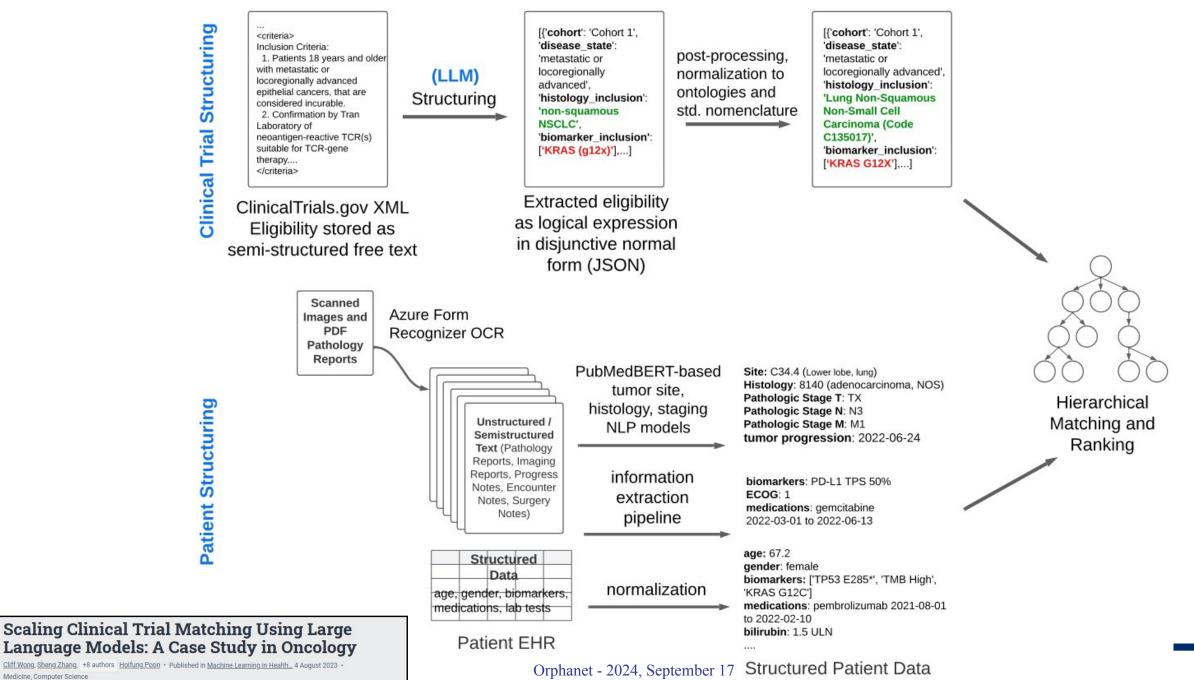
Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Languag	ge
D. Hamer, P. Schoor,       +1 author       Daniel Kapitan       • Published in arXiv.org       14 April 2023       • Computer Science, Medicine	Scaling Clinical Trial Matching Using Large
	Language Models: A Case Study in Oncology <u>Cliff Wong, Sheng Zhang</u> , +8 authors <u>Hoifung Poon</u> · Published in <u>Machine Learning in Health</u> 4 August 2023 ·
Transforming clinical trials: the emerging roles	of ine, computer Science
<b>large language models</b> Jong-Lyul Ghim, Sangzin Ahn • Published in <u>Translational and Clinical</u> 1 September 2023 • Medicine, Computer Science, Lingu	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
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	LLM for Patient-Trial Matching: Privacy-Aware Data Augmentation Towards Better Performance and Generalizability
Matching Patients to Clinical Trials with Large Language Models	Yuan, Ruixiang Tang, Xiaoqian Jiang, Xia Hu less • Published in <u>arXiv.org</u> 2023 • Computer Science, Medicine
Qiao Jin, Zifeng Wang, +2 authors Zhiyong Lu • Published in arXiv.org 27 July 2023 • Computer Science, Medicine	Zero-Shot Clinical Trial Patient Matching with
Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial	LLMS <u>Michael Wornow, A. Lozano</u> , +3 authors <u>Nigam H. Shah</u> • Published in <u>arXiv.org</u> 5 February 2024 • Computer Science, Medicine
Matching. <u>Jiayi Yuan, Ruixiang Tang</u> , +1 author Xia Hu • Published in AMIA Annual Symposium 24 March 2023 • Computer Science, Medicine • AMIA Annual Symposium proceedings. AMIA Symposium Orphane	t - 2024, September 17

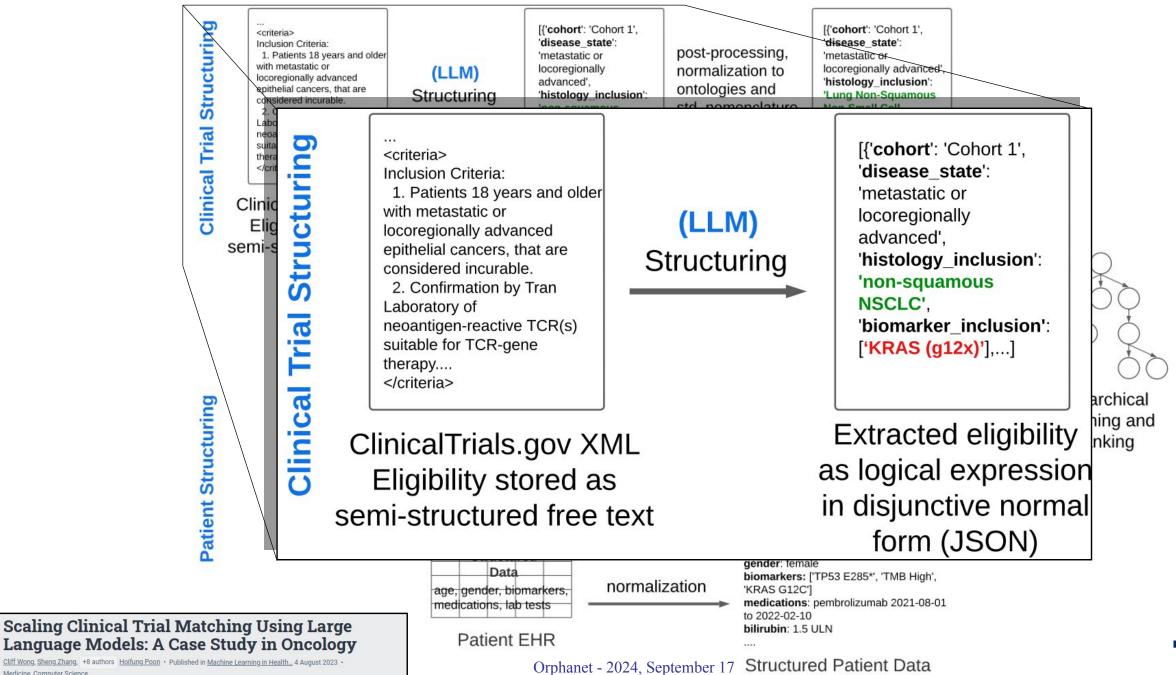
#### LLMs for patient selection : different possible strategies

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

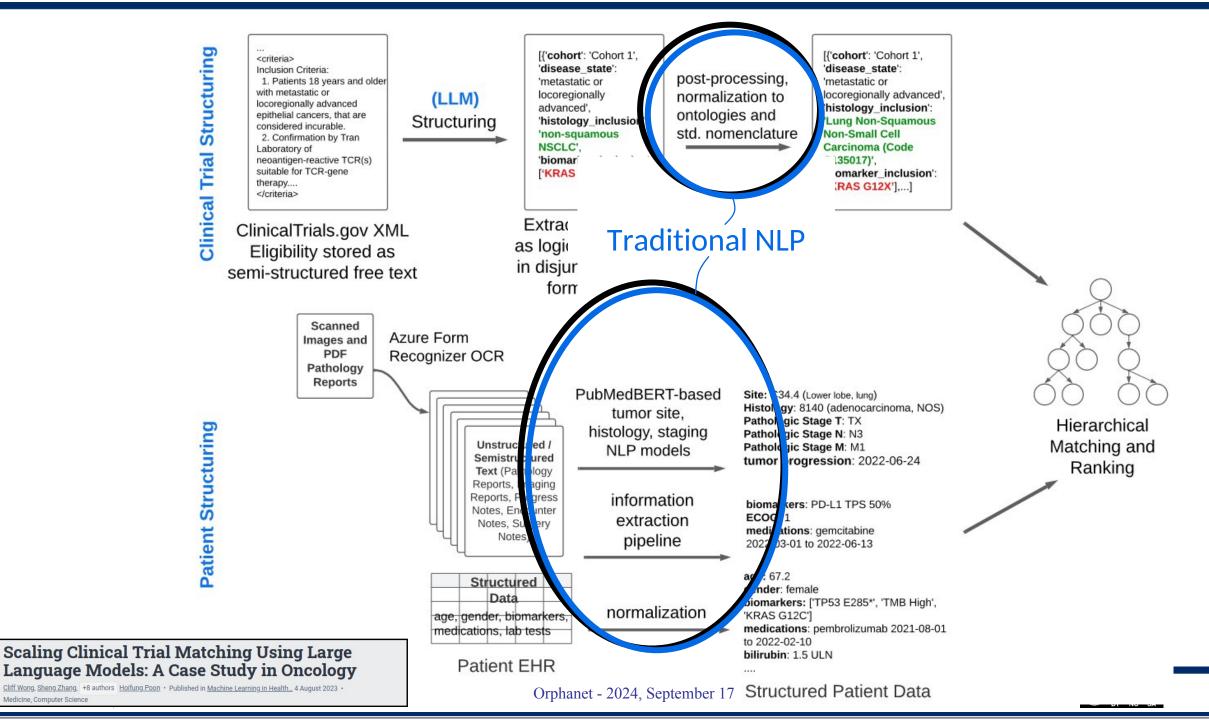








Cliff Wong, Sheng Zhang, +8 authors Holfung Poon • Published in Machine Learning in Health... 4 August 2023 Medicine, Computer Science



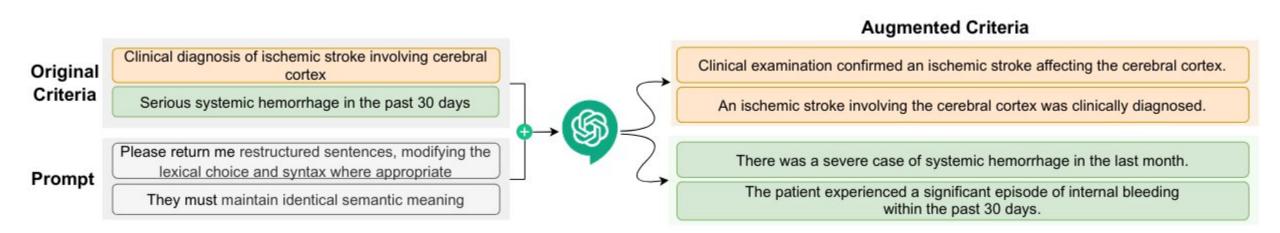
Medicine, Computer Science

# 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.

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

Orphanet - 2024, September 17



# 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

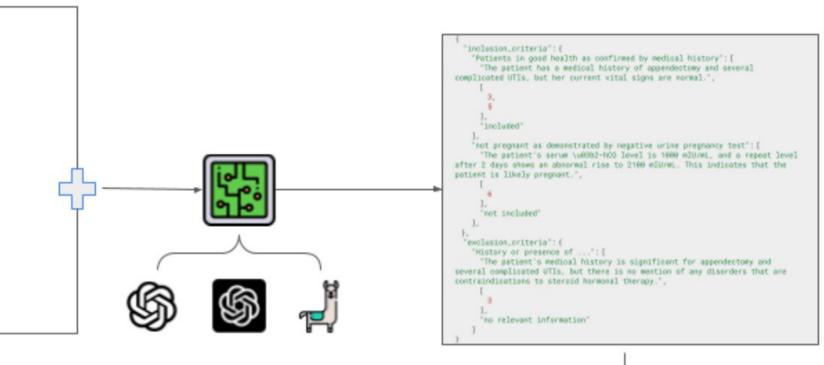




A 32-year-old woman comes to the hospital with vaginal spotting. Her last menstrual period was 10 weeks ago. Medical history is significant for appendectomy and several complicated UTIs. Vital signs are normal. Serum  $\beta$ -hCG level is 1800 mIU/mL, and a repeat level after 2 days shows an abnormal rise to ...

Summary : The purpose of this study is to compare the bleeding profile of ... Intervention : Drug: norelgestromin/ethinyl... Inclusion Criteria : Patients in good health as confirmed by medical history not pregnant as demonstrated by negative urine... Exclusion Criteria : History or presence of

Exclusion Criteria : History or presence of disorders commonly accepted as contraindications to steroid hormonal therapy including...



Distilling Large Language Models for Matching Patients to Clinical Trials

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

Orphanet - 2024, September 17



Ranking

# 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.

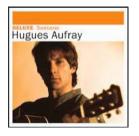
LLMs are **susceptible to generating misleading or incorrect information**, a phenomenon known as "hallucination", and **ensuring quality control may prove to be demanding**. The complexity and flexibility of LLMs correspondingly make validation regarding accuracy, safety, and clinical efficacy particularly 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.

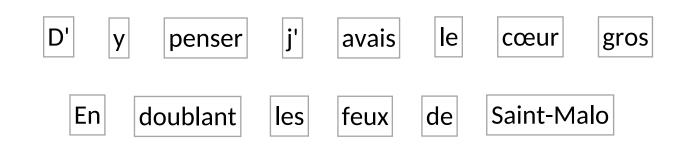
Transforming clinical trials: the emerging roles of large language models

Jong-Lyul Ghim, Sangzin Ahn • Published in Translational and Clinical... 1 September 2023 • Medicine, Computer Science Linguistics - 2024, September 17



## Words





Lebensversicherungsgesellschaftsangestellter

(Agglutinative language)

(employee of a life insurance company)

學而不思則罔,思而不學則殆

Orphanet - 2024, September 17



# "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



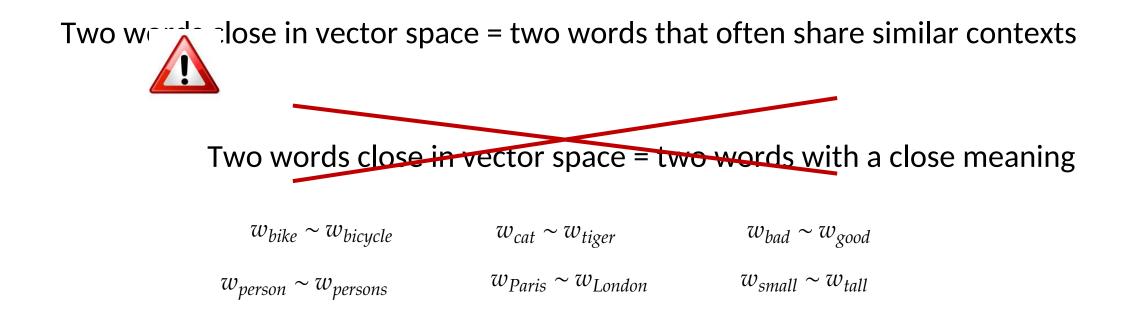
Tuebingen ≠ Tübingen, tomorrow ≠ tommorrow





## Dense representation (embeddings)











## **Dense 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

Arithmetic operations on word vectors

Depending on the method, see later



Orphanet - 2024, September 17



## **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

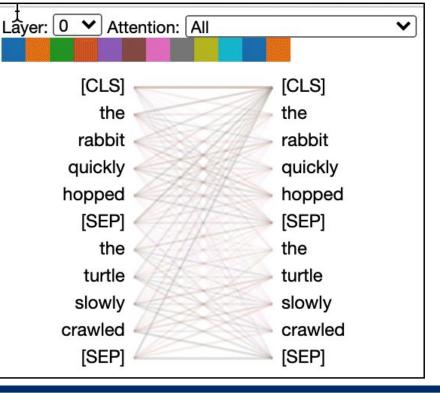






## **Contextual dense representation**

- 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 con
  - 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 represer (usually through an attention mechanism...)







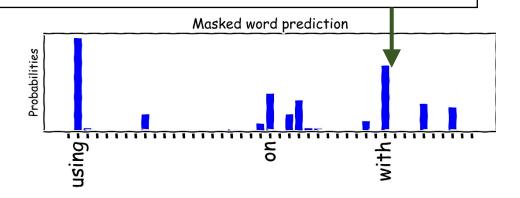
https://github.com/jessevig/bertviz/

# 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** 

## **GPT-like language model**

transformers-based model with multiple layers of attention







Orphanet - 2024, September 17

## 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

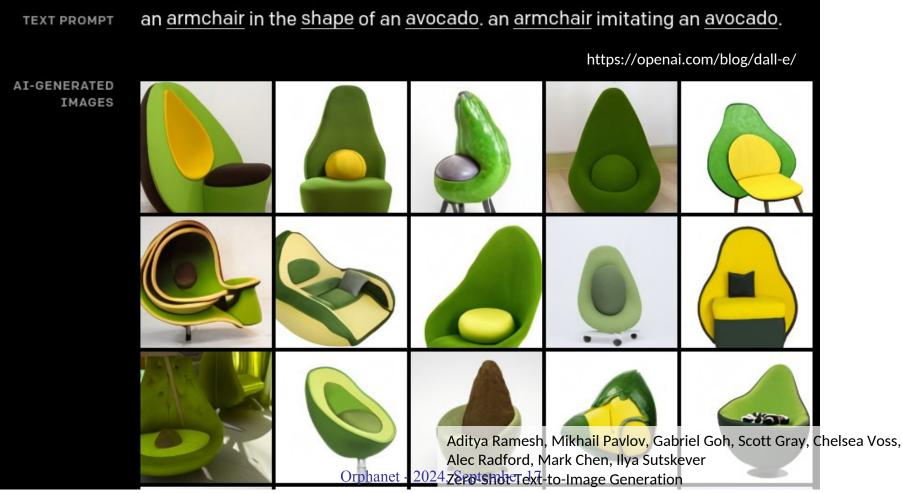
- A vocabulary of predefined size, composed of ngrams of characters
- Vocabulary chosen to maximize the frequency of ngrams
- Possibility of a multilingual tokenizer





Another advantage of dense representations:

they can be mixed with representations of something else



Another advantage of dense representations:

they can be mixed with representations of something else





Orphanet - 2024, September 17



### Another advantage of dense representations:

## they can be mixed with representations of something else



guacamole (90.1%) Ranked 1 out of 101 labels

### ✓ a photo of guacamole, a type of food

x a photo of ceviche, a type of food.

× a photo of edamame, a type of food.

× a photo of tuna tartare, a type of food.

× a photo of hummus, a type of food.



#### television studio (90.2%) Ranked 1 out of 397



> a photo of a podium indoor.
 × a photo of a conference room.
 × a photo of a lecture room.
 × a photo of a lecture room.

#### YOUTUBE-BB

airplane, person (89.0%) Ranked 1 out of 23



✓ a photo of a airplane.	2
× a photo of a <b>bird</b> .	
× a photo of a <b>bear</b> .	
× a photo of a <b>giraffe</b> .	100
∎ × a photo of a <b>car</b> .	

#### EUROSAT

annual crop land (12.9%) Ranked 4 out of 10



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

#### Alec Radford et al.

Learning Transferable Visual Models From Natural Language Supervision



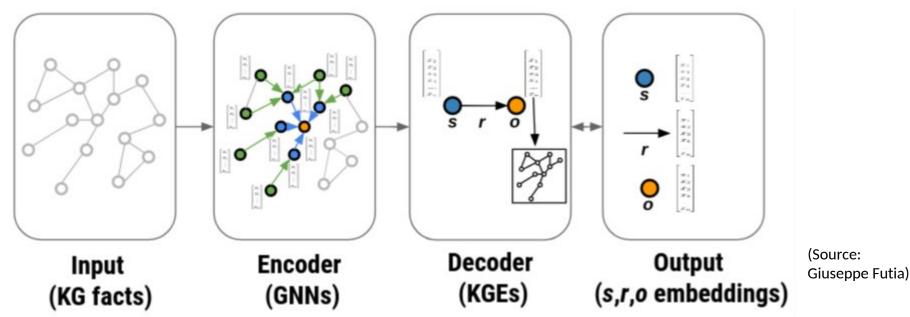
Orphanet - 2024, September 17



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





Q

SORBONNE UNIVERSITÉ



