

# Récentes avancées de l'inférence en langue naturelle pour les essais cliniques

Journée Santé et IA 2024

---

Mathilde AGUIAR, Pierre Zweigenbaum et Nona Naderi

Lundi 1<sup>er</sup> Juillet 2024

Laboratoire Interdisciplinaire des Sciences du Numérique – LISN, CNRS  
[mathilde.aguiar@lisn.upsaclay.fr](mailto:mathilde.aguiar@lisn.upsaclay.fr)

# Introduction

---

## CTR

### ELIGIBILITY - INCLUSION CRITERIA:

- Must be female with histologically confirmed breast cancer
- Stage II-IV disease
- ER and/or PR positive
- ECOG Performance Status 0-1
- Tumor must be present following core needle biopsy as determined by physical exam or radiographic evaluation.
- No prior treatment for current breast cancer. No other active malignancy is allowed. Adequately treated basal cell, squamous cell skin cancer, in situ cervical cancer, or any other cancer from which the patient has been disease-free for 5 years is permitted. Biphosphonates and palliative radiation for bone metastasis is permitted while on study

## STATEMENT

Adele is an 85 year old woman with Stage II histologically confirmed ER+ breast cancer with an ECOG of 0, she is eligible for the primary trial

## LABEL:

**ENTAILMENT**  
OR  
CONTRADICTION OR  
NEUTRAL

## COMMON-SENSE REASONING

H: 85 year old woman

P: Must be female

R: *Synonyms*

## NUMERICAL INFERENCE

H: Stage II

P: Stage II-IV disease

--

H: ECOG of 0

P: ECOG Performance Status  
0-1

R: *Within the intervals*

## CLINICAL INFERENCE

H: ER+

P: ER and/or PR positive

R: *ER+ and ER both relates  
to oestrogen*

--

H/P: histologically  
confirmed breast cancer

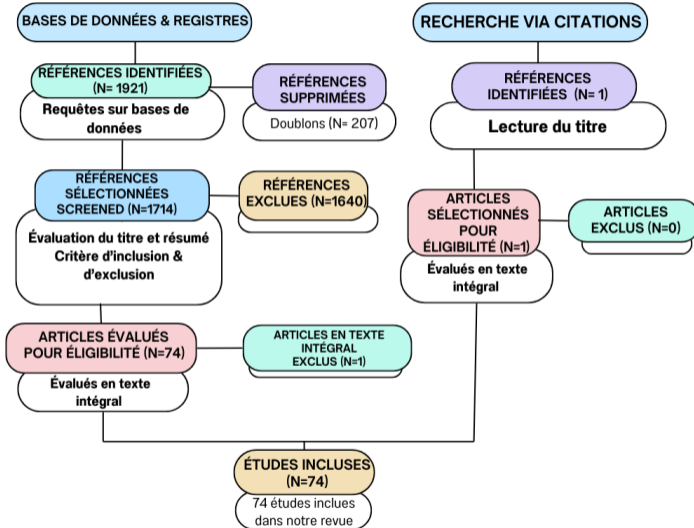
- Quelles sont les méthodes, jeux de données et approches disponibles pour la tâche d'ILN dans le domaine des essais cliniques ?
- Comment l'ILN pourrait-elle être bénéfique aux essais cliniques ?
- Quel sont les défis et travaux futurs potentiels pour l'ILN dans le domaine des essais cliniques ?

\*H = hypothesis, P = premise, R = reason

# Méthodes

---

# Méthode PRISMA



## Requêtes :

1. clinical AND "Natural Language Processing" AND "Natural Language Inference" OR NLI
2. clinical AND "Textual Entailment" OR TE
3. "Natural Language Inference"

## Critères :

- Inclusion :
  - Apprentissage profond (*Deep Learning*) OU Apprentissage automatique (*Machine Learning*)
  - Évaluation par des pairs
  - Publié en anglais
  - Publié en 2022 et 2023
- Exclusion :
  - Multi-modalité, traitement de l'image
  - Étude multi-tâches/non focalisée sur l'ILN

# Résultats

---



## L'ILN dans le domaine général

---

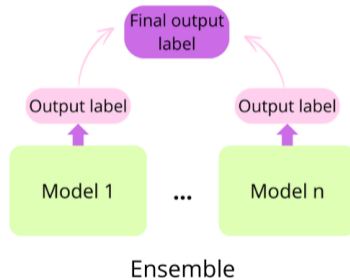
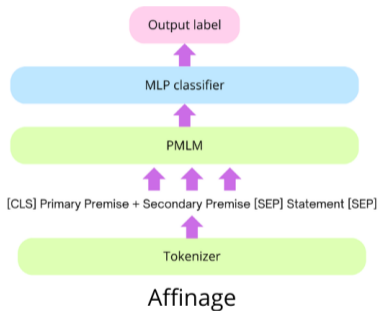
But : évaluation des systèmes développés

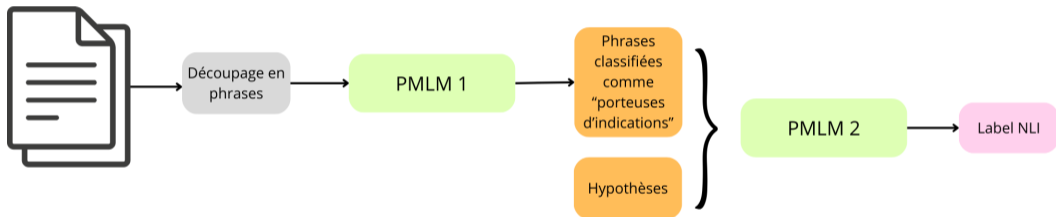
- Aspect linguistique particulier, par ex. : négation, implicature, etc.
- Capacité du modèle à raisonner sur de longs documents
- Capacité du modèle à raisonner sur un type de document (contrats, essais cliniques, etc.)

Composition des instances :

- **Prémisse** : de la phrase au document complet. Extraite de sources originales (site Web, articles de journaux, etc.) OU traduite automatiquement dans une langue cible à partir d'une tâche pré-existante dans une langue OU générée automatiquement (par GPT-3 par exemple)
- **Hypothèse** : généralement une phrase. Générée automatiquement OU manuellement par des annotateurs OU à l'aide de règles
- **Label** : *entailment, contradiction, neutral*

# Approches





Métriques d'évaluation : F1, rappel, précision et exactitude (*accuracy*)

- Base pour tâches de plus « haut niveau »: résumé automatique, systèmes de questions-réponses, etc.
- Via apprentissage par transfert
- Via apprentissage multi-tâches

# L'ILN dans le domaine des essais cliniques

---

BioNLI [Bastan et al., 2022] :

**Premise:**The outflow of **uracil** from the yeast *Saccharomyces cerevisiae* is known to be relatively fast in certain circumstances, to be retarded by **proton** conductors and to occur in strains lacking a **uracil proton** symport. In the present work, it was shown that **uracil** exit from washed yeast cells is an active process, creating a **uracil** gradient of the order of -80 mV relative to the surrounding medium. Glucose accelerated **uracil** exit, while retarding its entry. DNP or sodium azide each lowered the gradient to about -30 mV, simultaneously increasing the rate of **uracil** entry. They also lowered cellular ATP content. Manipulation of the external ionic conditions governing  $\Delta\mu_{H^+}$  at the plasma membrane had no detectable effect on **uracil** transport in yeast preparations thoroughly depleted of ATP.

**Consistent Hypothesis:**It was concluded that **uracil** exit is probably not driven by the **proton** gradient but may utilize ATP directly.

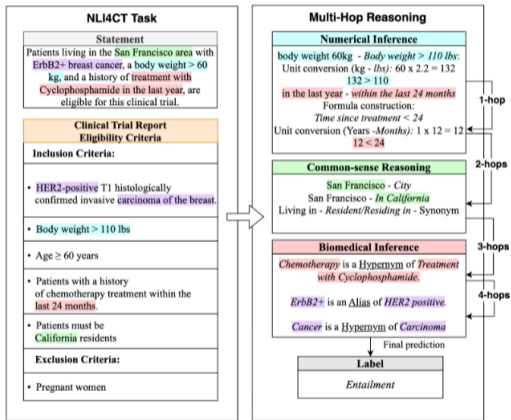
**Adversarial Hypothesis:**It is concluded that **uracil** exit from *S. cerevisiae* is an active process facilitated by a **proton** gradient and ATP.

- Construit automatiquement à partir de résumés de publications tirées depuis PubMed
- Prémisse : description détaillé d'une expérience scientifique
- Hypothèse : Phrase résumant l'expérience de la prémisse
- *entailment* et *contradiction*
- 8000+ instances



# Tâches et jeux de données - essais cliniques

NLI4CT [Jullien et al., 2023] :



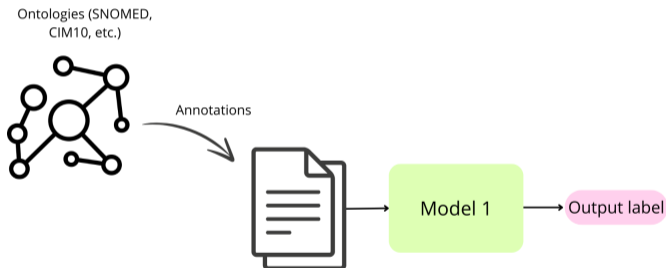
- Essais cliniques extraits de [clinicaltrials.gov](https://clinicaltrials.gov)
- Prémisse : Essai clinique avec les sections suivantes : intervention, résultats, effets secondaires et critères d'éligibilité
- Hypothèse : Créées par des experts. Balayant différents types d'inférences
- *entailment* et *contradiction*
- 2400 instances

## Défis :

- Vocabulaire spécifique au domaine clinique
- Jeux de données de taille restreinte

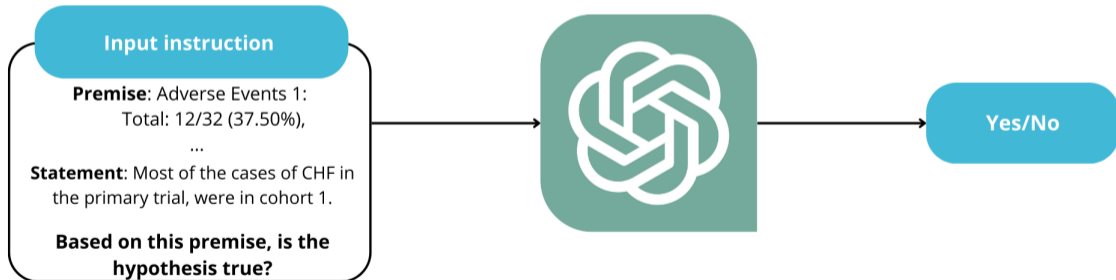
## Solutions possibles : Augmentation de données

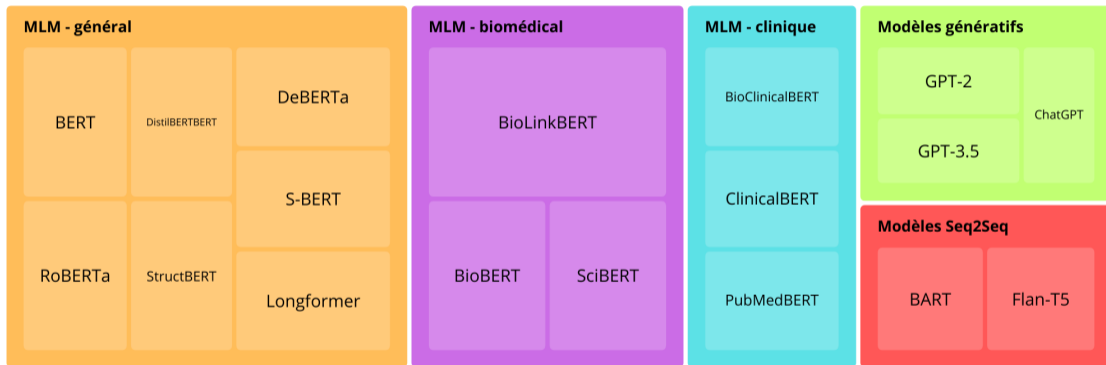
- Introduction de perturbations aléatoires
- Ajout d'annotations en utilisant des ontologies
- *Back translation*



# Approches

- Apprentissage par transfert
- Continuer le pré-apprentissage sur le type de document cible
- Instructions pour les LLM :





Comment l'ILN peut-elle être  
bénéfique pour les essais  
cliniques

---

## Modélisation des critères d'éligibilité

Initiation d'un essai clinique  
avec critères d'éligibilité



Recrutement selon  
critères d'éligibilité

**Aide au  
recrutement  
patient**



Base de patients avec dossier  
médicaux associés

**Modélisation des dossiers patients**



Déroulement des  
différentes phases de l'essai  
clinique sur la population  
cible



Publication des résultats

**Utilisation des  
publications pour  
entraîner des  
modèles**



Confronter la  
littérature

**Modéliser cette  
confrontation en  
tâche de NLI**



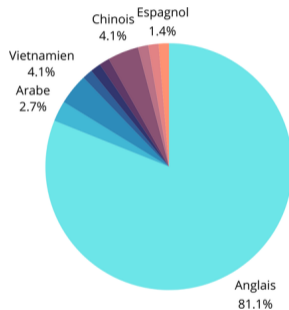
## Défis et perspectives

---

- Présence d'indices dans les hypothèses qui permettent au modèle de prédire le bon label en ne considérant que l'hypothèse et en ignorant la prémisse
- Phénomène résultant de la création des hypothèses via crowd-sourcing
- Par exemple l'emploi de la négation est souvent un indice d'une instance labéllisée « Contradiction »



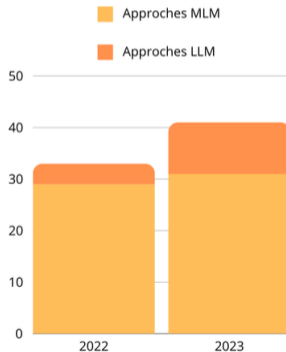
Répartition des langues traitées dans les études collectées :



Uniquement des jeux de données en anglais dans le domaine clinique 🤖

# Grand modèles de langue




- Tendence nouvelle, en particulier lors du défi partagé SemEval 2023 (et 2024)
- Atteignent des performances similaires aux MLM, voire les surpassent
- 🤔 quand est-il de leur explicabilité ?



## Conclusion

---

- Revue de 74 articles publiés en 2022 et 2023, en utilisant la méthode PRISMA.
- L'ILN pour les essais clinique est un domaine en plein essor, en particulier suite à l'organisation de la tâche partagée NLI4CT pour 2 années consécutives (2023 et 2024)
- Besoin de jeux de données dans des langues autres que l'anglais pour l'ILN dans le domaine médical et des essais cliniques
- Présence de biais (artefacts) dans les jeux de données

-  Bastan, M., Surdeanu, M., and Balasubramanian, N. (2022).  
**BioNLI: Generating a biomedical NLI dataset using lexico-semantic constraints for adversarial examples.**  
In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5093–5104, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
-  Jullien, M., Valentino, M., Frost, H., O'Regan, P., Landers, D., and Freitas, A. (2023).  
**NLI4CT: Multi-evidence natural language inference for clinical trial reports.**  
In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.
-  Liu, A., Swayamdipta, S., Smith, N. A., and Choi, Y. (2022).  
**Wanli: Worker and ai collaboration for natural language inference dataset creation.**

Merci !

# Exemple de jeu de données

WANLI [Liu et al., 2022], avec des instances générées automatiquement en utilisant GPT-3.5 :

<b>premise</b> string · lengths	<b>hypothesis</b> string · lengths	<b>gold</b> string · classes	<b>genre</b> string · classes
 5 590	 6 433	 3 values	 generated_... 3.6%
For the same reason, it is often said that, in the end, the only thing the economy's leading figures really care about is money.	The economy's leading figures are usually people like Boone Pickens who like to stay hidden.	neutral	generated_revised
But I was also angry that he had not told me that he was going to marry her.	I was angry that he had not told me that he was going to marry her.	neutral	generated_revised
As the sun set, the breeze from the Gulf of Mexico began to cool, and the water turned from a shimmering blue to a deep indigo.	The Gulf of Mexico was a shimmering blue.	entailment	generated_revised
But the worst of it was that I had no idea what was going on.	I did not know what was going on.	entailment	generated_revised
In the example of Figure 4-4, the ACI System consists of the following components, as shown in the simplified schematic of Figure 4-5.	The simplified schematic of Figure 4.5 shows the components that make up the ACI System.	entailment	generated_revised
It is not the case that I have no interest in politics.	I have no interest in politics.	contradiction	generated_revised
The series is a highly-successful series of popular games.	The series is highly-successful, but it's not a series of popular games.	contradiction	generated_revised
She says that her organization is in favor of full employment, but that is not all she has been saying.	She has been saying that her organization is in favor of full employment.	entailment	generated_revised
Under the proposed rule, the Department would be required to make a final determination on the application within 60 days of the date of...	The determination on the application is currently required to be made within 60 days of the date of submission.	contradiction	generated_revised
I have a mind to have a little fun with him, said the boy, who was a willing participant.	The boy was a willing participant.	entailment	generated_revised

# Résultats SemEval 2023 Tâche n°7

Work @ Team name	Approach	Generative/ Discriminative	Retrieval type	Pre-training Datasets	Task 1			Task 2		
					F1	Precision	Recall	F1	Precision	Recall
(Zhou et al., 2023) @THIFLY	MGNet, BiLSTM and SciFive model ensemble	G + D	Post	PubMed Abstract, PMC	0.856	0.856	0.856	0.853	0.811	0.898
(Kanakarajan and Sankarasubba, 2023) @Saama AI Research	Instruction-finetuned LLMs, Flan-T5	G + D	-	-	0.834	0.768	0.912	-	-	-
(Vladika and Matthes, 2023) @Sebis	Ensemble of a pipeline and joint system based on DeBERTa-v3	D	Pre	-	0.798	0.777	0.820	0.818	0.772	0.868
(Wang et al., 2023) @KnowComp	DeBERTa-v3-large.	D	-	-	0.764	0.757	0.772	-	-	-
(Chen et al., 2023) @NCUEE-NLP	Soft voting ensemble mechanism based on BioLink/BioBERT	D	Pre	MultiNLI, MedNLI, and SNLI	0.709	0.668	0.756	0.794	0.803	0.786
(Alameldin and Williamson, 2023) @Clemson NLP	GatorTron-BERT	D	Pre	UFHS notes, MIMIC-III and WikiText	0.705	0.654	0.764	0.806	0.802	0.811
(Rajamanickam and Rajaraman, 2023) @I <sup>2</sup> R	Evidence level inferences with T5	G + D	Pre	-	0.701	0.550	0.968	0.802	0.797	0.807
(Bevan et al., 2023) @MDC	PubMedBERT for evidence retrieval, and BioLinkBERT classifies entailment.	D	Pre	PubMed abstracts, PMC	0.695	0.668	0.724	0.804	0.814	0.795
(Zhao et al., 2023) @HW-TSC	Zero-shot ChatGPT for entailment and DeBERTaV3 for retrieval.	G + D	Post	-	0.679	0.592	0.796	0.842	0.816	0.871
(Palwa and Palwa, 2023) @BpHigh	Few-shot GPT-3.5 Davinci	G	-	-	0.679	0.523	0.968	-	-	-
(Feng et al., 2023) @YNU-HPCC	BioBERT, supervised contrastive learning, and back translation.	D	-	PubMed, PMC	0.679	0.621	0.748	-	-	-
(Alissa and Abdullah, 2023) @JUST-KM	Role-based Double Roberta-Large	D	-	-	0.670	0.529	0.912	-	-	-
(Noor Mohamed and Srinivasan, 2023) @SSNSheerinKavitha	Semantic Rule based Clinical Data Analysis, TF-IDF, and BM25	-	Post	-	0.667	0.500	1.00	0.572	0.542	0.606
(Correia Dias et al., 2023) @INF-UFRGS	EvidenceSCL using a modified PairSCL model and pre-trained Biomed RoBERTa checkpoints.	D	Pre	Semantic Scholar corpus	0.666	0.500	0.996	0.681	0.615	0.764
(Takehana et al., 2023) @Stanford MLab	Bio+Clinical/Distil/Bio Discharge Summary BERT, and ELECTRA Small ensemble	D	-	MIMIC-III, PubMed, PMC	0.662	0.575	0.780	-	-	-



# Résultats SemEval 2024 Tâche n°2

Work	F1	F	C	Average Score	Architecture	Inference Strategies	Fine-Tuning	Dataset Augmentation
FZI-WIM (Liu and Thoma, 2024)	<b>0.8</b>	0.9	0.73	0.81	Mixtral-8x7B-Instruct	CoT	Yes	GPT-4, bart-large-mnli Instruction Dataset
Lisbon Computational Linguists (Guimarães et al., 2024)	<b>0.8</b>	0.83	0.72	0.78	Mistral-7B-Instruct-v0.2	Zero-shot	Yes	Mistral-7B-Instruct-v0.2 dataset expansion
NYCU-NLP (Lee et al., 2024)	0.78	0.92	<b>0.81</b>	<b>0.84</b>	SOLAR (10.7B)	Zero-shot	Yes	OpenChat v3.5, Intervention Reduction
Edinburgh Clinical NLP (Gema et al., 2024)	0.78	<b>0.95</b>	0.78	<b>0.84</b>	GPT-4	Zero-shot	No	-
YNU-HPCC (Zhang et al., 2024)	0.77	0.67	0.73	0.72	DeBERTa-v3-large	Discriminative	Yes	MultiNLI, FeverNLI, ANLI, LingNLI, WANLI, Back Translation
BD-NLP (Nath and Samin, 2024)	0.77	0.79	0.76	0.77	DeBERTa-lg	Discriminative	Yes	-
CaresAI (Abdel-Salam et al., 2024)	0.77	0.76	0.75	0.76	Ensemble of DeBERTas	Discriminative	Yes	-
TüDuo (Smilga and Alabiad, 2024)	0.76	0.84	0.75	0.78	Flan-T5 XL	Few-shot	Yes	GPT-3.5-Turbo Instruction Dataset
RGAT (Chakraborty, 2024)	0.76	0.86	0.74	0.79	GPT-4	Zero-shot	No	-
DFKI-NLP (Verma and Raithele, 2024)	0.75	0.81	0.68	0.75	Mistral 7B	Zero-shot	Yes	Meta-Inventory dataset expansion, MedNLI
D-NLP (ALTINOK, 2024)	0.75	0.83	0.74	0.77	Gemini Pro	Zero-shot	No	-
LMU-BioNLP (Sun et al., 2024)	0.75	0.86	0.69	0.77	Mistral-7b	Zero-shot	Yes	GPT-3.5, GPT4 dataset expansion, and instruction tuning dataset
DKE-Research (Wang et al., 2024)	0.74	0.8	0.75	0.76	DeBERTa-l	Discriminative	Yes	GPT-3.5, TF-IDF dataset expansion
Puer (Dao et al., 2024)	0.72	0.59	0.64	0.65	BiLinkBERT-large	Discriminative	Yes	-
UniBuc (Micluța-Câmpeanu et al., 2024)	0.71	0.83	0.72	0.75	SOLAR 10B	few-shot	No	-
iML (Akkasi et al., 2024)	0.7	0.28	0.52	0.50	SciFive	Zero-shot	Yes	-

