TAL, IA et corpus médicaux

Ioana BUHNILA

Post-doctorante à ATILF, CNRS - Université de Lorraine

ioana.buhnila@univ-lorraine.fr

1. Présentation du projet de postdoc

1. Explorer les comptes-rendus des patients qui ont des tumeurs cérébrales de type gliome
2. Trouver des réponses à des questions de recherche médicales dans des articles scientifiques

- ATILF (Mathieu Constant)
- IECL (Marianne Clausel)
- CRAN (Hélène Dumond)
- CHRU Nancy (Luc Taillandier)

2. Motivation

1. Explorer les comptes-rendus des patients qui ont des tumeurs cérébrales de type gliome
2. Trouver des réponses à des questions de recherche médicales dans des articles scientifiques

- Aider à améliorer les traitements des patients
- Adapter les traitements des patients par rapport à leurs profils

médecine personnalisée

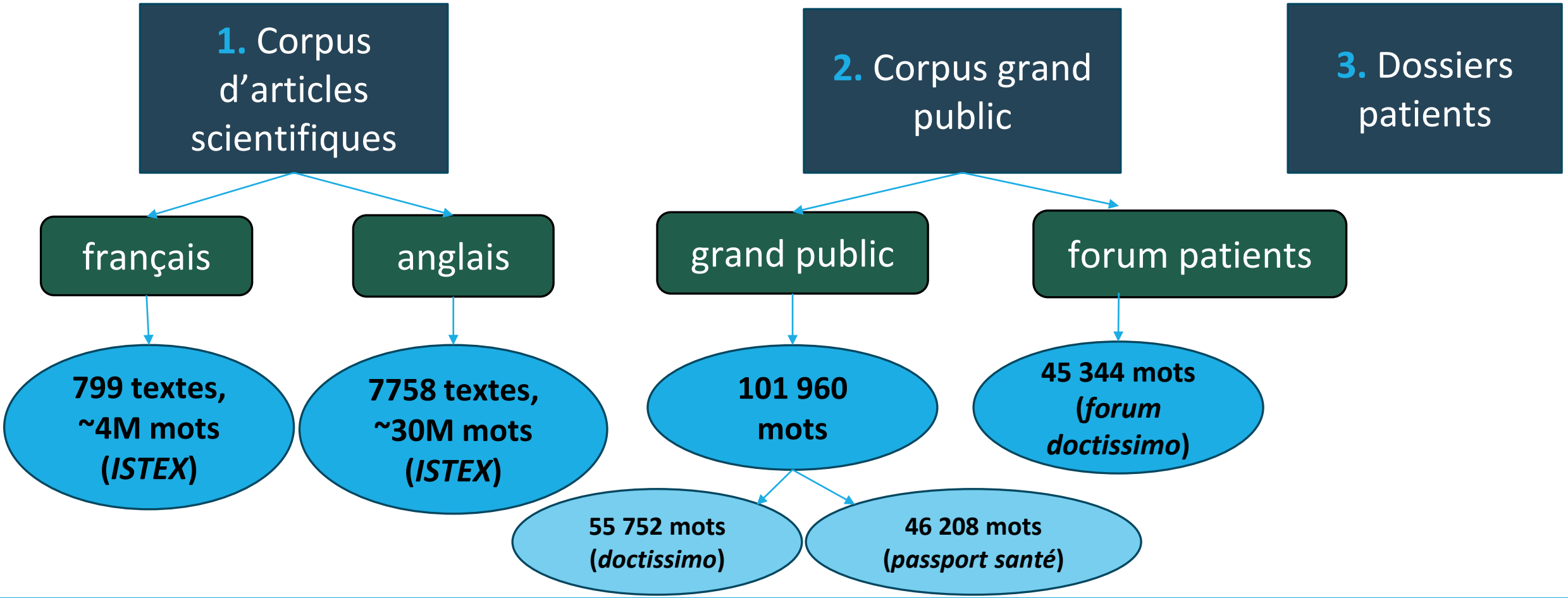
3. Contexte linguistique

1. **Le terme** - une unité lexicale de spécialité qui représente des connaissances spécifiques à un domaine du savoir (Costa, 2006)
2. **Entités nommées (NER)**
3. **Relations sémantiques (RE)**

4. Données

1. Articles scientifiques
2. Textes grand public
3. Dossiers des patients (CHRU Nancy)

4. Données


1. Corpus d'articles scientifiques
2. Corpus grand public
3. Dossiers patients
français
anglais
grand public
forum patients
799 textes, ~4M mots (ISTEX)
7758 textes, ~30M mots (ISTEX)
101 960 mots
45 344 mots (forum doctissimo)
55 752 mots (doctissimo)
46 208 mots (passport santé)

4. Données

3. Dossiers patients

Ressources textuelles

- Compte-rendu visite patient
- Compte-rendu IRM
- Bilan anatomopathologique
- Bilan orthophonique

5. Méthode(s) TAL

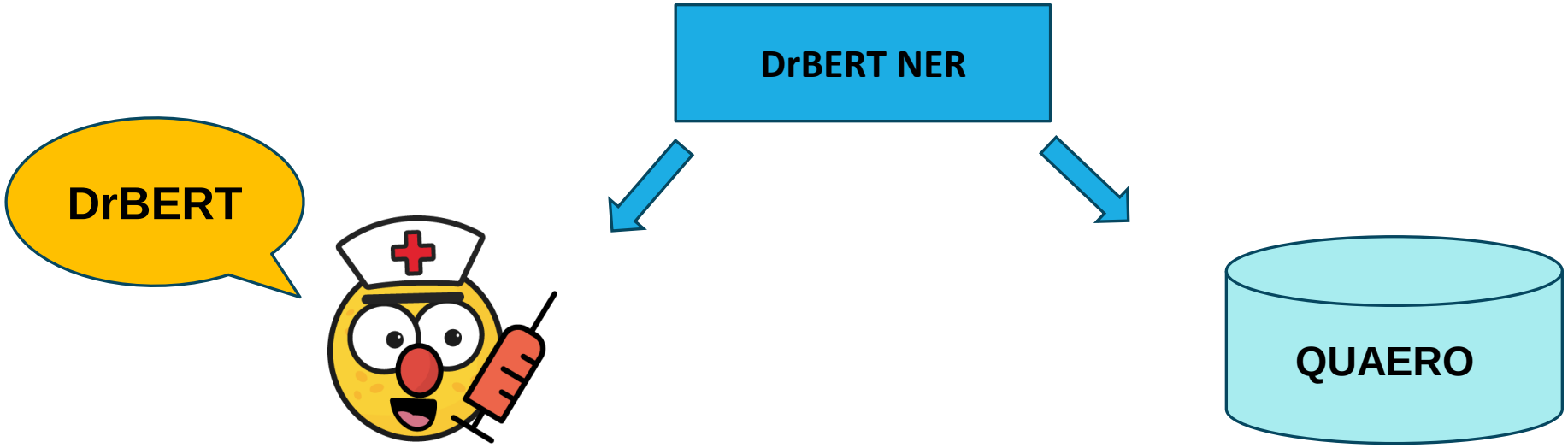
1. NER
2. RE
3. RAG
4. IA générative

5. Méthode(s) TAL

1. **NER**
2. RE
3. RAG
4. IA générative

5. Méthode(s) TAL

NER *(Named Entity Recognition)* = reconnaissance d'entités nommées d'un texte (PERS, ORG, LIEU, etc.)



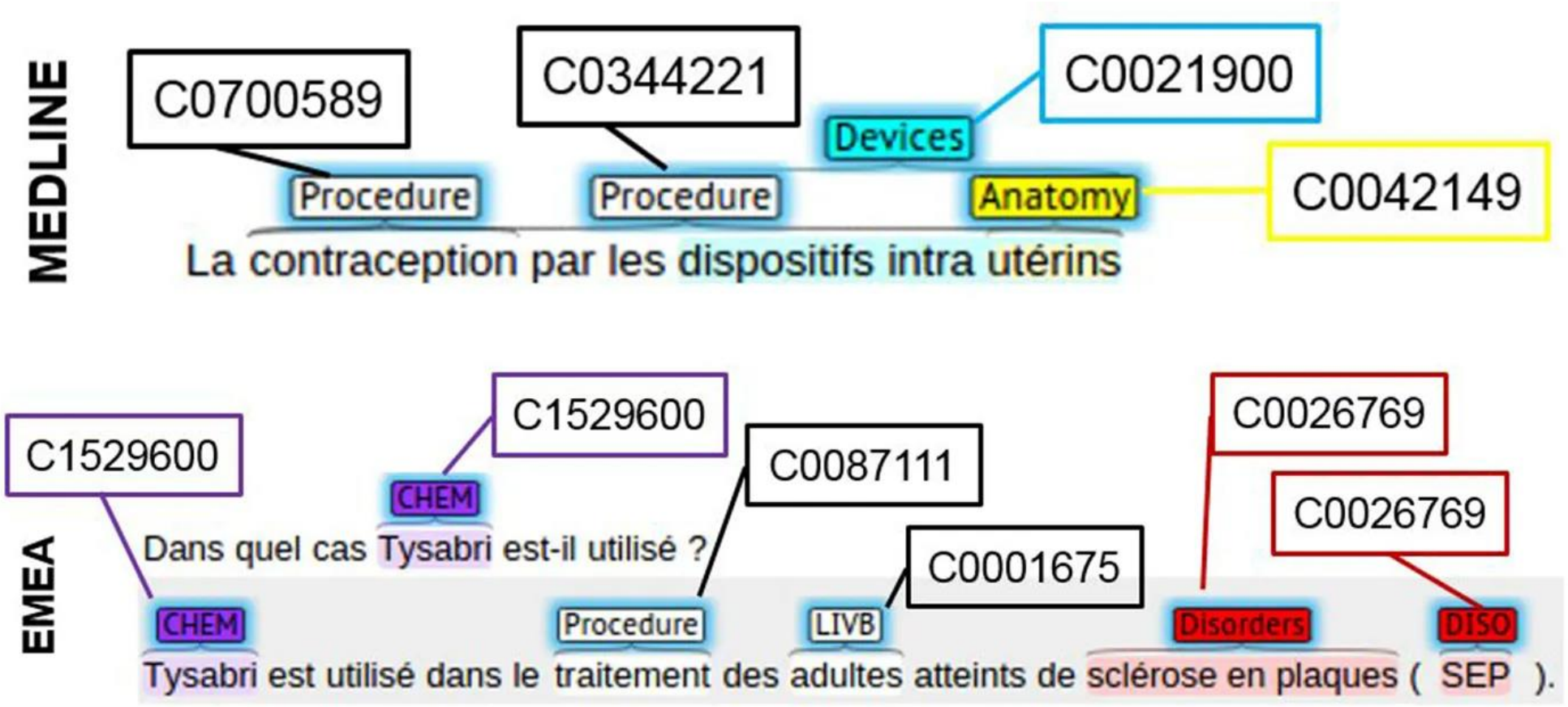
DrBERT (Labrak et al., 2023)

Modèle de langue (LM) pour le domaine médical en français

Basé sur le LM RoBERTa

Pré-entrainé sur le corpus médical en français NACHOS

QUAERO (Névéol et al., 2014)



DrBERT NER

- Exemple de résultat

```
462
463  phrase = Tysabri est utilisé dans le traitement des adultes atteints de sclérose en plaques ( SEP ).
464
465  DrBERT NER
466
467  [{'entity_group': 'CHEM', 'score': 0.96629834, 'word': 'T', 'start': 0, 'end': 1},
468   {'entity_group': 'CHEM', 'score': 0.9808131, 'word': 'ys', 'start': 1, 'end': 3},
469   {'entity_group': 'CHEM', 'score': 0.96141714, 'word': 'abri', 'start': 3, 'end': 7},
470   {'entity_group': 'PROC', 'score': 0.99979216, 'word': 'traitement', 'start': 27, 'end': 38},
471   {'entity_group': 'LIVB', 'score': 0.9994868, 'word': 'adultes', 'start': 42, 'end': 50},
472   {'entity_group': 'DISO', 'score': 0.999736, 'word': 'sclérose en plaques', 'start': 62, 'end': 82},
      {'entity_group': 'DISO', 'score': 0.97993124, 'word': 'SEP', 'start': 84, 'end': 88}]
473
```

5. Méthode(s) TAL

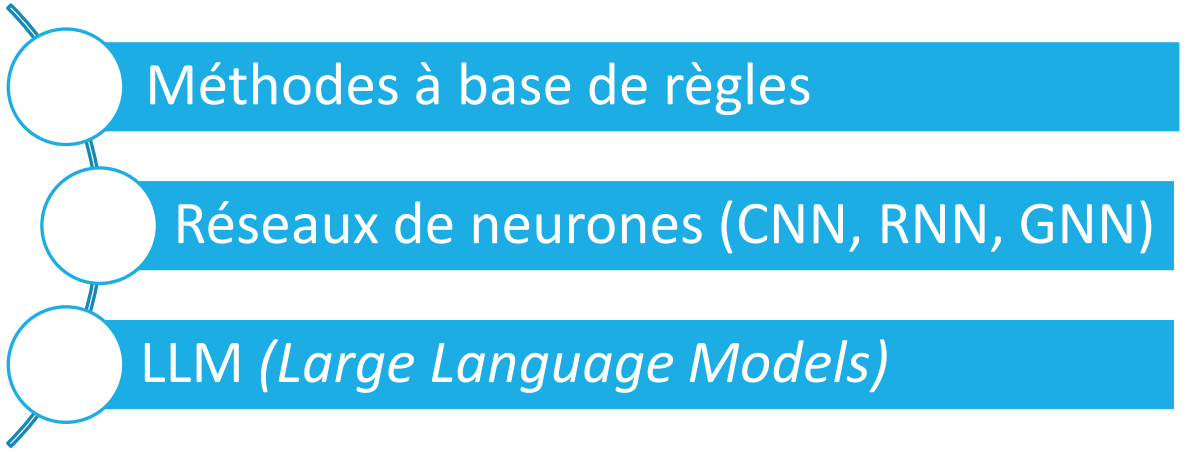
1. NER
2. **RE**
3. RAG
4. IA générative

5. Méthode(s) TAL

RE *(Relation Extraction)* =

extraction de relations sémantiques entre les entités nommées

(Detroja et al., 2023)



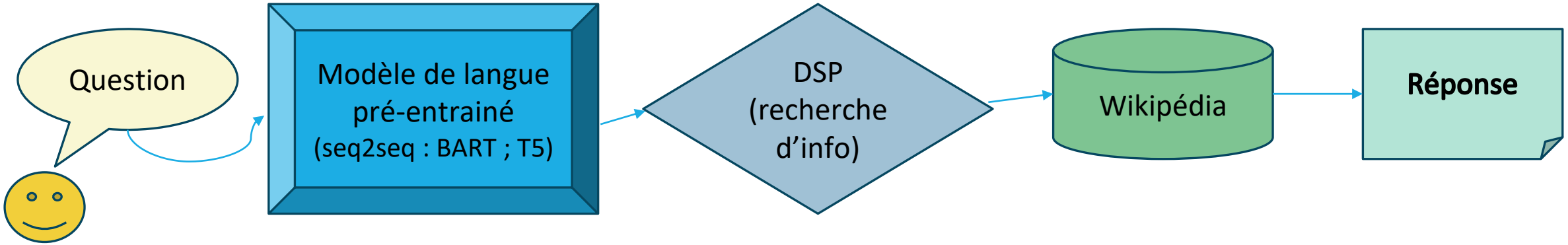
5. Méthode(s) TAL

1. NER
2. RE
3. **RAG**
4. IA générative

RAG

RAG *(Retrieval Augmented Generation)* = modèle qui combine la mémoire interne des modèles de langues avec une mémoire externe (base de connaissances) (Lewis et al., 2021)

- RAG-end2end (Siriwardhana et al., 2023) ; SELF-RAG (Asai et al., 2023) ; SELF-BioRAG (Jeong et al., 2024)



RAG

- Exemple de résultat

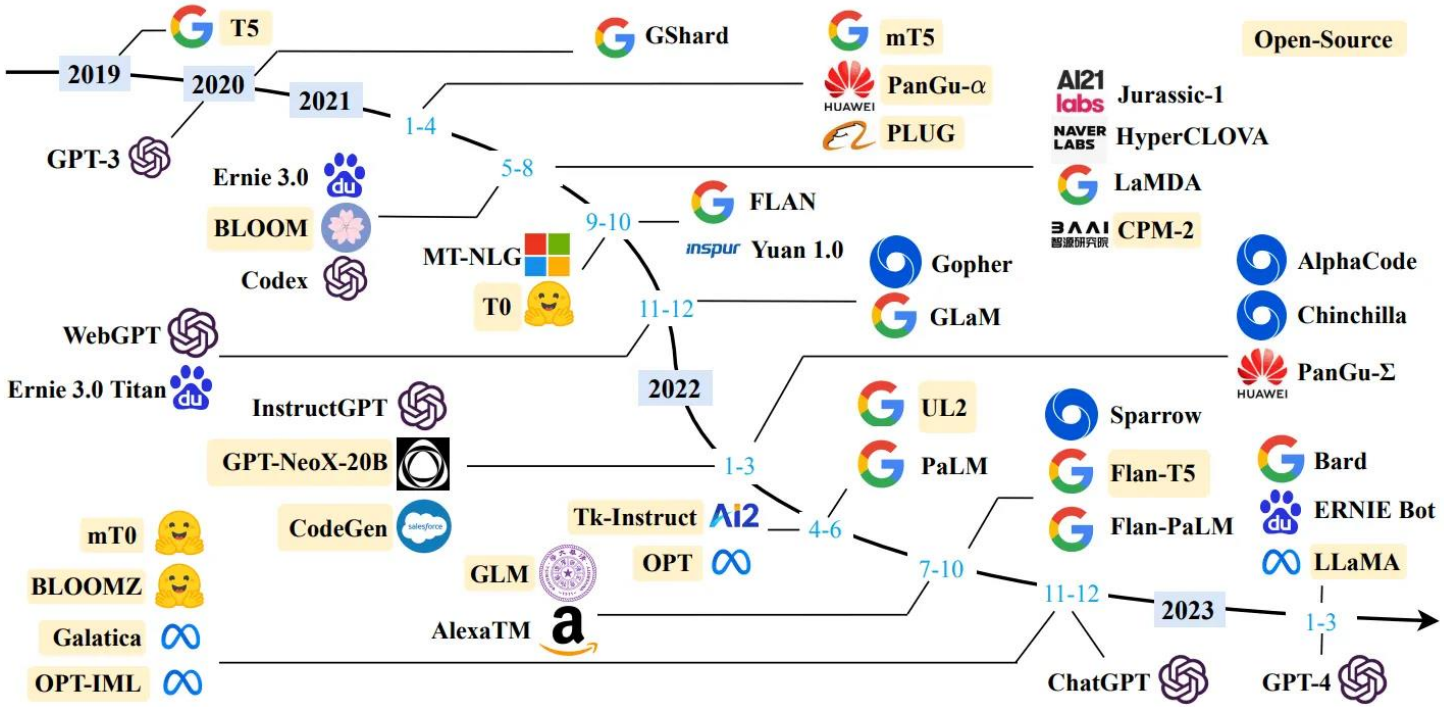
```python
from transformers import RagTokenizer, RagRetriever, RagTokenForGeneration

tokenizer = RagTokenizer.from_pretrained("facebook/rag-token-nq")
retriever = RagRetriever.from_pretrained("facebook/rag-token-nq", index_name="exact", use_dummy_dataset=True)
model = RagTokenForGeneration.from_pretrained("facebook/rag-token-nq", retriever=retriever)

input_dict = tokenizer.prepare_seq2seq_batch("What is a glioma?", return_tensors="pt")

generated = model.generate(input_ids=input_dict["input_ids"])
print(tokenizer.batch_decode(generated, skip_special_tokens=True)[0])

# Réponse : cancer
```

5. Méthode(s) TAL

1. NER
2. RE
3. RAG
4. **IA générative**

5. Méthode(s) TAL

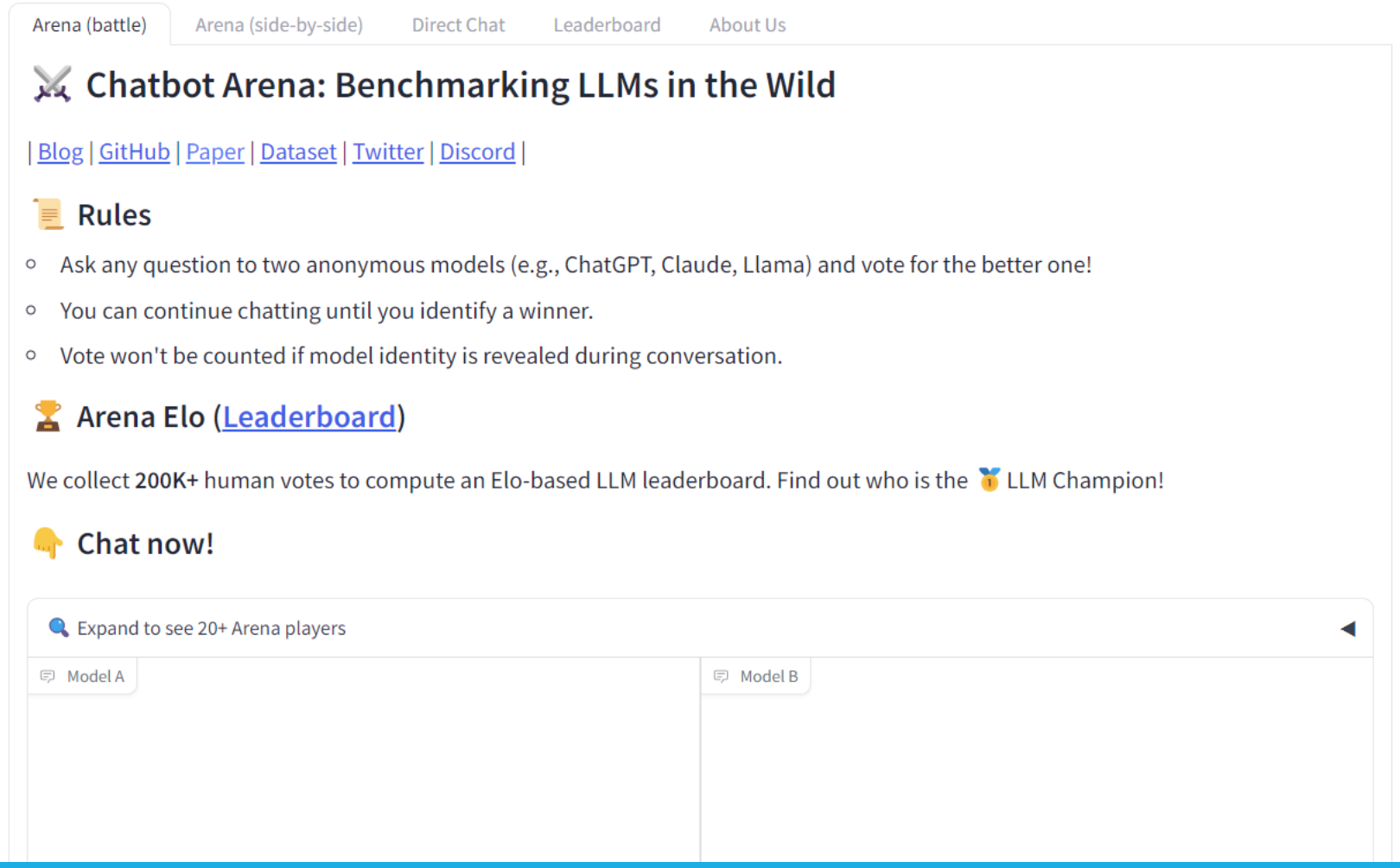
IA générative = modèle de langue pré-entrainé qui génère du texte



(Zhao et al., 2023)

Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

IA générative



UC Berkeley SkyLab

https://chat.lmsys.org/?model=koala-13b

6. Références bibliographiques

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Self-reflective Retrieval Augmented Generation. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Costa, R. (2005). Texte, terme et contexte. In *Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction*, pp. 79-88. Bruxelles, Belgique.

Detroja, K., Bhensdadia, C. K., & Bhatt, B. S. (2023). A Survey on Relation Extraction. *Intelligent Systems with Applications*, 200244.

Jeong, M., Sohn, J., Sung, M., & Kang, J. (2024). Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented Large Language Models. *arXiv preprint arXiv:2401.15269*.

Labrak, Y., Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille and Pierre-Antoine Gourraud. (2023). DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. (2014). The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing. In Proceedings of the Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing BioTxtM2014. 2014:24-30.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, *11*, 1-17.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Merci pour votre attention !