Grands modèles de langue : quelles perspectives à l'ère de la science ouverte ?

Didier Schwab

LIG – Université Grenoble Alpes
Open Science Days @ UGA 2025

25 novembre 2025

Domaine de recherche : Traitement Automatique des Langues et de la Parole (TALP)

Représentation, acquisition
et
exploitation
des connaissances
pour et par les données langagières

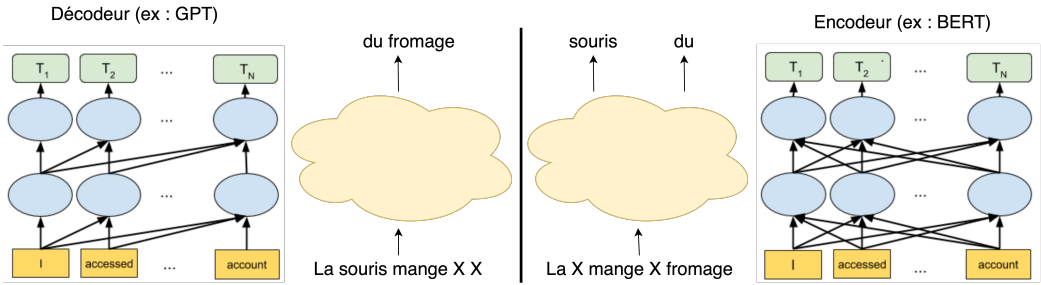
Traduction Automatique écrite, orale

Communication
Alternative & Augmentée

Applications médicales

Recherche d'information

Autoapprentissage : décodeurs vs encodeurs



- Type d'apprentissage non-supervisé (entrée : données dégradées ; sortie : données reconstituées)
- Modèles décodeurs : destinés à générer des textes cohérents à partir d'une amorce (prompt) – génératifs
- Modèles encodeurs : destinés à construire de bonnes représentations de mots, de phrases ou de documents – prédictifs

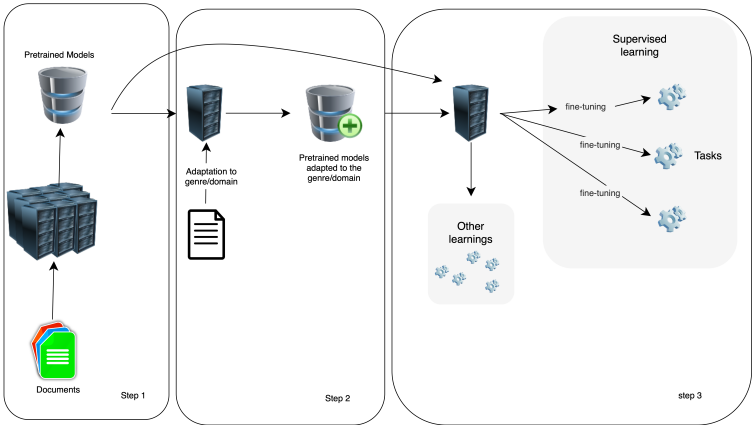
Données d'apprentissage (GPT-3)

Dataset	Quantité (jetons)	% des données d'apprentissage
Common Crawl (filtré)	410 milliards	60%
WebText2	19 milliards	22%
Books1	12 milliards	8%
Books2	55 milliards	8%
Wikipedia	3 milliards	3%

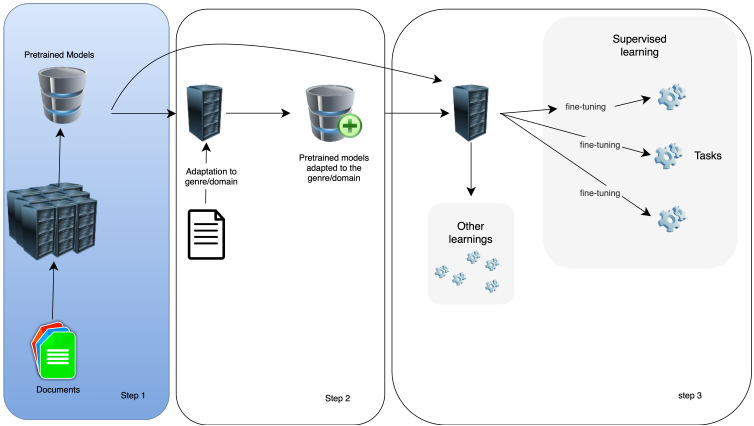
Table – Type de données utilisées pour entraîner GPT-3 [**?**]

Nous n'avons plus ces données pour les modèles OpenAI actuels (idem pour Gemini, Llama...)

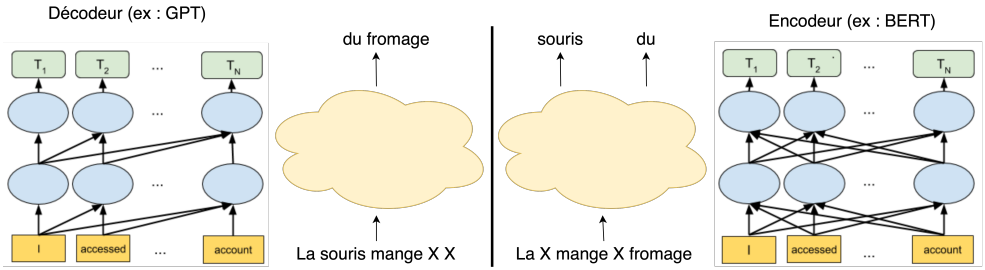
État de l'art de la chaîne de traitement en TALP



État de l'art de la chaîne de traitement en TALP



Apprentissage encodeurs – écrit – 2019. . .

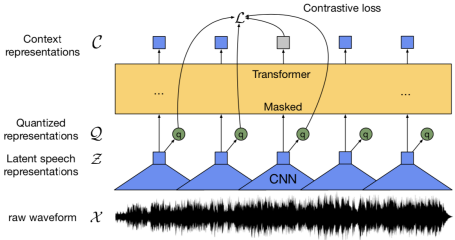


FlauBERT: Unsupervised Language Model Pre-training for French

Hang Le[1] Loïc Vial[1] Jibril Frej[1] Vincent Segonne[2] Maximin Coavoux[1]
Benjamin Lecouteux[1] Alexandre Allauzen[3] Benoît Crabbé[2] Laurent Besacier[1] Didier Schwab[1]

[1]Univ. Grenoble Alpes, CNRS, LIG [2]Université Paris Diderot [3]E.S.P.C.I, CNRS LAMSADE, PSL Research University
{thi-phuong-hang.le, loic.vial, jibril.frej}@univ-grenoble-alpes.fr
{maximin.coavoux, didier.schwab, benjamin.lecouteux, laurent.besacier}@univ-grenoble-alpes.fr
{vincent.segonne@etu, bcrabbe@linguist}.univ-paris-diderot.fr, alexandre.allauzen@espci.fr

Apprentissage encodeurs – oral – 2021. . .

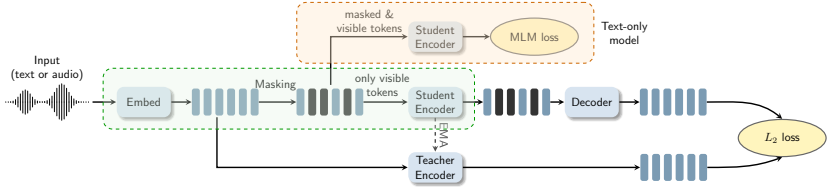


LeBenchmark 2.0: a Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech

Titouan Parcollet[a,b], Ha Nguyen[c], Solène Evain[d], Marcely Zanon Boito[f], Adrien Pupier[d], Salima Mdhaffar[c], Hang Le[d], Sina Alisamir[d], Natalia Tomashenko[c], Marco Dinarelli[d], Shucong Zhang[a], Alexandre Allauzen[e], Maximin Coavoux[d], Yannick Estève[c], Mickael Rouvier[c], Jerôme Goulian[d], Benjamin Lecouteux[d], François Portet[d], Solange Rossato[d], Fabien Ringeval[d], Didier Schwab[d], Laurent Besacier[f]

[a]*Samsung AI Center Cambridge, 50/60 Station Road, Cambridge, CB1 2JH, United Kingdom*
[b]*Department of Computer Science and Technology, University of Cambridge, 15 JJ Thomson Av., Cambridge, CB3 0FD, United Kingdom*
[c]*Laboratoire Informatique d'Avignon, Avignon Université, 339 Chem. des Meinajariès, Avignon, 84000, France*
[d]*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000, Grenoble, France*
[e]*ESPCI, CNRS LAMSADE, PSL Research University, France*
[f]*NAVER LABS Europe, France*

Apprentissage encodeurs/décodeurs – multimodal – 2024. . .



Pantagruel : Unified Self-Supervised Encoders for French Text and Speech

Phuong-Hang Le[1], Valentin Pelloin[2], Diandra Fabre[1], Solène Evain[1], Mohammed Ghennai[1], Maryem Bouziane[3], Arnault Chatelain[4], Aidan Mannion[1], Qianwen Guan[5], Kirill Milintsevich[2], Salima Mdhaffar[3], Nils Defauw[6], Shuyue Gu[5] Alexandre Audibert[1], Marco Dinarelli[1], Yannick Estève[3], Lorraine Goeuriot[1], Steffen Lalande[2], Nicolas Hervé[2] Maximin Coavoux[1], François Portet[1], Étienne Ollion[4], Marie Candito[5], Maxime Peyrard[1], Solange Rossato[1], Benjamin Lecouteux[1], Aurélie Nardy[7], Gilles Sérasset[1], Vincent Segonne[8], Didier Schwab[1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2] INA (Institut National de l'Audiovisuel), 4 Avenue de l'Europe, 94366 Bry-sur-Marne, France
[3] Avignon Université, LIA, France
[4] CREST (École Polytechnique, ENSAE, CNRS), 5 avenue Le Chatelier, 91120 Palaiseau, France
[5] LLF (Université Paris Cité and CNRS), UFRL Olympe de Gouges, 13 place Paul Ricoeur, 75013 Paris, France
[6] Univ. Grenoble Alpes, EFELIA-MIAI, IUT2 Grenoble, LIG, 38000 Grenoble, France
[7] Univ. Grenoble Alpes, Lidilem, 38000 Grenoble, France
[8] Université Bretagne Sud, CNRS, IRISA, France

Questions de recherche

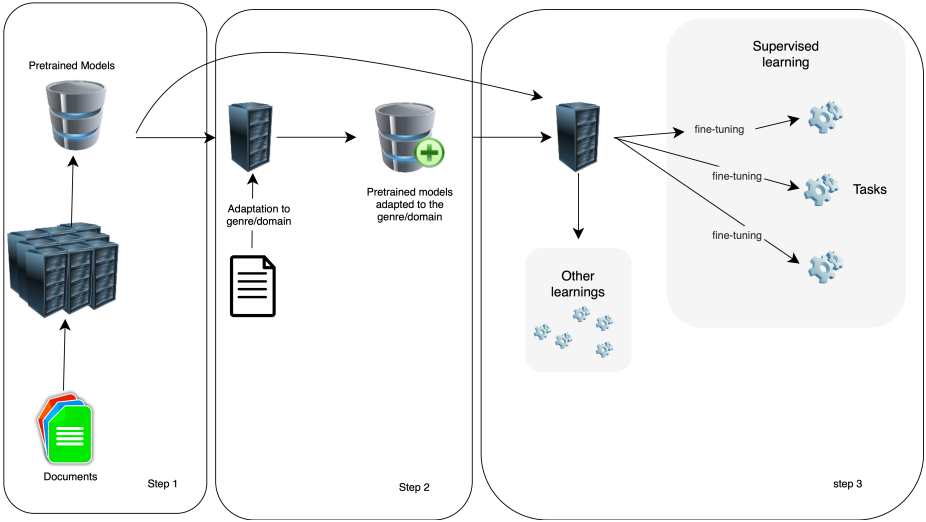
Prétraitement & Qualité des données

- **Impact du prétraitement sur les performances**
  Dans quelle mesure les choix de nettoyage, normalisation, segmentation, tokenisation influencent-ils les performances selon les modalités et tailles de corpus ?
- **Prétraitement et biais sociétaux**
  Quels effets le filtrage, le rebalancement démographique ou la sélection de sources ont-ils sur les biais présents dans les données et sur leur amplification ou atténuation dans les modèles ?

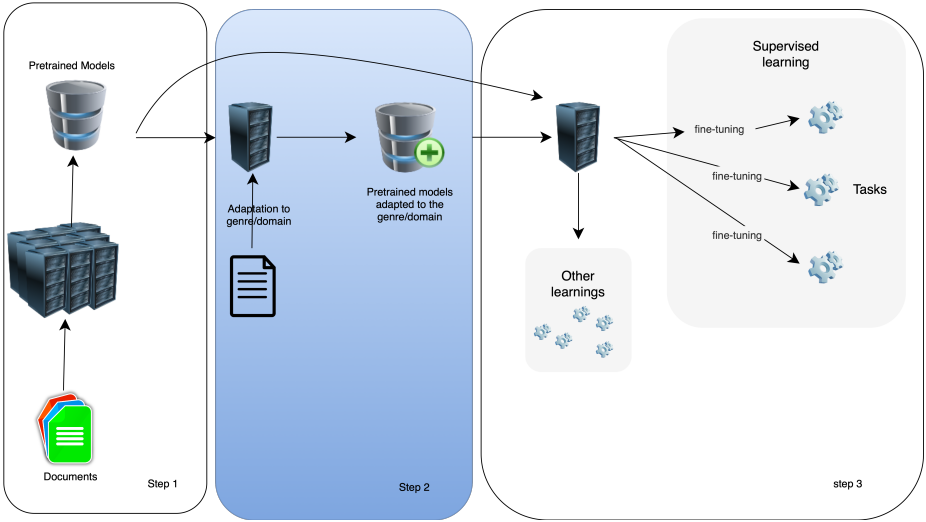
Modélisation et Alignement Multimodal

- **Alignement des modalités**
  Quelle est la meilleure manière d'aligner audio texte pictogrammes pour maximiser la qualité de l'espace latent commun ?
- **Évaluation de l'alignement**
  Peut-on définir des métriques robustes pour évaluer la qualité de cet alignement ?

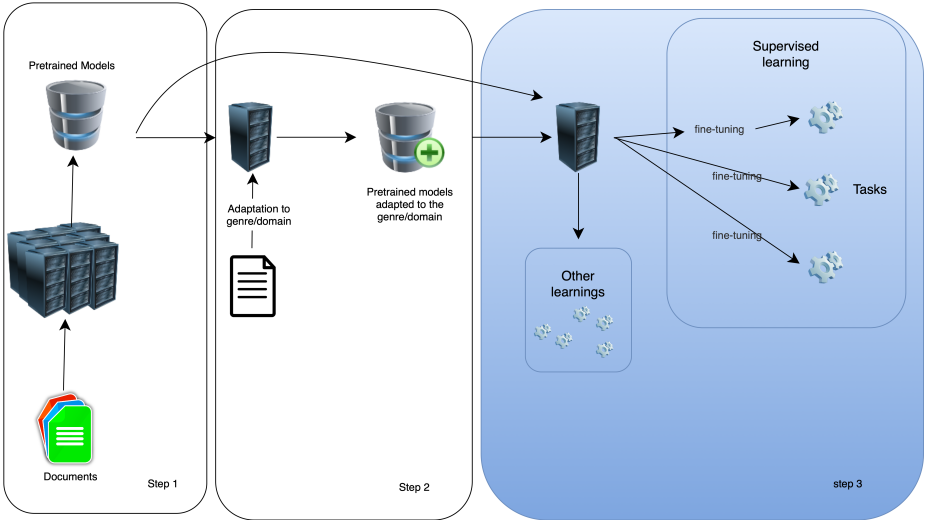
État de l'art de la chaîne de traitement en TALP



État de l'art de la chaîne de traitement en TALP



État de l'art de la chaîne de traitement en TALP



Tâches en aval : deux finalités complémentaires

1. Tâches « sur étagère » ouvertes

- Création de benchmark (unification des formats, code exécutable)
- **Objectif principal :** répondre aux *questions de recherche* définies précédemment
- Permettent une évaluation standardisée, comparable et reproductible.

2. Tâches créées sur mesure

- Conçues spécifiquement dans nos projets.
- **Objectif principal :** faire avancer un *sujet de thèse ou un projet scientifique ciblé*.
- Permettent d'explorer :
  - des compétences non couvertes par les benchmarks,
  - des besoins applicatifs réels (juridique, multimodalité, médical, ingénierie...).

Référentiel FLUE (2019) – inspiré par le référentiel GLUE (Wang et al., 2018)

- Classification de texte
- Paraphrase
- Inférence en langue naturelle (NLI)
- Analyse syntaxique et étiquetage morphosyntaxique
- Tâches de désambiguïsation du sens des mots

Dataset		Domaine	Train	Dev	Test
CLS-FR	Books	Avis de produits	2 000	-	2 000
	DVD		1 999	-	2 000
	Musique		1 998	-	2 000
PAWS-X-FR		Domaine général	49 401	1 992	1 985
XNLI-FR		Genres divers	392 702	2 490	5 010
French Treebank		Quotidien	14 759	1 235	2 541
Désambiguïsation lexicale des verbes		Genres divers	55 206	-	3 199
Désambiguïsation lexicale des noms		Genres divers	818 262	-	1 445

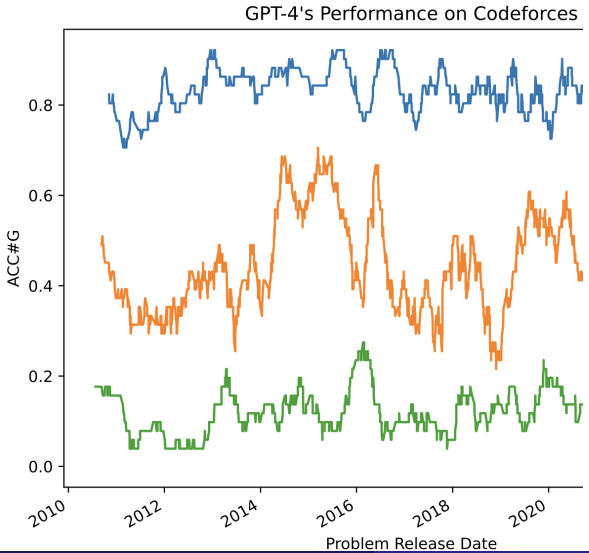
Table – FLUE : Évaluation de la compréhension de la langue française

Résultats sur FLUE (2019)

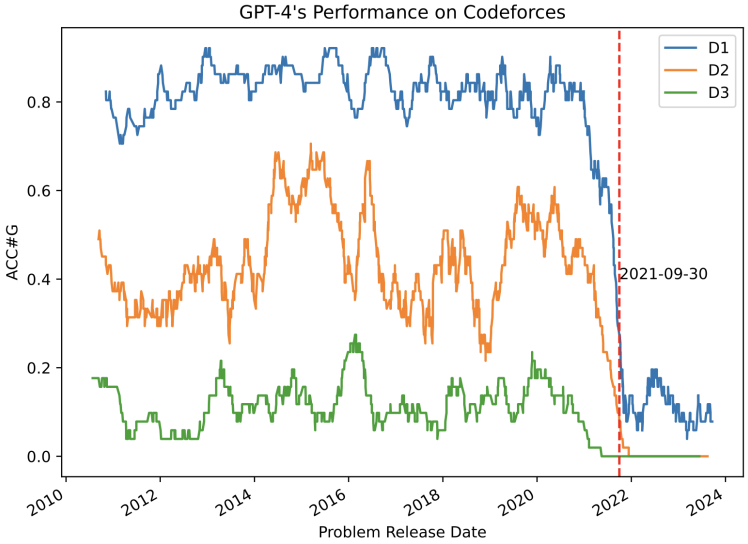
Task	Classification			Paraphrasing	NLI	Constituency		Dependency		Disambiguation	
Section	Livres	DVD	Musique							Nouns	Verbs
Measure	Acc.	Acc.	Acc.	Acc.	Acc.	F_1	POS	UAS	LAS	F_1	F_1
State-of-the-art	91.25[c]	89.55[c]	93.40[c]	66.20[d]	80.1/**85.2**[e]	87.4[a]		89.19[b]	85.86[b]	-	43.0[h]
Without pre-training	-	-	-			83.9	97.5	88.92	85.11	50.03	-
FastText	-	-	-			83.6	97.7	86.32	82.04	49.41	34.90
mBERT	86.15[c]	86.9[c]	86.65[c]	89.30[d]	76.9[f]	87.5	98.1	89.50	85.86	56.47	49.83
CamemBERT	92.30	93.00	94.85	**90.14**	81.2	88.4	**98.2**	91.37	88.13	56.06	50.02
FlauBERTbase	93.10	92.45	94.10	89.49	80.6	**89.1**	98.1	91.56	88.35	54.74	43.92
FlauBERTlarge	**95.00**	**94.10**	**95.85**	89.34	83.4	88.6	**98.2**	**91.61**	**88.47**	**57.85**	**50.48**

Table – Final results on FLUE. [a][Kitaev et al., 2019]. [b][Constant et al., 2013].
[c][Eisenschlos et al., 2019]. [d][Chen et al., 2017]. [e][Conneau et al., 2019]. [f][Martin et al., 2019].
[h][Segonne et al., 2019].

Performances de chatGPT [Huang, 2024]



Performances de chatGPT [Huang, 2024]



Contamination des données : un enjeu majeur pour l'évaluation des grands modèles de langue (en particulier décodeurs)

Qu'est-ce que la contamination ?

- Un modèle est dit **contaminé** lorsque des exemples de test apparaissent — même partiellement — dans ses données d'entraînement.
- Problème critique : le modèle peut réussir non pas par **compétence**, mais par **mémoire**.

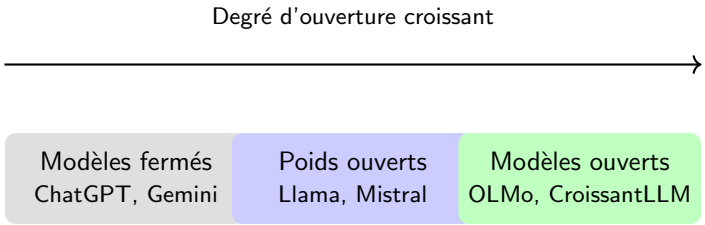
Pourquoi est-ce problématique ?

- **Surestimation massive des performances** sur les benchmarks classiques.
- Impossible d'évaluer la vraie **généralisation** ou la capacité de raisonnement.
- Impact sur la **science ouverte** : sans traçabilité des données, impossible de garantir une évaluation honnête.
- **Risque institutionnel** : modèles utilisés en décision, santé, droit, ingénierie.

LLM et science ouverte : où se joue l'ouverture ?

- **Données** : transparence des sources, licences, provenance, droits des auteurs, traçabilité du prétraitement, contrôle de la contamination.
- **Modèles** : publication des poids, description complète des hyperparamètres, versions stables et archivées, documentation des chaînes d'entraînement.
- **Code** : disponibilité des scripts de collecte, nettoyage, tokenisation, entraînement et évaluation ; reproductibilité totale des expériences.
- **Gouvernance** : qui choisit les données ? quels filtres ? quelles exclusions ? quels biais ? Comment garantir l'auditabilité et le contrôle institutionnel ?
- **Évaluation** : benchmarks non contaminés, renouvelés, versionnés ; règles claires sur l'usage ou non des modèles fermés dans les comparaisons.

Panorama des modèles de langue décodeurs



Modèles complètement ouverts : code + poids + données + pipeline

Exemples

- CroissantLLM, OLMo (AllenAI).

Avantages scientifiques

- **Transparence visée** : données d'entraînement publiées, pipeline de prétraitement documenté, code complet disponible.
- **Auditabilité maximale** : vérification des biais, des sources, des exclusions.
- **Reproductibilité forte** : possibilité de réentraîner et de vérifier les résultats.

Limites actuelles

- Performances encore inférieures aux modèles fermés ou poids ouverts.
- Taille souvent réduite (max : 32B)

Modèles poids ouverts : Llama, Mistral, Qwen...

Ce qui est ouvert

- Les **poids** sont accessibles : utilisation locale possible.
- Permet :
  - déploiement interne sécurisé,
  - adaptation (LoRA, fine-tuning),
  - intégration dans des pipelines reproductibles.

Limites importantes

- **Données d'entraînement non publiées** :
  - contamination non contrôlable,
  - impossible de connaître les biais, filtrages, exclusions.
- **Pipeline non ouvert** : impossible de reproduire l'entraînement ou les choix de nettoyage.
- **Reproductibilité partielle** : poids ouverts $\neq$ modèles ouverts.

Modèles complètement fermés : ChatGPT, Gemini, Claude...

Caractéristiques

- Accès uniquement via API ou interface Web.
- Modèle souvent non clairement identifié, version changeante, poids inaccessibles.
- Entraînement, données, filtrage et alignement entièrement opaques.

Problèmes majeurs

- **Confidentialité** : les conversations peuvent être collectées et réutilisées.
- **Non-reproductibilité** :
  - difficulté/impossibilité pour répéter une expérience,
  - le modèle peut être modifié sans préavis.
- **Opacité totale sur les données d'entraînement** :
- **Évaluation difficile à interpréter** : le modèle peut « connaître la réponse » via ses données.

Peut-on encore faire de la science avec les modèles actuels ?

La recherche n'a de valeur que si elle peut être comprise, reproduite, vérifiée et contestée.

- Peut-on publier des résultats fondés sur un modèle dont on ne peut ni documenter les données, ni répéter l'expérience ?
- Faut-il interdire l'usage de ces modèles fermés en recherche ? **Probablement pas** : pour les comprendre, il faut pouvoir les analyser, les tester, les déconstruire, les reproduire au plus juste.
- En revanche, il faut **cadrer strictement** leur usage :
  - justifier explicitement quand un modèle fermé est utilisé ;
  - exiger que les articles et les relectures identifient les évaluations non reproductibles.
  - accepter collectivement que les auteurs et autrices ≪ ne se comparent pas à ChatGPT ≫ ;

→ La science reste possible, mais seulement si nous fixons les règles.

Merci !

Didier Schwab

didier.schwab@univ-grenoble-alpes.fr

Références I

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017).
Enhanced lstm for natural language inference.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1657–1668.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019).
Unsupervised cross-lingual representation learning at scale.
arXiv preprint arXiv:1911.02116.

Constant, M., Candito, M., and Seddah, D. (2013).
The ligm-alpage architecture for the spmrl 2013 shared task : Multiword expression analysis and dependency parsing.
In *Proceedings of the EMNLP Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2013)*.

Références II

Eisenschlos, J., Ruder, S., Czapla, P., Kardas, M., Gugger, S., and Howard, J. (2019).
Multifit : Efficient multi-lingual language model fine-tuning.
In *Proceedings of the 2019 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Kitaev, N., Cao, S., and Klein, D. (2019).
Multilingual constituency parsing with self-attention and pre-training.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, É., Seddah, D., and Sagot, B. (2019).
CamemBERT : a Tasty French Language Model.
arXiv preprint arXiv:1911.03894.

Références III

Segonne, V., Candito, M., and Crabbé, B. (2019).
Using wiktionary as a resource for wsd : the case of french verbs.
In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 259–270.