Participation d'EDF R&D à DEFT 2023

M. Bothua, L. Hassani, M. Jubault, P. Suignard

Atelier du 05/06/2023

Plan de la présentation

DEFT 2023
Défi Fouille de Textes@TALN 2023

- EDF R&D
- Pourquoi participer à ce concours ?
- Les tâches et les méthodes utilisées
- Résultats obtenus
- Conclusion

EDF R&D

- Structure

✓ EDF R&D au service de toutes les entités du groupe EDF

✓ 3 centres de R&D principaux en France (**Chatou**, **Fontainebleau** et **Saclay**)

✓ Autres centres au UK, Allemagne, Chine et USA

✓ Environ 2000 personnes au total

EDF R&D

- Travaux sur le texte
  - ✓En appui aux différents métiers : EDF Commerce, Hydraulique, Eolien, Nucléaire, Enedis, RH, IT…
  - ✓En permanence : ~8 personnes, 1 doctorant, 2 stagiaires
  - ✓Thèmes : classification, clustering, orthographe, annotations, web sémantique, résumé, anonymisation de données, détection de nouveauté, plongements mots/documents…
  - ✓Sujets : mails, réclamations, comptes-rendus d'interventions, documents techniques, manuscrits anciens, conversations téléphoniques, réseaux sociaux, chatbots…
  - ✓Type de prestations : veille, développement, conseil, méthode, étude.

Pourquoi participer à ce concours ?

- Sujet complexe « réponse automatiquement à des questionnaires à choix multiples »
- Occasion de tester les « LLM » et leur surcouche conversationnelle
- Permet de se comparer/partager/discuter avec les autres équipes
- Emulation interne
- Reconnaissance interne
- Les résultats contribuent directement à EDF Commerce et à d'autres entités du groupe EDF

ChatGPT

- Tester la méthode/outil au-delà du *buzz* médiatique
- Mode « chat » ou « completion »
- « *Zéro-shot* »
- Modèle « `gpt-3.5-turbo` »
- Utilisation en mode « cURL »
- Premiers tests décevants. Importance de la manière de poser la question. (Liévin et al., 2022)
- `Q : Texte de la question \n A) Réponse 1 \n B) Réponse 2 \n C) Réponse 3 \n D) Réponse 4 \n E) Réponse 5.`

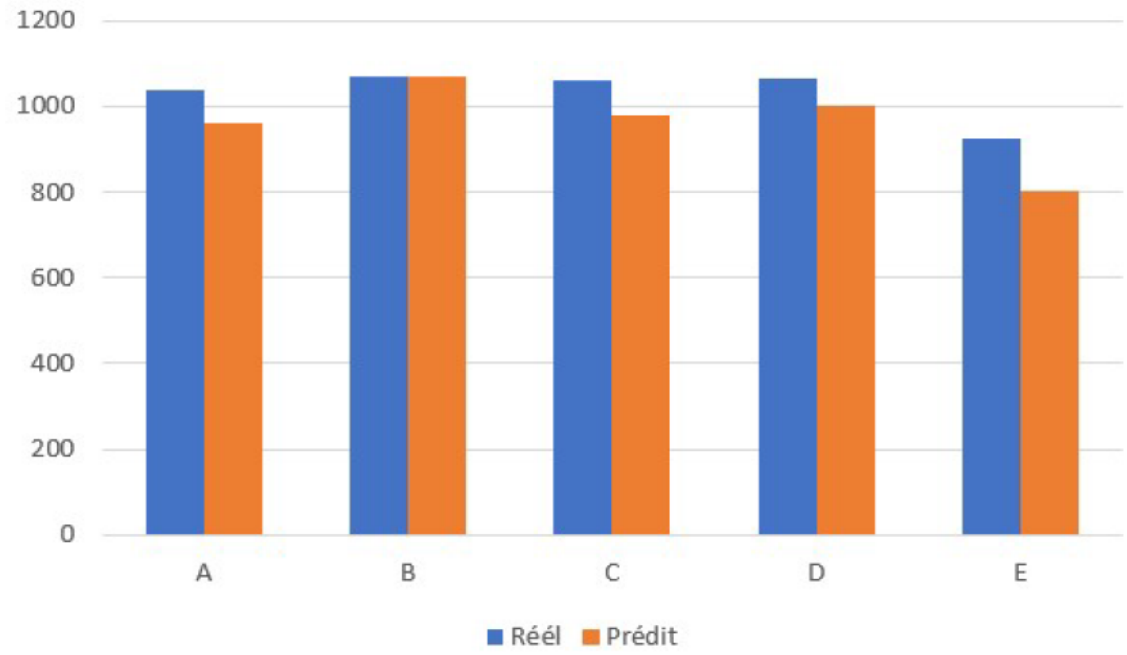
ChatGPT

- Sur le corpus d'apprentissage, réponses fournies par ChatGPT :

Format de réponse	Extraction	Nb de cas
B) Réponse 2 \n C) Réponse 3 (\n\n A) La réponse… est fausse)*	BC	2063/2171
toutes les (réponses|propositions|affirmations) sont (vraies|possibles|correctes|exactes)	ABCDE	37/2171
(Réponse :)* B et C sont exactes.\n \n Explication : ...	ABCDE (défaut)	71/2171
Autres types de réponses	ABCDE (défaut)	

ChatGPT

- Résultats obtenus sur le corpus d'apprentissage
  - ✓29,13% en EMR (Exact Match Ratio) et 58,26% pour Hamming
- Ventilation
  - ✓A,B,C et D réels quasiment égaux
  - ✓E réel un peu en dessous
  - ✓A peu près pareil sur le prédit
  - ✓E prédit encore inférieur



BloomZ

- Modèle Open-Source et gratuit entraîné sur 46 langues différentes dans le cadre du workshop collaboratif BigScience
- Modèle entraîné sur la tâche de **compréhension d'instruction en zero-shot**
- Test des modèles 1.1B et 7.1B en inférence > meilleurs résultats avec le plus gros modèle, qui comprends mieux les instructions
- Meilleurs résultats en one-shot lors des tests initiaux (avec un exemple qui contient plusieurs bonnes réponses)
- Même avec few-shot, il lui arrive de ne pas respecter le format de sortie demandé

BloomZ

- Test d'une méthode de « PEFT » (Parameter-Efficient Fine Tuning) sur le modèle BloomZ 1.1B : le **Prompt Tuning** [Lester et al., 2021]
- Résultats décevants : il n'améliore que très légèrement le Hamming Score, et pas du tout l'Exact Match Ratio
- Ajout de temps d'inférence + ajout de bruit au niveau de la génération

Résultats

- Tâche principale

Système	Hamming	EMR
ChatGPT	**64,40**	**46,46**
BloomZ - run1	26,34	14,63
BloomZ - run2	35,90	15,27
BloomZ - run3	37,93	12,70

Pour BloomZ :
- Run 1 : zero shot
- Run 2 et 3 : tests en one-shot, sans FT (avec des exemples différents).

- Tâche annexe

✓Nombre de réponses supposément justes : nombre de réponses considérées comme étant justes par ChatGPT

Système	Accuracy	F1-score macro
ChatGPT	65,92	44,36

Conclusion

- Participer à la campagne DEFT 2023, nous a permis de tester **ChatGPT** et **Bloom/BloomZ**

- **BloomZ** obtient des scores équivalents aux méthodes présentées dans (Labrak et al., 2022). **ChatGPT** présente des scores environ 2 fois supérieurs, ce qui prouve sa qualité et justifie l'engouement qu'il suscite.

- Ces méthodes, alliées aux mécanismes de « **Prompt Engineering** » sont très prometteuses pour le traitement des données textuelles au sein d'EDF.