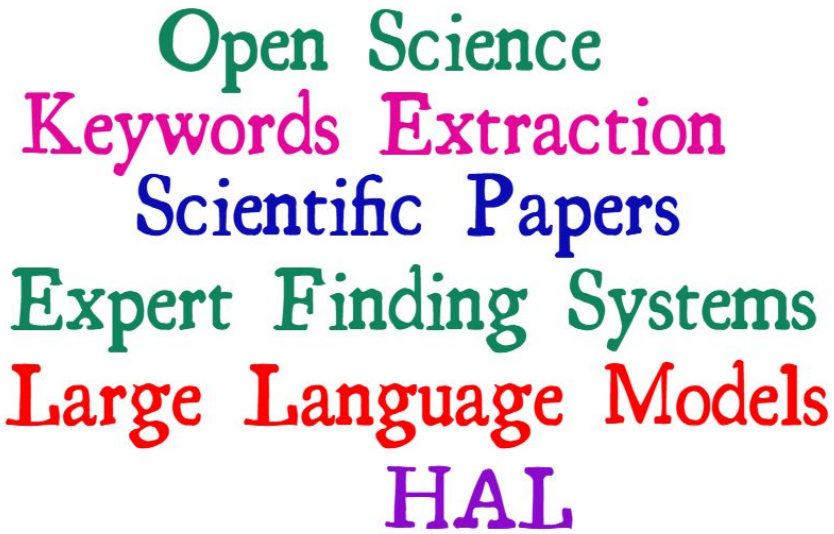
Extraction de mots clés à partir d'articles scientifiques: comparaison entre modèles traditionnels et modèles de langue

Nacef Ben Mansour et Motasem Alrahabi
Sorbonne Université

Colloque Ariane, Paris, 26 et 27 novembre 2024

Plan

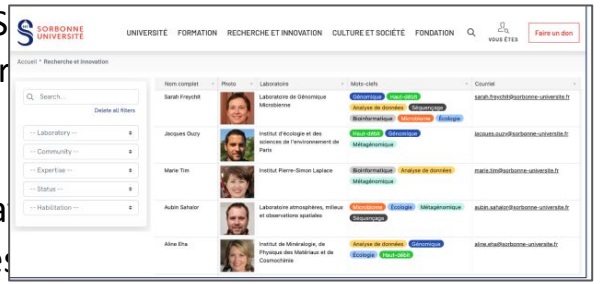
- Etat de l'art
- Approches
- Evaluation
- Discussion



Présentation

- Contexte général
  - Cette étude, initiée lors d'un stage à S
    dans un projet plus vaste visant à car
    des personnels et des laboratoires de SU.

    → renforcer les partenariats, les forma
    à l'aide de la structuration des données
    d'outils d'interrogation adaptés à la science ouverte.



Projet Expertise SU (prof. Stéphane Le Crom)

- L'extraction automatique des mots-clés est au cœur de ce projet en cours: elle permet de relier chaque expertise à une liste de mots-clés représentatifs et pertinents.
- Objectif du stage: comparer les performances des outils existants pour l'extraction des mots clés à partir de la base de données HAL.

Extraction des mots-clés: état de l'art

- L'extraction de mots-clés permet de sélectionner les termes les plus importants et représentatifs d'un document ou d'un corpus.
- Cette technique permet de condenser un texte, d’améliorer la recherche d’information et d’analyser les tendances émergentes.
- Distinguer entre méthodes extractives qui sélectionnent des mots-clés existants dans le texte, et les méthodes génératives qui créent ou reformulent des mots-clés.

Contexte et état de l'art

- Les approches classiques
  - Heuristiques: utilisation de mesures comme TF-IDF pour pondérer les mots selon leur fréquence.
  - Statistiques et probabilistes: comme les chaînes de Markov.
  - Règles linguistiques : analyse syntaxique et extraction basée sur des syntagmes nominaux ou des caractéristiques linguistiques (YAKE!).
  - Apprentissage supervisé et non supervisé : Modèles entraînés sur des données annotées ou techniques exploitant des graphes (TextRank, PositionRank).
- Bien qu'efficaces, ces approches manquent de nuance contextuelle.

Contexte et état de l'art

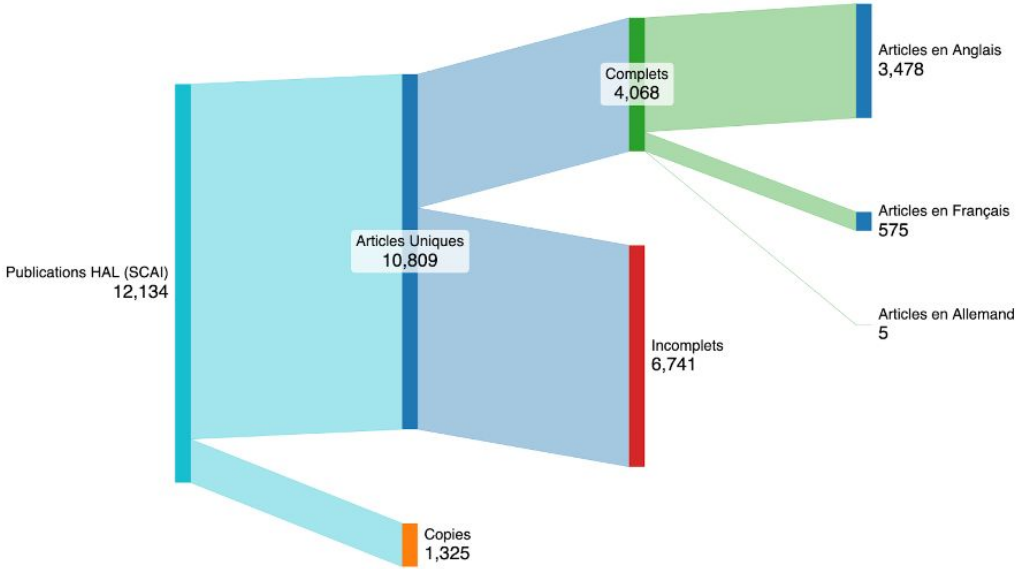
- Les approches modernes (basées sur les réseaux neuronaux et les LLMs)
  - Méthodes basées sur les motifs et la similarité contextuelle : des approches comme PatternRank et KeyBERT utilisent des embeddings contextuels et des motifs syntaxiques (PoS) pour évaluer la proximité des mots-clés avec le texte source.
  - Extraction Zero-shot, Few-shot et Fine-tuning : Ces modèles permettent une génération de mots-clés sans entraînement (zero-shot), avec peu d'exemples annotés (few-shot) ou via un fine-tuning sur des corpus spécifiques.
- Tendances émergentes: approches hybrides, adaptation linguistique et modélisation thématique (BERTopic), etc.

Méthodologie

- Extraction des résumés et des mots-clés d'auteurs depuis HAL.
  - Considérer que les mots clés d'auteurs comme référence
- Mise en œuvre d'approches classiques et modernes:
  - avec une utilisation zero-shot pour les LLMs.
  - avec deux configurations: résumés seuls ou résumés combinés aux titres.
- Évaluation de la pertinence et de la précision des mots-clés extraits (P, R, F1).
- Analyse des résultats via des statistiques descriptives et des analyses qualitatives.

Données

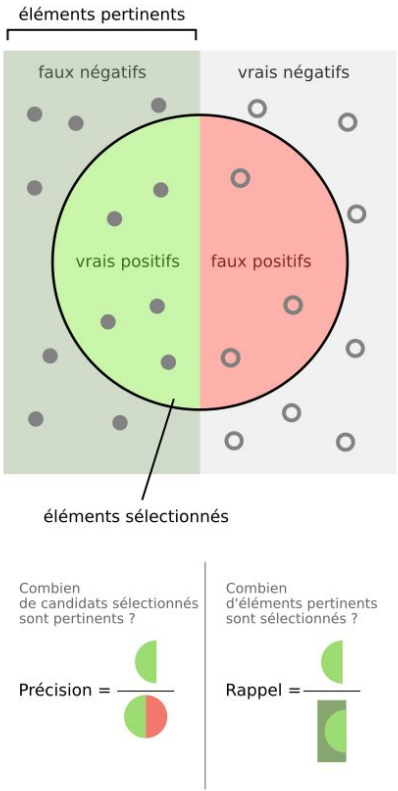
- Source des données : [Plateforme HAL](), une archive ouverte pour la diffusion des travaux de recherche scientifique (216k publications SU → échantillon SCAI, ≈ 12k)
- Normalisation : tous les titres, mots-clés et résumés ont été convertis en minuscules.



Evaluation

- Comparaison des mots-clés générés avec ceux fournis par les auteurs sur HAL. Métrique utilisée : F1-Score, moyenne harmonique entre la précision et le rappel:

$$F = 2 \cdot \frac{(\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})}$$



Résultats: approches classiques

- Ces résultats montrent une performance disparate, avec des écarts allant jusqu'à 60% entre les modèles les plus faibles (TextRank) et les plus performants (PositionRank, suivi de MultipartiteRank).

Graph / Stat with title

Modèle	Precision	Recall	F1 Score
kw_by_pos_rank	0.062	0.115	0.080
kw_by_mp_rank	0.062	0.113	0.079
kw_by_topic_rank	0.059	0.108	0.076
kw_by_single_rank	0.053	0.098	0.068
kw_by_yake	0.053	0.098	0.068
kw_by_text_rank	0.039	0.072	0.050

Graph / Stat Without title

Modèle	Precision	Recall	F1 Score
kw_by_pos_rank	0.056	0.103	0.072
kw_by_mp_rank	0.056	0.103	0.072
kw_by_topic_rank	0.053	0.096	0.068
kw_by_yake	0.052	0.096	0.067
kw_by_single_rank	0.045	0.083	0.058
kw_by_text_rank	0.036	0.066	0.046

Résultats: approches modernes (embeddings)

- Les résultats pour KeyBERT selon les variantes MMR et MSum sont très similaires, ce qui indique des performances quasiment équivalentes.
- L'inclusion du titre n'a pas d'effet significatif sur les résultats.

Keybert with title

Modèle	Precision	Recall	F1 Score
kw_by_keybert	0.058	0.081	0.067
kw_by_keybert_mmr_msum	0.052	0.073	0.061

Keybert without title

Modèle	Precision	Recall	F1 Score
kw_by_keybert	0.056	0.078	0.065
kw_by_keybert_mmr_msum	0.050	0.070	0.058

Résultats: approches modernes (LLMs)

- Les résultats montrent une variabilité importante, avec des performances allant du simple au triple. L'inclusion des titres améliore les performances d'environ 10 %.

LLM with title

Modèle	Precision	Recall	F1 Score
llama3-70b-8192	0.132	0.245	0.163
claude-3-haiku-20240307	0.130	0.218	0.154
llama3-8b-8192	0.147	0.181	0.151
gpt4o	0.075	0.222	0.108
claude-instant-1.2	0.073	0.183	0.097
kw_openai_gpt3.5	0.089	0.094	0.087
open-mixtral-8x7b	0.057	0.188	0.083
open-mistral-7b	0.050	0.199	0.077
gemma-7b-it	0.051	0.079	0.059

LLM without title

Modèle	Precision	Recall	F1 Score
llama3-70b-8192	0.120	0.224	0.148
claude-3-haiku-20240307	0.120	0.204	0.143
llama3-8b-8192	0.136	0.172	0.142
gpt4o	0.071	0.206	0.101
claude-instant-1.2	0.066	0.171	0.088
kw_openai_gpt3.5	0.086	0.089	0.083
open-mistral-7b	0.047	0.176	0.070
open-mixtral-8x7b	0.048	0.156	0.069
gemma-7b-it	0.052	0.081	0.060

Evaluation : Exact Matching *vs* Fuzzy Matching

- Exact Matching : Une correspondance stricte où les mots-clés extraits sont exactement identiques à ceux attendus. Example:

$$"chat" = "chat"$$

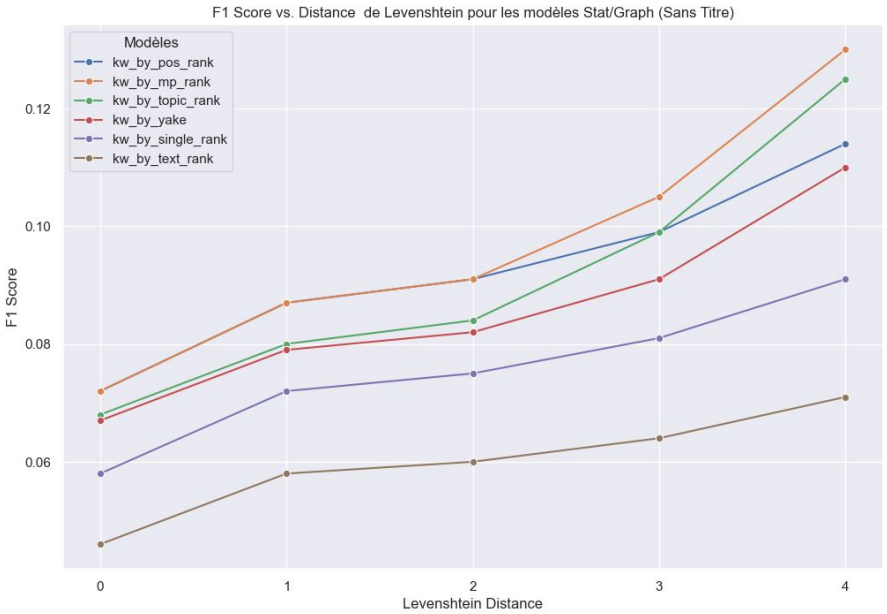
$$"chat" \neq "chats"$$

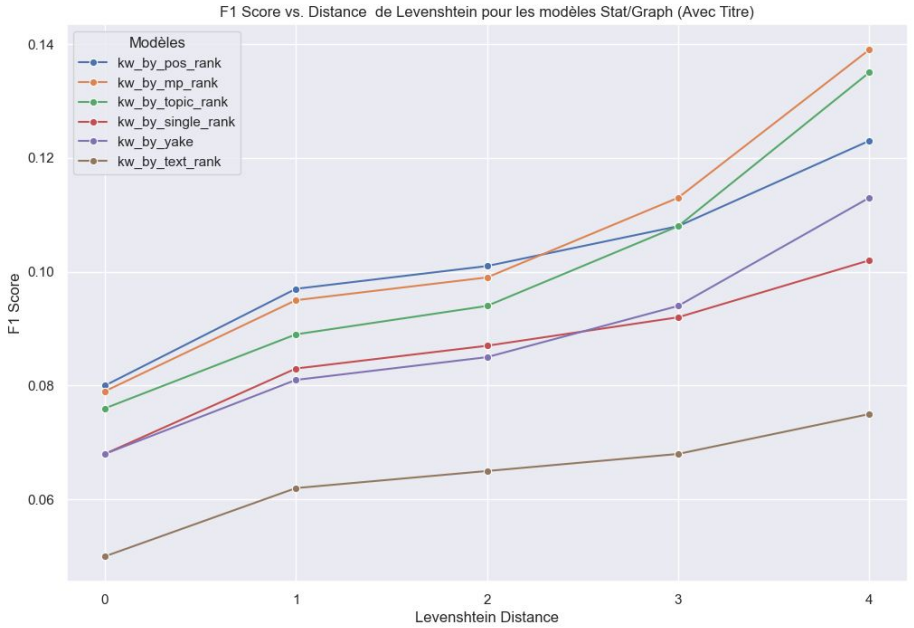
- Fuzzy Matching : Une correspondance flexible basé sur des différents algorithmes : Distance de Jaro-Winkler, Distance de Levenshtein, embeddings. Example :

$$"chat" \approx "chats"$$

avec la distance de Levenshtein

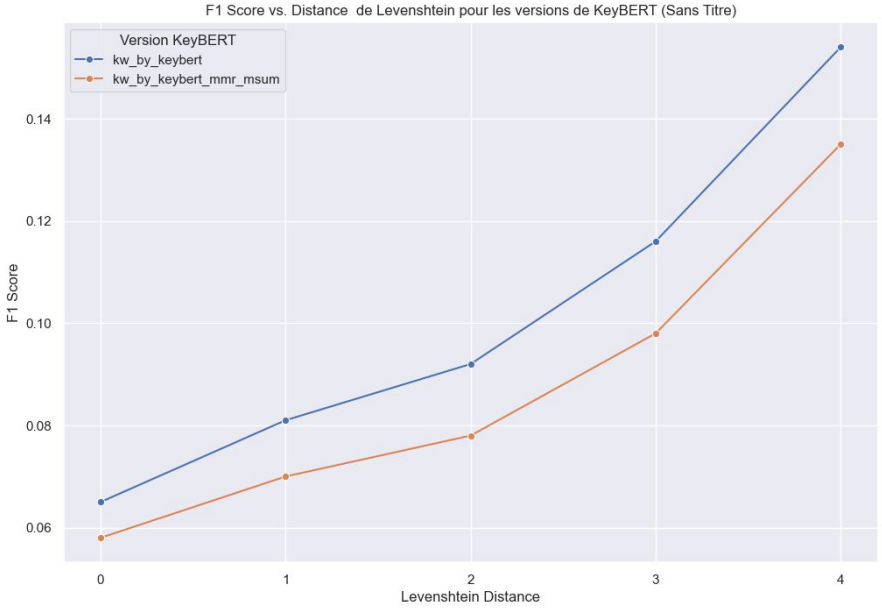
Résultats: approches classiques

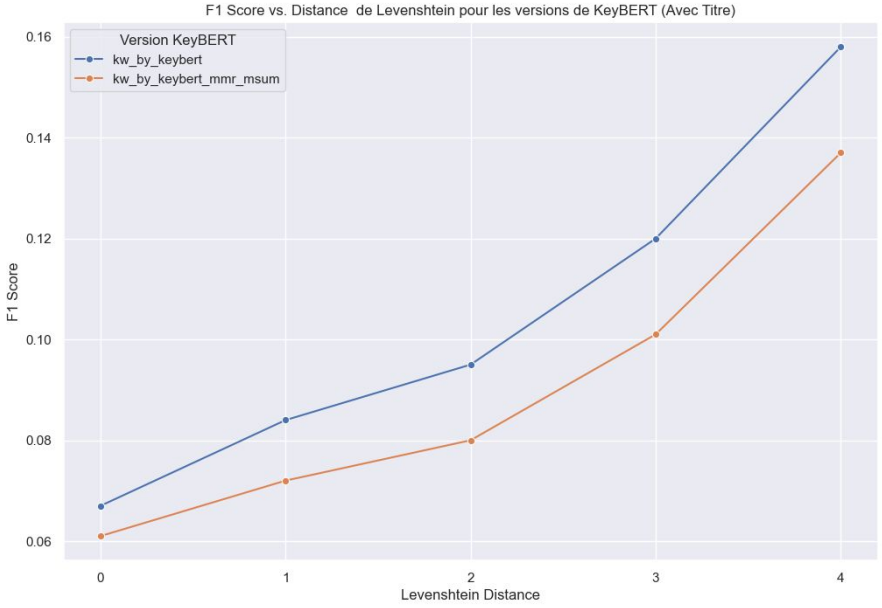




avec la distance de Levenshtein

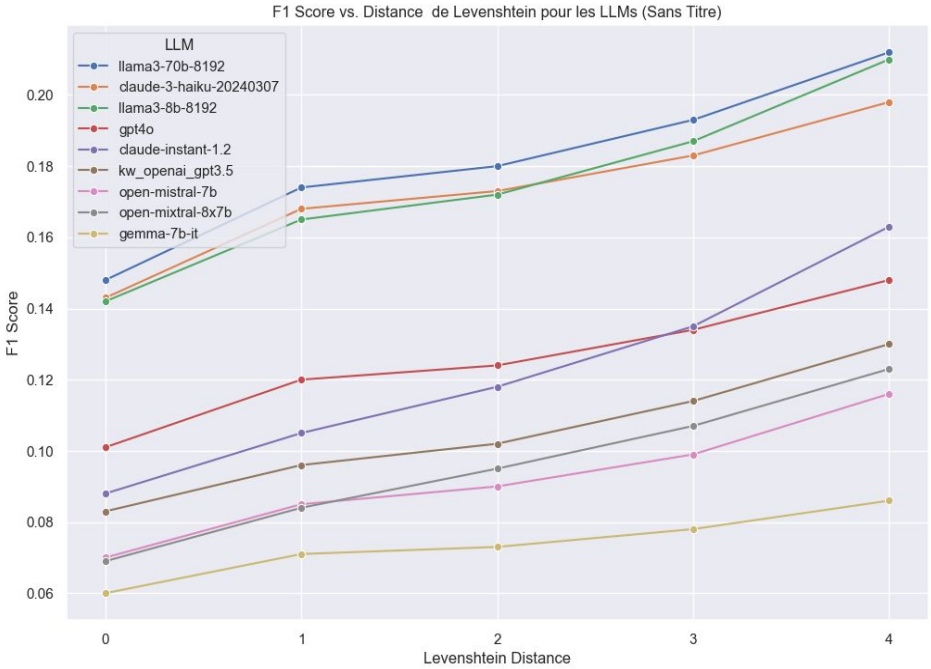
Résultats: approches modernes (embeddings)

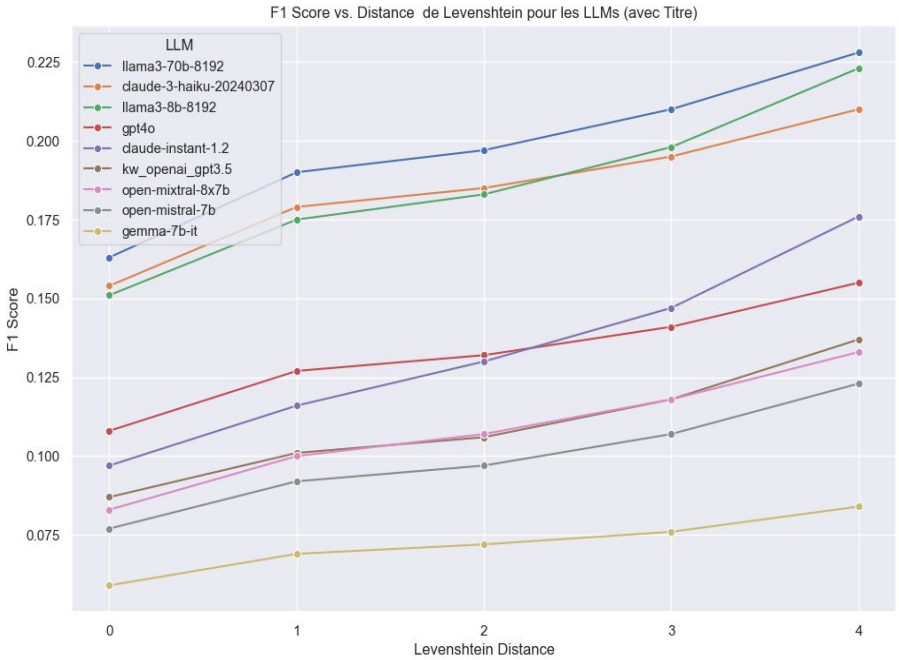




avec la distance de Levenshtein

Résultats: approches modernes (LLMs)





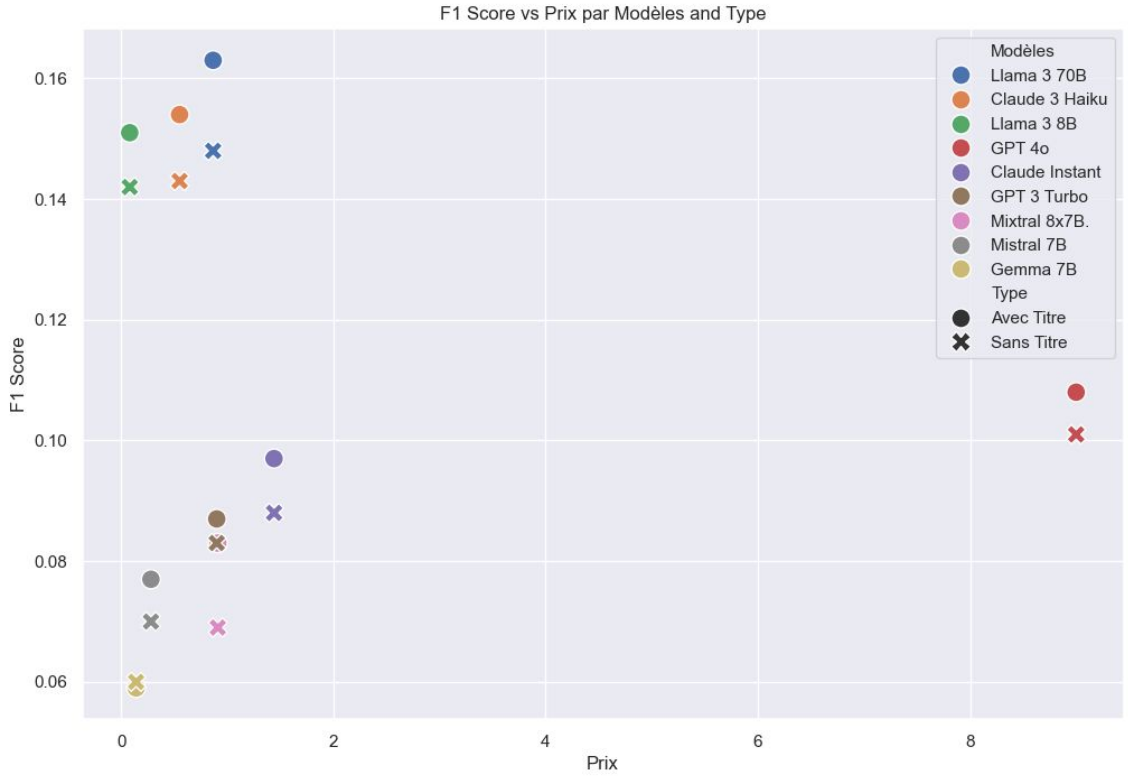
LLMs et coût par token (Token Efficiency Score)

- On introduit une nouvelle métrique ***TES*** combinant les performances des LLMs et les coûts, tout en pénalisant les coûts.

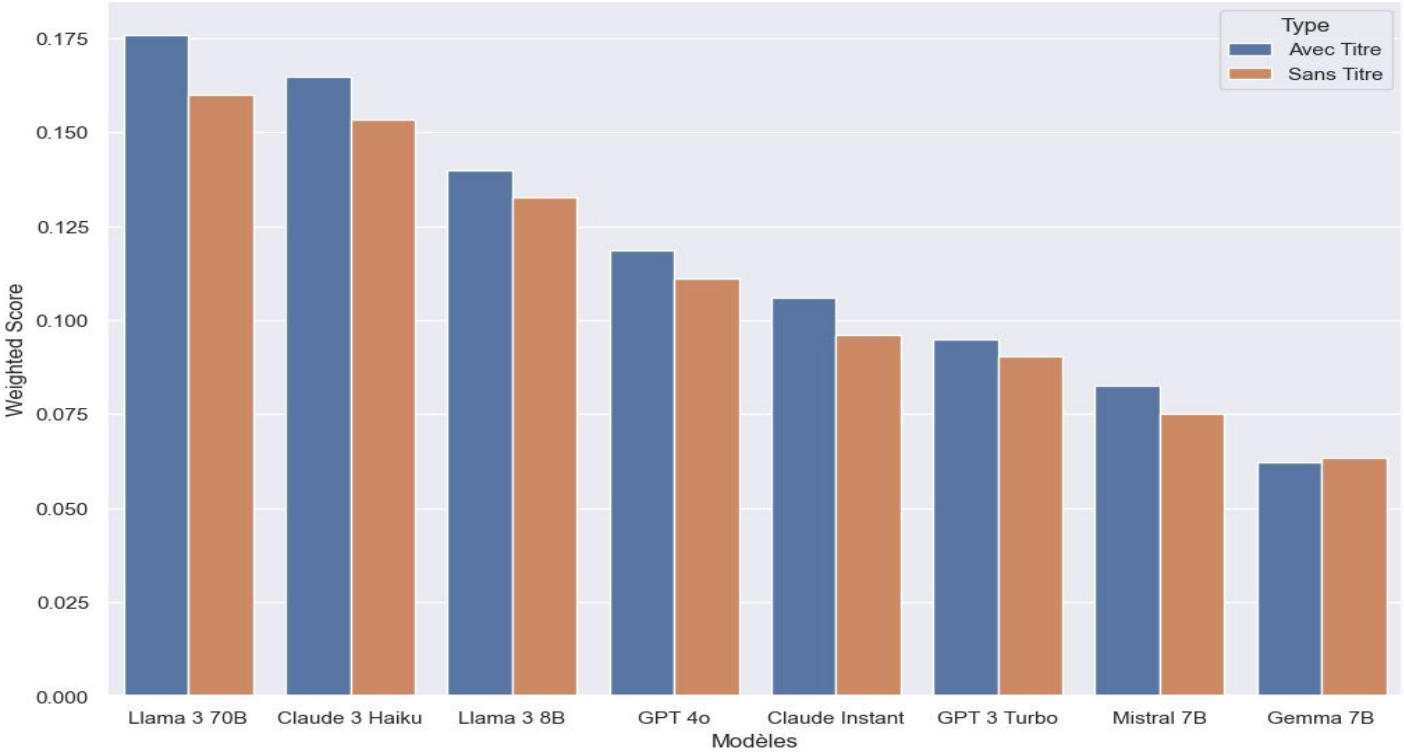
$$\mathrm{TES} = \frac{(1+\alpha) \times F_1 \times \mathrm{Prix}}{\alpha \times \mathrm{Prix} + F_1}$$

avec α = 10 ici.

LLMs et coût par token



LLMs et coût par token (Token Efficiency Score)



Conclusion

- Apports:
  - Les LLMs surpassent les méthodes classiques en précision et pertinence, même en zero-shot.
  - La TES montre que les modèles les moins coûteux offrent le meilleur compromis qualité/coût.
  - L’intégration des titres améliore les scores F1 sans surcharger les tokens.
  - Code disponible: https://github.com/obtic-sorbonne/keywords

Conclusion

- Limites:
  - Fonctionnement des LLMs comme des "boîtes noires".
  - La sensibilité aux prompts entraîne des résultats instables et des ajustements fréquents.
  - Les prompts détaillés augmentent les coûts sans garantie d'amélioration significative.
  - Consistance du corpus HAL :
    - De nombreux mots-clés fournis par les auteurs ne figurent pas dans les résumés ou titres.
    - Mots-clés absents…
    - Auto-promotion, mise à jour, validation humaine?

Conclusion

- Perspectives:
  - Analyser les textes complets pour extraire des mots-clés plus riches.
  - Calibrer les prompts pour affiner les résultats sans alourdir les requêtes.
  - Fine-tuner un LLM sur un nouveau "Gold standard".

Merci ! des questions?