LLMs dans le domaine Juridique

PRESENTE PAR AMINATA THIOUNE

Plan

I- REVUE DES ARTICLES

II - LAW DATASETS

III - VALIDATIONS

IV - TESTS

I - REVUE DES ARTICLES

1.LawBench: Benchmarking Legal Knowledge of Large Language Models

2.Adapting Large Language Models via Reading Comprehension

3.SaulLM-7B: A pioneering Large Language Model for Law

4. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering

5. RAG vs Finetuning — Which Is the Best Tool to Boost Your LLM Application?

I - REVUE DES ARTICLES

1 . LawBench: Benchmarking Legal Knowledge of Large Language Models

- Publié le 28 Septembre 2023

- Evaluation des performances des grands modèles de langage (LLM) dans le domaine juridique

- Développement d'un benchmark, LawBench, pour évaluer les capacités des LLMs dans le domaine juridique.

- Définition de trois niveaux cognitifs juridiques pour évaluer les LLMs :
    - mémorisation des connaissances juridiques,
    - compréhension des textes juridiques,
    - application des connaissances juridiques.

I - REVUE DES ARTICLES

1. LawBench: Benchmarking Legal Knowledge of Large Language Models

- Jeu de données :

Cognitive Level	ID	Task	Data Source	Metric	Type
Legal Knowledge Memorization	1-1	Article Recitation	FLK	Rouge-L	Generation
	1-2	Knowledge Question Answering	JEC_QA	Accuracy	SLC
Legal Knowledge Understanding	2-1	Document Proofreading	CAIL2022	F0.5	Generation
	2-2	Dispute Focus Identification	LAIC2021	F1	MLC
	2-3	Marital Disputes Identification	AIStudio	F1	MLC
	2-4	Issue Topic Identification	CrimeKgAssitant	Accuracy	SLC
	2-5	Reading Comprehension	CAIL2019	rc-F1	Extraction
	2-6	Named-Entity Recognition	CAIL2022	soft-F1	Extraction
	2-7	Opinion Summarization	CAIL2021	Rouge-L	Generation
	2-8	Argument Mining	CAIL2022	Accuracy	SLC
	2-9	Event Detection	LEVEN	F1	MLC
	2-10	Trigger Word Extraction	LEVEN	soft-F1	Extraction
Legal Knowledge Applying	3-1	Fact-based Article Prediction	CAIL2018	F1	MLC
	3-2	Scene-based Article Prediction	LawGPT	Rouge-L	Generation
	3-3	Charge Prediction	CAIL2018	F1	MLC
	3-4	Prison Term Prediction w.o. Article	CAIL2018	nLog-distance	Regression
	3-5	Prison Term Prediction w. Article	CAIL2018	nLog-distance	Regression
	3-6	Case Analysis	JEC_QA	Accuracy	SLC
	3-7	Criminal Damages Calculation	LAIC2021	Accuracy	Regression
	3-8	Consultation	hualv.com	Rouge-L	Generation

I - REVUE DES ARTICLES

1. LawBench: Benchmarking Legal Knowledge of Large Language Models

- Jeu de données :
    - FLK - Droit financier
    - JEC_QA - Questions et réponses juridiques
    - CAIL2022 - Général (divers domaines juridiques)
    - LAIC2021 - Général (divers domaines juridiques)
    - AIStudio - Général (divers domaines juridiques)
    - CrimeKgAssitant - Droit criminel
    - CAIL2019 - Général (divers domaines juridiques)
    - LEVEN - Général (divers domaines juridiques)
    - CAIL2018 - Général (divers domaines juridiques)
    - LawGPT - Général (divers domaines juridiques)

I - REVUE DES ARTICLES

1. LawBench: Benchmarking Legal Knowledge of Large Language Models

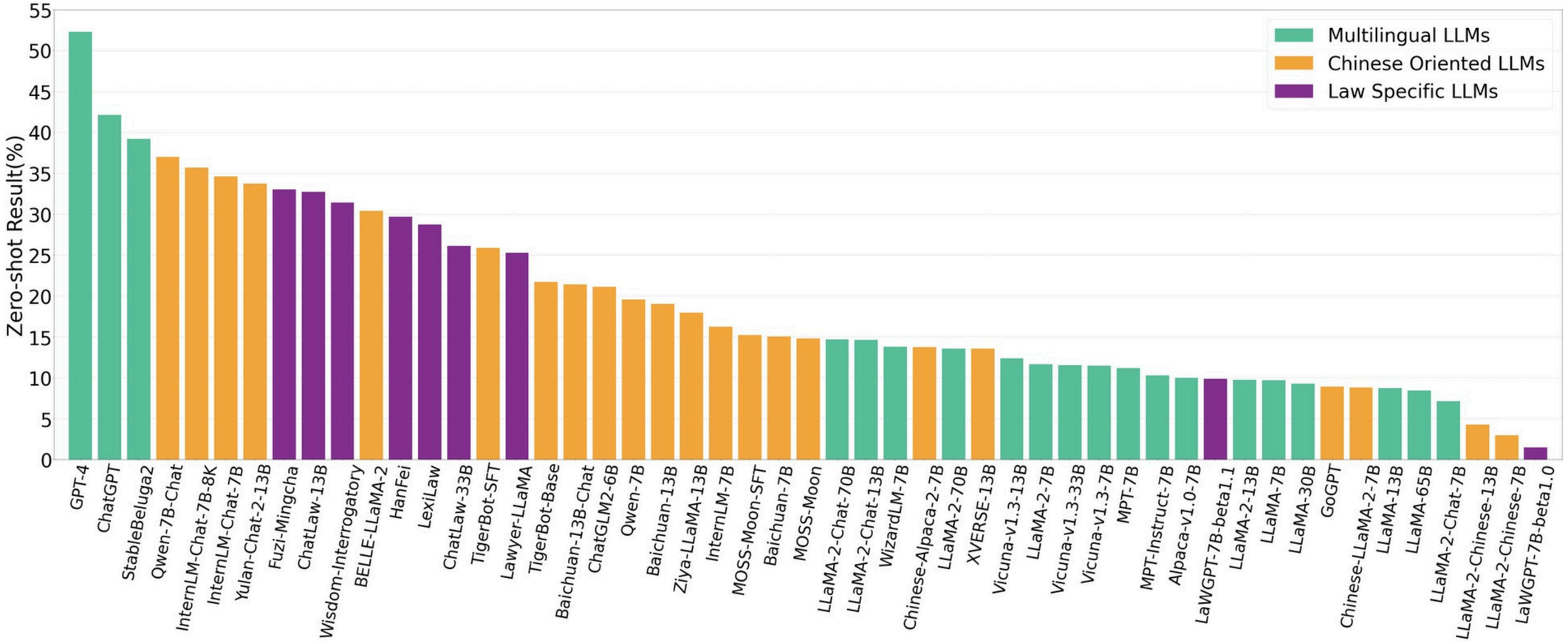
- Modèles comparés :

Model	Parameters	SFT	RLHF	Access	BaseModel
Multilingual LLMs					
MPT	7B	✗	✗	Weights	-
MPT-Instruct	7B	✓	✗	Weights	MPT-7B
LLaMA	7/13/30/65B	✗	✗	Weights	-
LLaMA-2	7/13/70B	✓	✗	Weights	-
LLaMA-2-Chat	7/13/70B	✓	✓	Weights	LLaMA-2-7/13/70B
Alpaca-v1.0	7B	✓	✗	Weights	LLaMA-7B
Vicuna-v1.3	7/13/33B	✓	✗	Weights	LLaMA-7/13/33B
WizardLM	7B	✓	✗	Weights	LLaMA-7B
StableBeluga2	70B	✓	✗	Weights	LLaMA-2-70B
ChatGPT	N/A	✓	✓	API	-
GPT-4	N/A	✓	✓	API	-
Chinese-oriented LLMs					
MOSS-Moon	16B	✗	✗	Weights	-
MOSS-Moon-SFT	16B	✓	✗	Weights	MOSS-Moon
TigerBot-Base	7B	✗	✗	Weights	-
TigerBot-SFT	7B	✓	✗	Weights	TigerBot-Base
GoGPT	7B	✓	✗	Weights	LLaMA-7B
ChatGLM2	6B	✓	✗	Weights	ChatGLM
Ziya-LLaMA	13B	✓	✓	Weights	LLaMA-13B
Baichuan	7/13B	✗	✗	Weights	-
Baichuan-13B-Chat	13B	✓	✗	Weights	Baichuan-13B
XVERSE	13B	✗	✗	Weights	-
InternLM	7B	✗	✗	Weights	-
InternLM-Chat	7B	✓	✗	Weights	InternLM-7B
InternLM-Chat-8K	7B	✓	✗	Weights	InternLM-7B
Qwen	7B	✗	✗	Weights	-
Qwen-Chat	7B	✓	✗	Weights	Qwen-7B
Yulan-Chat-2	13B	✓	✗	Weights	LLaMA-2-13B
BELLE-LLaMA-2	13B	✓	✗	Weights	LLaMA-2-13B
Chinese-LLaMA-2	7B	✓	✗	Weights	LLaMA-2-7B
Chinese-Alpaca-2	7B	✓	✗	Weights	LLaMA-2-7B
LLaMA-2-Chinese	7/13B	✓	✗	Weights	LLaMA-2-7/13B
Legal Specific LLMs					
HanFei	7B	✓	✗	Weights	HanFei
LaWGPT-7B-beta1.0	7B	✓	✗	Weights	Chinese-LLaMA
LaWGPT-7B-beta1.1	7B	✓	✗	Weights	Chinese-alpaca-plus-7B
LexiLaw	6B	✓	✗	Weights	ChatGLM-6B
Wisdom-Interrogatory	7B	✓	✗	Weights	Baichuan-7B
Fuzi-Mingcha	6B	✓	✗	Weights	ChatGLM-6B
Lawyer-LLaMA	13B	✓	✗	Weights	LLaMA
ChatLaw	13/33B	✓	✗	Weights	Ziya-LLaMA-13B/Anima-33B

I - REVUE DES ARTICLES

1 . LawBench: Benchmarking Legal Knowledge of Large Language Models

- Performance moyenne (zero-shot) de 51 LLMs évalués sur LawBench.



I - REVUE DES ARTICLES

2 . Adapting Large Language Models via Reading Comprehension

- Publié le 21 Février 2024,
- Modèle adapté en transformant des corpus bruts spécifiques à un domaine en textes de compréhension de lecture,
- Fine-tuning du modèle de base LLaMA-7B avec des instructions générales.

I - REVUE DES ARTICLES

2 . Adapting Large Language Models via Reading Comprehension

- Jeu de données
  - Biomedecine : PubMedQA, ChemProt, MQP, RCT, USMLE.
  - Finance : ConvFinQA, FPB, FiQA SA, Headline, NER.
  - Droit :
    - SCOTUS : Décisions de la Cour suprême des États-Unis, analyse et prédiction judiciaire.
    - CaseHOLD : Prédiction des décisions juridiques, analyse des documents juridiques.
    - UNFAIR-ToS : Identification des termes abusifs dans les accords de service, droits des consommateurs et conformité légale.

I - REVUE DES ARTICLES

2 . Adapting Large Language Models via Reading Comprehension

- performance des modèles sur des tâches spécifiques à des domaines particuliers (biomédecine, finance et droit) dans des évaluations de prompting

Biomedicine	PubMedQA	ChemProt	MQP	RCT	UMSLE	AVERAGE
LLaMA-7B	59.6	31.4	50.7	45.1	34.5	44.2
DAPT-7B	52.6	26.6	49.2	46.6	33.5	41.7
MedAlpaca-7B	58.6	**39.0**	50.7	40.8	**36.7**	45.1
AdaptLLM-7B	**63.3**	35.2	**54.4**	**50.4**	33.1	**47.3**
LLaMA-13B	59.6	42.8	49.3	**56.7**	34.7	48.6
DAPT-13B	51.1	38.0	49.0	50.9	34.6	44.7
MedAlpaca-13B	60.7	38.4	57.4	51.3	**41.2**	49.8
AdaptLLM-13B	**66.0**	**47.6**	**73.0**	50.4	34.0	**54.2**

Finance	ConvFinQA	FPB	FiQA SA	Headline	NER	AVERAGE
BloombergGPT-50B	43.4	51.1	75.1	82.2	60.8	62.5
LLaMA-7B	29.2	55.9	69.2	77.7	**61.1**	58.6
DAPT-7B	29.6	55.3	64.9	77.5	60.6	57.6
AdaptLLM-7B	**41.5**	**62.5**	**72.1**	**81.4**	59.3	**63.4**

Law	SCOTUS		CaseHOLD		UNFAIR-ToS	AVERAGE
	mic-F1	mac-F1	mic-F1	mac-F1		
GPT-J-6B	15.9	13.6	**34.9**	**34.9**	79.8	35.9
DAPT-6B	10.1	10.5	34.6	34.6	**84.9**	35.0
LexGPT-6B	16.9	7.7	27.0	27.0	81.9	32.1
AdaptLLM-6B	**18.8**	**20.1**	34.7	34.7	80.0	**37.7**
LLaMA-7B	28.3	10.8	32.9	32.9	65.8	34.2
DAPT-7B	25.0	9.8	34.2	34.2	72.0	35.0
AdaptLLM-7B	**30.0**	**17.8**	**35.1**	**35.1**	**74.4**	**38.5**

I - REVUE DES ARTICLES

2 . Adapting Large Language Models via Reading Comprehension

- Impact des configurations de données sur la performance de l'évaluation du prompting

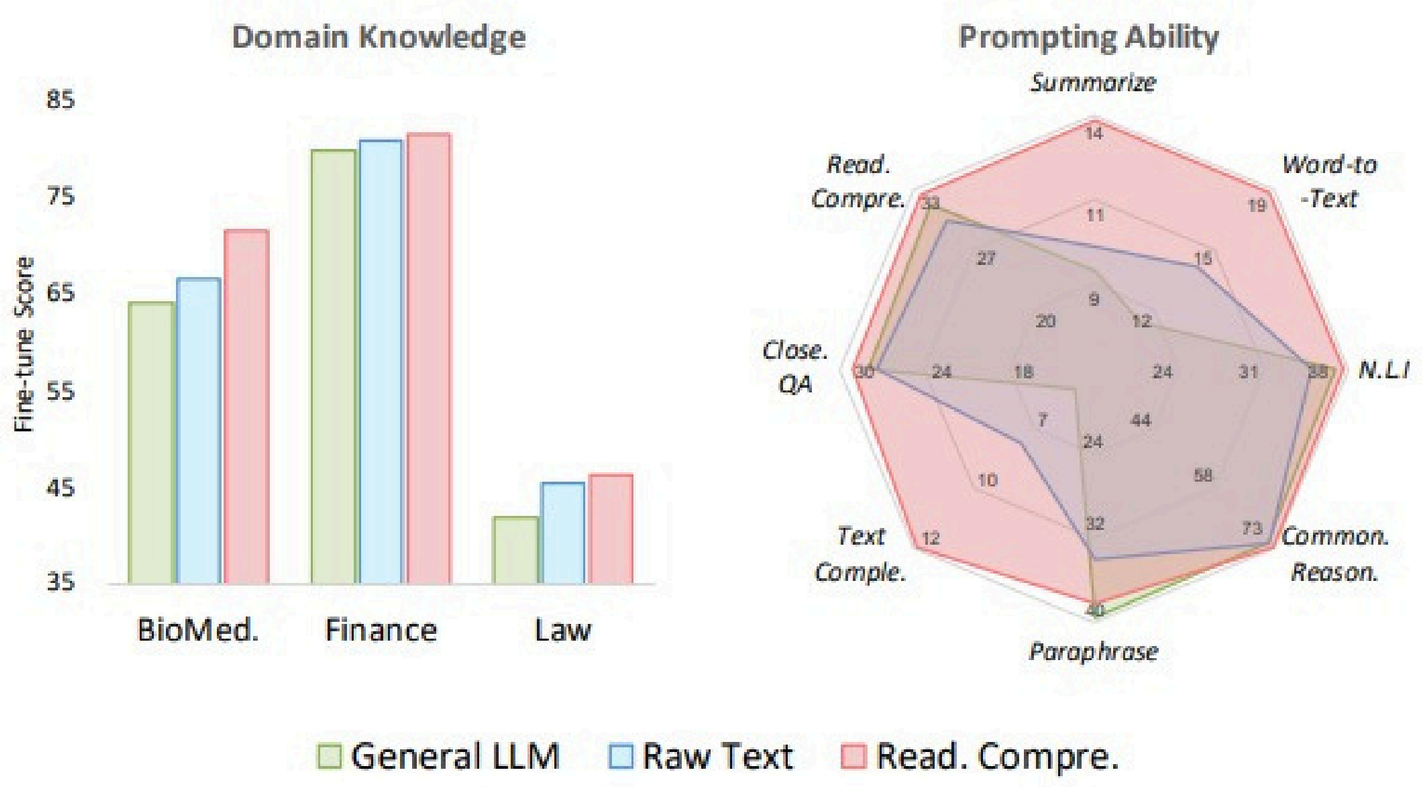
Data	Raw Text	Read. Compre.	Gen. Ins.	Raw. + Gen. Ins.	Read. + Gen. Ins.
BioMed.	41.7	44.3	43.3	44.8	**47.3**
Finance	57.6	60.0	62.2	61.7	**63.4**
Law	35.0	37.0	37.8	34.7	**38.5**

  - Raw Text : Corpus bruts spécifiques au domaine pour l'entraînement.
  - Read. Compre. : Corpus bruts convertis en textes de compréhension de lecture
  - Gen. Ins. : Instructions générales
  - Raw. + Gen. Ins. et Read. + Gen. Ins. : Combinaisons variées de corpus bruts et d'instructions générales.

I - REVUE DES ARTICLES

2 . Adapting Large Language Models via Reading Comprehension

- Évaluation du fine-tuning et du prompting sur des tâches spécifiques au domaine et générales



I - REVUE DES ARTICLES

3. SaulLM-7B: A pioneering Large Language Model for Law

- Publié le 07 Mars 2024
- Modèle décodeur spécifique au domaine juridique
- Fine-tuning avec l'architecture Mistral 7B

I - REVUE DES ARTICLES

3. SaulLM-7B: A pioneering Large Language Model for Law

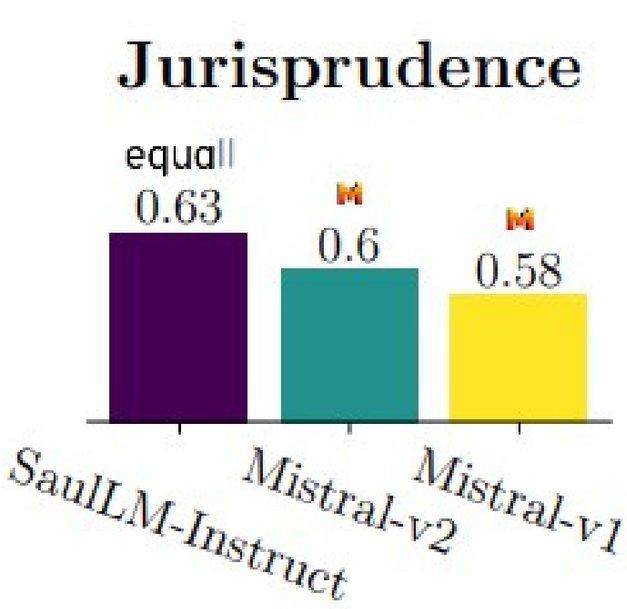
- Jeu de données :

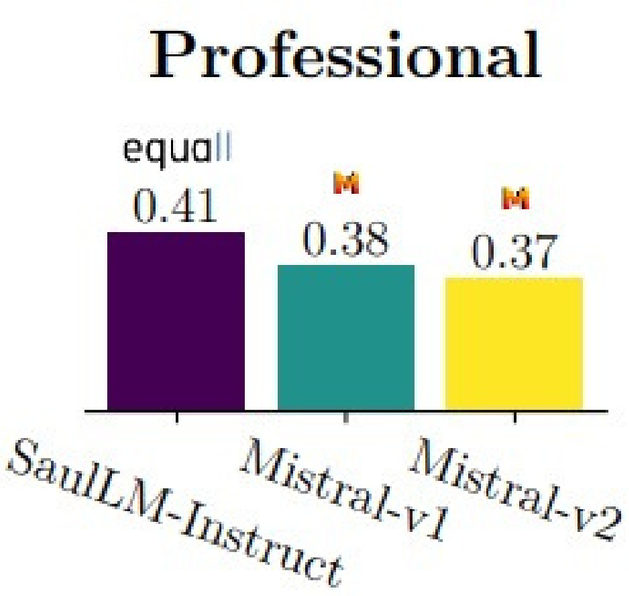
Name	Tokens	
FreeLaw[4]	15B	Juridique
EDGAR[5]	5B	Juridique (informations financières)
English MultiLegal Pile[6]	50B	Juridique
English EuroParl (Koehn, 2005)	6B	Juridique (législation européenne)
GovInfo[7] Statutes, Opinions & Codes	11B	Juridique (législation et avis gouvernementaux)
Law Stack Exchange[8]	19M	Juridique (forum de questions-réponses)
Commercial Open Australian Legal Corpus[9]	0.5B	Juridique (législation commerciale)
EU Legislation[10]	315M	Juridique (législation européenne)
UK Legislation[11]	190M	Juridique (législation britannique)
Court Transcripts[12]	350M	Juridique (transcriptions de procès)
UPSTO[13]	4.7B	Juridique
Total	94B	

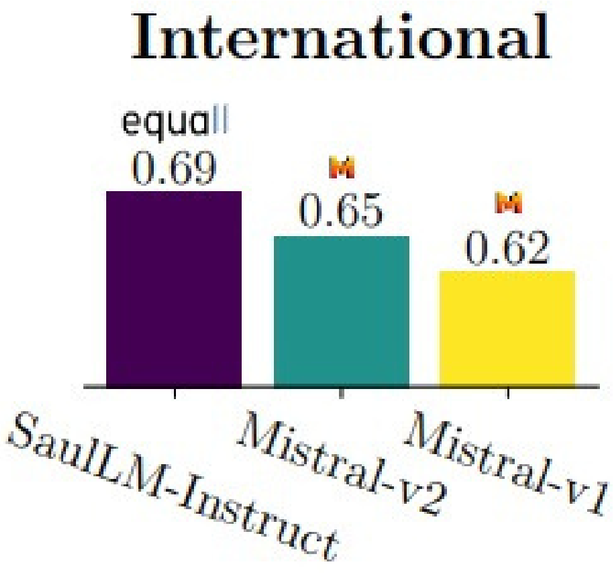
I - REVUE DES ARTICLES

3. SaulLM-7B: A pioneering Large Language Model for Law

- Evaluation des modèles sur Legal-MMLU



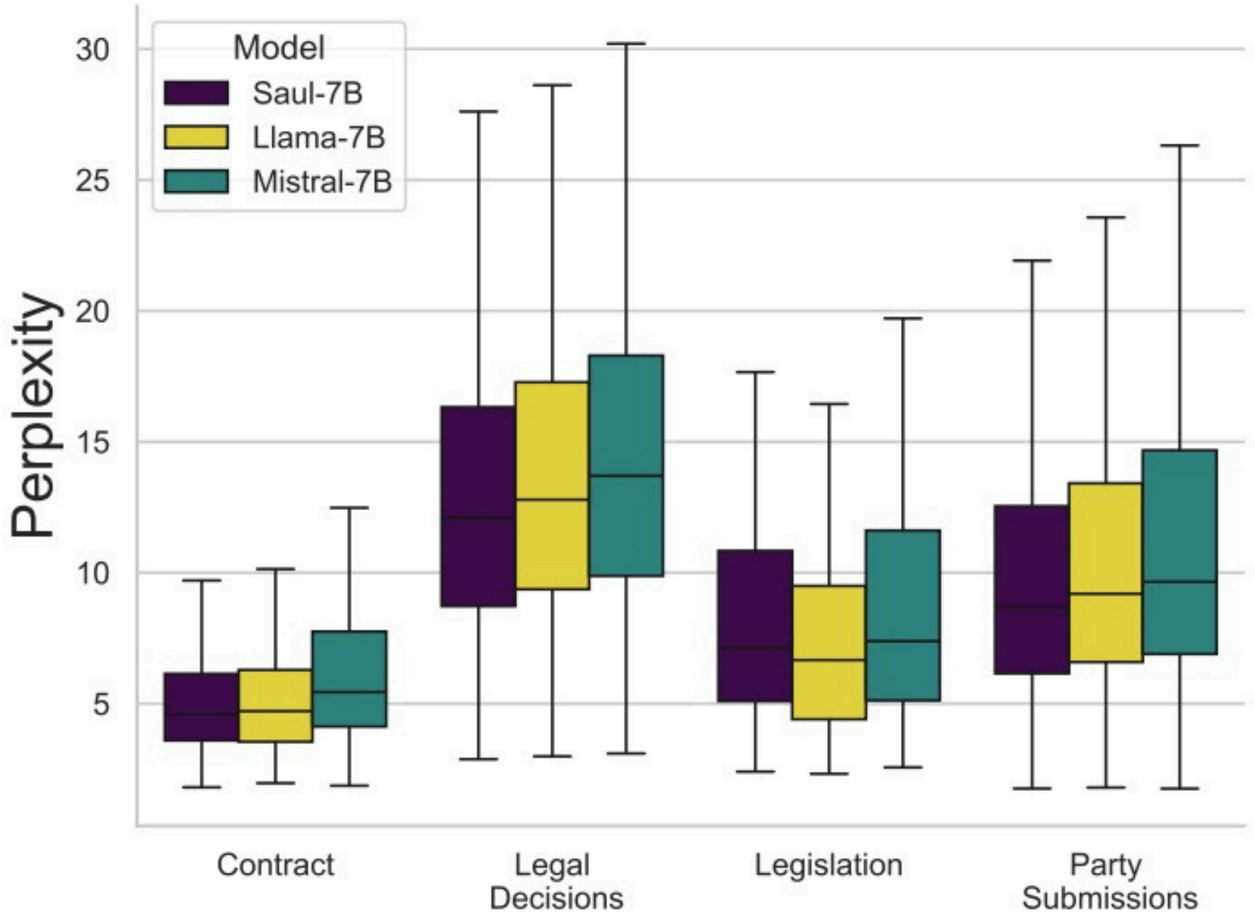




I - REVUE DES ARTICLES

3. SaulLM-7B: A pioneering Large Language Model for Law

- Analyse de la perplexité :



I - REVUE DES ARTICLES

4 .CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering

- Publié le 04 Avril 2024

- Présentation de CBR-RAG : intégration du Raisonnement Basé sur les Cas (CBR) avec l Génération Augmentée par la Recherche (RAG) pour optimiser les réponses juridique des LLMs.

- Utilisation des modèles BERT, AnglEBERT et LegalBERT pour générer des embeddings textuels spécialisés en droit, essentiels à la comparaison de cas dans CBR-RAG.

I - REVUE DES ARTICLES

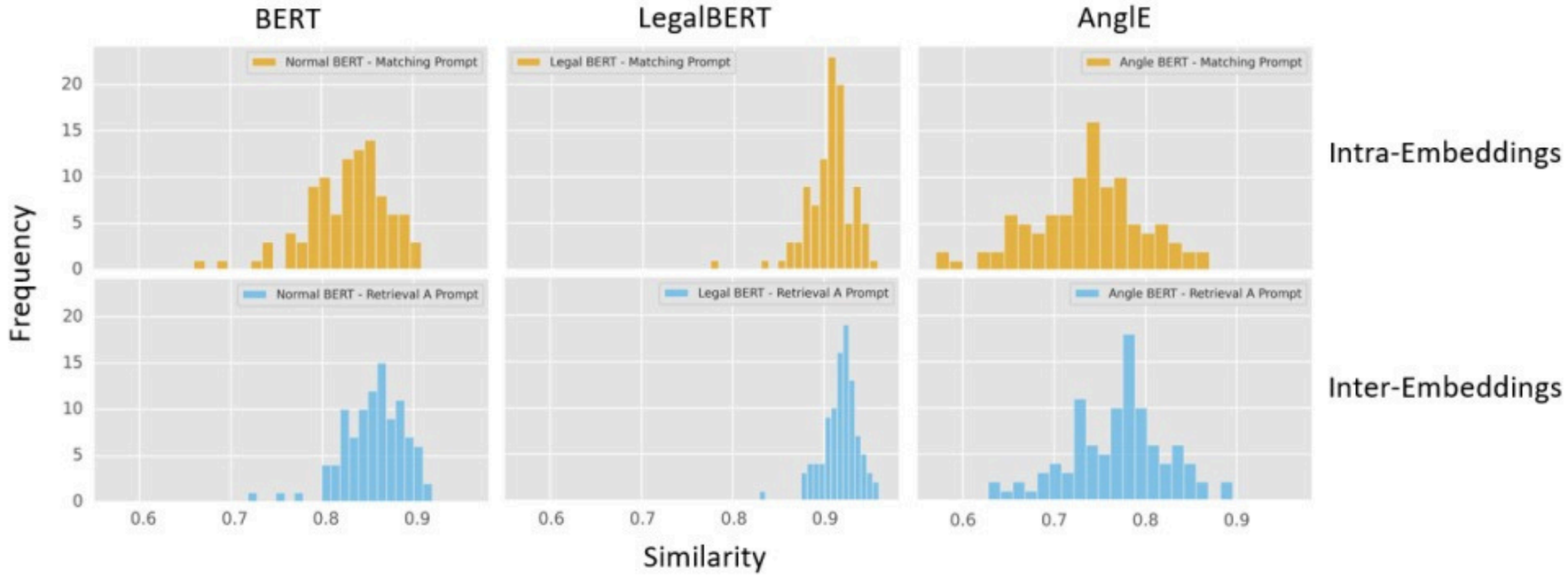
4 .CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering

- Jeu de données utilisé :
    - ALQA (Australian Legal Question-Answering) : est un jeu de données composé de 2 124 paires question-réponse générées par des LLM à partir du corpus Australian Open Legal Corpus.

I - REVUE DES ARTICLES

4 .CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering

- Distribution de la Similarité Cosinus pour les Embeddings Intra- et Inter- basés sur BERT, LegalBERT et AnglEBERT



I - REVUE DES ARTICLES

4 .CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering

- Evaluation des algorithmes hybrides en utilisant les scores de similarité cosinus dans différentes conditions de support de contexte et de nombre de contextes (k)

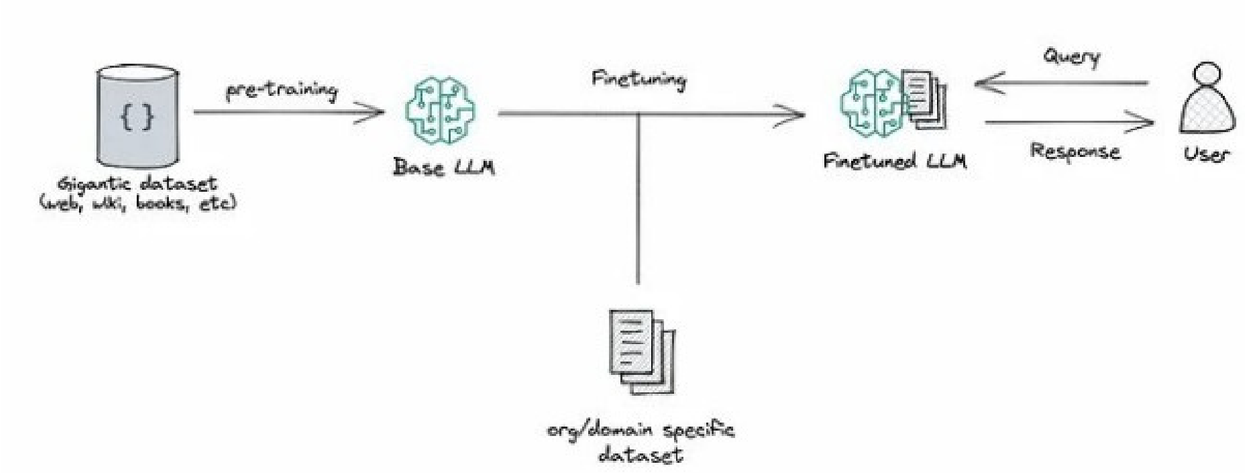
Table 4: Cosine scores for hybrid algorithms

		No Context	*Support*	*Full Case*
$k = 0$	No-RAG	0.8967		
$k = 1$	Hybrid BERT	-	0.8986	**0.9068**
	Hybrid LegalBERT	-	0.9020	0.9043
	Hybrid AnglEBERT	-	0.9121	0.9074
$k = 3$	Hybrid BERT		0.9007	0.8998
	Hybrid LegalBERT	-	0.9034	**0.9045**
	Hybrid AnglEBERT	-	0.9092	***0.9141**

I - REVUE DES ARTICLES

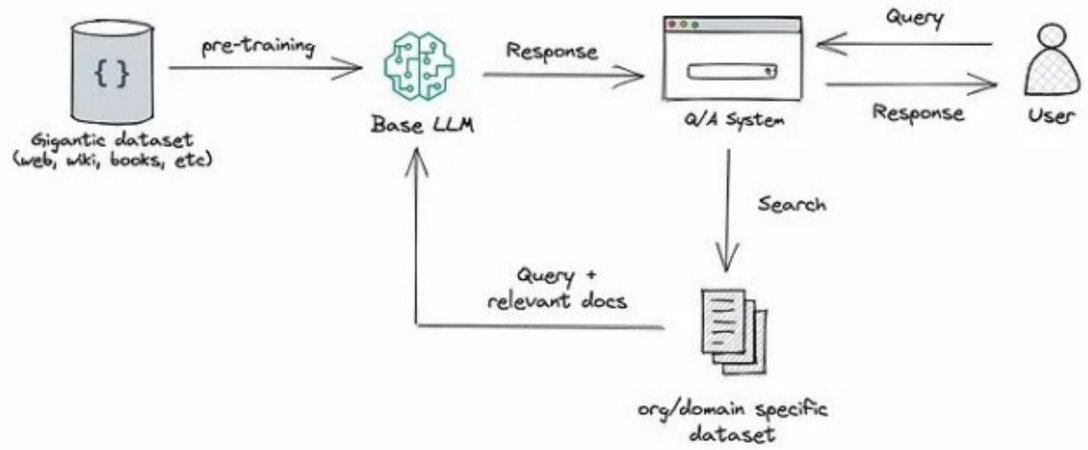
5 . RAG vs Finetuning — Which Is the Best Tool to Boost Your LLM Application?

Fine-tuning



vs

RAG



I - REVUE DES ARTICLES

5 . RAG vs Finetuning — Which Is the Best Tool to Boost Your LLM Application?

Dimensions	RAG	Fine-tuning
Besoin de connaissances externes	✓	✗
Adaptation du modèle	✗	✓
Minimisation des hallucinations	✓	✓
Disponibilité des données d'entraînement	✗	✗
Données dynamiques	✓	✗
Transparence/Interprétabilité	✓	✗

I - REVUE DES ARTICLES

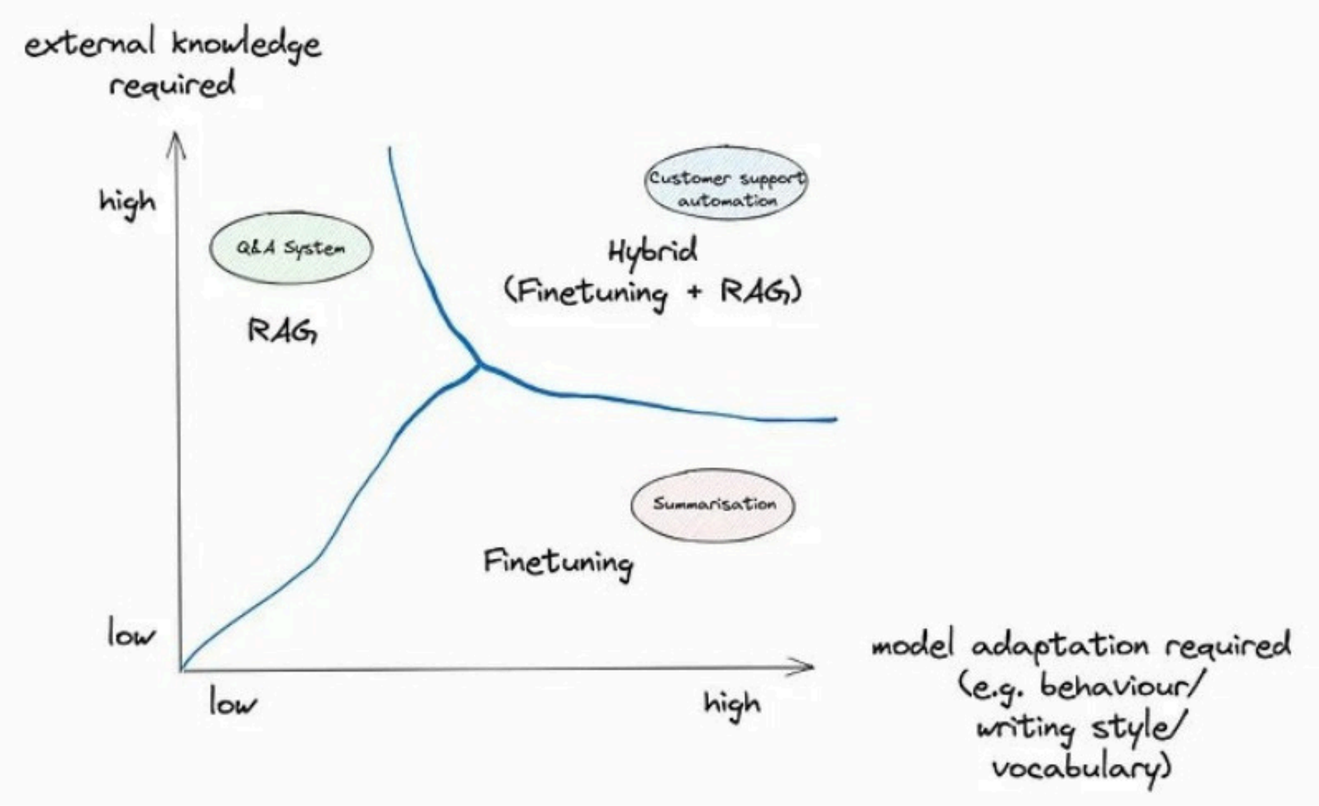
5 . RAG vs Finetuning — Which Is the Best Tool to Boost Your LLM Application?

- Cas d’utilisation :
    - Résumés : Fine-tuning pour le style, RAG ou hybride pour les données dynamiques.
    - Système de questions/réponses sur les connaissances organisationnelles : RAG pour l'accès dynamique, fine-tuning pour le style.
    - Automatisation de l'assistance à la clientèle (chatbots automatisés ou solutions d'assistance) : une approche hybride combinant fine-tuning et RAG est optimale pour offrir un support complet et cohérent.

I - REVUE DES ARTICLES

5 . RAG vs Finetuning — Which Is the Best Tool to Boost Your LLM Application?

- Fine-tuning + RAG



II - LAW DATASETS

- **COLD** (Collaborative Open Legal Data) contient 800,000 articles incluant des textes de lois, décisions de justice, règlements et articles doctrinaux de diverses juridictions.
- **EUR-LEX** : base de données de législation de l'UE offrant un accès aux textes juridiques et aux décisions des institutions européennes.
- **LEDGAR** : système de gestion de la réglementation conçu pour suivre et gérer les réglementations et législations.
- **UE Legislation** : compilation de textes législatifs et directives adoptés par les institutions de l'Union européenne.

II - LAW DATASETS

- **UNFAIR-ToS** : jeu de données sur les termes de service (ToS) potentiellement injustes dans les contrats.
- **CEDH** : décisions et jurisprudences de la Cour européenne des droits de l'homme en droit international des droits de l'homme.
- **Pile of Law** : collection de divers documents juridiques provenant de différentes juridictions et domaines du droit.
- **LegiFrance** : plateforme de diffusion officielle du droit français, incluant lois, décrets et ordonnances.

III - VALIDATIONS

- **DeepEval :**
  - **G-Eval** :évaluation globale de la qualité et de la performance d'un modèle de langage.
  - **Summarization** : capacité à condenser un texte long en une version plus courte tout en conservant les informations essentielles.
  - **Hallucination** : présence d'informations incorrectes ou inventées dans les réponses générées par le modèle.
  - **Faithfulness** : exactitude et fidélité des réponses par rapport au texte source.
  - **Contextual Relevancy** : pertinence des réponses par rapport au contexte fourni.

III -VALIDATIONS

- **Answer Relevancy :** pertinence et adéquation des réponses à une question posée.
- **Contextual Recall :** capacité à récupérer et utiliser efficacement les informations pertinentes du contexte.
- **Contextual Precision** : précision des informations extraites et utilisées du contexte.
- **RAGAS** : cadre d'évaluation basé sur la robustesse, la précision, la généralisation, l'adaptabilité et la scalabilité d'un modèle.
- **Bias** : évalue des préjugés ou partialités présents dans les réponses du modèle.
- **Toxicity** : mesure de la présence de langage offensant ou inapproprié dans les réponses générées.

III -VALIDATIONS

- Openai Evals : un outil qui permet de créer, exécuter et analyser des évaluations de modèles de langage en utilisant l'API d'OpenAI pour mesurer leur performance et qualité

- Exact match : une mesure qui évalue la proportion de réponses générées par un modèle correspondant exactement aux réponses de référence

Références :

- Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z., Ge, J. (2023). "Lawbench: Benchmarking legal knowledge of large language models." arXiv preprint arXiv:2309.16289.
- Cheng, D., Huang, S., Wei, F. (2023). "Adapting large language models via reading comprehension." arXiv preprint arXiv:2309.09530.
- Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., Morgado, S., et al. (2024). "Saullm-7b: A pioneering large language model for law." arXiv preprint arXiv:2403.03883.
- Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., Weerasinghe, R., Liret, A., Fleisch, B. (2024). "CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering." arXiv preprint arXiv:2404.04302.
- https://utfs.io/f/4c369a70-d163-422f-9bdd-75202ef456c2-96opok.pdf