

# Présentation pour l'audition au poste PR 251555

Sciences des Données et Intelligence Artificielle



par Julien VELCIN, PR en Informatique  
Université Lumière Lyon 2, Laboratoire ERIC



# Plan de la présentation

- Parcours antérieur
  - Vue d'ensemble de ma carrière
  - Parcours antérieur : enseignement
  - Parcours antérieur : recherche
  - Prises de responsabilités collectives
- Projet d'intégration
  - Positionnement général
  - Projet enseignement
  - Projet recherche
  - Intégration ECL / LIRIS

# Parcours antérieur

Julien Velcin, candidat au poste PR ECL-LIRIS

# Vue d'ensemble

	Parcours	Thématiques	Faits notables
1998	<ul style="list-style-type: none"><li>• Licence math puis informatique</li><li>• Maîtrise informatique</li></ul>		
2002	<ul style="list-style-type: none"><li>• DEA IARFA (Intelligence Artificielle)</li><li>• Thèse en Intelligence Artificielle</li><li>• ATER informatique</li></ul>	IA généraliste (machine learning, agents, computer vision, ingénierie des connaissances...) Clustering, optim. multicritère Topic models	
2005			
2007	<ul style="list-style-type: none"><li>• MCF informatique</li></ul> 	Text mining Social media analysis Fouille d'opinion Modèles temporels Modélisation bayésienne	Responsable M2 ECD (7 ans)
2015	<ul style="list-style-type: none"><li>• MCF HDR « Contributions à la science des données : Fouille de données textuelles appliquée à l'analyse des médias sociaux »</li></ul>	Word/doc embedding Apprentissage de représentations	Coordinateur ANR ImagiWeb Directeur de l'équipe DMD (3 ans) Délégation CNRS au LIRMM
2018	<ul style="list-style-type: none"><li>• PU informatique (2<sup>ème</sup> classe)</li></ul>	Modèles Transformers GNNs LLMs	Responsable L3 IDS (4 ans) Coordinateur du pôle HuNIS (4 ans) Coordinateur Master informatique Responsable M1 (depuis 3 ans) Dir. adjoint de l'ICOM (depuis 1 an) Président d'un comité ANR
2023	<ul style="list-style-type: none"><li>• PU informatique (1<sup>ère</sup> classe)</li></ul>		

# Activité d'enseignement

Type de cours	Aperçu du contenu	Public	Profil du poste
<b>Intelligence artificielle</b>	Notions fondamentales et historique, logique des propositions, résolution de problèmes, approches à base de règles, introduction au machine learning	M1 Info (2011-16), L3 MIASHS IDS (2016-2023), L2 Info (2022-aj.)	Intelligence Artificielle
<b>Traitement automatique de la langue</b>	Approches quantitatives, représentation vectorielle (creuse, dense), recherche d'information, modèles de langue, machine learning, LLMs	M2 ECD puis DM puis MALIA (2011-2022), M1 DMKM (2010-2016), M2 HN (2018-aj.), DU Big Data (2014-2021), ED InfoMath (2016-2018)	Science des données, IA
<b>Programmation</b>	Programmation orientée objet, Java puis Python, un peu de C++	M1 Info (2007-aj.), M1 IDSM (2007-2024), L3 Info (2022-2023)	Programmation orientée objet
<b>Deep learning</b>	Notions fondamentales, principales architectures, éléments d'optimisation	M2 DM puis MALIA (2016-aj.), UdL (2020-2024)	Science des données, IA
<b>Analyse des réseaux d'information</b>	Rappel de théorie des graphes, techniques d'analyse des graphes (ex. clustering), combiner avec l'information textuelle, deep learning pour les graphes	M2 MALIA et MIASHS (2022-aj.)	Science des données, IA
<b>Projets</b>	Suivi de projets, stages, alternants	L3 Info, M1 Info, M2 MALIA	

# Activité recherche : problématiques 1/2

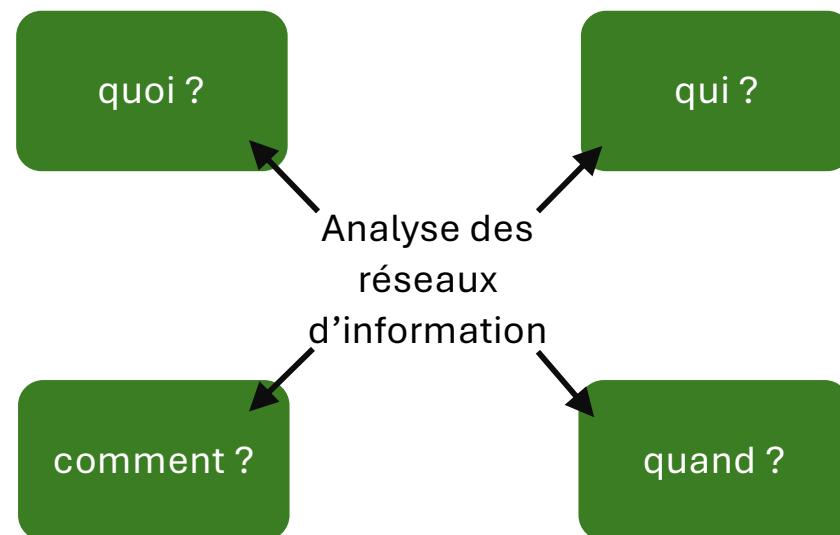
## Modèles thématiques

(thèses de M.A. Rizou, M. Dermouche\*, C. Christophe\*), **analyse des chemins et interaction de l'information**

(thèses de C.H. Despointes\*, G. Poux-Médard)

## Analyse d'opinion

(thèses de A. Stavrianou, M. Dermouche\*)



**Analyse des rôles** (thèses de M. Forestier, A. Lumbreras\*), **représentation d'auteurs et recherche d'experts**

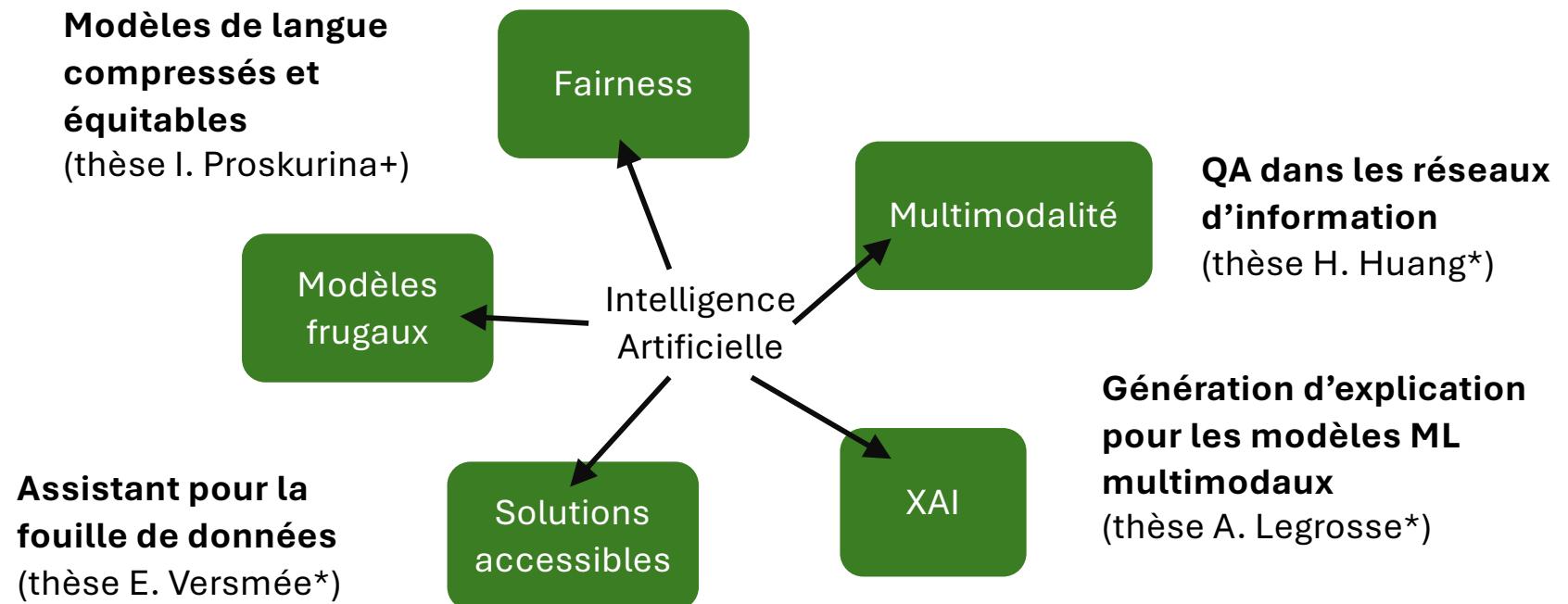
(thèses de R. Brochier\*, A. Gourru, E. Terreau+)

## Développement de modèles temporels

(thèses de M. Dermouche\*, C. Christophe, A. Gourru, E. Terreau+)

Projets : ANR ImagiWeb, ANR LIFRANUM (+), thèses CIFRE (\*)

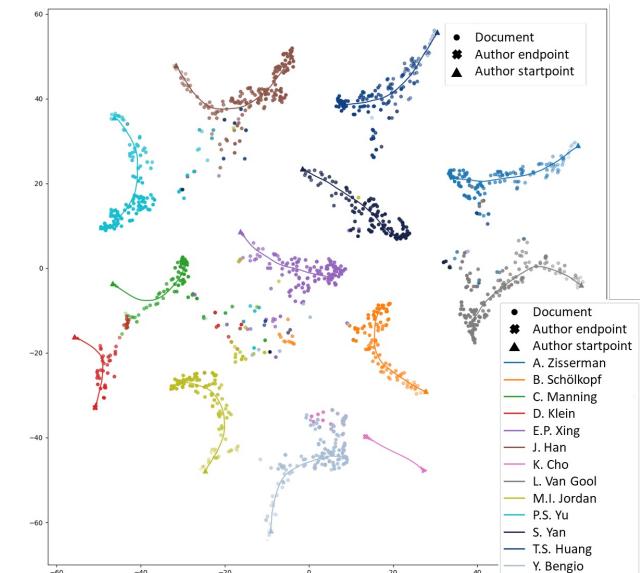
# Activité recherche : problématiques 2/2



Projets : ANR DIKé (+), thèses CIFRE (\*)

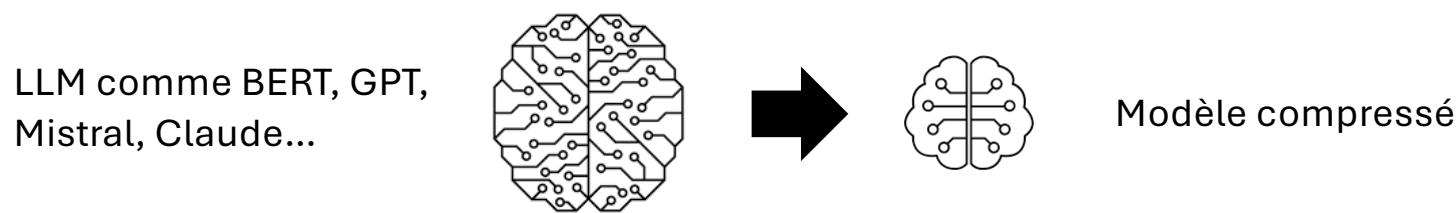
# Exemple (1) : représentation d'auteurs

- Problématique : construire des espaces de représentations latentes pour représenter les auteurs de textes
- Plusieurs thèses : R. brochier (CIFRE), A. Gourru (bourse ministérielle), E. Terreau (ANR LIFRANUM)
- Quelques contributions :
  - représentations de nœuds dans des corpus structurés (WWW 2019)
  - modèles temporels de trajectoire d'auteurs (IJCAI 2022, IDA 2024)



## Exemple (2) : biais dans les LLMs 1/2

- Cadre du projet ANR DIKé et thèse d'I. Proskurina
- Compression des modèles de langue :



- Techniques de compression : pruning, distillation, quantization (arrondi sur l'encodage des paramètres)
- Les modèles compressés sont susceptibles d'exacerber les biais déjà présents dans les données et les modèles initiaux

## Exemple (2) : biais dans les LLMs 2/2

- Dans ce contexte, nous avons :
  - expérimenté le pruning avec des modèles un peu anciens de type encodeur et proposé une solution pour débiaiser qui prend en compte une annotation fine sur les mots qui comptent réellement (IDA 2023) ;
  - évalué l'impact de la quantization sur la calibration de LLMs récents et mis en évidence certains phénomènes intéressants (NAACL 2024) ;
  - modifié l'algorithme de quantization GPTQ afin de compresser des LLMs tout en diminuant les biais liés à l'équité de ces modèles (travail en cours).
- En parallèle, nous nous intéressons à l'encodage des valeurs morales dans les LLMs (NAACL 2025) et à la détection de la haine implicite (collaboration en cours avec Naver Labs)

# Projets et collaborations

Nom du projet	Type	Collaboration	Année
ImagiWeb	ANR	2 laboratoires académiques (LIA, CEPEL), 3 entreprises (AMI Software, EDF, XRCE)	2012-2015
Text mining pour l'étude des réseaux sociaux	inter <sup>ale</sup>	S. Trausan-Matu (UPB, Roumanie)	2009-2018
Interactive document clustering	inter <sup>ale</sup>	E. Milios (Dalhousie Univ., Canada)	2015-2021
Solutions « human in the loop » en machine learning	inter <sup>ale</sup>	I. Davidson (UCDavis, US)	2017-aj.uj.
LIFRANUM	ANR	MARGE, BnF	2020-2024
TIGA	PIA	Métropole, UrbaLyon	2020-2025
DIKé	ANR	LabHC, Naver Labs	2022-2025
Trévoux	local	ICAR, LIRIS	2022-aj.uj.

+ 8 thèses CIFRE (dont 5 soutenues) : AMI Software, Technicolor, EDF, DSRT, Worldline, Infologics

# Publications (résumé)

Portée internationale		Portée nationale	
Revues internationales	15 <sup>1</sup>	Revues nationales	4
Conférences internationales	47 <sup>2</sup>	Conférences nationales	13
Ateliers internationaux	12	Ateliers nationaux	5

<sup>1</sup> dont 9 classées Q1 ou Q2 dans Scimago

<sup>2</sup> dont 23 classées A ou A\* dans le CORE

- When Quantization Affects Confidence of Large Language Models? I. Proskurina, L. Brun, G. Metzler, J. Velcin. North American Chapter of the Association for Computational Linguistics (Findings of the **NAACL**), 2024. CORE A
- Dynamic Mixed Membership Stochastic Block Model for Weighted Labeled Networks. G. Poux-Médard, J. Velcin, S. Loudcher. International ACM SIGIR Conference (**SIGIR**), 2023. CORE A\*
- Dynamic Gaussian Embedding of Authors. A. Gourru, J. Velcin, C. Gravier, J. Jacques. The Web Conference (formerly known as **WWW**), 2022. CORE A\*
- Serialized Interacting Mixed Membership Stochastic Block Model. G. Poux-Médard, J. Velcin, S. Loudcher. IEEE International Conference on Data Mining (**ICDM**), 2022. CORE A\*
- Monitoring geometrical properties of word embeddings for detecting the emergence of new topics. C. Christophe, J. Velcin, J. Cugliari, P. Suignard, M. Boumghar. Conference on Empirical Methods in Natural Language (**EMNLP**), 2021. CORE A\*

# Responsabilités collectives

- **Gestion des formations**

- Responsable pédagogique M2 ECD (7 ans), L3 IDS (4 ans), M1 Informatique (depuis 2022)
- Coordination de la mention de Master Informatique (depuis 2022)

- **Implication dans la composante de formation**

- Élu au conseil de l'UFR FSEG (4 ans) puis au CA de l'ICOM (depuis 3 ans)
- Directeur adjoint de l'ICOM (depuis 1 an)

- **Implication dans le laboratoire**

- Responsable de l'équipe DMD (3 ans entre 2012 et 2015 puis 1 an en 2020)

- **Implication dans l'établissement**

- Président du GEI 26-27-61-71 (depuis 3 ans)
- Participation au dépôt du projet de cluster IA AILyS (2023) + dépôt du projet AMI CMA InART (2025)
- Co-animation du pôle de spécialité HuNIS (4 ans entre 2020 et 2024)
- Élu à la CR (et donc au CAC) entre 2013 et 2018, puis depuis 2025
- Référent IA à Lyon 2 (situation en pause)

- **Au niveau national**

- Président d'un comité ANR (depuis septembre)

# Projet d'intégration

Julien Velcin, candidat au poste PR ECL-LIRIS

# Mon positionnement général

- Développement de solutions d'**IA performantes et éthiques** :
  - Bien fondées et efficaces  
modèles bien définis, sachant prendre en compte des notions d'incertitude, implémentés dans des algorithmes efficaces
  - Accessibles  
à la portée de toutes et tous, qui s'adapte aux usages (cf. travail pluridisciplinaire), implémentées dans des solutions simples d'accès et ouvertes
  - De confiance  
robustes aux biais (cf. travail sur la fairness ou sur la calibration des modèles), dont on peu expliquer le fonctionnement et/ou les sorties (lien XAI)
  - Frugales  
solutions peu coûteuses, légères (cf. travaux sur la compression), utilisées à bon escient (cf. travaux en lien avec le méta-learning)

# Projet d'enseignement (ECL)

- Implication dans l'offre de formation :
  - Algo / programmation Python / EDA / machine learning (BS Data Science for Responsible Business)
  - Algo / programmation Python / programmation Web (cursus ingénieur, A1)
  - Data Science (cursus ingénieur, A3)
- Opportunités de nouveaux enseignements :
  - Découverte de l'IA (cursus ingénieur, A2)
  - Analyse des réseaux d'information (MOD ou MSO au S9)
  - Résumé automatique de corpus (PE, 1<sup>ère</sup> année; PaR, 2<sup>ème</sup> année)
- Quelques défis
  - Susciter l'intérêt et maintenir la motivation des élèves
    - choix des sujets (originalité, défi, lien avec la société), utilisation d'un projet fil rouge
  - Adapter nos modalités d'enseignement et d'évaluation aux nouveaux outils d'IA
    - codage assisté, modifier ce qu'on évalue, importance des tests, évaluations courtes
  - Éveiller la curiosité pour les recherches menées au laboratoire
    - cas d'étude / projets en lien avec les activités du labo, partage de problématiques, stages « blancs » ouverts aux propositions

# Projet recherche (LIRIS / Imagine)

Développer des solutions d'IA performantes et éthiques pour :

- modèles de machine learning multimodaux
  - combiner le traitement du texte et l'image pour la manipulation en robotique
  - développer des modèles compressés pouvant être embarqués
  - exploration de collections de documents complexes
  - prise en compte de l'information géométrique dans les images
- intégration de l'humain, interactions avec le langage naturel
  - modèles de langue pour le traitement du mouvement
  - développement d'agents contrôlés par les utilisateurs
  - assistants conversationnels pour le traitement des données
  - intégrer des critères d'acceptabilité des solutions, en lien avec leur coût mais aussi des questions d'ordre éthique

# Intégration ECL / LIRIS

- Implication dans le département math-info d'ECL
  - Participation au travail collectif sur l'évaluation de nos pratiques pédagogiques
  - Participation au fonctionnement des formations
  - Prise de responsabilité à moyen terme en fonction des besoins
- Intégration au LIRIS
  - Intégration à l'équipe Imagine (E. Dellandrea, L. Chen, V. Eglin, L. Tougne...)
  - Collaboration avec les autres équipes du LIRIS (R. Vuillemot, R. Chalon, E. Lavoué, A. Tabard de SICAL ; L. Moncla et R. Cazabet de DM2L)
  - Prise de responsabilité à court / moyen terme en fonction des besoins
- Autres collaborations en ML et IA
  - Laboratoires associés à ECL, comme LMFA, INL, ICJ
- Développement stratégique de l'Intelligence Artificielle

# Synthèse du CV pour le poste PR 251555 - 0141

## • **Enseignement**

- **thématisques** : Intelligence artificielle, traitement automatique de la langue, programmation, machine learning (notamment deep learning), analyse des réseaux d'information
- **public** : niveaux variés (L, M, D, pro), informatique mais aussi formations pluridisciplinaires (M2 HN, Master MIASHS), étudiants internationaux (Master DMKM, Master IDSM)

## • **Recherche**

- **mots-clefs** : IA, machine learning, data science, NLP, social media analysis, digital humanities
- **encadrements** : 15 thèses (11 soutenues), 5 post-docs, stages de M2R
- **publications** : 62 internationales, dont 15 articles de revue (dont DMKD, ESWA) et 47 conférences internationales (dont IJCAI, SIGIR, WWW, EMNLP, NAACL, ECML-PKDD)
- **collaborations** : académiques nationales (LabHC, LIRIS, LIA, LIRMM...), industrielles (AMI Software, EDF, Technicolor, DSRT, Wordline, Infologics), internationales (UCDavis, USA)

## • **Responsabilités collectives**

- **Pédagogique** : responsable du M2 ECD (7 ans), L3 IDS (4 ans), M1 Informatique (depuis 2022)
- **Recherche** : coordinateur de projets recherche (ANR ImagiWeb), directeur de l'équipe DMD (3 ans)
- **Lyon 2** : élu à différents conseils (CR, CA de l'ICOM), directeur adjoint de l'ICOM (1 an), impliqué dans plusieurs initiatives liées à l'IA (pôle de spécialité HuNIS, projet AILyS, AMI CMA InART)
- **National** : Président d'un comité ANR

# **MATERIEL ADDITIONNEL**

# Exemple de cours

Julien Velcin, candidat au poste PR ECL-LIRIS

# Découverte de l'IA (cursus ingénieur, S8)

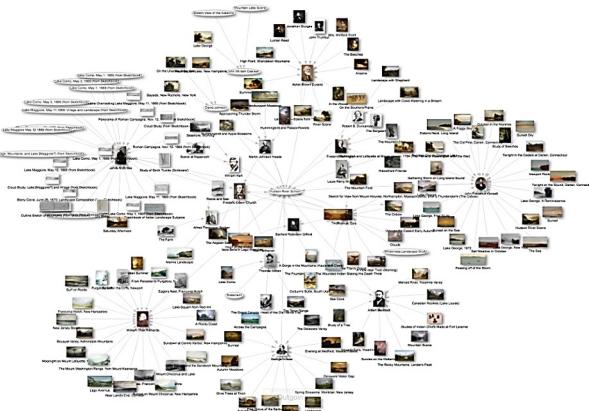
- **Objectif** : donner des éléments de culture générale sur l'IA mais aussi développer quelques réalisations en travaux pratiques
- **Déroulement des séances** :
  - En CM : introduction et historique, principaux concepts (quelles définitions ? distinction entre les systèmes à base de règles et le ML), fondement de logique (notion d'inférence), enjeux de l'IA aujourd'hui
  - En TP : chatbot, agent dans un labyrinthe
  - Exemple de projet : développement d'un jeu de société
- **Lien avec d'autres enseignements** : algorithmique (manipulation de graphes), RO, machine learning

# Représenter les proximités entre auteurs

Julien Velcin, candidat au poste PR ECL-LIRIS

# Représenter les proximités entre auteurs

- Deux exemples illustratifs :
  - [An ocean of books](#)
  - [Hudson River School](#) artists  
(explorer le [graphe sémantique](#))
- Ces méthodes emploient généralement la structure des données (par ex. les liens entre les pages Wikipedia)
- Comment faire en se basant sur le *contenu* textuel ?



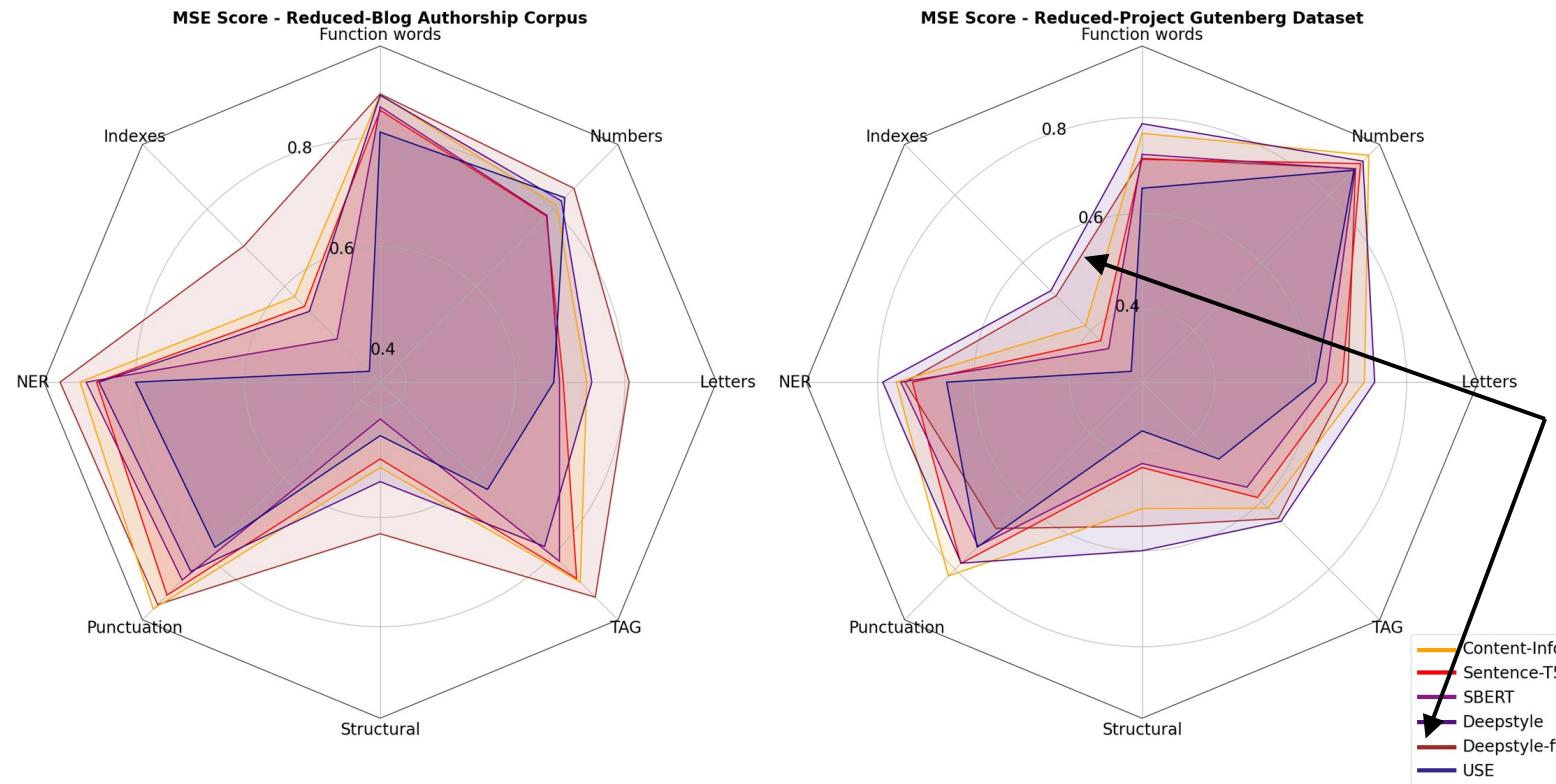
# Mesurer le style littéraire (Terreau et al., 2021)

- Suivant la littérature sur le sujet, nous nous basons sur **303 descripteurs stylistiques** :

Catégories	Exemples	Nombre de marqueurs
Lettres	Fréquences de lettre	26
Nombre	Fréquences de nombre	11
Structurel	Longueur moyenne des mots, Hapax Legomena, ...	9
Ponctuation	Fréquences des signes de ponctuation	36
Mots outils	Fréquences des mots outils (does, once, doing, ...)	153
Tag	Fréquences des POS-tag	43
Ner	Fréquences des entités nommées	18
Index	Index de lisibilité et de complexité	7

- On va évaluer à quel point les représentations apprises par les modèles *capturent* ces différentes mesures

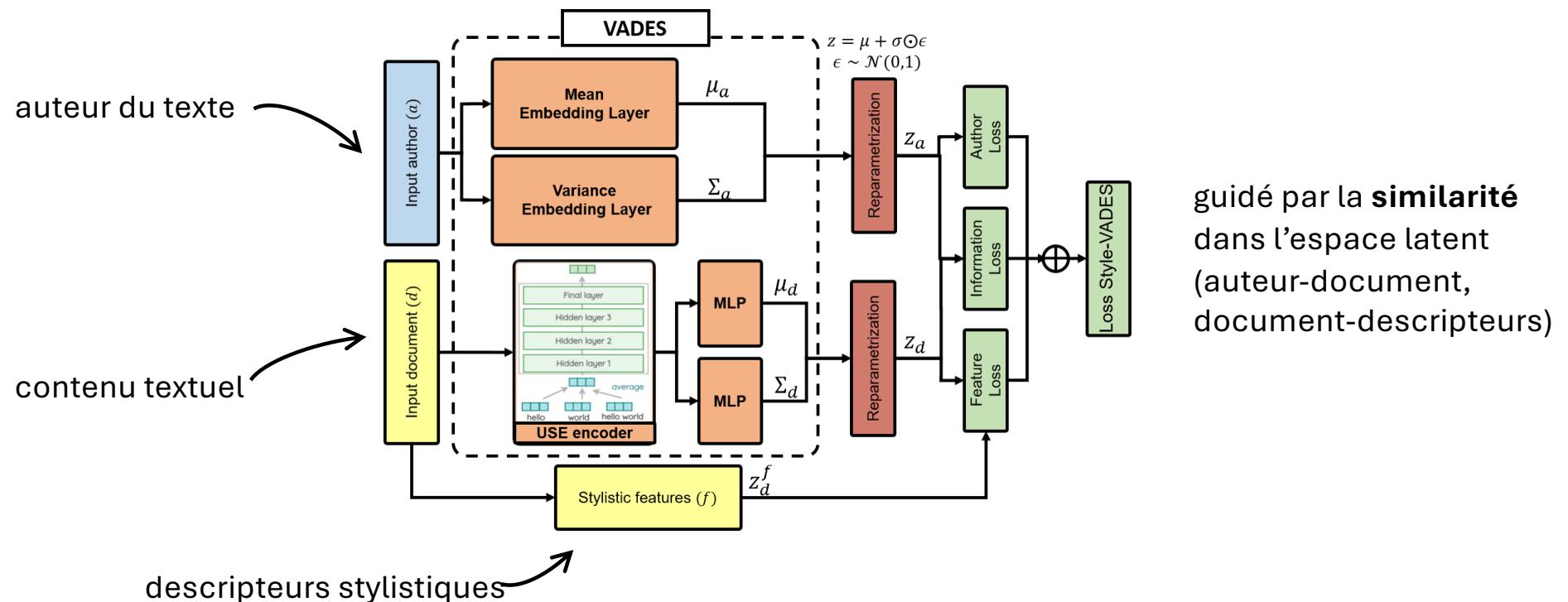
# Comparaison des modèles



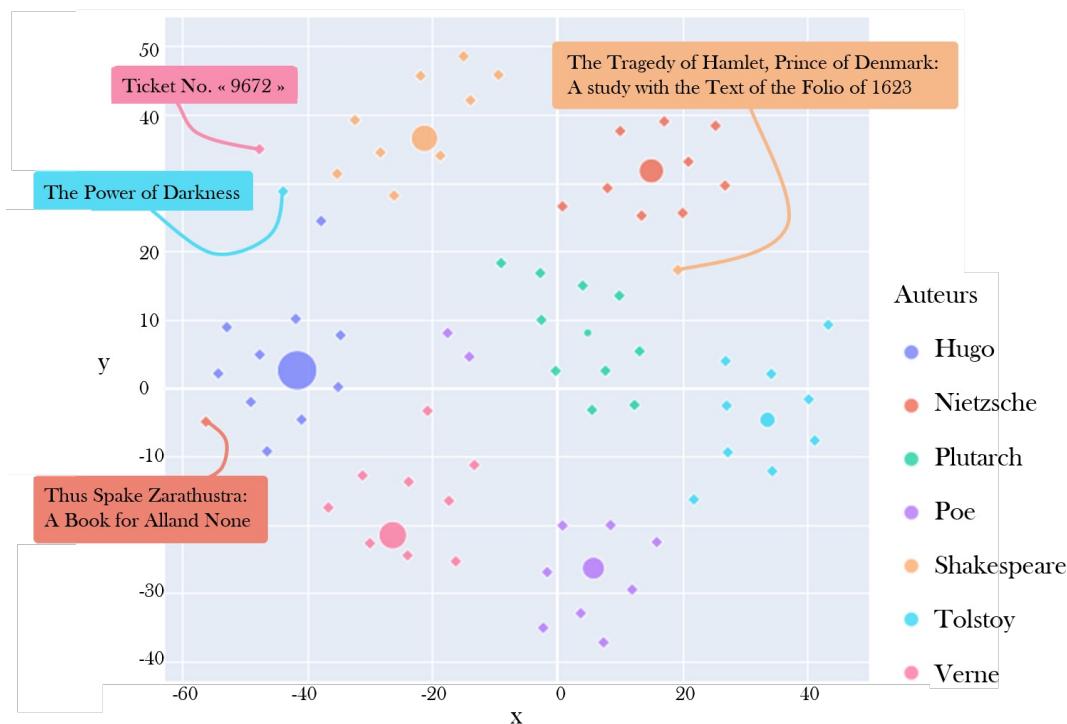
Deepstyle-ft est  
« affiné » sur la  
tâche d'attribution  
d'auteurs

# VADES : modèle de représentation des auteurs

(Terreau et al., arXiv 2024)

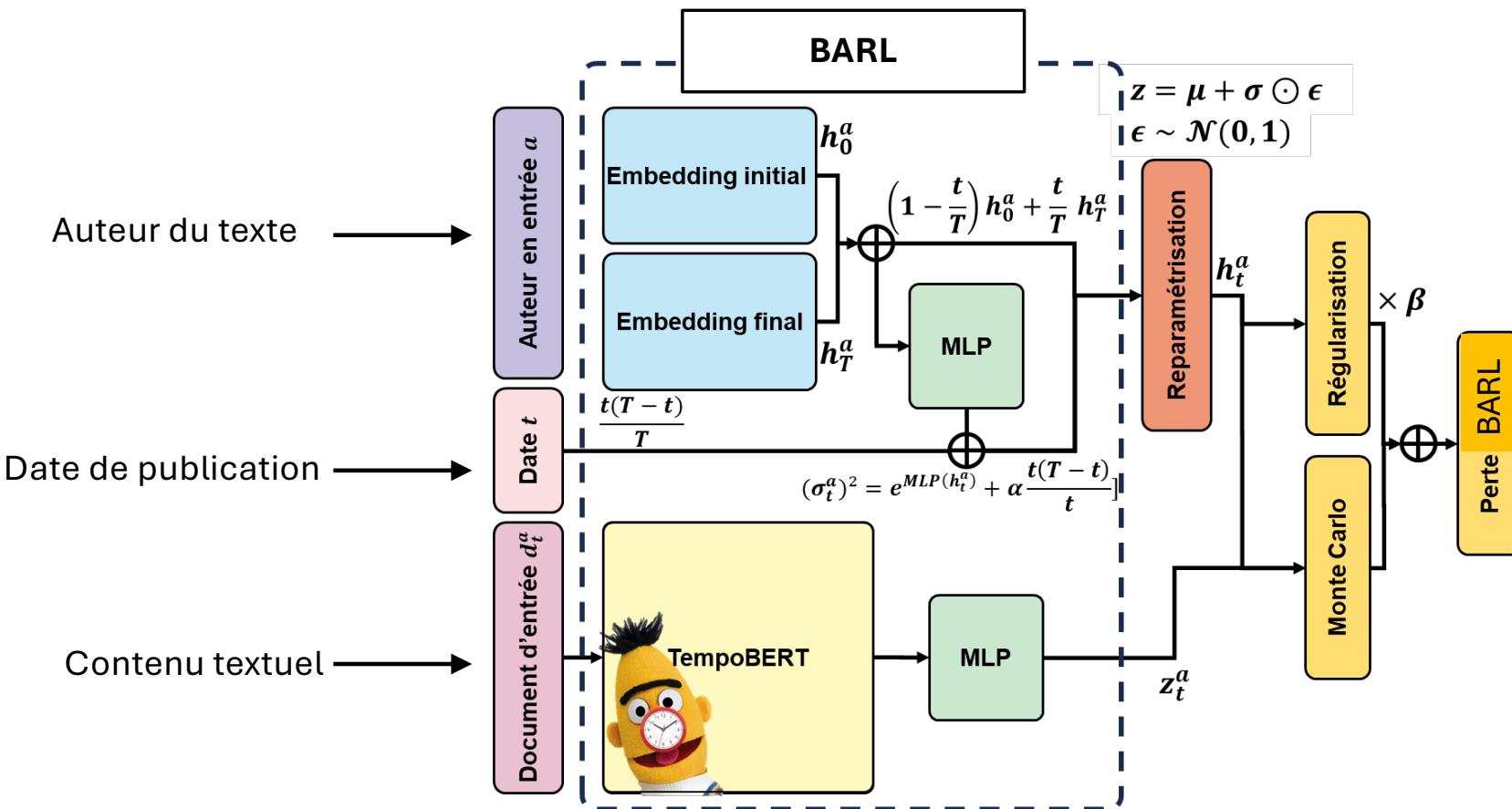


# Application à l'analyse du style littéraire



Ici, il s'agit d'un extrait de données sont tirées du [Projet Gutenberg](#)

# BARL : modèle pour apprendre des représentations temporelles (Terreau & Velcin, 2024)

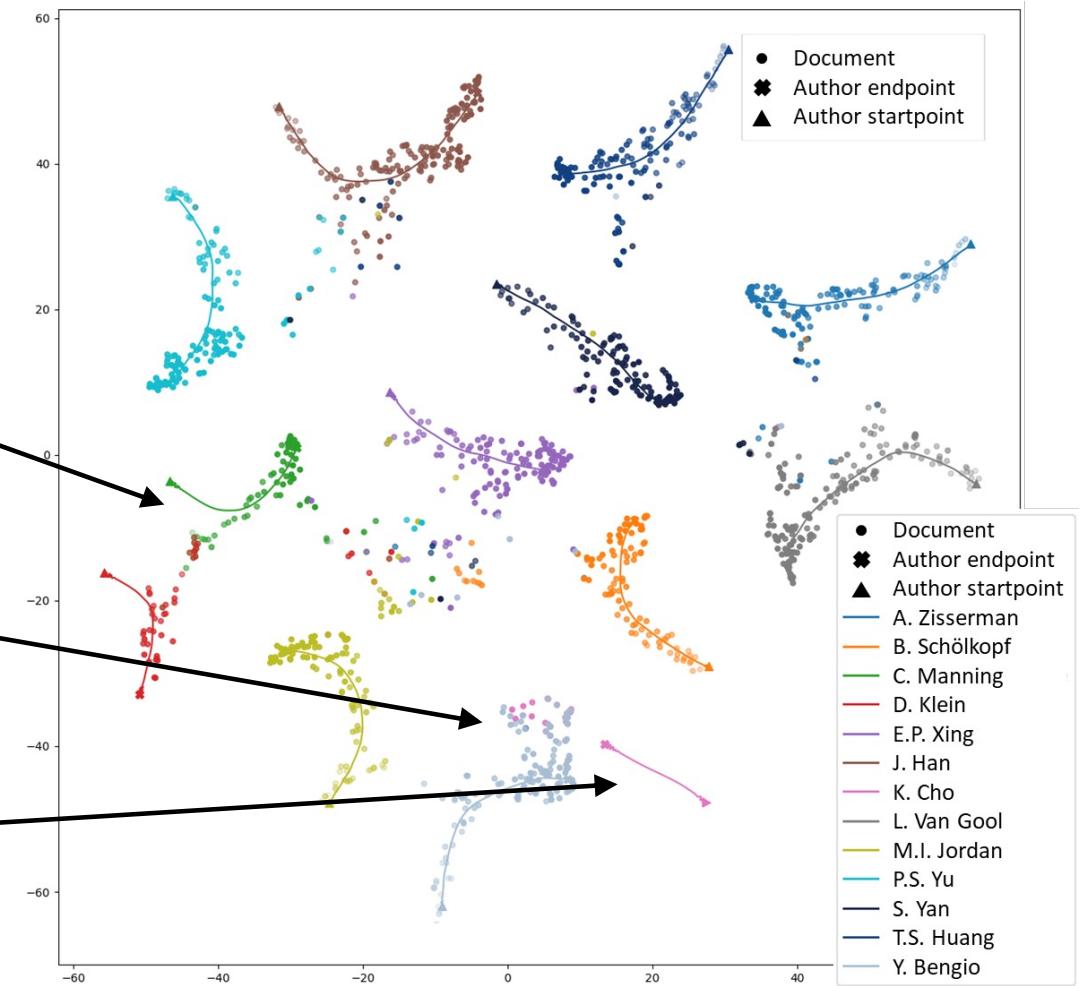


# Visualiser les trajectoires dans l'espace latent

trajectoires proches

diversification  
dans le temps

période de publication courte



Il s'agit ici de données issues de bases de données bibliographiques (*Semantic Scholar*).

# Compression impacts hate speech detection models

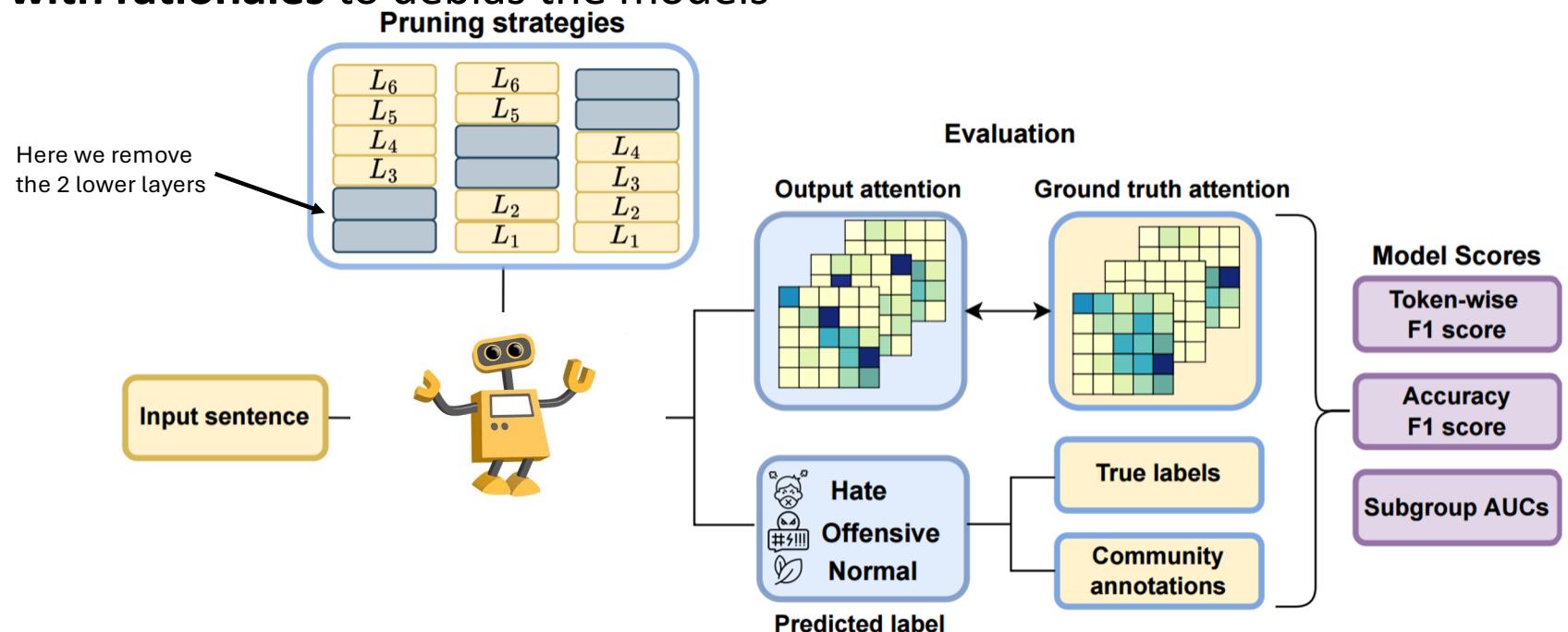
Julien Velcin, candidat au poste PR ECL-LIRIS

# The other side of compression: Measuring and combating bias in pruned transformers

- Early work based on **simple encoder-based models** and **pruning**, presented at IDA in 2023 (Proskurina et al, 2023)
- We measure identity-based **bias** in pruned Transformer LMs (eg., BERT)
- We study **which group** of encoder **layers** (bottom, middle or upper) can be efficiently pruned without biased outcomes
- We propose **word-level supervision** as a debiasing method

# Methodology

- 1) **Prune Transformer** (BERT, DistillBERT, RoBERTa, DistillRoBERTa)
- 2) **Fine-tune** Transformer on hate speech classification task (with **HateXplain**)
- 3) **Evaluate** performance, bias
- 4) **Fine-Tune with rationales** to debias the models



# Results: Compressed LMs are prone to bias

Model	Layers	F1 score	Token F1 score	Count Signif Target Classes		
				Subgroup	BNSP	BPSN
BERT	12/12	67.28±0.13	48.58±3.28	-	-	-
	10/12	65.31±0.17	38.35±4.11	2	0	1
	8/12	64.82±0.15	32.57±4.06	2	0	2
	6/12	63.46±0.21	34.4±3.87	4	0	2
DistilBERT	6/6	66.19±0.44	43.31±3.42	-	-	-
	5/6	66.08±0.62	42.77±4.13	0	0	0
	4/6	65.66±0.51	42.1±3.98	3	0	1
	3/6	64.31±0.83	39.81±4.22	3	1	2
RoBERTa	12/12	83.42±0.4	46.64±3.51	-	-	-
	10/12	81.46±0.41	39.37±4.61	4	2	2
	8/12	78.67±0.58	38.49±4.23	6	3	4
	6/12	77.08±0.33	24.47±4.08	6	5	5
DistilRoBERTa	6/6	82.02±0.36	42.08±5.24	-	-	-
	5/6	81.08±0.4	33.2±4.75	3	0	2
	4/6	77.06±0.48	32.76±5.21	3	2	4
	3/6	74.05±0.43	32.6±4.61	6	5	6

Performance of original and pruned models on HATEXPLAIN test set

We count how many times:

$\beta_0$  full model

$\beta_c$ , compressed model

$$H_1 : \beta_0^t - \beta_0 \neq \beta_c^t - \beta_c,$$

F1 for group t

overall F1

number of groups  
with a significant  
difference in term of  
classification

Some groups in  
HateXPlain:

- Men
- Women,
- African,
- Arabs,
- Asians,
- Caucasian,
- ...

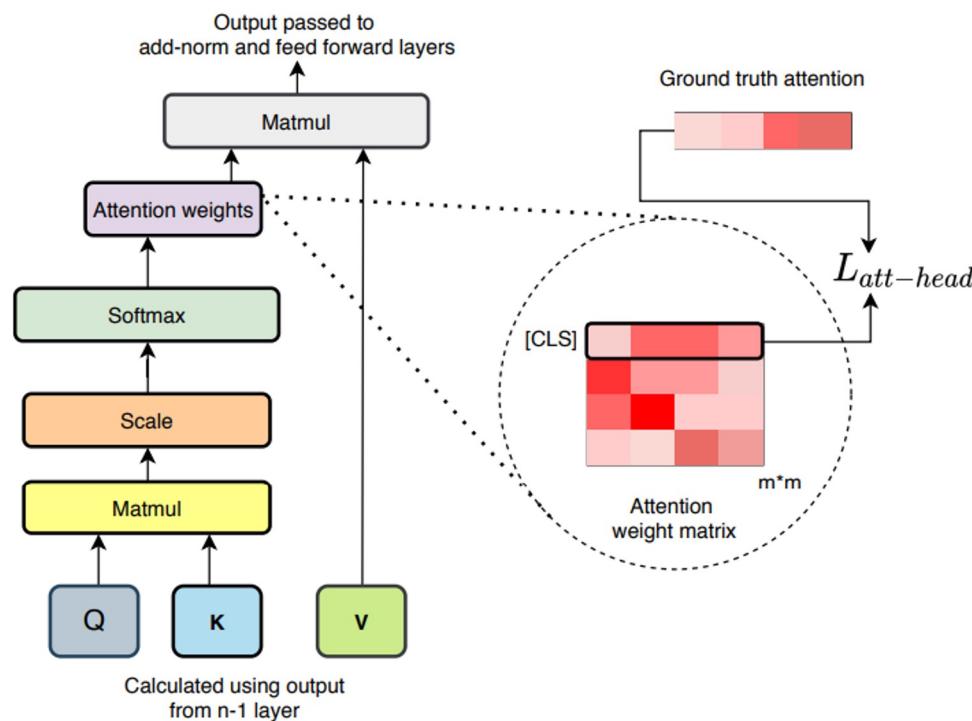
# Results: Compressed LMs rely on unimportant tokens

Model	Layers	F1 score	Token F1 score	Count Signif Target Classes		
				Subgroup	BNSP	BPSN
BERT	12/12	67.28±0.13	48.58±3.28	-	-	-
	10/12	65.31±0.17	38.35±4.11	2	0	1
	8/12	64.82±0.15	32.57±4.06	2	0	2
	6/12	63.46±0.21	34.4±3.87	4	0	2
DistilBERT	6/6	66.19±0.44	43.31±3.42	-	-	-
	5/6	66.08±0.62	42.77±4.13	0	0	0
	4/6	65.66±0.51	42.1±3.98	3	0	1
	3/6	64.31±0.84	39.81±4.22	3	1	2
RoBERTa	12/12	83.42±0.41	46.64±3.51	-	-	-
	10/12	81.46±0.41	39.37±4.61	4	2	2
	8/12	78.67±0.58	38.49±4.23	6	3	4
	6/12	77.08±0.33	24.47±4.08	6	5	5
DistilRoBERTa	6/6	82.02±0.36	42.08±5.24	-	-	-
	5/6	81.08±0.41	33.2±4.75	3	0	2
	4/6	77.06±0.48	32.76±5.21	3	2	4
	3/6	74.05±0.43	32.6±4.61	6	5	6

Performance of original and pruned models on HATEXPLAIN test set

# Solution: Supervised Attention learning

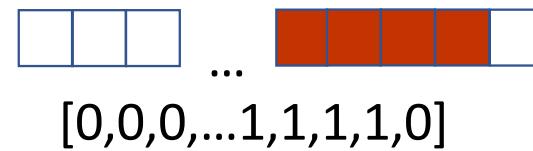
$$Loss_{\Sigma} = Loss_{pred} + \lambda Loss_{attn}$$



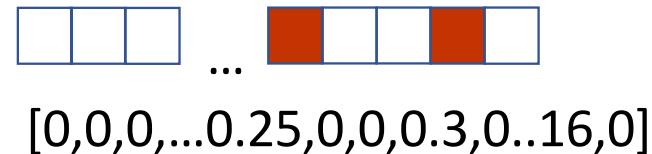
<user>: I got a guilty pleasure and it is country music and hillbilly movies and tv shows about rednecks hunting in the woods... trailer<sup>ab</sup> trash<sup>abc</sup>  
poor<sup>c</sup> plump<sup>c</sup> thing<sup>c</sup>

<sup>a</sup>Annotator 1: Target labels: Economic, Caucasian  
<sup>b</sup>Annotator 2: Target labels: Economic  
<sup>c</sup>Annotator 3: Target labels: Caucasian

## True Rationales



## Predicted Rationales (via attention maps)



# Results: Fine-tuning with attention loss compensates for fairness loss

Model	$\lambda$	F1 score	Token F1 score	Subgroup AUC
BERT (6/12)	0	$63.46 \pm 0.21$	$34.4 \pm 3.87$	$0.59 \pm 0.01$
	0.01	$65.12 \pm 0.38$	$36.3 \pm 4.01$	$0.707 \pm 0.11$
	0.1	$65.92 \pm 0.24$	$39.26 \pm 3.91$	$0.784 \pm 0.07$
	1	$66.61 \pm 0.17$	$45.54 \pm 3.29$	$0.803 \pm 0.12$
DistilBERT (3/6)	0	$64.31 \pm 0.83$	$39.81 \pm 4.22$	$0.768 \pm 0.24$
	0.01	$64.35 \pm 0.51$	$40.4 \pm 3.04$	$0.748 \pm 0.16$
	0.1	$65.11 \pm 0.7$	$41.03 \pm 3.28$	$0.794 \pm 0.31$
	1	$66.71 \pm 0.22$	$42.67 \pm 3.14$	$0.796 \pm 0.28$
RoBERTa (6/12)	0	$77.08 \pm 0.33$	$24.47 \pm 4.08$	$0.519 \pm 0.21$
	0.01	$80.86 \pm 0.22$	$33.19 \pm 3.28$	$0.612 \pm 0.29$
	0.1	$78.58 \pm 0.23$	$36.49 \pm 4.11$	$0.681 \pm 0.17$
	1	$82.38 \pm 0.26$	$40.52 \pm 3.81$	$0.691 \pm 0.14$
DistilRoBERTa (3/6)	0	$71.05 \pm 0.43$	$32.6 \pm 4.61$	$0.62 \pm 0.08$
	0.01	$79.14 \pm 0.47$	$34.41 \pm 4.11$	$0.634 \pm 0.04$
	0.1	$81.25 \pm 0.33$	$36.51 \pm 3.5$	$0.635 \pm 0.08$
	1	$81.96 \pm 0.51$	$43.02 \pm 4.14$	$0.65 \pm 0.09$

$$Loss_{\Sigma} = Loss_{pred} + \lambda Loss_{attn}$$

Performance and fairness scores  
(Subgroup AUC) of models trained  
with word-level supervision

BERT Subgroup AUC scores

- .59 - without attention supervision
- .80 - with attention supervision

\* $\lambda = 0$  - non-supervised attention learning

# Conclusion on this work

- We conducted two chains of experiments to analyze the effect of Transformer LMs **pruning** in the context of **hate speech classification** tasks (with and without attention supervision)
- We compare **both fairness and performance loss** for pruned BERT, RoBERTa, and their distilled versions
- We show and statistically prove that **removing any layer** from Transformer LMs **results in fairness loss** even when the performance loss could be negligible
- We conducted supervised attention-learning experiments that help to **reduce bias in pruned models**

# Compression impacts model calibration

Julien Velcin, candidat au poste PR ECL-LIRIS

# Our contribution (Proskurina et al., 2024)

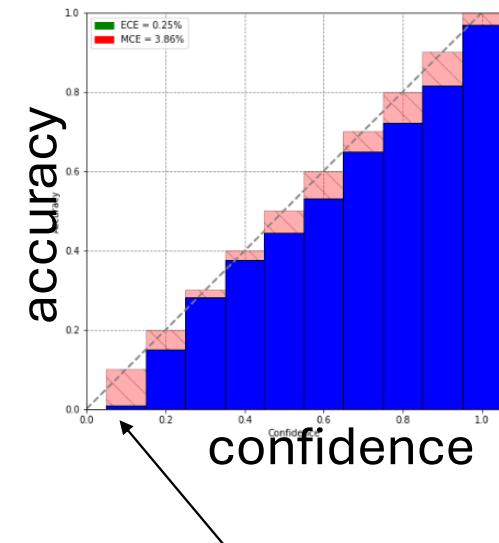
- We investigate how quantization with **GPTQ** (Frantar et al., 2023) influences the **calibration** and **confidence** of LLMs
  - => well calibrated models ensure that the model outputs probabilities (prediction) are well aligned with the confidence of the models
- We assess the confidence alignment between compressed and full-precision LLMs **at scale** (ie various model sizes)
- We provide some **explanations** to the quantization loss from the initial confidence perspective

# Calibration and (post-training) quantization

- **Good calibration:** model output = prediction confidence
- **Compression (quantization):** we rely on GPTQ where we want to find a quantized version of weight  $\hat{W}_l^*$  to minimize the mean squared error:

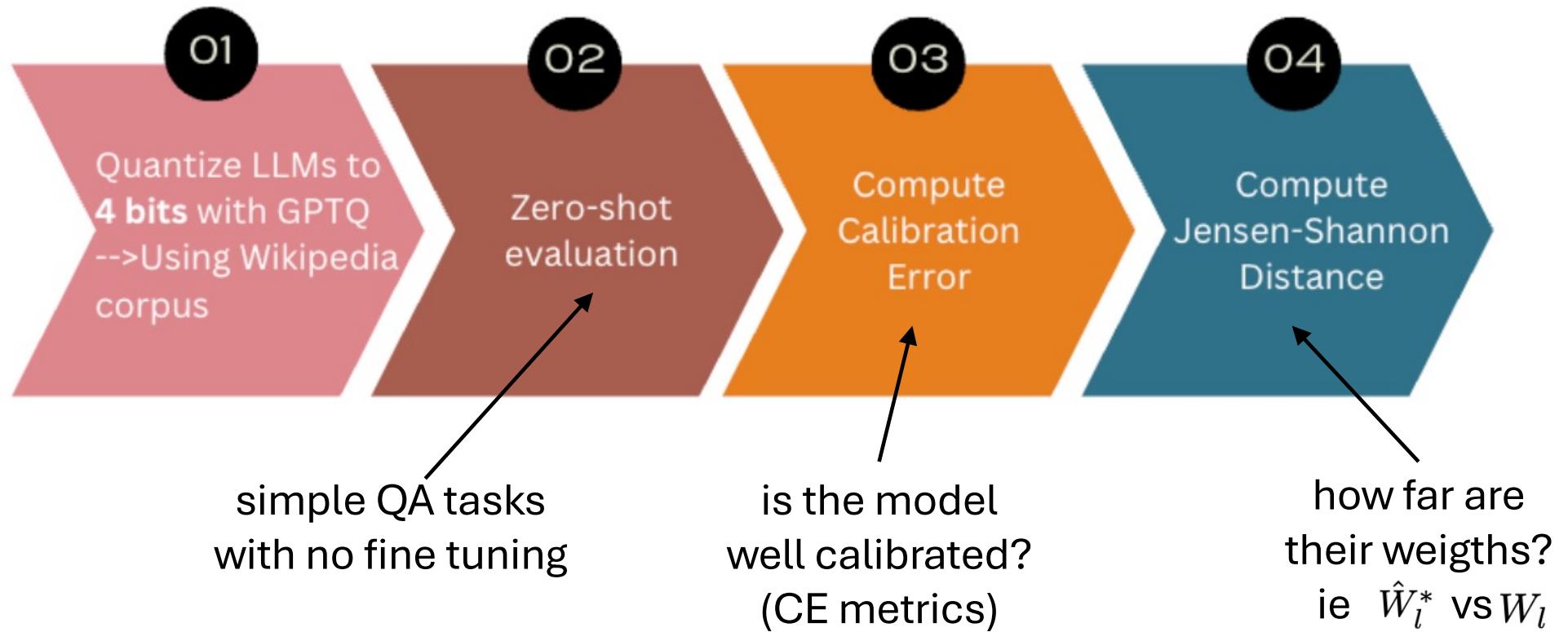
$$\hat{W}_l^* = \operatorname{argmin}_{\hat{W}_l} \|\hat{W}_l X - W_l X\|_2^2$$

↑  
quantized weights      ↑  
initial weights



in this bucket, we expect 10% of examples are predicted as classe + (here, binary classification)

# Zero-shot Question Answering: pipeline



# Data and baselines

- Data: Six standard commonsense reasoning tasks:
  - question answering involving reading comprehension (BoolQ)
  - natural text entailment (XStory-En, HellaSwag)
  - science fact knowledge (ARC, OBQA)
  - physical commonsense (PIQA)
- Baselines: causal (auto-regressive) LLMs:
  - BLOOM (560M, 1B1, 1B7, 3B, and 7B1 parameters)
  - OPT (125M, 350M, 1B3, 2B7, 6B7, and 13B)
  - Mistral-7B
  - LLaMA-7B

# Results: Quantization amplifies CE

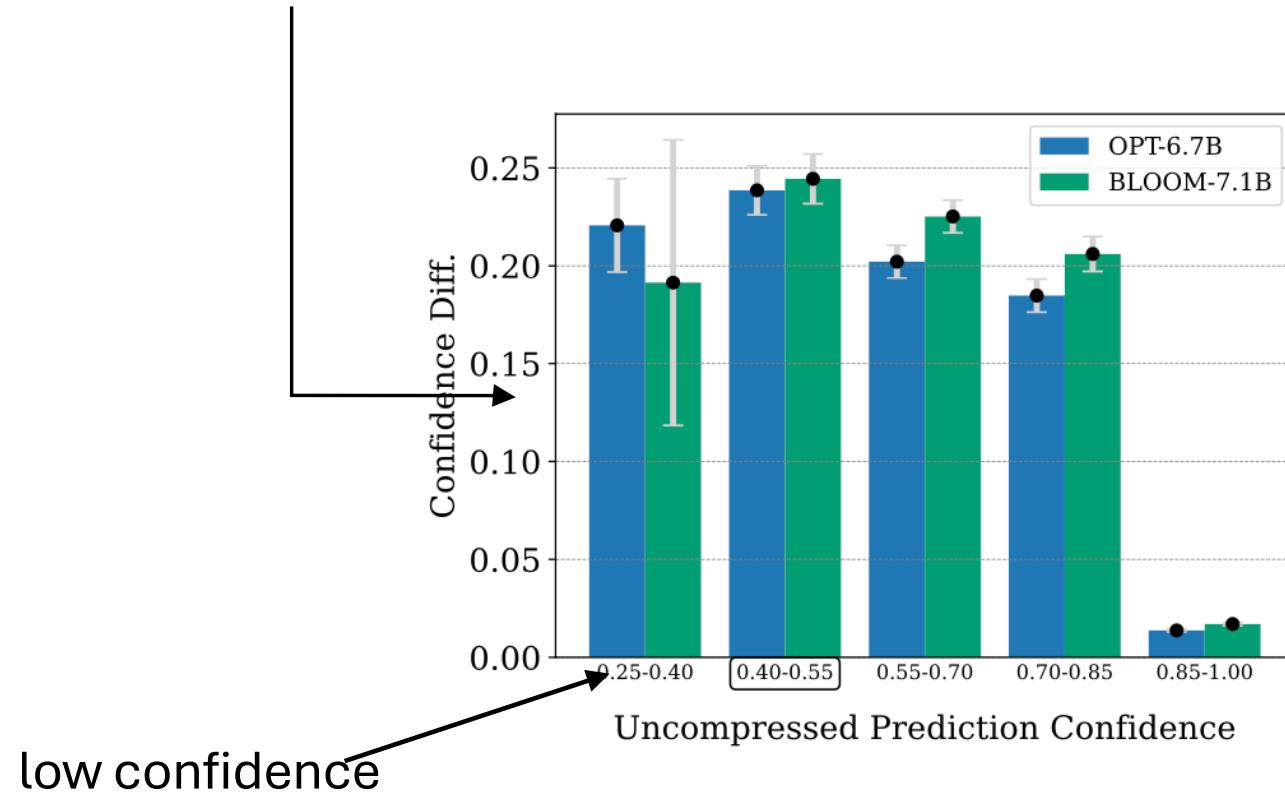
The general trend is that quantization **amplifies** the pre-existing high calibration error present in the models before compression

M	Acc. ↑	CE↓
ArcEasy	81.10 ↓ 1.18	7.94 ↑ 0.83
BoolQ	83.61 ↓ 0.86	38.62 ↑ 3.13
HellaSwag	61.30 ↓ 1.53	34.3 ↑ 1.29
OpenBookQA	32.60 ↓ 0.40	45.24 ↑ 2.08
PiQA	80.83 ↓ 0.65	45.24 ↓ 0.4
Xstory	78.89 ↓ 0.27	4.78 ↓ 0.08

Table 1: Zero-shot accuracy scores (Acc.) and calibration error (CE)

# Results: Quantization affects low-confidence samples

After quantization, **confidence shift is larger** for samples with initial low confidence



# Results at scale: Differences decrease with model size

Distances between original and compressed LLMs **decrease** as the model size scales up

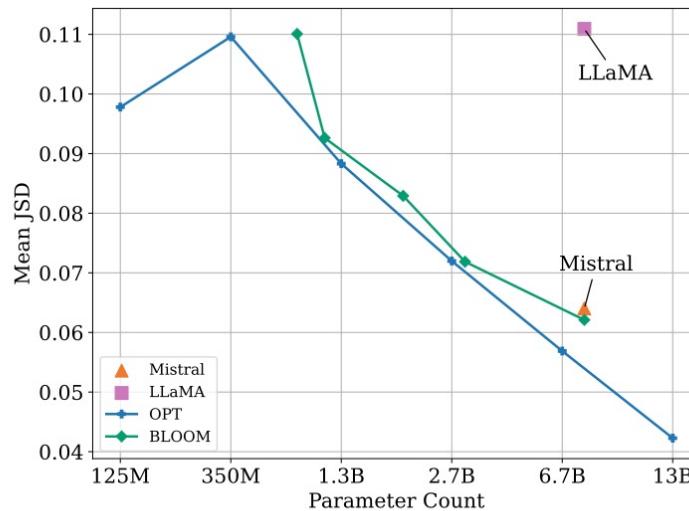


Figure 2: Mean Jensen-Shannon distances between full and quantized LLMs across benchmarks. The distances depict dissimilarities in true-class probability distributions.

## Conclusion on this work

- Impact of quantization **on the confidence and calibration** of LLMs
- Quantization leads to an **increase in calibration error** and statistically significant changes in confidence levels for correct predictions
- **Confidence change bigger** when models unconfident before quantization
- Need to **focus on calibrating LLMs**, specifically on uncertain examples

# Taking the bias term into account in quantization (ongoing work)

- Find  $W_c$  the best quantized weight matrix:

$$\mathbf{W}_c = \arg \min_{\mathbf{W}'} \left\| \mathbf{W}\mathbf{X}_0 - \mathbf{W}'\mathbf{X}_0 \right\|_2^2 + \left\| \mathbf{W}\mathbf{X}_1 - \mathbf{W}'\mathbf{X}_1 \right\|_2^2 + \alpha \left\| \mathbf{W}'(\mathbf{X}_0 - \mathbf{X}_1) \right\|_2^2.$$

usual loss terms (see GPTQ)

minimize the biases

- The optimization problem is not trivial and we derived solution to do it column-wise efficiently
  - Experiments show that it's possible to quantize *and* reduce biases