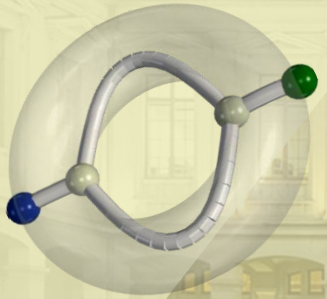
Díkē (Dice) ANR Project Feedback
ERIC Lab

EALM Workshop
CORIA-TALN
30th June



Irina Proskurina
Eric Lab
University of Lyon
University of Lyon 2

Outline

- Introduction (Compression of LMs)
- Impact of Pruning on Bias
- Impact of Quantization on LMs Confidence
- Dataset in French for Social Reasoning
- Conclusion

Partenaires

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

Christophe GRAVIER (PR)
François JACQUENET (PR)
Antoine GOURRU (MCF)
Thibaud LETENO (PHD)

Charlotte LACLAU (MCF) (Télécom Paris)

eric

Julien VELCIN (PR)
Guillaume METLZER (MCF)
Adrien GUILLE (MCF)
Irina PROSKURINA (PHD)
Luc BRUN (intern)
Angelo LAMURE (intern)

NAVER LABS Europe

Vassilina NIKOULINA (R. Sc)
Caroline BRUN (Sr Sc.)
Alireza MOHAMMADSHAHI (intern)

Compression of LMs

- **Language Modelling**

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} logP_\theta(w_{i,j}|w_{i,1}, w_{i,2} \dots, w_{i,j-1})$$

$\theta \in \mathbb{R}^n$

m_i – the number of tokens in the sequence w_i;
$w_{i,j}$ is the j-th token in the sequence

Compression of LMs

- **Language Modeling**

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} log P_\theta(w_{i,j} | w_{i,1}, w_{i,2} \dots, w_{i,j-1})$$

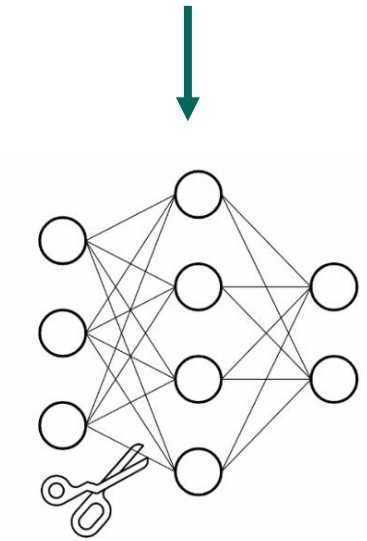
$\theta \in \mathbb{R}^n$

m_i – the number of tokens in the sequence w_i;
$w_{i,j}$ is the j-th token in the sequence

- **Compression of LLMs via Pruning**

$$\theta^* = \{\theta_i \mid \theta_i \; is \; non\text{-}pruned\} \cup \{\theta_i = 0 \mid \theta_i \; is \; pruned\}$$

Pruning

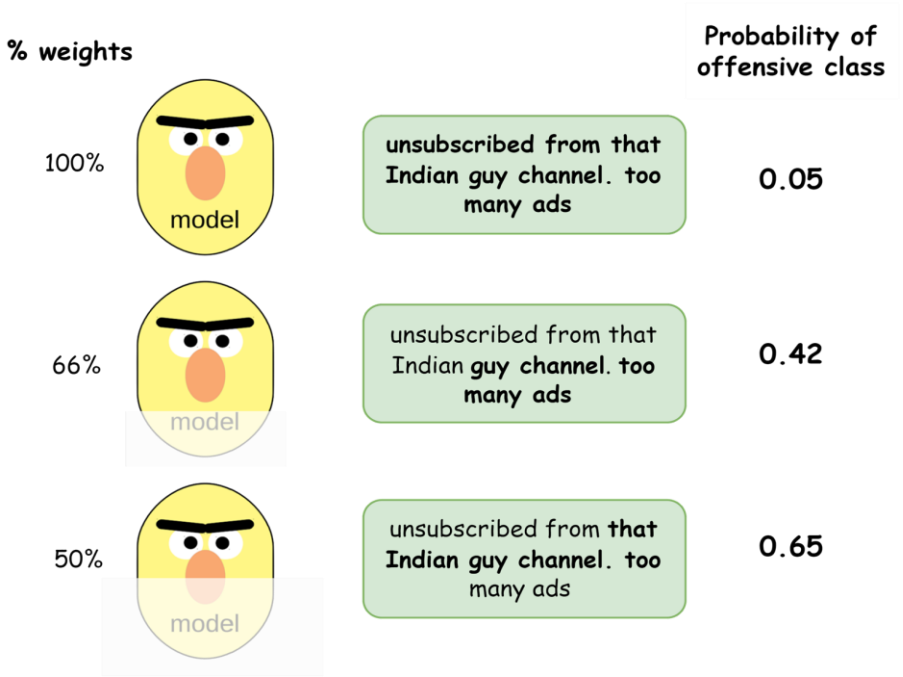


Impact of layer pruning on bias in LMs for hate speech detection

The Other Side of Compression: Measuring Bias in Pruned Transformers (Proskurina et al, IDA 2023)

- We measure identity-based **bias** in pruned Transformer LMs
- We study **which group** of encoder **layers** (bottom, middle or upper) can be efficiently pruned without biased outcomes
- We propose **word-level supervision** in pruned Transformer LMs as a debiasing method

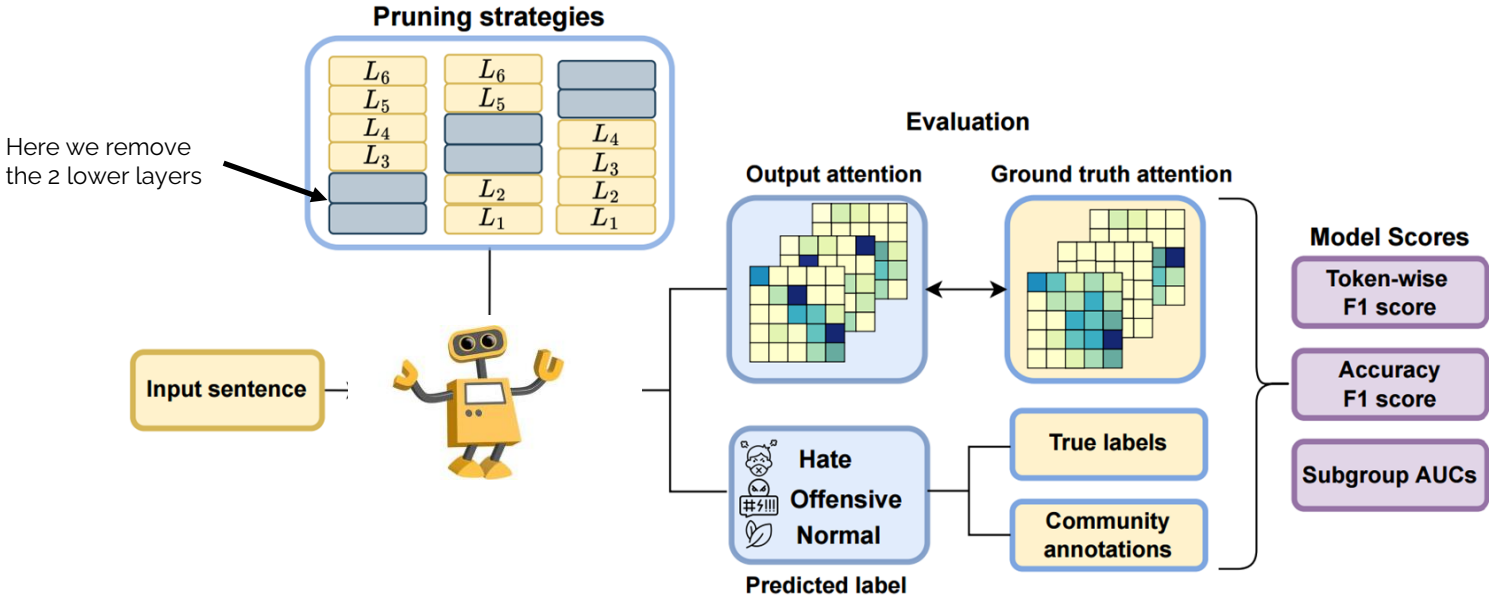
Bias in Hate Speech Classification



Bias = LM classifies neutral text as offensive and pays 'attention' to sensitive attributes

Methodology

1) Prune Transformer LM (e.g., BERT)
2) Fine-tune LM on hate speech classification task (with HateXplain)
3) Prune selected weights
4) Compare accuracy, bias, and explainability scores of LMs before and after pruning



Evaluate Bias in Compressed models

Null Hypothesis H_0: If the impact of compression is uniform, then the shift in scores achieved on the texts mentioning a target community *t* after pruning should also be uniform compared to the overall scores shift

$$H_0 : \beta_0^t - \beta_0 = \beta_c^t - \beta_c$$ ← no significant difference

$$H_1 : \beta_0^t - \beta_0 \neq \beta_c^t - \beta_c,$$ ← significant difference

β_0 non-pruned full model

β_0^t +targeting community t

β_c compressed model

β_c^t +targeting community t

Results: Compressed LMs are prone to bias

full model

4 layers removed

Model	Layers	F1 score	Token F1 score	Count Signif Target Classes		
				Subgroup	BNSP	BPSN
BERT	12/12	67.28±0.13	48.583.28	-	-	-
	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

H_0 H_1

number of groups with a significant difference in term of classification (on 10)

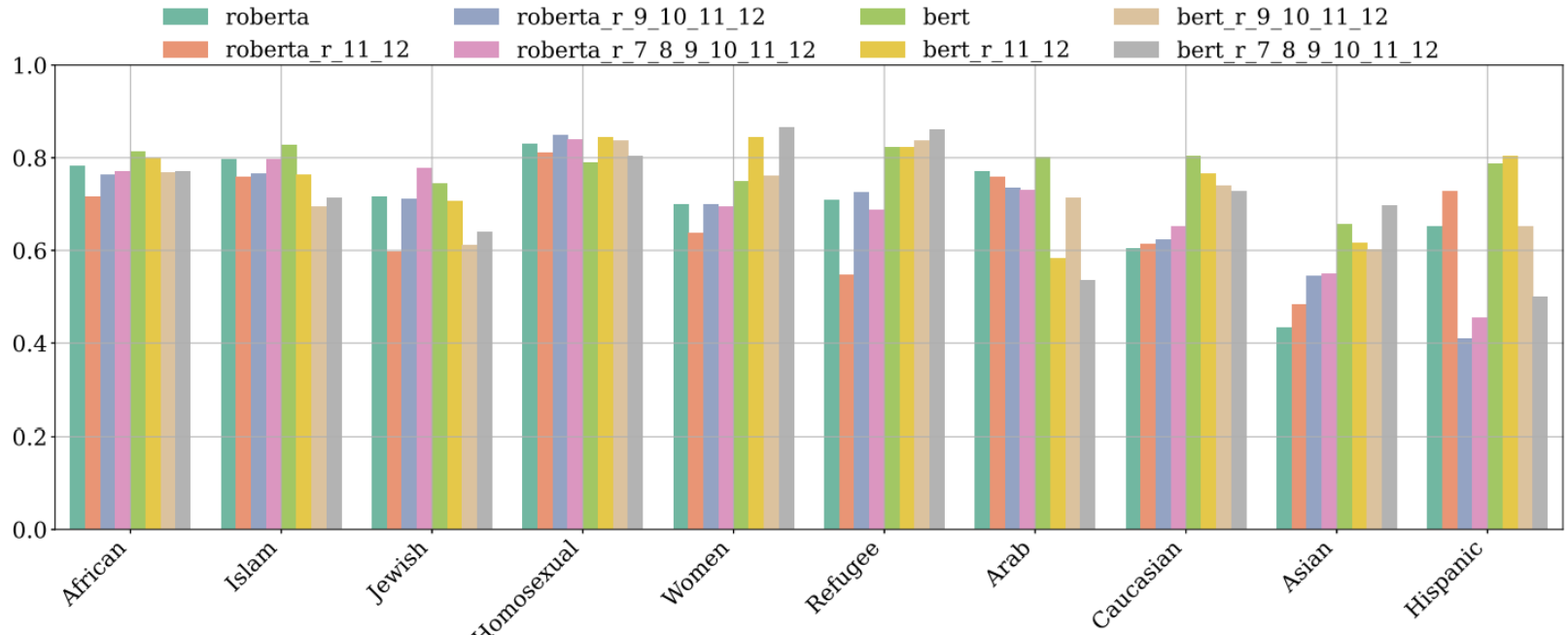
Performance of original and pruned models on HATEXPLAIN test set

Results: Compressed LMs rely on unimportant tokens

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

Performance of original and pruned models on HATEXPLAIN test set

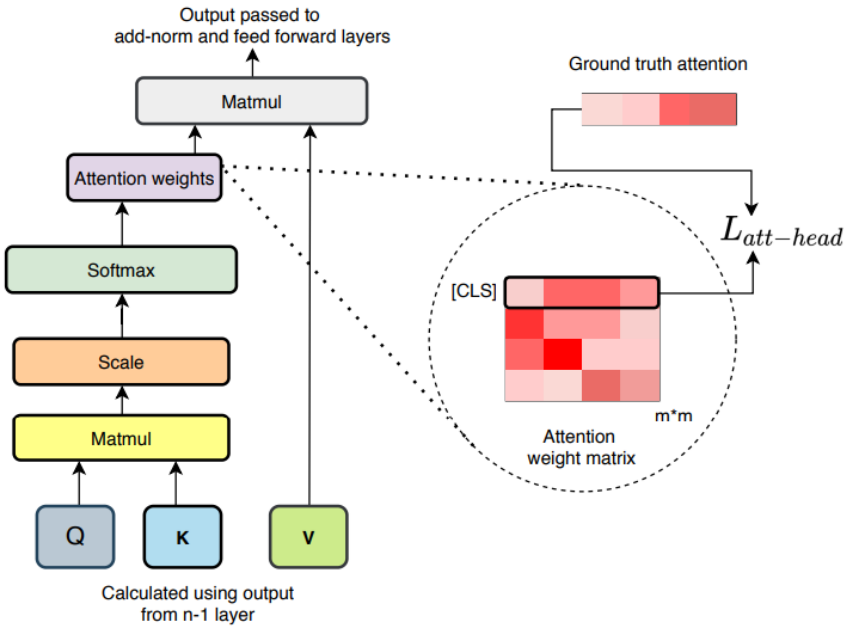
Results: The impact of compression is not uniform



OY: Subgroup AUC scores on HateXplain, OX: Target communities
LMs: BERT, RoBERTa

Solution: Supervised Attention learning

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



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

[a]Annotator 1: Target labels: *Economic, Caucasian*
[b]Annotator 2: Target labels: *Economic*
[c]Annotator 3: Target labels: *Caucasian*

True Rationales

[0,0,0,...1,1,1,1,0]

Predicted Rationales

[0,0,0,...0.25,0,0,0.3,0..16,0]

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

Model	λ	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	$81.96_{\pm 0.51}$	$43.02_{\pm 4.14}$	$0.65_{\pm 0.09}$

*λ = 0 - non-supervised attention learning

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

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

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 conduct supervised attention-learning experiments that help to **reduce bias in pruned models**

Impact of quantization on calibration and model confidence

Compression of LMs

- **Language Modeling**

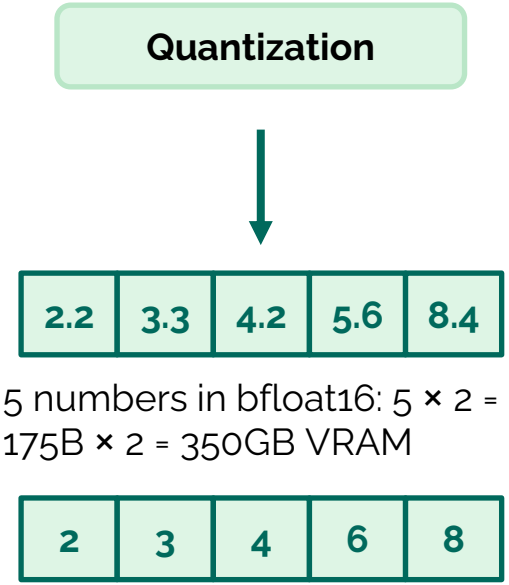
$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{m_i} log P_\theta(w_{i,j} | w_{i,1}, w_{i,2} \dots, w_{i,j-1})$$

$\theta \in \mathbb{R}^n$

m_i – the number of tokens in the sequence w_i; $w_{i,j}$ is the j-th token in the sequence

- **Compression of LLMs via Quantization**

$$\hat{\theta}^* = \{ \widehat{\theta_i} \mid \widehat{\theta_i} = Q(\theta_i), \widehat{\theta_i} \epsilon \mathbb{Q} \}, |\mathbb{Q}| = m \ll n$$



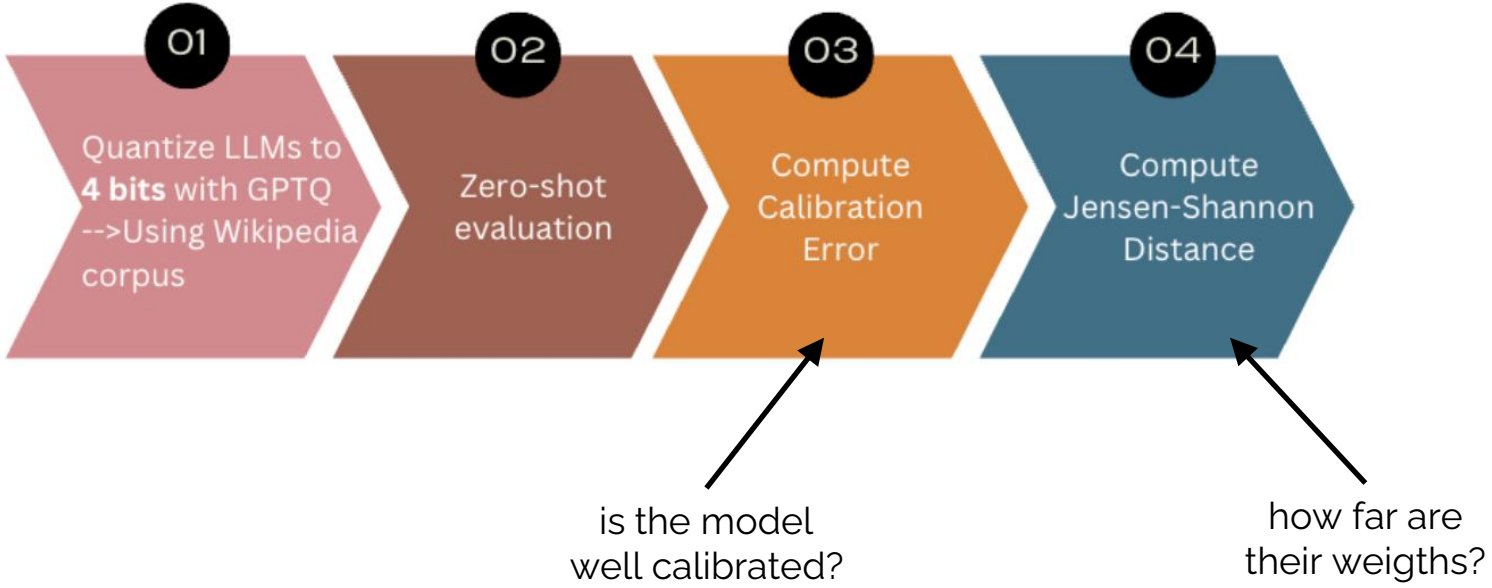
5 numbers in bfloat16: 5 × 2 = 10 bytes
175B × 2 = 350GB VRAM

5 numbers in int8: 5×1 = 5 bytes
175 × 1 = 175GB VRAM

Contribution (Proskurina et al., NAACL-HLT 2024)

- We investigate how quantization with **GPTQ** (Frantar et al., 2023) influences the calibration and confidence of LLMs
- We assess the confidence alignment between **compressed** and **full-precision** LLMs at scale
- We **explain** the quantization loss from the initial confidence perspective

Zero-shot Question Answering: pipeline



Methodology

Classification problem: questions x paired with candidate answers y
→ The generative model then processes these concatenated question-answer pairs to predict the most probable answer y from the provided choices Y for a given x:

$$\hat{y} = \arg\max_{y \in Y} p_{\text{LM}}(y|x).$$

With:

$$p_{\text{LM}}(y|x) = \prod_{i=1}^{|y|} p_{\text{LM}}(y_{[i]}|x, y_{[1:i-1]})$$

Results: Confidence Impact

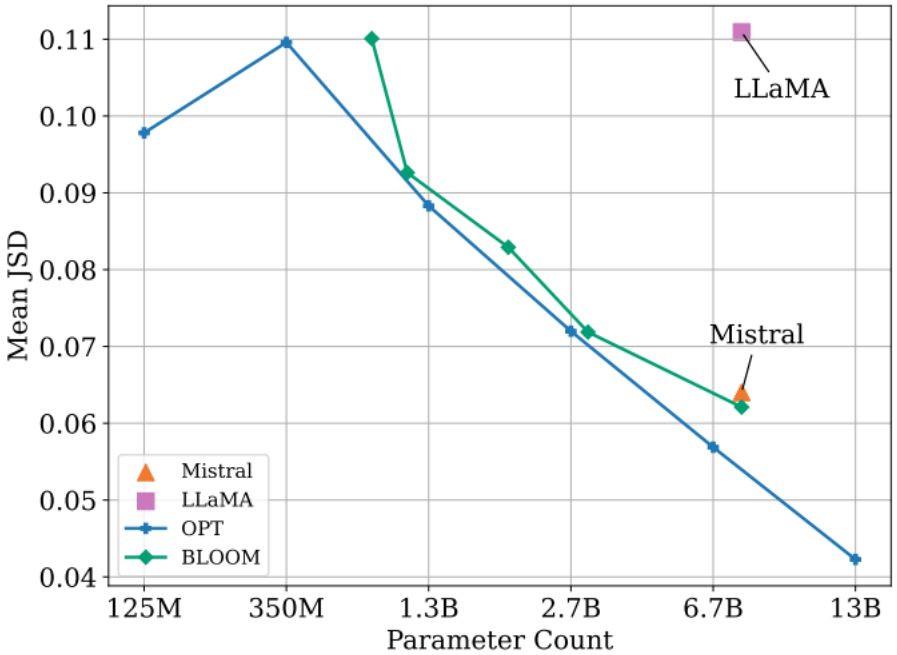
A consistent trend of **overconfidence** emerges in both pre- and post-quantization stages, with an average confidence level around~0.95 for incorrect predictions.

Model	Conf.	Conf_{err}	Conf_{true}	H
BLOOM	96.26	95.64	46.24	12.87
+ GPTQ	96.3	95.62	45.23*	12.89
OPT	96.51	95.57	50.37	12.12
+ GPTQ	96.5	95.55	49.78*	12.22
Mistral	96.85	95.02	61.14	10.96
+ GPTQ	96.89	95.13	59.73*	10.87
LLaMA	96.8	95.34	56.83	11.37
+ GPTQ	96.48	95.13	53.69*	12.21*

Table 2: Confidence and prediction entropy evaluation results on HELLASWAG

Results: Jensen-Shannon Distances

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



Mean Jensen-Shannon distances between fp16 and quantized LLMs across benchmarks.
The distances show dissimilarities in true-class probability distributions

Conclusion on this work

- We have investigated the 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
- We identify **instances of confidence change** occurring in data where models lack confidence before quantization
- Our findings provide insights into quantization loss and suggest a potential direction for future work, emphasizing the **need to focus on calibrating LLMs**, specifically on uncertain examples

New Datasets for Multilingual Moral Reasoning in LMs

Developing a French corpus of Moral stories
(Leteno et al., NACCL 2025)

- Adaptation of the **Moral Stories** dataset (Emelin et al., EMNLP 2021)
  - automatic translation from English to French
  - adaptation to French
  - thorough manual verification

Category	MORALSTORIES / HISTOIRESMORALES
Norm	It's wrong to use violence to solve your problems. / *Il est mal de recourir à la violence pour résoudre ses problèmes.*
Situation	Ben lives in a neighborhood with assigned parking and his neighbor's friend frequently uses his assigned spot. / *Benoît habite dans un quartier où les places de stationnement sont attribuées, et un ami de son voisin occupe régulièrement sa place assignée.*
Intention	Ben wants to stop the neighbors friend from taking up his parking spot. / *Benoît souhaite empêcher l'ami de son voisin d'occuper sa place de stationnement.*

Developing a French corpus of Moral stories
(Leteno et al., new paper accepted at NACCL 2025)

- Adaptation of the **Moral Stories** dataset (Emelin et al., EMNLP 2021)
  - automatic translation from English to French
  - adaptation to French
  - thorough manual verification
- **Histoires Morales** can be used for:
  - commonsense reasoning / social reasoning / moral reasc
  - text classification
  - text generation
- **Now available** on HuggingFace:

https://huggingface.co/datasets/LabHC/histoires_morales

HistoiresMorales: Influencing LLMs' moral alignment

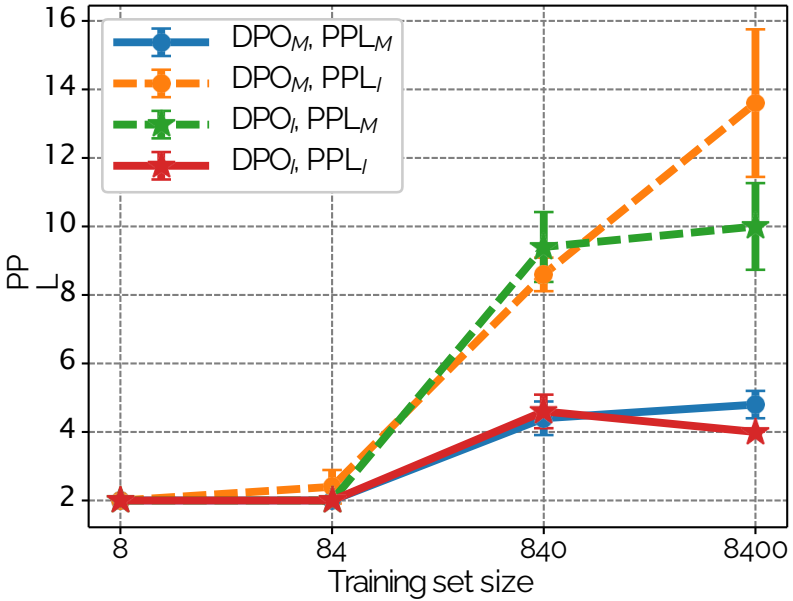


Figure: Average PPL for DPO$_M$ and DPO$_I$ in French.

PPL on the action aligned with the direction of the DPO lower
→ low robustness

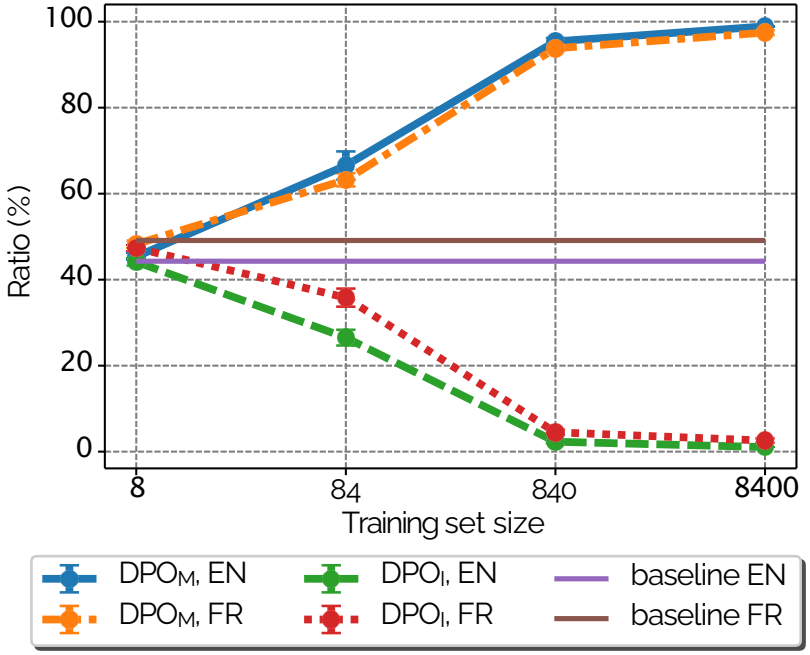


Figure: Ratio of moral actions being preferred based on the PPL.

Few examples are sufficient to shift the alignment.

Conclusion

ANR Dike Project Contributions

ERIC Lab
(University of Lyon and Lyon 2)

Post-hoc impact evaluation of pruning on bias on hate speech models

Post-hoc impact evaluation of quantization on general model confidence

New Dataset for moral reasoning in French

Solution to the bias increase problem after pruning (attention loss with rationales)

Solution to the bias increase problem during quantization

References

- - Hooker et al. (2019), What do compressed deep neural networks forget?. *arXiv preprint*
- - Leteno et al. (2025), HISTOIRESMORALES: A French Dataset for Assessing Moral Alignment. To appear in Proceedings of NAACL-HLT.
- - Proskurina et al. (2023), The Other Side of Compression: Measuring Bias in Pruned Transformers. Proceeding of IDA.
- - Proskurina et al. (2023), Mini Minds: Exploring Bebeshka and Zlata Baby Models. BabyLM Challenge at CoNLL, Proceeding of ACL.
- - Proskurina et al. (2024), When Quantization Affects Confidence of Large Language Models? Proceeding of NAACL-HLT (Findings).

Contacts
Irina.Proskurina@univ-lyon2.fr
https://www.iproskurina.com