Auditing and mitigating biases in (compressed) LMs

Julien Velcin
ERIC Lab – Université Lumière Lyon 2
https://eric.univ-lyon2.fr/jvelcin/

MBZUAI France Lab - Workshop 2/12

Outline

- DIKé project and introduction to main concepts
- Impact of compression on hate speech detection models
- Impact of compression on calibration and confidence
- Participation to the BabyLM challenge, and beyond
- Conclusion and future work

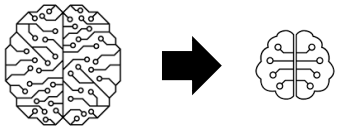
The DIKé project

Julien Velcin, ERIC Lab, Université Lumière Lyon 2
MBZUAI France Lab - Workshop 2/12

DIKé project https://www.anr-dike.fr

- Funded by ANR (AAPG 2021, 2022-2025)
- Partners
  - Laboratoire Hubert Curien (LabHC), Université Jean Monnet
  - Laboratoire ERIC, Université Lumière Lyon 2
  - Naver Labs
- Objectives
  - evaluation of **biases** in LMs (in particular, fairness)
  - English but also **French datasets** for fairness and ethics of NLP systems
  - **new** compressed, fairer language models

LLM compression (1)



- Transformers are **over-parametrized** (#parameters >> #data)
- Can we just **train smaller** Transformers? No (lottery ticket hypothesis)
- **Mobile** devices

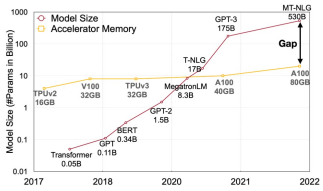
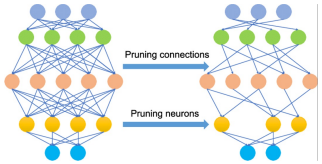


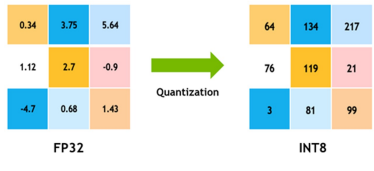
Figure 1: The model size of large language models is developing at a faster pace than the GPU memory in recent years, leading to a big gap between the supply and demand for memory. Quantization and model compression techniques can help bridge the gap.

(Xiao et al, 2023)

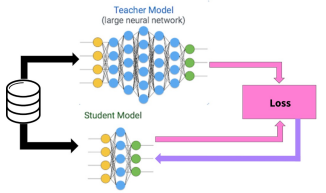
LLM compression (2)



i) Model pruning



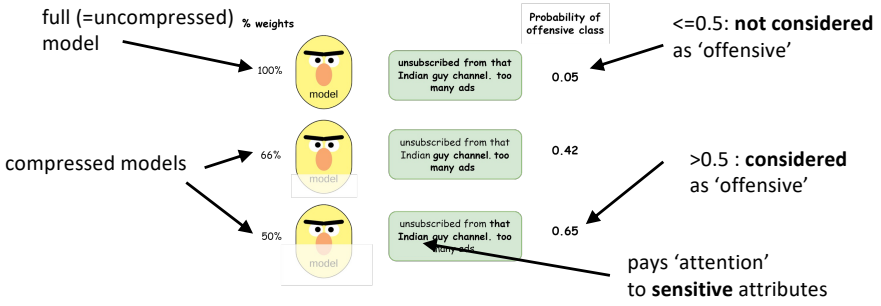
iii) Quantization



ii) Knowledge distillation

Bias and Fairness

- Recent works study the **link between compression and fairness** (Hooker et al., 2019; Gonçalves et al., 2023)



Compression impacts hate speech detection models

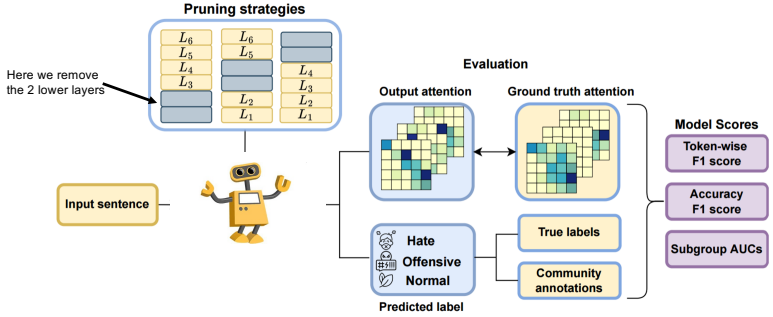
Julien Velcin, ERIC Lab, Université Lumière Lyon 2
MBZUAI France Lab - Workshop 2/12

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

full model — 8/12 : 4 layers removed

Model	Layers	F1 score	Token F1 score	Count Signif Target Classes: Subgroup	BNSP	BPSN
BERT	12/12	67.28±0.13	48.58±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

We count how many times:

β_0 full model — β_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,
...

Performance of original and pruned models on HATEXPLAIN test set

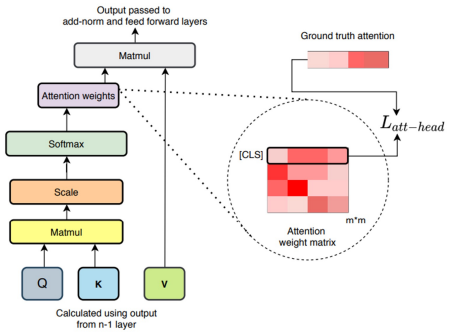
Results: Compressed LMs rely on unimportant tokens

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

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 (via **attention maps**)

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

Compression impacts model calibration

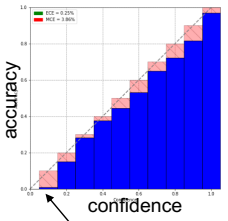
Julien Velcin, ERIC Lab, Université Lumière Lyon 2
MBZUAI France Lab - Workshop 2/12

Contribution (Proskurina et al., 2024)

- Goal of **calibration** is to ensure that the model outputs probabilities (prediction) that are **well aligned** with the **confidence** of the models
- 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** (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



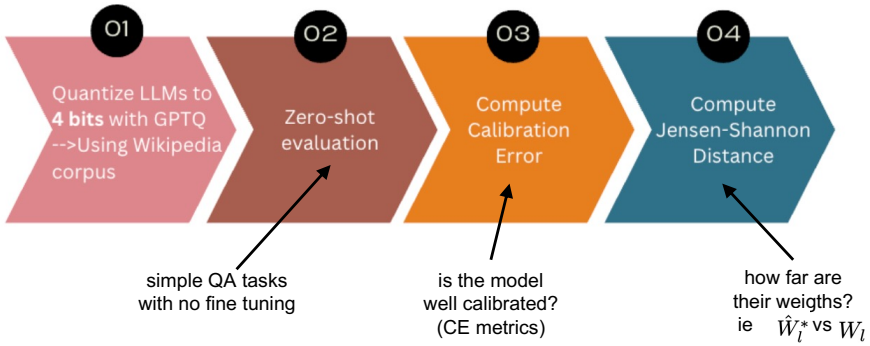
in this bucket, we expect 10% of examples are predicted as classe + (here, binary classification)

- **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:

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

quantized weights — initial weights

Zero-shot Question Answering: pipeline



simple QA tasks with no fine tuning

is the model well calibrated? (CE metrics)

how far are their weigths? ie $\hat{W}_l^*$ vs W_l

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

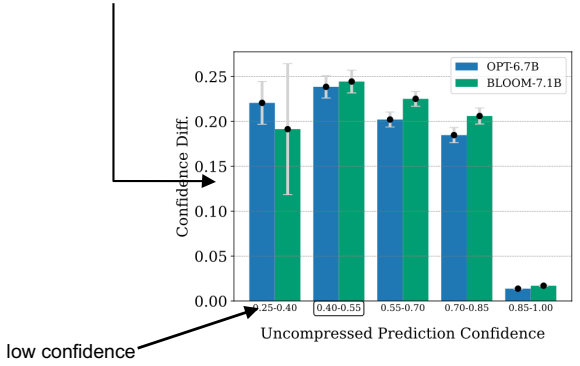
The general trend is that quantization **amplifies** the pre-existing high calibration error present in the models before compression

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

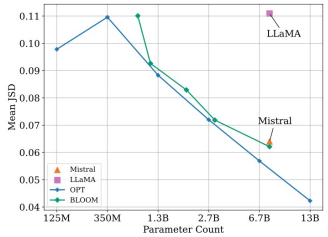


Figure 2: Mean Jensen-Shannon distances between full and quantized LLMs across benchmarks. The distances depict dissimilarities in true-class probability distributions.

Conclusion on this work

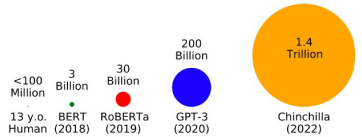
- Impact of quantization **on the confidence and calibration** of LLMs
- Quantization leads to an **increase in calibration error** and statistically significant changes in confidence levels for correct predictions
- **Confidence change bigger** when models unconfident before quantization
- Need to **focus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

BabyLM challenge and beyond

Julien Velcin, ERIC Lab, Université Lumière Lyon 2
MBZUAI France Lab - Workshop 2/12

BabyLM Challenge

Sample-efficient pretraining on a developmentally plausible corpus



- Pre-training models from scratch on corpus of the **children-like vocabulary**
- Main focus of the competition: language acquisition, cognitive modelling
- Our interest within the objectives of Diké project: evaluation of **ethics** in models which have 'seen' only children's literature

Our participation to the challenge

(Proskurina et al., BabyLM@CoNLL 2023)

- Two models presented to BabyLM Workshop, co-located with CoNLL 2023: **Bebeshka** (RoBERTa-based, 16M) and **Zlata** (GPT-based, 66M), pretrained on STRICT-SMALL

Model	CoLA MCC	SST-2 Acc.	MRPC F1	QQP F1	MNLI Acc.	$MNLI_{mm}$ Acc.	QNLI Acc.	RTE Acc.	BoolQ Acc.	MultiRC Acc.	WSC Acc.
OPT	15.2	81.9	72.5	60.4	57.6	60.0	61.5	60.0	63.3	55.2	60.2
RoBERTa	**25.8**	**87.0**	79.2	73.7	**73.2**	**74.0**	**77.0**	**61.6**	**66.3**	**61.4**	61.4
T5	11.3	78.1	**80.5**	66.2	48.0	50.3	62.0	49.4	66.0	47.1	61.4
Bebeshka	0.11	81.3	73.5	66.4	58.7	62.0	59.0	45.4	63.9	48.7	61.4
Zlata	0.05	81.7	77.6	65.9	61.9	63.9	61.7	56.6	65.3	53.8	**61.5**
Bebeshka-2	24.5	83.5	77.7	**77.3**	65.4	66.9	64.0	56.6	60.2	46.9	61.4

with full precision

Table 4: Evaluation results on GLUE and SuperGLUE (BoolQ, MultiRC, WSC) benchmark datasets. We report metrics suggested in the shared task evaluation pipeline and baselines. The best score is in bold, and the second-best score is underlined.

An interesting observation on moral judgement

We did additional experiments on the **ETHICS benchmark**

(Hendrycks et al., 2021)

Model	Justice	Deontology	Virtue	Utilitarianism	Commonsense
RoBERTa-large (355M)	56.7	60.3	53.0	**79.5**	**90.4**
GPT-3 few-shot (175B)	15.2	15.9	18.2	73.7	73.3
Bebeshka (16M)	**64.6**	**71.4**	**74.1**	69.0	-
Zlata few-shot (66M)	50.7	49.6	72.0	50.3	53.3

Table 5: Accuracy scores on ETHICS benchmark. LMs trained on STRICT-SMALL corpus reach results close to the large model baselines reported by Hendrycks et al., 2020. We do not report results for the fine-tuning tasks which require the maximum sequence length exceeding the one of an LM. The best score is in bold, and the second-best score is underlined.

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 reasoning
  - text classification
  - text generation
- **Preprint:** https://huggingface.co/papers/2501.17117
- **Now available** on HuggingFace:
  https://huggingface.co/datasets/LabHC/histoires_morales

Conclusion

Julien Velcin, ERIC Lab, Université Lumière Lyon 2
MBZUAI France Lab - Workshop 2/12

Contributions of the DIKé project (so far)

Projet DIKÉ

- SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages (**EMNLP 2022**)
- What Do Compressed Multilingual Machine Translation Models Forget? (**EMNLP 2022, findings**)
- An Investigation of Structures Responsible for Gender Bias in BERT and DistilBERT (**IDA 2023**)
- The Other Side of Compression: Measuring Bias in Pruned Transformers (**IDA 2023**)
- Mini Minds: Exploring Bebeshka and Zlata Baby Models (**CoNLL 2023, BabyLM challenge**)
- Fair Text Classification with Wasserstein Independence (**EMNLP 2023**)
- FrenchToxicityPrompts: a Large Benchmark for Evaluating and Mitigating Toxicity in French Texts (**LREC-COLING 2024, TRAC workshop**)
- When Quantization Affects Confidence of Large Language Models? (**NAACL 2024, findings**)
- HISTOIRESMORALES: A French Dataset for Assessing Moral Alignment (**NAACL 2025, to appear**)

DIKé consortium

LABORATOIRE HUBERT CURIEN

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)

Some challenges ahead

- Detection of **harmful** language generations (immoral behaviour, hate speech, and beyond)
- In particular, **implicit** hate speech detection
- **Mitigating** harmful generations in quantized models

Additional References

- Frantar et al. (2023), GPTQ: Accurate post-training quantization for generative pre-trained transformers, ICRL
- Hendrycks et al. (2021), Aligning AI with shared human values, ICLR
- Hooker et al. (2019), What do compressed deep neural networks forget? *arXiv preprint*
- Mathew et al. (2021), Hatexplain: A benchmark dataset for explainable hate speech detection, AAAI
- Ramesh et al. (2023), A comparative study on the impact of model compression techniques on fairness in language models, ACL
- Xiao et al. (2023), Smoothquant: Accurate and efficient post-training quantization for large language models, ICML

Thank you

Email: Julien.Velcin@univ-lyon2.fr
Website: https://eric.univ-lyon2.fr/jvelcin/
DIKé project website: https://www.anr-dike.fr

Projet DIKÉ