Adaptation of language models to Orange's domains

Camille Barboule, Viet-Phi Huynh, Adrien Bufort,
Yoan Chabot, Thomas Labbé, Gwénolé Lecorvé, Ghislain Putois
camille.barboule@orange.com

Why ?

- Master the specificities of the Telco field (standards, definitions, etc.)
- Master the specificities of Orange (equipment, offers, procedures, tools, etc.)

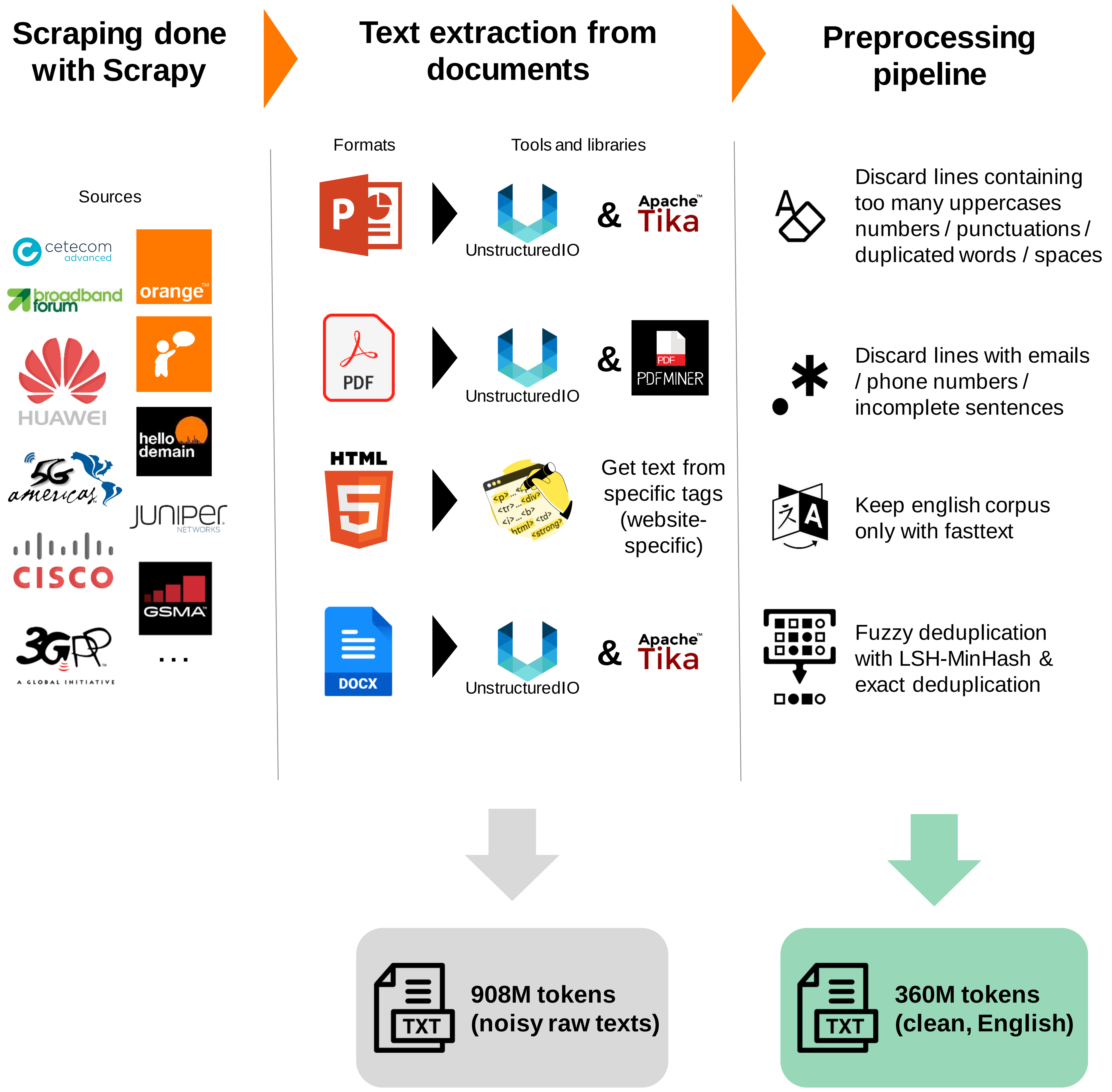
→ **Applicative goal:** better understanding of natural language and generation for internal use cases (technicians, customer assistants, etc.)

Specificities are expressed through new terms, acronyms or context-specific usage of usual words

Experiments & Methodology

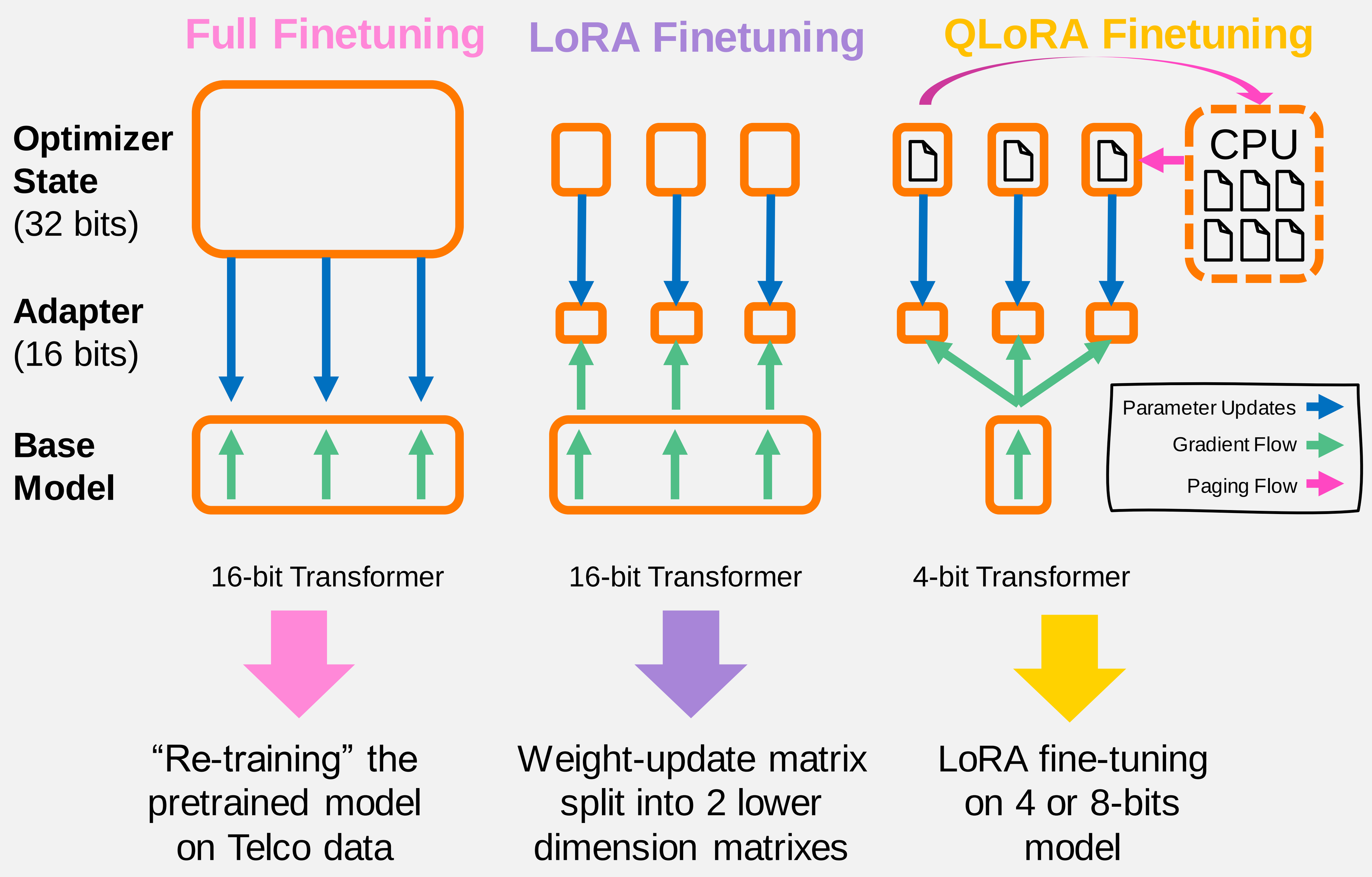
- **What ?**
  - Fine-tune foundation pretrained models on a Telco data
  - Evaluate the adapted models on Orange use cases
- **How ?**
  1. Data Preparation (corpus extraction & preproc.)
  2. Fine-tuning using different methods
  3. Perplexity measurement

1 Data Preparation



2 Fine-tuning methods

Task = Pretraining task (i.e. prediction of last token)



Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.

Experiments

- Test of these 3 fine-tuning methods making a variation of batch sizes, weights precisions, model types
- No vocabulary update (preliminary experiments showed that this is not effective)
- Encoder models were tested (RoBERTa) but results are only reported here on decoder/auto-regressive models (Falcon-1b).

3 First results

- **Experiments on model tiiuae/falcon-rw-1b**
- **Intrinsic measure: perplexity on in-domain texts. Low perplexity means the model models well the text.**

Batch size	Weight precision	Finetuning type	Number of trainable parameters	RAM (GB)↓ (GPU A100)	Perplexity↓ (test set texts)
		None			41.38
8	fp32	Full fine-tuning	1.3B	29.0	27.41
8	fp32	LoRA	1.6M	16.9	32.93
8	fp16	LoRA	1.6M	15.8	32.91
8	8bits	QLoRA	1.6M	12.6	33.08
8	4bits	QLoRA	1.6M	11.9	33.38
4	fp32	Full fine-tuning	1.3B	25.3	27.44
4	fp32	LoRA	1.6M	11.7	33.97
4	fp16	LoRA	1.6M	11.2	33.96
4	8bits	QLoRA	1.6M	7.9	34.97
4	4bits	QLoRA	1.6M	7.3	35.27

✓ Fine-tuning the model on a Telco corpus improved the perplexity of the model on Telco sentences (from 41.38 perplexity to 27.41 for full fine-tuning)

✓ The batch size has a great importance on the fine-tuning efficience, and LoRA fine-tuning is degrading the result compared to full fine-tuning. Between LoRA & QLoRA, there is no big differences however.

- **Other measures on Orange internal knowledge**

Example of question-answering before fine-tuning

> Who is Steve Jarrett for Data and AI at Orange ? Steve Jarrett is the Chief Data Scientist at Orange. He is also the co-founder of the Orange Data Science Lab.

Example of question-answering after fine-tuning with LoRA

> Who is Steve Jarrett for Data and AI at Orange ? Steve Jarrett is the Head of Data and AI at Orange. He is responsible for the development of Orange's data strategy and the implemtation of its AI strategy.

- **Perspectives**
  - Further evaluations with quantitative results incoming
  - Interraction with knowledge graphs
  - Tests with bigger models (trade-off between model size & performance to define)

Journée GdR TAL, 7 septembre 2023, Nantes