



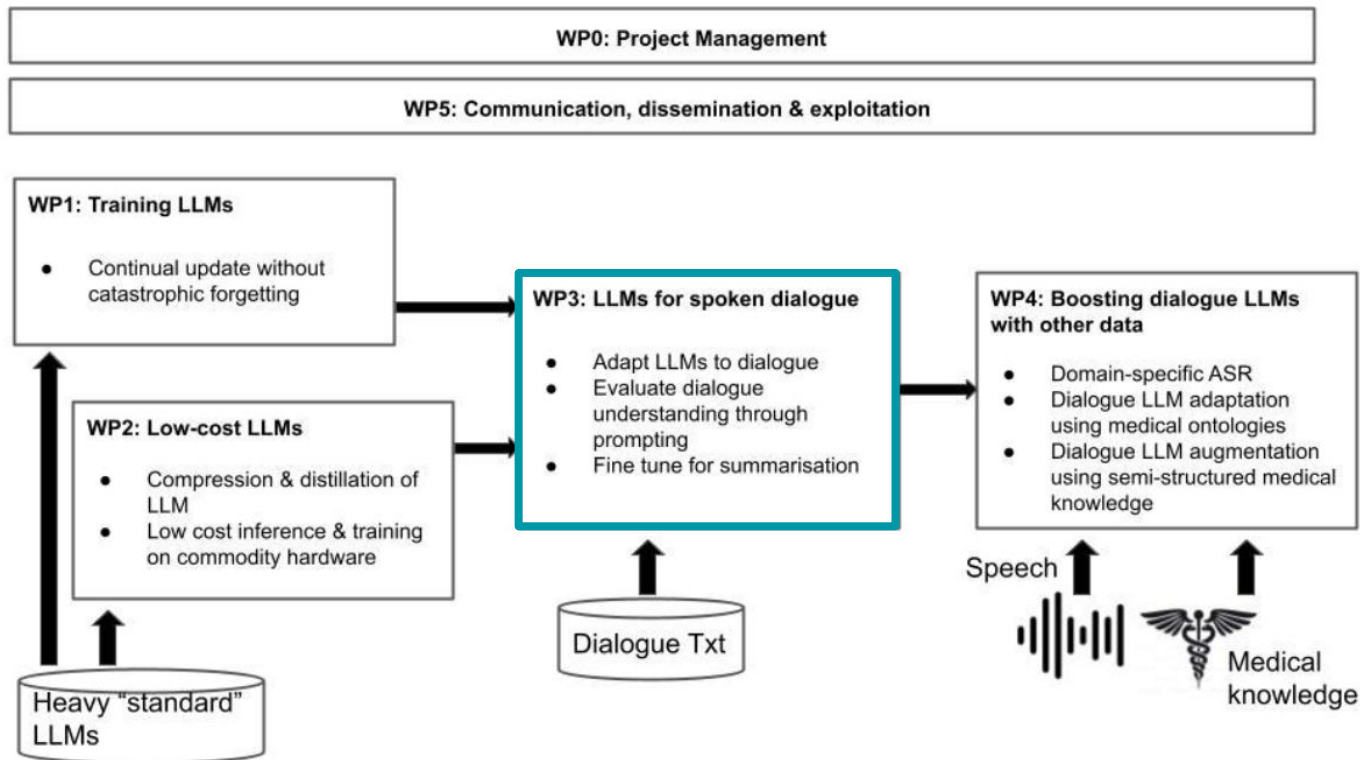
LLM4All WP3: LLMs for spoken dialogue

Julie Hunter, Jérôme Louradour, Damien Simon, Virgile Rennard

05/09/2024

LLM4All Year 1 meeting - 10/23/2024

Shifting to spoken dialogue



Two dimensions

- **Evaluate impact of continued training on spoken dialogue:** Will continuing pretraining on dialogue data make for a better foundation model for downstream tasks requiring dialogue understanding?
- **Evaluate impact of model size on summarization and querying:** Is one task harder for smaller models than the other?

Adaptation to dialogue

Context: oral conversation

A natural means of communication

meetings, medical appointments,
client service calls, lectures, webinars

...but with a special character

- Spontaneity: disfluencies, sparse language, specific vocab and syntax
- Multiparty: need to follow and understand dialogue interactions

[speaker001:] Ok, c'est quoi le plus, le souvenir, le premier souvenir, le plus clair que tu as en tête, de nous deux ?

[speaker002:] Le souvenir ?

[speaker001:] Ton premier souvenir de nous en fait.

[speaker002:] Le premier souvenir de nous le plus clair que j'ai, euh, je pense que c'est en Algérie. Je crois que c'est en Algérie. Et je sais pas, si tu te souviens. On était sur la place. Et, euh, et on courait. On jouait, comme des...

A pretrained dialogue model

Train a model specifically for dialogue in hopes of improving downstream:

- Comprehension: needed for summarization, querying and analysis
- Generation: needed for natural exchanges with dialogue assistants

Meeting and
conversation
summarization

Querying of
transcripts

Analysis of
medical calls and
documents

Language-specific training

- Mimic conversational style
- Target a specific (non-English) language
- Improve culture-specific understanding and responses
- Use limited but well-curated data

Pour faire un boeuf
bourguignon il faut

commencer par un bon vin.
To make a beef
bourguignon you must
start with a good wine.

- Julia Child

####JULIA CHILD

General approach

1. Start with a pretrained language model (falcon, llama, mistral, ...)
2. Continue training with dialogue-specific data

Two new models

claire-falcon

claire-mistral

[Intervenant 1:] il y a été combien de temps à l'école maternelle ?

[Intervenant 2:] lui il a commencé à trois ans et demi il était il est rentré à la grande école à cinq ans et demi du fait qu'il est de du mois de décembre

[Intervenant 1:] oui et l'année dernière quand vous l'avez fait inscrire il était encore à l'école maternelle ?

[Intervenant 2:] ah non il était à déjà à sa première année de grande école

Corpora for French dialogue

Conversation type	Words	Sampling Weight	Corpus
Parliamentary Proceedings	135 M	35%	Assemblée Nationale
Theater	16M	18%	Théâtre Classique, Théâtre Gratuit
Interviews	6.4M	29%	TCOF, CFPP, CFPB, ACSYNT, PFC, Valibel
Free conversation	2.2M	10%	CRFP, OFROM, CID, Rhapsodie, ParisStories, PFC, CLAPI, C-ORAL-ROM, ESLO
Meetings	1.2M	5%	SUMM-RE, LinTO, ORFEO
Debates	402K	< 2%	FreD, ESLO
Assistance	159k	< 1%	ORFEO, UBS, OTG, ESLO
Presentations	86k	< 0.5%	Valibel, LinTO, ESLO

Data preparation

Heterogeneous data

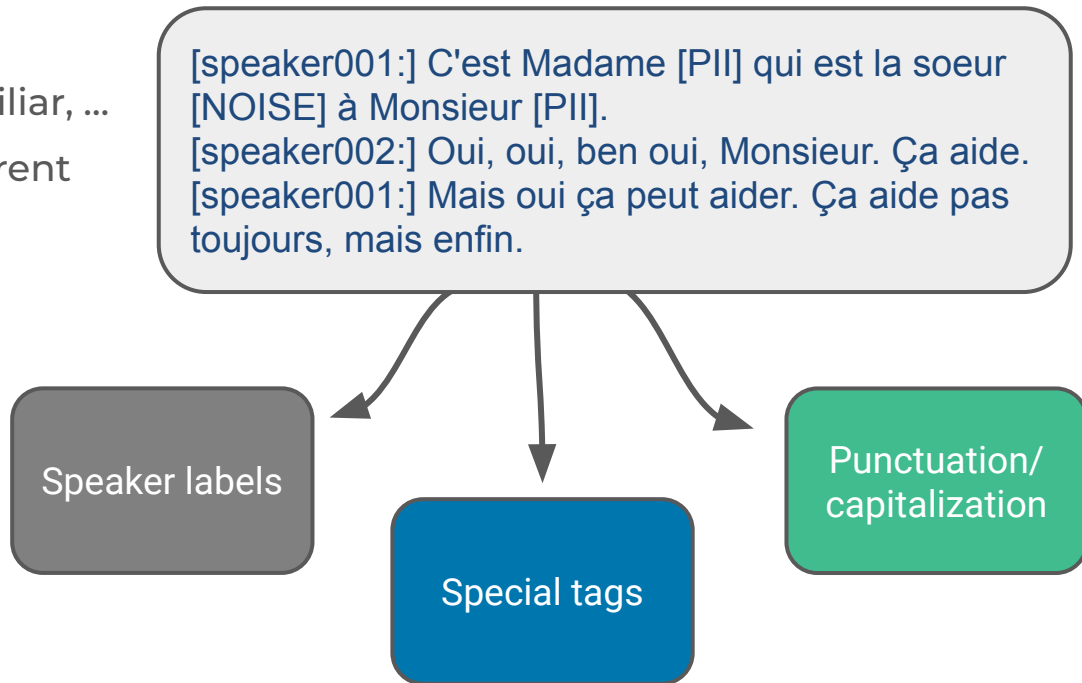
- More or less conversational, familiar, ...
- With/without punctuation, different annotation conventions

Standardisation

- Diarization
- Anonymization
- Cleaning

Chunking (max 2048/4096 tokens)

- Around speaker turns



Data augmentation

Random Speaker names, Case, Punctuation marks

[speaker001:] C'est Madame [PII] qui m'envoie.
[speaker002:] Oui... [NOISE] Que veut-elle?
[speaker001:] Savoir où en est son contrat!

Normalization

[Intervenant 1:] C'est Madame **Ronald** qui m'envoie.
[Intervenant 2:] Oui... **■** Que veut-elle?
[Intervenant 1:] Savoir où en est son contrat!

[**Michelle Tate**.:] C'est Madame Ronald qui m'envoie.
[**Lawrence Foster**.:] Oui... Que veut-elle?
[**Michelle Tate**.:] Savoir où en est son contrat!

- C'est Madame Ronald qui m'envoie.
- Oui... Que veut-elle?
- Savoir où en est son contrat!

[Michelle Tate:] **c'est madame ronald** qui m'envoie.
[Lawrence Foster:] **oui... q**ue veut-elle?
[Michelle Tate:] **s**avoir où en est son contrat!

- **c'est madame ronald** qui m'envoie.
- **oui... q**ue veut-elle?
- **s**avoir où en est son contrat!

[Intervenant 1:] **c'est madame ronald** qui m'envoie.
[Intervenant 2:] **oui... q**ue veut-elle?
[Intervenant 1:] **s**avoir où en est son contrat!

[Michelle Tate:] c'est madame ronald qui m'envoie **■**
[Lawrence Foster:] oui **■** que veut-elle **■**
[Michelle Tate:] savoir où en est son contrat **■**

- c'est madame ronald qui m'envoie **■**
- oui **■** que veut-elle **■**
- savoir où en est son contrat **■**

[Intervenant 1:] c'est madame ronald qui m'envoie **■**
[Intervenant 2:] oui **■** que veut-elle **■**
[Intervenant 1:] savoir où en est son contrat **■**

Data augmentation

Speaker anonymization

[Jean:] C'est Madame [PII] qui m'envoie.
[Marie:] Oui... [NOISE] Que veut-elle?
[Jean:] Savoir où en est son contrat!

Normalization

[Jean:] C'est Madame **Ronald** qui m'envoie.
[Marie:] Oui... **■** Que veut-elle?
[Jean:] Savoir où en est son contrat!

[Intervenant 1:] C'est Madame Ronald qui m'envoie.
[Intervenant 2:] Oui... Que veut-elle?
[Intervenant 1:] Savoir où en est son contrat!

- C'est Madame Ronald qui m'envoie.
- Oui... Que veut-elle?
- Savoir où en est son contrat!

[Intervenant 1:] **c'est madame ronald** qui m'envoie.
[Intervenant 2:] **oui... que** veut-elle?
[Intervenant 1:] **savoir** où en est son contrat!

- **c'est madame ronald** qui m'envoie.
- **oui... que** veut-elle?
- **savoir** où en est son contrat!

[Jean:] **c'est madame ronald** qui m'envoie.
[Marie:] **oui... que** veut-elle?
[Jean:] **savoir** où en est son contrat!

[Intervenant 1:] c'est madame ronald qui m'envoie **■**
[Intervenant 2:] oui **■** que veut-elle **■**
[Intervenant 1:] savoir où en est son contrat **■**

- c'est madame ronald qui m'envoie **■**
- oui **■** que veut-elle **■**
- savoir où en est son contrat **■**

[Jean:] c'est madame ronald qui m'envoie **■**
[Marie:] oui **■** que veut-elle **■**
[Jean:] savoir où en est son contrat **■**

The Claire Models

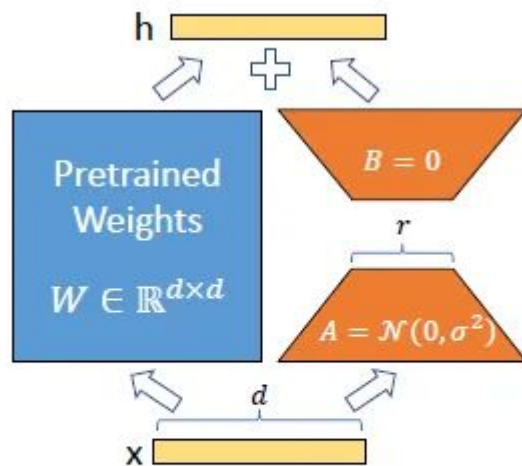
Training

Base model:

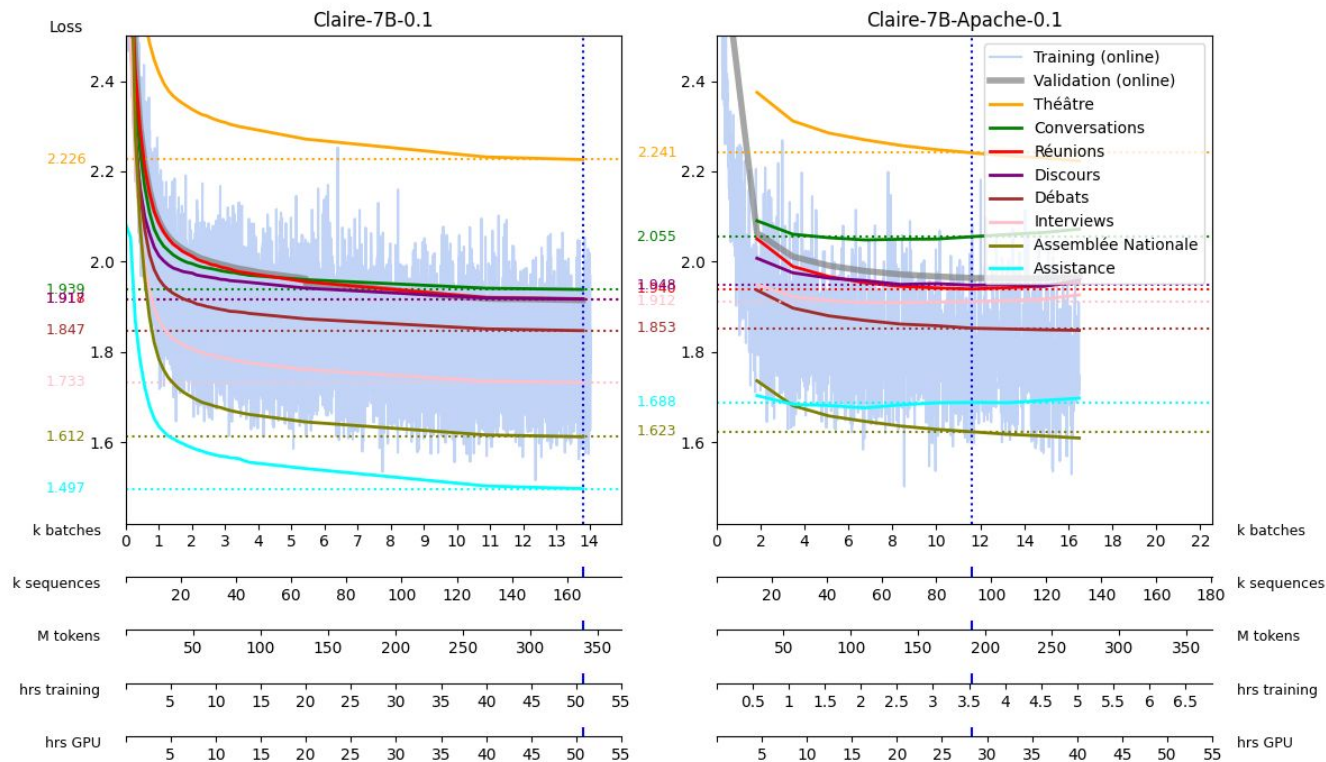
- falcon-7b: non GAFAM, open-source
- mistral-7b: same but released later
- without instruction-tuning

Training:

- LoRA
- Multi-GPU: FSDP
- Jean-Zay: 1 to 8 A100s
- 1 hour for 8M tokens



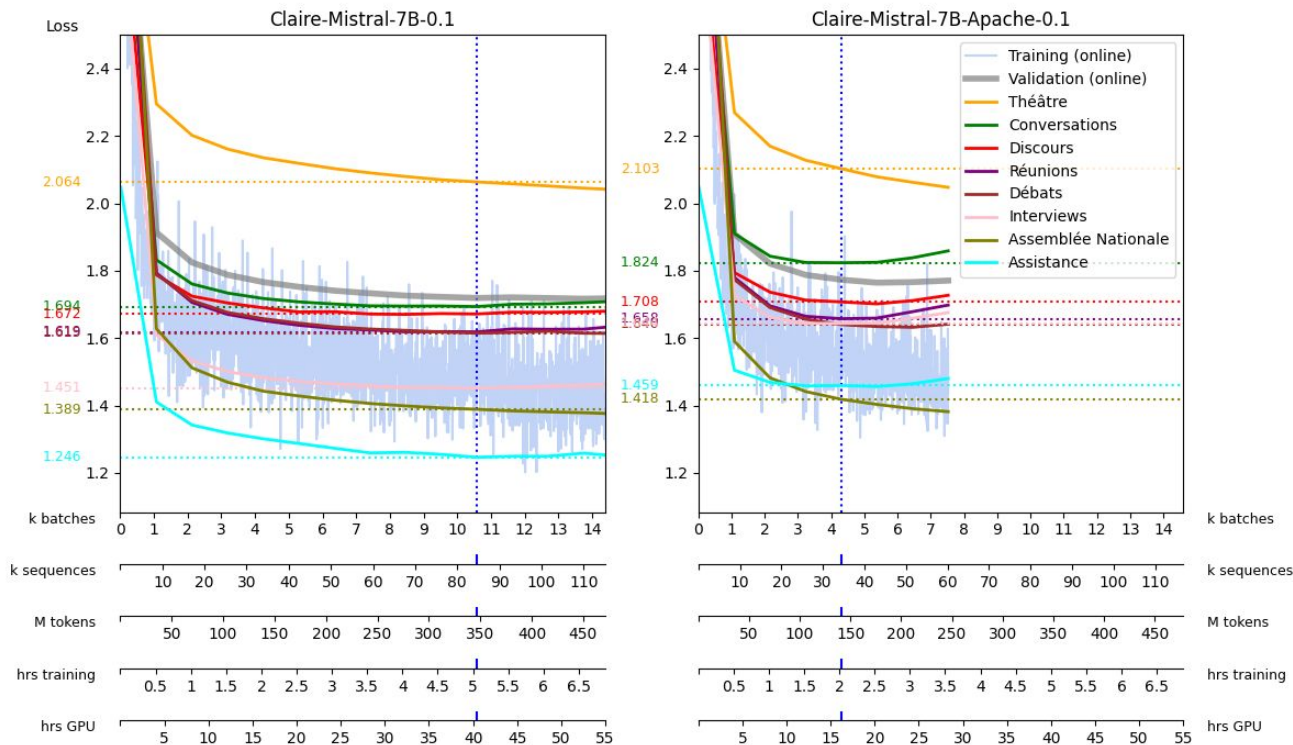
Convergence (online & offline)



Claire-Falcon : 50h × 1GPU

Claire-Falcon (Apache): 5h × 8GPU

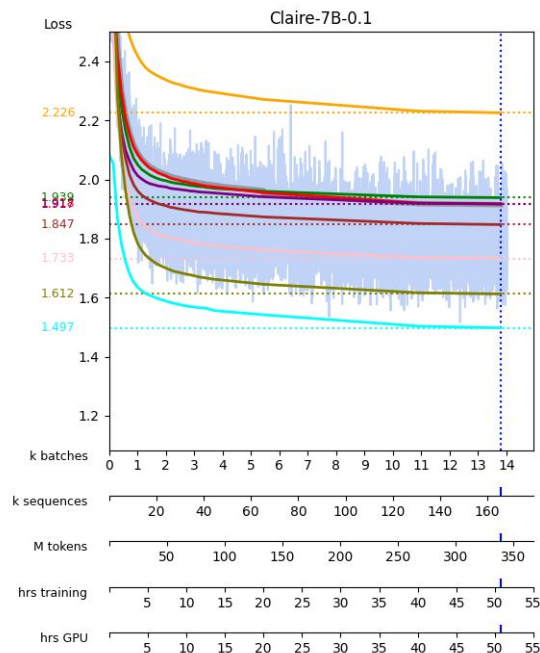
Convergence (online & offline)



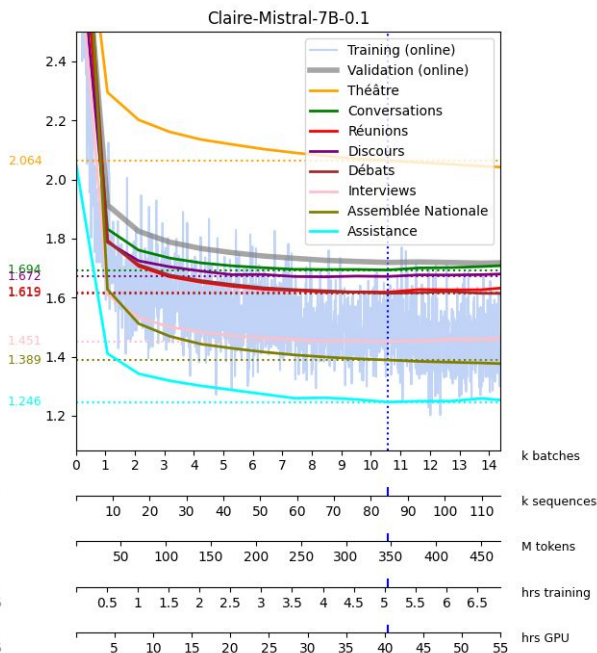
Claire-Mistral : 14h × 1GPU

Claire-Mistral (Apache): 7h × 8GPU

Comparing models



Claire-Falcon



Claire-Mistral

Model perplexity:

Is it a good indicator of model quality?

- different tokens
- different context length

Human evaluation

Three categories (13 questions):

1. **Interaction:** can the model capture the kind of back and forth exchanges that we have in conversation and make it look like two people are trying to talk to each other?
2. **Fluency:** is the language normal for spoken language and is it human-like?
3. **Relevance:** does it directly engage with the prompt and develop a coherent conversation?

Example of an “Interaction” question

En cas de dialogue avec plusieurs échanges, semble-t-il que les interlocuteurs cherchent à engager une discussion (en se tutoyant/vouvoyant directement, répondant aux questions, utilisant des expressions conversationnelles comme 'oui, c'est vrai', etc.) ?

☐ Oui

☐ Non

☐ Il n'y a pas d'échange

Human evaluation

“Culture” prompts

Prompt: Dis-moi, où est-ce que je peux trouver des cèpes ?

- En général, on les trouve un peu avant la Toussaint.
- Et comment on les reconnaît, ces champignons?
- Ils sont très différents.
- Comment tu peux le savoir?
- Parce que, en fait, je suis champignon.
- Oh, dis donc.
- Oui, c'est moi qui t'ai parlé.
- Tu es vraiment un champignon?
- Mais oui.
- Tu es un champignon et tu es un homme?
- Je suis un homme et je suis un champignon.

“Casual” prompts

Prompt: [Dominique:] Dis moi, quels sont tes hobbies ou centres d'intérêt ?

[Camille:] J'adore la lecture. Ça me permet de m'évader.

[Dominique:] Tu peux me parler de tes lectures?

[Camille:] Je n'ai pas trop de centre d'intérêt. Je lis tout, de la science-fiction à la poésie, des thrillers aux livres historiques. Je me laisse pas enfermer dans une seule catégorie. Ça m'arrive de lire un bouquin et après de passer à un autre qui m'intéresse tout autant.

[Dominique:] Ok, et tu as une bibliothèque?

[Camille:] Non. Je lis en ligne.

Human evaluation

Two rounds (80 examples tested each time):

1. monologue vs dialogue; culture prompts
2. different dialogue styles; culture + casual prompts

Main takeaways:

- **Claire** > **Claire Mistral** > **Mistral**
- **Claire** > **Falcon** for **culture prompts**
- **Claire** >= **Falcon** for dialogue prompts
- **Claire Mistral** far above **Mistral** – **Claire Mistral** sticks to French!

Links, Demo

- Model(s)

Main model (license CC-BB-NC-SA)	huggingface.co/OpenLLM-France/Claire-7B-0.1	 
Main model in GGUF formats	huggingface.co/TheBloke/Claire-7B-0.1-GGUF	
Variant model, license Apache	huggingface.co/OpenLLM-France/Claire-7B-Apache-0.1	
Demo (simulated chat) 	huggingface.co/spaces/OpenLLM-France/Claire-Chat-0.1	

- Dataset & Code

Full dataset	huggingface.co/datasets/OpenLLM-France/Claire-Dialogue-French-0.1	
Paper (about the dataset)	arxiv.org/abs/2311.16840 “The Claire French Dialogue Dataset” huggingface.co/papers/2311.16840 (with links to related assets)	
Databases survey	github.com/OpenLLM-France/Claire-datasets	 
Code to train	github.com/OpenLLM-France/Lit-Claire	

Data for prompting and fine-tuning

Two tasks

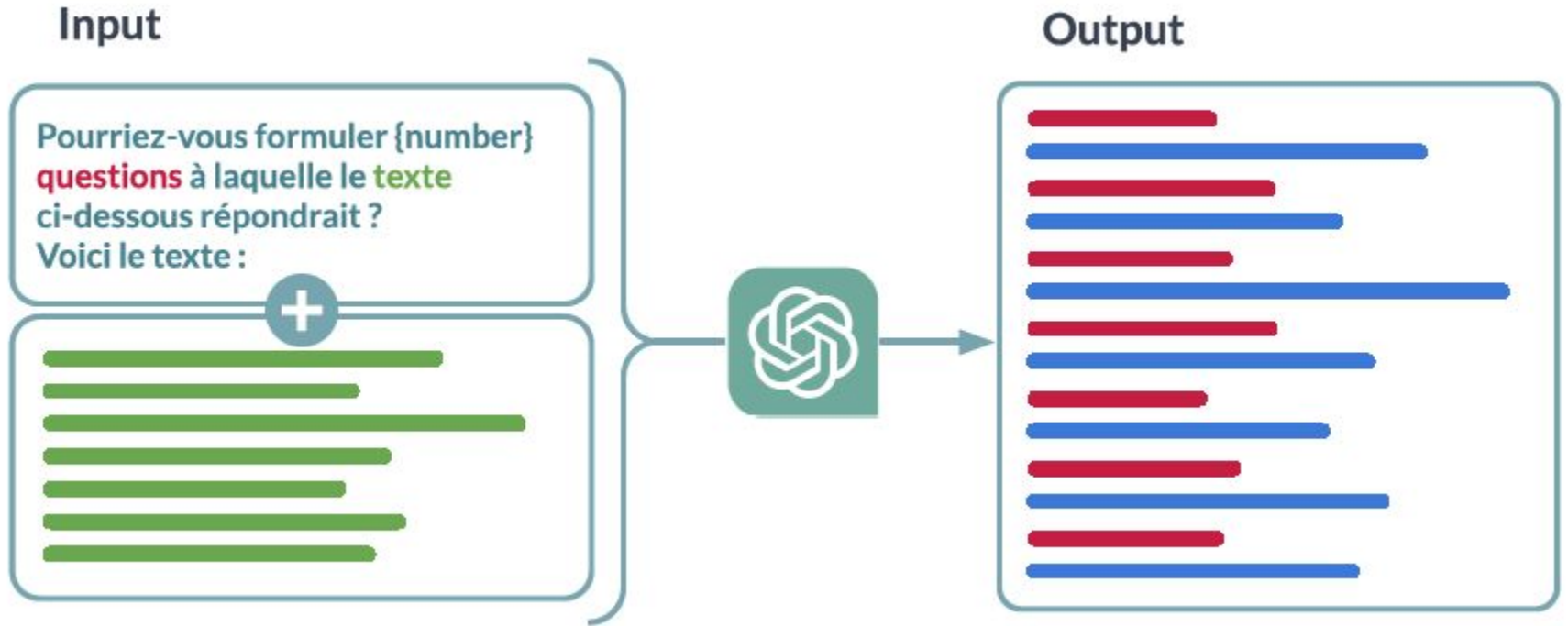
1. **Abstractive summarization**: locate and synthesize the most important points of a conversation
2. **Query/response**: ask the model questions about a conversation transcript

The FREDSum Corpus

Some FREDSUM characteristics :

- **144** Transcriptions of French political debate
- **432** abstractive summary (3 per debate)
- **288** extractive summary (2 per debate)
- Community files for mapping abstractive sentences to sets of extracted sentences

Query/response generation



An example

1. Quel est le projet de François Hollande concernant l'indexation du SMIC ?

Il souhaite indexer le SMIC sur la croissance.

2. Comment François Hollande envisage-t-il de baisser les dépenses contraintes ?

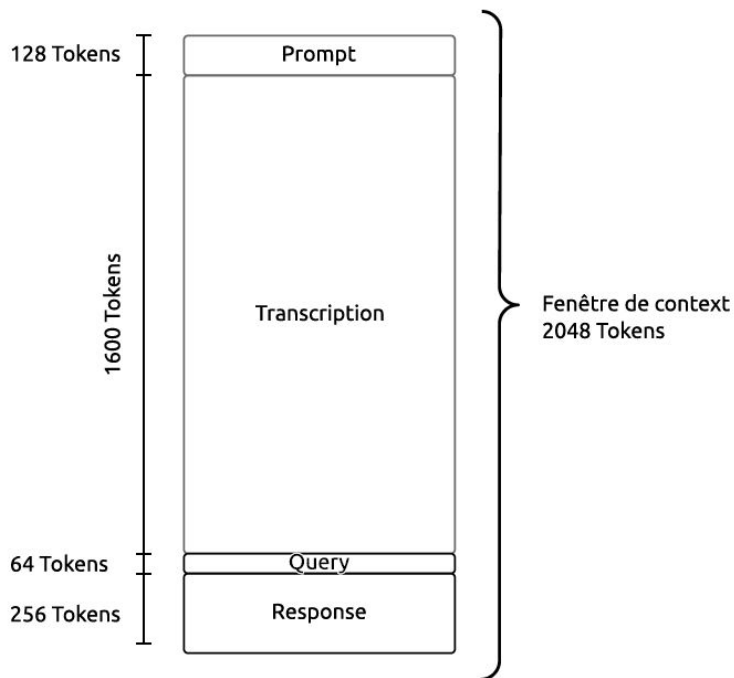
Il souhaite baisser les dépenses contraintes en instaurant un forfait de base pour que les consommateurs paient le même prix sur un volume de leur consommation, avec une progressivité au-delà.

The summary (just in case)

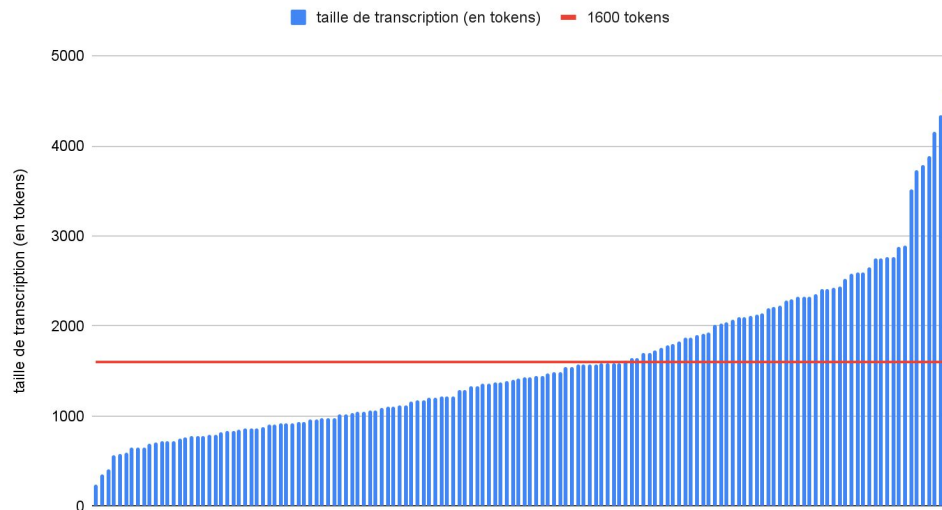
Dans ce débat, François Hollande et Nicolas Sarkozy discutent du pouvoir d'achat. François Hollande explique qu'il souhaite indexer le SMIC sur la croissance, et baisser les dépenses contraintes. François Hollande explique vouloir instaurer un forfait de base pour que les consommateurs paient le même prix sur un volume de leur consommation, et qu'il y ait une progressivité au delà. François Hollande souhaite mettre en place la TIPP flottante sur le prix des carburants afin que l'état ne gagne pas d'argent sur la hausse des prix. François Hollande souhaite aussi augmenter l'allocation de rentrée scolaire. Nicolas Sarkozy explique que la France a déjà payé les heures de chômage pareil, et que les heures supplémentaires sont la meilleure manière d'augmenter le pouvoir d'achat, alors que François Hollande souhaite les supprimer. Nicolas Sarkozy critique l'augmentation du SMIC, qui n'aiderait pas la majeure partie des gens. Nicolas Sarkozy explique qu'il est impossible de bloquer le prix de l'essence que l'on achète ailleurs, et que tous blocage sera payé par le contribuable. Nicolas Sarkozy explique que la baisse du prix du carburant est trop coûteuse, et qu'il est mieux d'encourager les Français à acheter des voitures qui consomment moins. Nicolas Sarkozy critique aussi le coût des réformes de François Hollande. François Hollande explique qu'il est obligatoire d'aider les Français qui vivent dans des départements ruraux avec le prix des carburants, et que la TIPP flottante ne coûtera rien à l'état. Nicolas Sarkozy critique aussi le souhaite de fermeture des centrales nucléaires de François Hollande, qui est la meilleure manière d'économiser sur l'énergie. François Hollande explique ne vouloir fermer qu'une seule centrale sur son quinquennat.

Context window limitations

Some FREDSUM transcriptions are too big for 2048 token size context window



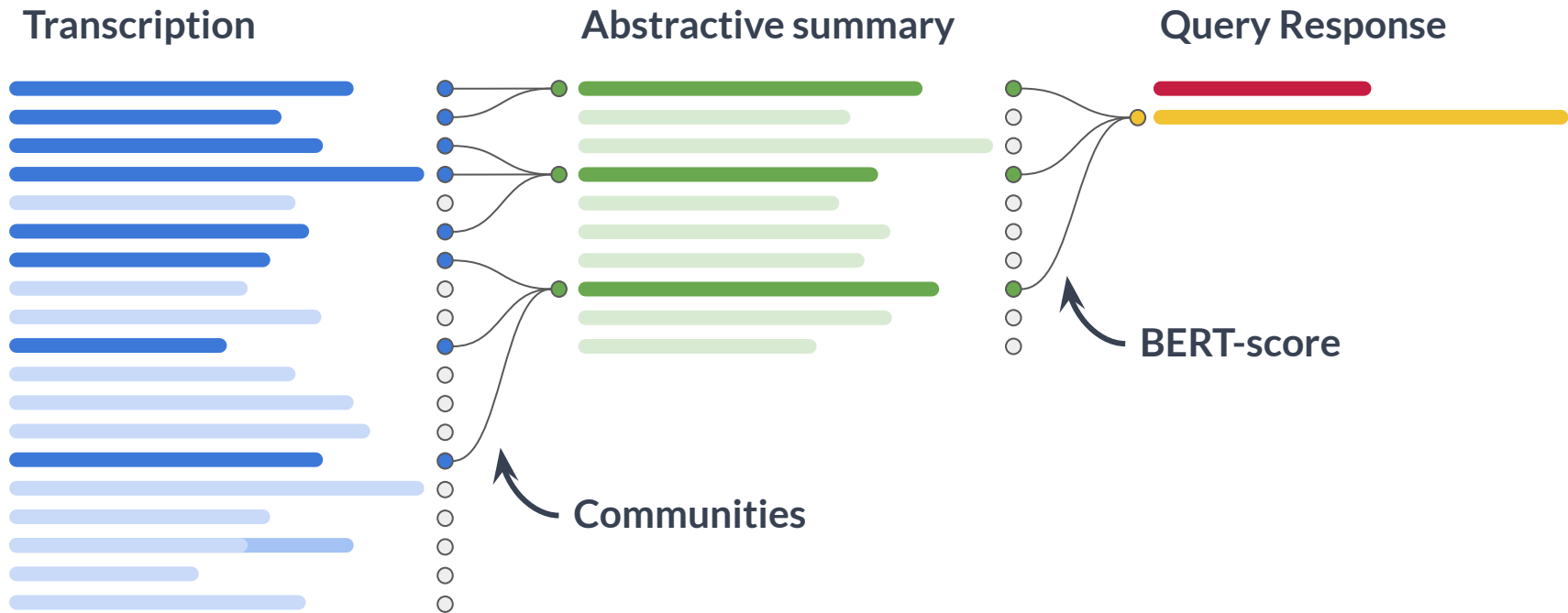
Taille des transcriptions de Fredsum en Tokens



Context window solution

Context windows issue :

Step 1 : How to select the parts of the transcript that contain the elements needed to answer the question ?



First results

Comparison - Mistral FT, Claire FT

	R1	R2	RL	Bertscore
Mistral FT	0.26	0.11	0.13	0.65
Claire FT	0.21	0.10	0.10	0.64

“Clémentine Autain commence par exprimer son soutien à l’augmentation des salaires des enseignants, souhaitant les relever de 10 points. Elle explique que le pouvoir d’achat des enseignants a considérablement diminué sur les 20 dernières années. Jordan Bardella explique de son côté que l’état s’effondre, et que l’école s’effondre avec.”

Comparison - Mistral IFT, Claire IFT

	R1	R2	RL	F1
Mistral	0.16	0.04	0.12	0.62
Mistral FT	0.38	0.22	0.34	0.75
Claire FT	0.38	0.21	0.33	0.76

CLAIRE-M FT : Autain_Attal_Bardella-Partie_6_1_Education

QUERY :

Quelle est la proposition de Clémentine Autain pour revaloriser les enseignants ?

TARGET :

- Clémentine Autain propose d'augmenter de dix points le salaire des enseignants en finançant cela par une augmentation des impôts des 10% des plus riches.

RESPONSE :

- Elle souhaite revaloriser les enseignants de 10 points.
- Elle souhaite financer cette revalorisation en augmentant l'impôt des 10% les plus riches

MISTRAL FT : Debat_Zemmour_Melenchon-Partie_2_1_Baisse_Du_Co2

QUERY :

Quelle alternative à l'énergie nucléaire propose Jean Luc Mélenchon ?

TARGET :

- Il propose de sortir progressivement du nucléaire et de travailler sur les énergies renouvelables comme les hydroliennes et l'énergie géothermique.

RESPONSE :

- Jean Luc Mélenchon propose d'abandonner le nucléaire et de passer à l'énergie géothermique, de la sobriété et de l'éolien.

Without fine-tuning

MARINE LE PEN était l'invitée de BFMTV ce jeudi matin. Elle était invitée sur le plateau de BFMTV pour discuter de la laïcité. MARINE LE PEN a répondu à la présidente du Front national sur le voile dans l'espace public.

Conclusions

Conclusions

- ✓ Dialogue pretraining dataset
- ✓ Dialogue model (continued pretraining)
- ✓ Q/R dataset for FREDSUM

Next steps



Q/R + summaries for SUMM-RE (more disfluencies)



Evaluation on smaller models (FREDSum)



Claire training of smaller model?