Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Claire: Large Language Models for Spontaneous French Dialogue

Jérôme Louradour[1] Julie Hunter[1] Ismaïl Harrando[1]
Guokan Shang[2] Virgile Rennard[1] Jean-Pierre Lorré[1]
(1) LINAGORA, Paris and Toulouse
(2) MBZUAI, Paris
{jlouradour, jhunter, iharrando, vrennard, jplorre}@linagora.com,
guokan.shang@mbzuai.ac.ae

RÉSUMÉ

Nous présentons la famille de modèles Claire, une collection de modèles de langage conçus pour améliorer les tâches nécessitant la compréhension des conversations parlées, tel que le résumé de réunions. Nos modèles résultent de la poursuite du pré-entraînement de deux modèles de base exclusivement sur des transcriptions de conversations et des pièces de théâtre. Aussi nous nous concentrons sur les données en français afin de contrebalancer l'accent mis sur l'anglais dans la plupart des corpus d'apprentissage. Cet article décrit le corpus utilisé, l'entraînement des modèles ainsi que leur évaluation. Les modèles, les données et le code qui en résultent sont publiés sous licences ouvertes, et partagés sur Hugging Face et GitHub.

ABSTRACT

We present the Claire family of language models, a collection of foundation models designed to serve as the basis for further fine-tuning on downstream tasks, such as meeting summarization, that require understanding of spoken conversation. Our models result from continuing the pretraining of two foundation models exclusively on conversation transcripts and theater plays. We focus on French dialogue data in an effort to offset the English-heavy focus of much training corpora. This paper describes the data sets and their preparation, as well as model training and evaluation. The resulting models, data and code are released under open licenses and shared on Hugging Face and GitHub.

MOTS-CLÉS : Modèles de langue, Français, Dialogue, Parole spontanée, Pré-entrainement.

KEYWORDS: Language models, French, Dialogue, Spontaneous speech, Continual pretraining.

1 Introduction

A lot of information is shared through spoken conversation – in meetings, medical appointments, assistance calls, lectures, to give just a few examples. Due to the spontaneous nature of such interactions, the kind of language we employ in these contexts can differ considerably from that found in text documents used to train large language models (LLMs). We hesitate, repeat ourselves, revise our wording in mid-sentence, leading to utterances that are ungrammatical by written standards. We might employ slightly different vocabulary and even syntactic constructions.

(1) A : Ok, c'est quoi le plus, le souvenir, le premier souvenir, le plus clair que tu as en tête, de nous deux ?
B : Le souvenir ?
A : Ton premier souvenir de nous en fait.
B : Le premier souvenir de nous le plus clair que j'ai, euh, je pense que c'est en Algérie. Je crois que

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 530–548.

c’est en Algérie. Et je sais pas, si tu te souviens. On était sur la place. Et, euh, et on courait.

In (1), it takes multiple lines and two turn changes for the speakers to establish that the question under discussion is simply, “Quel est ton premier souvenir de nous ?”. A reformulates their question multiple times in the first line, leading to constructions such as “le souvenir le premier souvenir” that we would not expect to find in a grammar book. There are filled pauses with expressions like “euh”, which do not appear in written text. And we have focus constructions such as “Le premier souvenir de nous le plus clair que j’ai, je pense que c’est...” that are characteristic of spoken language. A further contrast with most written documents is that conversation involves multiple speakers.

One might hypothesize that a foundation model that has been trained to be sensitive to the kinds of interactions that we see in spontaneous conversations might serve as a better base for downstream fine-tuning on language generation and understanding tasks that focus on natural conversation. A model for meeting summarization or transcript querying, for example, would need to be able to navigate these types of interactions. It might be interesting to create chatbots that could produce more natural and informal language to improve the feel of interaction.

In this paper, we describe the first step that we took to addressing our hypothesis : the creation of foundation models trained on transcripts of spontaneous conversation and theater plays. Given what would inevitably be a small dataset, we decided to start with foundation models and continue pretraining—that is, to continue training with an autoregressive language modeling objective—rather than training dialogue models from scratch. We were curious to see how the additional French training impacted the quality of French responses and also how conversational we could get our models to be without specific chat-style training. For the latter reason, we chose pure foundation models that had not received specific instruction or chat training. The result was two models : Claire-7B, based on Falcon-7B, [1] and Claire-Mistral-7B, based on Mistral-7B.

We also decided to focus on French. A drawback of many current LLMs is that most of their training data is in English, which is a double cause for concern. Not only might these models struggle to produce high-quality responses in other languages, but they will also have a poor representation of non-anglophone cultures. Consider this example of Mistral-7B [2] completion :

prompt :	Pour faire un boeuf bourguignon il faut
Mistral completion :	commencer par un bon vin. To make a beef bourguignon you must start with a good wine. – Julia Child

What we see is that the first line of the response is in French but then the model quickly switches to English—a phenomenon that we observed frequently with Mistral (see Section 5). What’s more, the response cites Julia Child—a chef made famous in the States for introducing French cuisine to an American audience, but hardly known in France. Ideally, a French speaker who turns to an LLM to ask a question about French culture (such as French cuisine !) would be able to receive a response in French from a French source.

The resulting Claire models are quite conversational in French and even have a touch of humor. This is an example output from Claire [3] :

1. https://huggingface.co/tiiuae/falcon-7b
2. https://huggingface.co/mistralai/Mistral-7B-v0.1
3. Because all of the models that we discuss are 7B size, we drop the “7B” mention for the rest of the paper.

prompt :	– Dis-moi, où est-ce que je peux trouver des cèpes ?
Claire completion :	– 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.

We emphasize that our models are foundation models trained to produce extended sequences of dialogue in one go, not to engage in back and forth chat with a human, and that it is expected for them to imitate spoken dialogue features, including disfluencies.

In what follows, we present our dataset and data preparation approach as well as an overview of our model training and evaluation approach.

2 Data

While the majority of data used to train high-profile LLMs comes from English (LLaMA2, for example, uses 89.7% English data Touvron *et al.* (2023)), the large-scale datasets tend to include some data from other languages, especially web-crawled data. ROOTS (Responsible Open-science Open-collaboration Text Sources Laurençon *et al.*, 2023), which BigScience assembled for training BLOOM (Scao *et al.*, 2023), contains 1.6TB of data from 59 languages, including French.[4] RefinedWeb (Penedo *et al.*, 2023), OSCAR (Suárez *et al.*, 2019), and RedPajama (Computer, 2023) are filtered and refined versions of CommonCrawl data dumps[5] prepared specifically for LLM training. These datasets do not concentrate on dialogue data as we did for our models, however.

On the dialogue side, there have been recent efforts to assemble collections of dialogue datasets that can be used to train conversational AI agents, though these focus on English. DialogStudio (Zhang *et al.*, 2023) regroups some well-known English spoken dialogue corpora, such as AMI (McCowan *et al.*, 2005), ICSI (Janin *et al.*, 2003) and MediaSum (Zhu *et al.*, 2021). It also contains a variety of short, written dialogues collected through crowdsourcing. Other similar but smaller-scale collections include InstructDial (Gupta *et al.*, 2022) and ParlAI (Miller *et al.*, 2017).

Our dataset (see Hunter *et al.* (2023) and https://huggingface.co/datasets/OpenLLM-France/Claire-Dialogue-French-0.1) is broken down in Table 1. It benefited considerably from other open-data initiatives such as *Ortolang*[6] (Plate-forme d'outils et de ressources linguistiques pour un traitement optimisé de la langue française) and Orféo (Outils et Ressources sur le Français Ecrit et Oral)[7]. Projects such as the CEFC[8] (Corpus d'Etude pour le Français Contemporain) and Parole Publique

4. https://huggingface.co/bigscience-data
5. https://commoncrawl.org/
6. https://www.ortolang.fr/fr/accueil/
7. Orféo platform : http://ortolang107.inist.fr/
8. https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/home/index.html

(Nicolas *et al.*, 2002) also propose both text and oral corpora, and often offer standardized formats.

Conversation type	**M Words**	**Samp. Weight**	**Constituent datasets**
Parliamentary Proceedings	135	35 %	Assemblée Nationale
Theater	16	18 %	Théâtre Classique, Théâtre Gratuit
Interviews	6.4	29 %	TCOF, CFPP, CFPB, ACSYNT, PFC, Valibel
Free conversation	2.2	10 %	CRFP, OFROM, CID, CLAPI, Rhapsodie, ParisStories, PFC, C-ORAL-ROM, ESLO
Meetings	1.2	5 %	SUMM-RE, LinTO, ORFEO
Debates	0.402	< 2 %	FreD, ESLO
Assistance	0.159	< 1 %	ORFEO, UBS, OTG, ESLO
Presentations	0.086	< 0.5 %	Valibel, LinTO, ESLO
Total	160	100 %	

TABLE 1 – A breakdown of sub-corpora in our training dataset and their sources. The sub-corpora are organized into categories and listed in order of size starting from the largest, at 135 million words, to the smallest, at 86 thousand words. Citations for the original datasets are in Appendix C.

As reflected in Table 1, we grouped sub-corpora of our dataset into categories to reflect the types of interactions that we might expect them to contain. The categories are based on descriptions provided by those who distributed the corpora, where possible, and some manual verification by authors of this paper. While this approach is far from fool-proof, we hypothesized that having a rough categorization would not only help ensure that our model had seen a diverse distribution of interactions but would also offer the possibility of selecting portions of our dataset that might be more pertinent than others for a specific downstream task. Note that dividing by category led us to split up certain corpora, such as ESLO, ORFEO, LinTO, and PFC.

For parliamentary proceedings, we expect the conversations to be slightly more formal and to include multiple participants. Theater plays imitate multiparty spoken conversation to some extent, but have their own polished, theatrical style. In interviews, we expect more question/answer sequences, whereas free conversations are generally unguided discussions. The meeting category includes both real and simulated meetings and are expected to contain more structured interactions, though arguably less structured than those found in debates, where turns are controlled to a large extent by a moderator and questions are posed in a formal manner. Finally, assistance interactions involve one person asking for information from another in a professional context while presentations involve longer monological sequences, frequently accompanied by question/answer sequences.

The second column of Table 1 reflects the number of words in the corpora for each category before data augmentation, and the third column shows the sampling weights that we chose for each category in order to balance the different types of interactions seen by our models. Each sampling weight represents the chance of taking a random sequence from the corresponding augmented and tokenized sub-corpus to constitute a training batch. The final weights were determined by assigning custom penalties to texts from parliamentary proceedings and from theater plays. These datasets make up around 95% of the diarized transcripts/scripts that we were able to collect ; penalizing them allowed us to increase the impact of less formal conversational data which, while more representative of what we were searching for, was also harder to find in the form of transcribed and diarized transcripts.

3 Data preparation and augmentation

As our datasets came from a variety of sources, we had to contend with diverse data formats, and preparing a high-quality dataset for training often required dataset-specific solutions. Some corpora came with punctuation, while others did not. Different corpora adopted different annotation conventions for background noises, laughter, and other sounds. We attempted to standardize these annotations using a few tags like [NOISE], as in (2), so that they could be easily found later if desired. The [PII] tag, which stands for "Personally Identifying Information," marks anonymized content.

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

For training, we opted to remove the [NOISE] labels and to replace the [PII] tags with random names,[9] yielding results like (3) :

(3) [speaker001 :] C'est Madame Ronald qui m'envoie.

Because part of our aim was to train a model to represent the interactive turn taking characteristic of conversation, it was necessary that all transcripts contain speaker labels. Recovering these labels was not always straightforward and sometimes required going into separate files to recombine this information with the transcript. We also standardized the style of the speaker label so that it could be easily identified for data augmentation, as explained below. In the end, we opted for a format in which speaker names were encased in square brackets and followed by a colon before the closing bracket as shown in example (2).

With our conventions in place for marking non-verbal sounds, anonymized content and speaker labels, we were able to exploit the square brackets to augment our data. When dealing with machine learning models, we always have a choice : we can normalize the data so that the model is insensitive to things it should not be sensitive to, or we can augment the data to increase model robustness, without adding a preprocessing, normalization brick. We chose the latter option.

First, we generated multiple variants for speaker labels by either enclosing the label in brackets, as described above and shown in (4-a) and (4-b), or by marking speaker changes with dashes as in (4-c).

(4) a. [Intervenant :] C'est Madame Ronald qui m'envoie.
b. [Michelle Tate :] C'est Madame Ronald qui m'envoie.
c. – C'est Madame Ronald qui m'envoie.

The bracketed variants contained either "Intervenant" or a proper name. The former was chosen as a replacement for "speaker" after preliminary tests with Falcon suggested that this label helped the model return responses in French. The option of using a dash (-) in conversations with only two speakers also came from experiments with Falcon and Mistral in which both models produced dialogues formatted in this way. For augmentation, we used the "Intervenant" labels to add an anonymous alternative for corpora that used names and we used randomly generated proper names to add a named alternative for datasets that were anonymized. The choice between using a first and last name or only a first name throughout a transcript added another dimension for augmentation.

Next, we played with changing case and removing punctuation. Ultimately, we wanted to design our models to be robust to different transcript formats, keeping in mind that these transcripts, many of which will be produced by ASR systems, may need to handle transcripts without case or punctuation.

9. While [PII] tags can indicate other types of censored content, such as addresses, our observations suggested that they usually indicate proper names and the majority of our datasets do not distinguish different types of censored content. We judged that the benefit of including this data outweighed the risk of replacing an address with a proper name during data augmentation.

Combining these three options for data augmentation—speaker labels, case and punctuation—left us with up to nine different formats that we could pull from for training. (In some cases, as when a dataset did not contain punctuation or case, we did not have all nine options of course.)

A final step in the data preparation pipeline involved cutting individual documents from our data set into smaller chunks that would fit into the context windows of Falcon and Mistral, which are limited to 2048 and 4096 tokens, respectively. To do this, we split tokenized documents so that each training sample begins with the start of a speech turn.

4 Model training

Because many of the original corpora in our training set were shared for research purposes only, we published Claire-Falcon and Claire-Mistral under a non-commercial license (CC BY-SA-NC 4.0). Simultaneously, however, we released two models under an Apache 2.0 license, Claire-Apache and Claire-Mistral-Apache, trained only on the datasets from our corpus that allow commercial use.

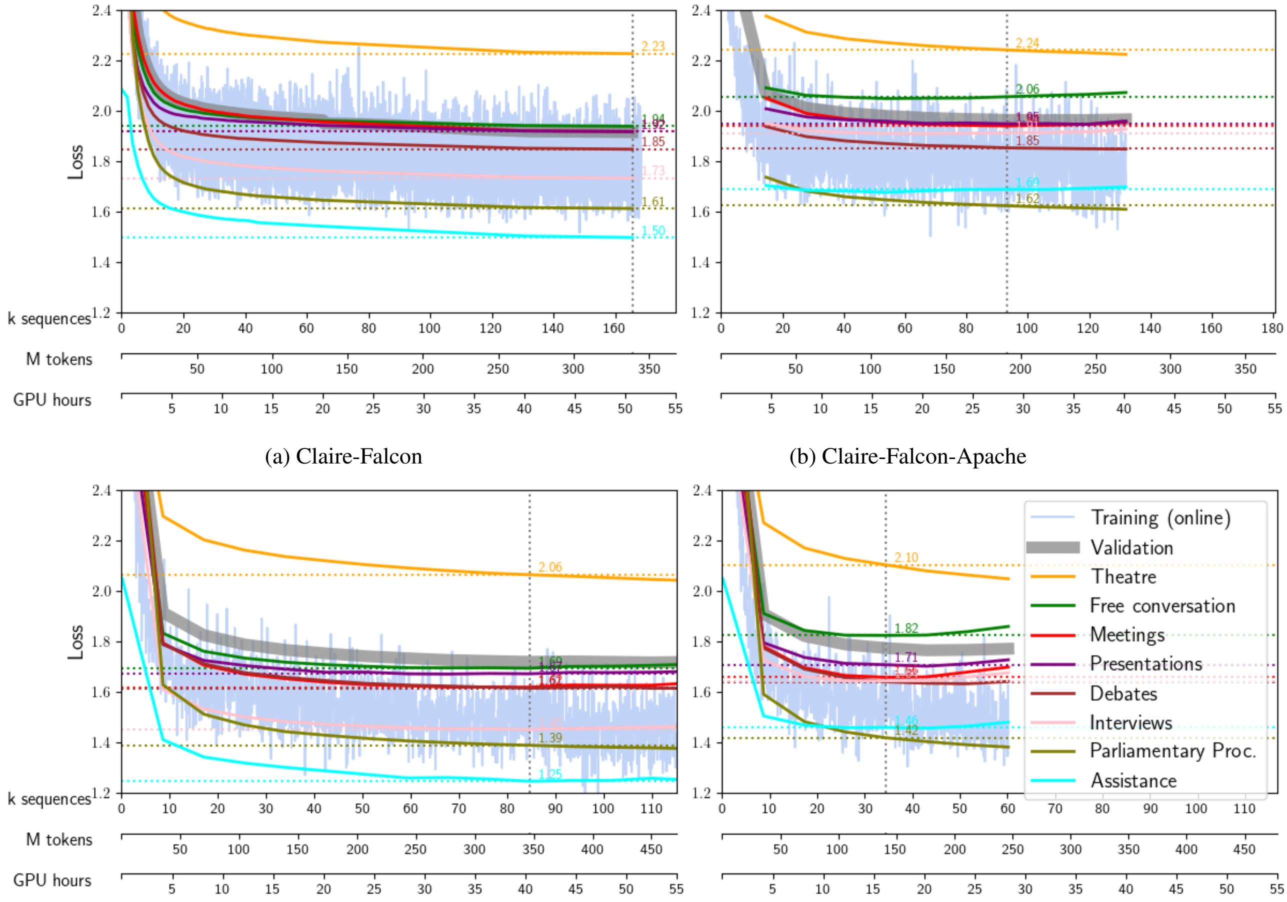
To train our Claire models, we used LoRA (Hu *et al.*, 2021) with bfloat16 precision. LoRA is a lightweight technique that greatly reduces the number of parameters to be trained, making training more efficient. Because training with LoRA enforces the change in model parameters to be of low rank, an additional advantage of this method is that it greatly reduces the chances of catastrophic forgetting, in which a pretrained model forgets what it learned from its previous training. While LoRA is often associated with fine-tuning, we note that we applied it to all layers of the original models and used it to continue unsupervised training with an auto-regressive objective, and so our task was closer to pretraining than traditional fine-tuning. Our hyperparameter configuration was : LoRA with rank $r = 16$ and $\alpha = 32$, AdamW optimizer with a learning rate of $1e^{-4}$, batch size of 128, dropout ratio of 0.05, weight decay regularization factor of 0.01, and gradient clipping of norm 1.

Training was carried out on the Jean Zay supercomputer run by GENCI (Grand Equipement National De Calcul Intensif) and installed at IDRIS, the national computer center put in place for the CNRS (Centre national de la recherche scientifique). To take advantage of Jean Zay's multi-GPU nodes, we used Fully Sharded Data Parallel (FSDP) (Xu *et al.*, 2020; Zhao *et al.*, 2023), which shards a model's parameters, gradients and optimizer states across different workers instead of making a full copy of these states on each GPU, greatly facilitating multi-GPU training. We trained with 8 A100 GPUs [10], each with 80GB of memory and processed tokens at a rate between 7 and 8 million tokens per hour.

Figure 1 shows the convergence curves for the different variants of the model. The monitored loss is cross-entropy, which is in the case of multi-class classification (with discrete targets that are text tokens here) equivalent to the average Negative Log-Likelihood. Note that perplexity of the model is then the exponentiated loss. Convergence curves show how fast that loss decreases on several hold-out validation subsets. The noisy background curve in light blue is the cost on the training samples estimated in real time during training.

The first scale indicates the number of sequences that the model has seen up to a certain point, where a sequence can be a whole conversation or a proper part of a conversation when context size limitations forced to split the conversation. The second scale shows the number of training (subword) tokens, including padding tokens that are ignored in the loss function : this total number of tokens simply equals the number of sequences multiplied by the training context size (2048 for Falcon, 4096 for

10. only Claire-Falcon was trained on a single GPU, with a slightly different batch size (132 instead of 128).



(a) Claire-Falcon (b) Claire-Falcon-Apache

(c) Claire-Mistral (d) Claire-Mistral-Apache

FIGURE 1 – Convergence curves for the different models.

Mistral). The last scale measures training GPU hours. With around 50H GPU hours, training a Claire model consumes a total energy around 25 kWh on the Jean Zay supercomputer, equivalent to around 1.5kg C02 eq following the estimation of (Luccioni *et al.*, 2022).

The thick grey line is the online validation curve that we used for early stopping ; that is, training is automatically stopped when the online validation loss does not decrease for some time. The colored lines show the loss curves for our data by subcategories. To get these results, we produced model checkpoints every hour and then tested the different types of data at the various checkpoints. The legend in the figure on the right sorts the data sets by level of learning difficulty. We see that the theater pieces were the most difficult to learn, probably because they were quite heterogeneous in nature, ranging from old French up through modern plays. After that come conversations, then meetings, speeches, debates, interviews, the National Assembly, which is our biggest dataset, and then call-center type dialogues, which seem to be the easiest.

Figure 1 shows that overfitting appears earlier and more prominently for the Apache models, which is explained by the fact that they did not see all of the data presented in Section 2. It might also seem to suggest that that Claire-Mistral (bottom left) learns more efficiently than Claire-Falcon (top left) and converges to a lower loss. However, the evaluation results presented in Section 5 reveal the opposite, showing that average token perplexity is not relevant to compare models with different tokenization and different context sizes.

5 Human evaluation

Evaluation of causal foundation models—which have only been trained with an autoregressive training objective and not fine-tuned to follow instructions or perform a specific task—is a nebulous affair. The only gold data against which model output can be compared is data held out from the training set for validation, which is just a diverse collection of (punctuated) sequences of words. For this reason, foundation models are often evaluated for perplexity, or how well they learn to match a probability distribution over word sequences with the actual probability distribution of the validation set.

This does not mean that foundation models should not be evaluated using standard benchmarks, of course : such evaluation can facilitate the comparison between a model's behavior before and after fine-tuning and can be used to shed light on how different training corpora can impact the model. It's just that the output of these evaluations should be taken with a grain of salt.

In our case, we saw that our Claire models responded to their continued training and wanted to get a clearer idea of how. In particular, we wanted to see to what extent the dialogue data encouraged a style of language that appears interactive, conversational and spontaneous. Standard benchmarks, however, tend to focus on knowledge and reasoning-based tasks, including general knowledge tasks (e.g., SQuADv2 ; Rajpurkar *et al.*, 2016), natural language inference (e.g., HellaSwag ; Zellers *et al.*, 2019), coding (e.g., MBPP ; Austin *et al.*, 2021), maths (GSM8K ; Cobbe *et al.*, 2021), and occasionally a combination of multiple domains (e.g., MMLU from Hendrycks *et al.* (2021) and BIG-bench from bench authors (2023)). Not even MT-Bench (Zheng *et al.*, 2023), which assesses chat interactions, focuses on a model's ability to generate spoken-style conversational interactions.

There are also limited options for benchmarks in French. One approach is to translate the existing resources into other languages, as is done by Bactrian-X (Li *et al.*, 2023), Taco (Upadhayay & Behzadan, 2023) and MT-Bench-Fr [11], though it has been noted that the translation process can compromise the quality of the dataset, and thus of the evaluation. Another approach is to use the high-quality but limited evaluation resources available for a specific target language as Bawden *et al.* (2024) do in their evaluation of Bloom (BigScience Workshop, 2022) using hand-picked French datasets. These evaluation suites still do not target dialogue dynamics however.

To evaluate our Claire models, we developed our own approach to human evaluation that targets conversational abilities, evaluating model output along three dimensions : Interaction, Fluency and Relevance. For Interaction, our main interest was to determine not whether the model produced coherent language but rather if it succeeded in acting like it was engaging in conversation. We looked for evidence of turn taking, direct addressing of the "other speaker(s)", and expressions used to smooth conversation such as "well yeah". (5) shows an example question from the Interaction dimension.

(5) 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.) ?

Fluency questions focused on such factors as whether the model consistently output French, whether the French was acceptable for oral conversation, and more generally, whether the model output seemed human. For Relevance, evaluators were asked to judge whether the response actually addressed the prompt, whether the response stayed more or less on topic, whether it was logically coherent.

We carried out two evaluation campaigns. In both, each evaluator was asked to review a set of five surveys, where each survey included one prompt together with the generated responses from the

11. https://huggingface.co/datasets/bofenghuang/mt-bench-french

four models that we wished to compare : Falcon, Mistral, Claire-Falcon and Claire-Mistral. For each generated output, evaluators were asked 13 questions (covering the three dimensions) and additionally asked to rank the four generated outputs. We created 8 distinct groups of surveys, leading to 40 unique surveys for each campaign. Each survey was reviewed by two people.

For each campaign, we made a list of 10 prompts and varied the form of the speaker label along two dimensions. In the first campaign, we used either the [Intervenant :] label or no speaker label at all, keeping a 50/50 distribution in our surveys. The idea here was two-fold : use monologue-style prompts to test whether the Claire models would output dialogue even without dialogue-specific prompting and also minimize the bias towards our models, which were trained on [Intervenant :] labels. In the second campaign, all prompts included speaker labels. Here we opted for a 50/50 distribution between labels with brackets and those without. The former category included [Intervenant :] labels but also proper names in brackets ; the latter included turns marked with dashes and proper names without brackets. The idea again was to make sure that half of the prompt variants were in a style that had been seen by the Claire models while the other half remained neutral for the base models. We also randomly selected two generated responses per prompt variant (per model) so that each of the 10 prompts was seen in four different ways (2 speaker-label styles x 2 samples) throughout the surveys.

In the second survey, we varied the nature of the prompt. In the first round, our ten prompts (a mixture of interrogative and declarative prompts) were what we call “culture prompts”. These, like *Dis-moi, où est-ce que je peux trouver des cèpes ?*, were designed to discuss concepts that were assumed to be better represented in French training data than in English in order to target the models’ cultural training as well. In the second round, we kept five culture prompts that we felt were the most natural from the first round and added five “casual” prompts in which we asked the model generic questions such as “What will you cook for us today ?” or “What are your favorite French novels ?”. This shows Claire-Falcon’s response to the casual prompt :

prompt :	[Dominique :]	Dis moi, quels sont tes hobbies ou centres d’intérêt ?
Claire completion :	[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.

Our two campaigns of evaluation revealed certain interesting tendencies (see Appendix B for more details). First, while Mistral had a clear tendency to output English or switch to code or other data formats when prompted in French, our Claire-Mistral model did not, suggesting that a small amount of additional French training was sufficient to correct the problem.

Claire-Falcon was consistently rated above Claire-Mistral, while Mistral was consistently rated below the other three models. Note that this ranking of Mistral runs contrary to what one might expect from the perplexity results in Figure 1 in the Appendix, underscoring our warning from Section 4 that perplexity does not permit a direct comparison of the models.

In both rounds of evaluation, the question that correlated the most with a model’s overall preference rank was “Does the output seem human ?", which underscores the difficulty of evaluating foundation models on a task like spontaneous dialogue generation. In second place, positive responses to whether the model generated excessive disfluencies were negatively correlated with model preference, even though such disfluencies are common in spoken language and ultimately quite “human”.

The comparison of Claire-Falcon and Falcon was more complex. In the first round, Claire-Falcon was the clear winner, but this advantage disappeared in the second round. In the end, we traced the difference in results to the nature of the prompts : the results for Claire-Falcon and Falcon on culture prompts were consistent in the two rounds of evaluation. However, Falcon outperforms Claire-Falcon on casual prompts, which were introduced only in round two, explaining why Claire-Falcon was roughly tied with Falcon in the second round. These results show how much LLM performance can be influenced by subtle, and often seemingly unexplainable, differences in prompts.

Because of the incredible cost of human annotation, we decided to carry out our evaluation experiments using GPT-3.5 and GPT-4 as evaluators, following the *LLM-as-a-Judge* evaluation paradigm. Unfortunately, testing on December 14 and 15, 2023, we found that both models, and especially GPT-3.5, were unable to differentiate the performance of the models we were evaluating, especially on subjective questions, adding to the already existing body of evidence that LLMs are unable to match or replace human evaluators when it comes to assessing human preference (Koo *et al.*, 2023).

6 Conclusion

We have presented the Claire family of language models, a set of models designed to focus on dialogue dynamics with the aim of improving performance on downstream tasks involving spoken dialogue understanding. This work complements that of Pelloin *et al.* (2022), though our datasets are largely manually transcribed and include more spontaneous spoken language (as opposed to language from TV or radio). Our models were produced by continuing the training of the Falcon and Mistral 7B foundation models on conversation transcripts and theater plays. A key feature of our models, apart from the dialogue component, is their focus on French, contributing to a growing trend to emphasize multi-linguality in language models (Faysse *et al.*, 2024; Groeneveld *et al.*, 2024; Scao *et al.*, 2023). Our two principal models are released under a CC BY-SA-NC 4.0 license [12] while the other two, trained on a subset of our data that allows commercial use, have an Apache 2.0 license [13]. The GitHub repository with code used to train our Claire models is also available publicly [14]. We are currently working on extending the approach presented in this paper to train a bilingual English-French dialogue model with the objective of fine-tuning the resulting model for meeting summarization in order to test the impact of our dialogue pretraining on downstream tasks involving dialogues.

7 Acknowledgements

This work was granted access to the HPC resources of IDRIS under the GENCI Grant 2023-AD011014561. It was supported by the ANR projects SUMM-RE (ANR-20-CE23-0017) and LLM4ALL (ANR-23-IAS1-0008-02) and the European Union project Cortex2 (101070192).

12. Claire-Falcon model is available at https://huggingface.co/OpenLLM-France/Claire-7B-0.1 and Claire-Mistral at https://huggingface.co/OpenLLM-France/Claire-Mistral-7B-0.1

13. Claire-Falcon with Apache license is available at https://huggingface.co/OpenLLM-France/Claire-7B-Apache-0.1 and Claire-Mistral with Apache license at https://huggingface.co/OpenLLM-France/Claire-Mistral-7B-Apache-0.1

14. https://github.com/OpenLLM-France/Lit-Claire

Références

ANTOINE J.-Y., GOULIAN J., VILLANEAU J. & LE TALLEC M. (2009). Word order phenomena in spoken french : a study on four corpora of task-oriented dialogue and its consequences on language processing. In *Proc. Corpus Linguitics*.

ANTOINE J.-Y., LETELLIER-ZARSHENAS S., NICOLAS P. & SCHADLE I. (2002). Corpus OTG et ECOLE_MASSY : vers la constitution d'un collection de corpus francophones de dialogue oral diffusés librement. *Actes TALN 2002*.

ASSEMBLÉE NATIONALE (2023). https ://www.assemblee-nationale.fr/.

ATILF (2020). TCOF : Traitement de corpus oraux en français. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

AUSTIN J., ODENA A., NYE M., BOSMA M., MICHALEWSKI H., DOHAN D., JIANG E., CAI C., TERRY M., LE Q. *et al.* (2021). Program synthesis with large language models. *arXiv preprint arXiv :2108.07732*.

AVANZI M. (2012). *L'interface prosodie/syntaxe en français : dislocations, incises et asyndètes*. Gramm-R. P.I.E. Peter Lang.

AVANZI M., BÉGUELIN M.-J., CORMINBOEUF G., FEDERICA D. & JOHNSEN L.-A. (2012–2023). Corpus OFROM – corpus oral de français de suisse romande. Université de Neuchâtel.

AVANZI M., SIMON A.-C., GOLDMAN J.-P. & AUCHLIN A. (2010). C-PROM. un corpus de français parlé annoté pour l'étude des proéminences. *Actes des 23èmes journées d'étude sur la parole*.

BALDAUF-QUILLIATRE H., COLÓN DE CARVAJAL I., ETIENNE C., JOUIN-CHARDON E., TESTON-BONNARD S. & TRAVERSO V. (2016). CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes. *Corpus*, **15**, 165–194. DOI : 10.4000/corpus.2991, HAL : halshs-01316283.

BAWDEN R., BOURFOUNE H., CABOT B., CASSEREAU N., CORNETTE P., NAGUIB M., NÉVÉOL A. & YVON F. (2024). Les modèles Bloom pour le traitement automatique de la langue française. working paper or preprint.

BENCH AUTHORS B. (2023). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

BERTRAND R., BLACHE P., ESPESSER R., FERRÉ G., MEUNIER C., PRIEGO-VALVERDE B. & RAUZY S. (2008). Le CID - corpus of interactional data - annotation et exploitation multimodale de parole conversationnelle. *Revue TAL : traitement automatique des langues*, **49**(3), pp.105–134.

BIGSCIENCE WORKSHOP (2022). BLOOM (revision 4ab0472). DOI : 10.57967/hf/0003.

BLACHE P., BERTRAND R., BRUNO E., BIGI B., ESPESSER R., FERRÉ G., GUARDIOLA M., HIRST D., TAN N., CELA E., MARTIN J.-C., RAUZY S., MOREL M.-A., MURISASCO E. & NESTERENKO I. (2010). Multimodal annotation of conversational data. In N. XUE & M. POESIO, Éds., *Proceedings of the Fourth Linguistic Annotation Workshop*, p. 186–191, Uppsala, Sweden : Association for Computational Linguistics.

BRANCA-ROSOFF S., FLEURY S., LEFEUVRE F. & PIRES M. (2013). Corpus de français parlé parisien des années 2000 (CFPP).

CARRUTHERS J. (2008). Annotating an oral corpus using the text encoding initiative. methodology, problems, solutions. *Journal of French Language Studies*, **18**(1), 103–119. DOI : 10.1017/S0959269507003183.

CLESTHIA (2018). Cfpp2000. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

COBBE K., KOSARAJU V., BAVARIAN M., CHEN M., JUN H., KAISER L., PLAPPERT M., TWOREK J., HILTON J., NAKANO R., HESSE C. & SCHULMAN J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv :2110.14168*.

COGNITION, LANGUE, LANGAGES, ERGONOMIE (CLLE) (2013). ACSYNT. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

COMPUTER T. (2023). Redpajama : an open dataset for training large language models.

CRESTI E., DO NASCIMENTO F. B., SANDOVAL A. M., VERONIS J., MARTIN P. & CHOUKRI K. (2004). The C-ORAL-ROM CORPUS. a multilingual resource of spontaneous speech for Romance languages. In M. T. LINO, M. F. XAVIER, F. FERREIRA, R. COSTA & R. SILVA, Éds., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA).

DEBAISIEUX J.-M., BENZITOUN C. & DEULOFEU H.-J. (2016). Le projet ORFEO : Un corpus d'études pour le français contemporain. *Corpus*, **15**, 91–114. DOI : 10.4000/corpus.2936, HAL : hal-01449600.

DISTER A., FRANCARD M., HAMBYE P. & SIMON A. C. (2007). Du corpus à la banque de données. du son, des textes et des métadonnées. l'évolution de banque de données textuelles orales VALIBEL (1989-2006). *Cahiers de Linguistique*, **33**(2), 113–129.

DISTER A. & LABEAU E. (2017). Le corpus de français parlé à bruxelles : origines, hypothèses, développements et prédictions. *Cahiers AFLS*, **21**(1).

DURAND J., LAKS B. & LYCHE C. (2009). Le projet PFC (Phonologie du Français Contemporain) : une source de données primaires structurées. *Phonologie, variation et accents du français*, p. 19–61.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral disponible : le corpus d'orléans 1968-2012 [a large available oral corpus : Orleans corpus 1968-2012]. *Traitement Automatique des Langues*, **52**(3), 17–46.

FAYSSE M., FERNANDES P., GUERREIRO N., LOISON A., ALVES D., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P. *et al.* (2024). Croissantllm : A truly bilingual french-english language model. *arXiv preprint arXiv :2402.00786*.

FISCHER F., BÖRNER I., GÖBEL M., HECHTL A., KITTEL C., MILLING C. & TRILCKE P. (2019). Programmable Corpora : Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019 : "Complexities", Utrecht, July 9–12, 2019* : Utrecht University. DOI : 10.5281/zenodo.4284002.

GRAVELLIER L., HUNTER J., MULLER P., PELLEGRINI T. & FERRANÉ I. (2021). Weakly supervised discourse segmentation for multiparty oral conversations. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1381–1392, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.104.

GROENEVELD D., BELTAGY I., WALSH P., BHAGIA A., KINNEY R., TAFJORD O., JHA A. H., IVISON H., MAGNUSSON I., WANG Y. *et al.* (2024). Olmo : Accelerating the science of language models. *arXiv preprint arXiv :2402.00838*.

GUPTA P., JIAO C., YEH Y.-T., MEHRI S., ESKENAZI M. & BIGHAM J. (2022). InstructDial : Improving zero and few-shot generalization in dialogue through instruction tuning. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 505–525, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.33.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. *CoRR*, **abs/2106.09685**.

HUNTER J., LOURADOUR J., RENNARD V., HARRANDO I., SHANG G. & LORRÉ J.-P. (2023). The claire french dialogue dataset. *arXiv preprint arXiv :2311.16840*.

JANIN A., BARON D., EDWARDS J., ELLIS D., GELBART D., MORGAN N., PESKIN B., PFAU T., SHRIBERG E., STOLCKE A. *et al.* (2003). The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, volume 1, p. I–I : IEEE.

KAHANE S., CARON B., STRICKLAND E. & GERDES K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora : A proposal. In D. DAKOTA, K. EVANG & S. KÜBLER, Éds., *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, p. 35–47, Sofia, Bulgaria : Association for Computational Linguistics.

KOO R., LEE M., RAHEJA V., PARK J. I., KIM Z. M. & KANG D. (2023). Benchmarking cognitive biases in large language models as evaluators.

LACHERET A. (2003). *La prosodie des circonstants en français parlé*. Volume 85 de (Collection Linguistique). Peeters. ISBN : 90-429-1414-9 (Peeters Leuven). - 2-87723-771-0 (Peeters France)., HAL : halshs-00349268.

LAURENÇON H., SAULNIER L., WANG T., AKIKI C., DEL MORAL A. V., SCAO T. L., WERRA L. V., MOU C., PONFERRADA E. G., NGUYEN H., FROHBERG J., ŠAŠKO M., LHOEST Q., MCMILLAN-MAJOR A., DUPONT G., BIDERMAN S., ROGERS A., ALLAL L. B., TONI F. D., PISTILLI G., NGUYEN O., NIKPOOR S., MASOUD M., COLOMBO P., DE LA ROSA J., VILLEGAS P., THRUSH T., LONGPRE S., NAGEL S., WEBER L., MUÑOZ M., ZHU J., STRIEN D. V., ALYAFEAI Z., ALMUBARAK K., VU M. C., GONZALEZ-DIOS I., SOROA A., LO K., DEY M., SUAREZ P. O., GOKASLAN A., BOSE S., ADELANI D., PHAN L., TRAN H., YU I., PAI S., CHIM J., LEPERCQ V., ILIC S., MITCHELL M., LUCCIONI S. A. & JERNITE Y. (2023). The BigScience ROOTS Corpus : A 1.6TB Composite Multilingual Dataset.

LI H., KOTO F., WU M., AJI A. F. & BALDWIN T. (2023). Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation.

LPL (2021). Transcriptions du corpus CID. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

LUCCIONI A. S., VIGUIER S. & LIGOZAT A.-L. (2022). Estimating the carbon footprint of bloom, a 176b parameter language model.

MCCOWAN I., CARLETTA J., KRAAIJ W., ASHBY S., BOURBAN S., FLYNN M., GUILLEMOT M., HAIN T., KADLEC J., KARAISKOS V., KRONENTHAL M., LATHOUD G., LINCOLN M., LISOWSKA MASSON A., POST W., REIDSMA D. & WELLNER P. (2005). The AMI meeting corpus. *International Conference on Methods and Techniques in Behavioral Research*.

MERTENS P. (1987). L'intonation du français. *De la description linguistique à la reconnaissance automatique. unpublished doctoral dissertation, Catholic University of Leuven, Belgium*.

MILLER A., FENG W., BATRA D., BORDES A., FISCH A., LU J., PARIKH D. & WESTON J. (2017). ParlAI : A dialog research software platform. In L. SPECIA, M. POST & M. PAUL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 79–84, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-2014.

MILLING C., FISCHER F. & (EDS.) M. G. (2021). French Drama Corpus (FreDraCor) : A TEI P5 Version of Paul Fièvre's "Théâtre Classique" Corpus. https://github.com/dracor-org/fredracor.

MODYCO & RUG (2017). PFC - phonologie du français contemporain. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

NICOLAS P., LETELLIER-ZARSHENAS S., SCHADLE I., ANTOINE J. & CAELEN J. (2002). Towards a large corpus of spoken dialogue in french that will be freely available : the "parole publique" project and its first realisations. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain* : European Language Resources Association.

PELLOIN V., DARY F., HERVÉ N., FAVRE B., CAMELIN N., LAURENT A. & BESACIER L. (2022). Asr-generated text for language model pre-training applied to speech tasks. *arXiv preprint arXiv :2207.01893*.

PENEDO G., MALARTIC Q., HESSLOW D., COJOCARU R., CAPPELLI A., ALOBEIDLI H., PANNIER B., ALMAZROUEI E. & LAUNAY J. (2023). The RefinedWeb dataset for Falcon LLM : outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv :2306.01116*.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). SQuAD : 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, p. arXiv :1606.05250.

RENNARD V., SHANG G., GRARI D., HUNTER J. & VAZIRGIANNIS M. (2023). FREDSum : A dialogue summarization corpus for french political debates. In *Findings of Empirical Methods in Natural Language Processsing (EMNLP)*.

RHAPSODIE (2015). https ://rhapsodie.modyco.fr/propriete-intellectuelle/.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M., TOW J., RUSH A. M., BIDERMAN S., WEBSON A., AMMANAMANCHI P. S., WANG T., SAGOT B., MUENNIGHOFF N., DEL MORAL A. V., RUWASE O., BAWDEN R., BEKMAN S., MCMILLAN-MAJOR A., BELTAGY I., NGUYEN H., SAULNIER L., TAN S., SUAREZ P. O., SANH V., LAURENÇON H., JERNITE Y., LAUNAY J., MITCHELL M., RAFFEL C., ... 357 AUTHORS ... & WOLF T. (2023). Bloom : A 176b-parameter open-access multilingual language model.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)* : Leibniz-Institut für Deutsche Sprache.

THÉÂTRE CLASSIQUE (2022). http ://www.theatre-classique.fr/.

THÉÂTRE GRATUIT (2023). https ://theatregratuit.com/.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

TUTIN A. & GROSSMAN F. (2014). L'écrit scientifique : du lexique au discours. *Autour de Scientext*, p. 27–44.

UPADHAYAY B. & BEHZADAN V. (2023). Taco : Enhancing cross-lingual transfer for low-resource languages in llms through translation-assisted chain-of-thought processes. *arXiv preprint arXiv :2311.10797*.

XU Y., LEE H., CHEN D., CHOI H., HECHTMAN B. & WANG S. (2020). Automatic cross-replica sharding of weight update in data-parallel training.

YAMASAKI H., LOURADOUR J., HUNTER J. & PRÉVOT L. (2023). Transcribing and aligning conversational speech : A hybrid pipeline applied to french conversations. In *Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

ZELLERS R., HOLTZMAN A., BISK Y., FARHADI A. & CHOI Y. (2019). Hellaswag : Can a machine really finish your sentence ? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

ZHANG J., QIAN K., LIU Z., HEINECKE S., MENG R., LIU Y., YU Z., SAVARESE S. & XIONG C. (2023). DialogStudio : Towards richest and most diverse unified dataset collection for conversational AI. *arXiv preprint arXiv :2307.10172*.

ZHAO Y., GU A., VARMA R., LUO L., HUANG C.-C., XU M., WRIGHT L., SHOJANAZERI H., OTT M., SHLEIFER S., DESMAISON A., BALIOGLU C., DAMANIA P., NGUYEN B., CHAUHAN G., HAO Y., MATHEWS A. & LI S. (2023). Pytorch fsdp : Experiences on scaling fully sharded data parallel.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E., ZHANG H., GONZALEZ J. E. & STOICA I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

ZHU C., LIU Y., MEI J. & ZENG M. (2021). Mediasum : A large-scale media interview dataset for dialogue summarization. *CoRR*, **abs/2103.06410**.

ÉQUIPE DELIC (2004). Autour du corpus de référence du français parlé. *Recherches sur le français parlé*, **18**, 265.

A Survey questions from the second evaluation campaign

Here is a list of the survey questions corresponding to the second round of evaluation. Most of the questions have a yes/no format, with a third option proposed if the question does not apply or to quantify some characteristic (e.g. for 2.d, the options are "Plusieurs erreurs"/"Quelques erreurs"/"Aucune erreur").

1. Interaction
   (a) En présence de “tours” délimités par des tirets (“-”) ou des étiquettes telles que “[Nom 1 :]” ou “Prénom :”, ces marqueurs sont-ils positionnés de manière intuitive et, pour les étiquettes, dans un ordre plausible ? (Ignorez les erreurs mineures de format, e.g., [Intervenant 2) :])
   (b) En excluant les étiquettes de locuteurs, la réponse semble-t-elle davantage être extraite d’une conversation ou d’un document écrit ? (La cohérence de la réponse n’est pas primordiale ici ; l’accent est mis sur le style généré par le modèle.)
   (c) En cas de dialogue avec plusieurs échanges, semble-t-il que les interlocuteurs cherchent à engager une discussion (en se tutoyant/vousvoyant directement, répondant aux questions, utilisant des expressions conversationnelles comme ’oui, c’est vrai’, etc.) ?
2. Fluidité
   (a) La réponse bascule-t-elle vers l’anglais ou un autre type de langage ?
   (b) La réponse présente-t-elle des répétitions non motivées ? (Par exemple, le modèle semble-t-il répéter de manière robotique les mêmes expressions, sans justification ? Les répétitions naturelles de la conversation sont considérées comme justifiées.)
   (c) Le niveau de disfluences (telles que “euh”, “hm”, “quoi”) vous semble-t-il naturel par rapport à une conversation orale ? (s’il y en a pas, répondez “oui”)
   (d) En dehors des disfluences et des répétitions robotiques, la partie restante de la réponse est-elle formulée dans un français parlé correct ? (Veuillez noter que les constructions typiques du français parlé, telles que “y a pas” ou “ton père, il est où ?”, sont acceptables.)
   (e) Dans l’ensemble, avez-vous l’impression que la conversation semble humaine (par opposition à “générée par l’IA”) ?
3. Pertinence
   (a) La suite de la conversation, répond-elle de manière spécifique au début de la conversation proposé, même si elle est incorrecte sur le plan factuel ? (Cela s’oppose à une suite qui pourrait être tout aussi valable pour un début de conversation bien différent.)
   (b) Y a-t-il des changements de sujet ou de ton qui attirent particulièrement l’attention ? (Ignorez les changements légers de sujet courants dans la conversation spontanée ; un passage à un sujet très différent avant d’avoir traité le premier pourrait être considéré comme frappant.)
   (c) Le niveau de disfluences (telles que “euh”, “hm”, “quoi”) vous semble-t-il naturel par rapport à une conversation orale ? La conversation semble-t-elle stagnante sur un sujet (plutôt que de progresser comme elle le devrait) ?
   (d) En dehors des disfluences et des répétitions robotiques, la partie restante de la réponse est-elle formulée dans un français parlé correct ? La conversation contient-elle des stagnante ?
   (e) Le modèle semble-t-il inventer des choses de manière exagérée ? (On exclut les détails à vérifier dans une encyclopédie ; l’accent est mis sur les détails qui semblent inventés et perturbent la cohérence de la conversation.)

B Evaluation results

As explained in Section 5 and illustrated in Figure 2, Mistral (red) had a clear tendency to switch to English, code or another data format when prompted in French. This tendency disappeared for the Claire-Mistral model (orange). The results in Figure 2 are from our second evaluation campaign and indicate a score out of 80, as 80 was the number of surveys conducted.

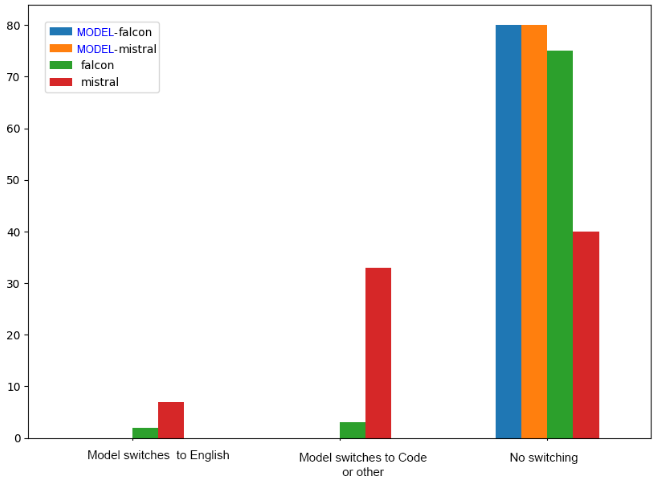


FIGURE 2 – The tendency of each model to stick to French when prompted in French

Results from both campaigns showed a clear overall preference for Claire-Falcon over Claire-Mistral, with Mistral coming it at fourth place. See Figure 3.

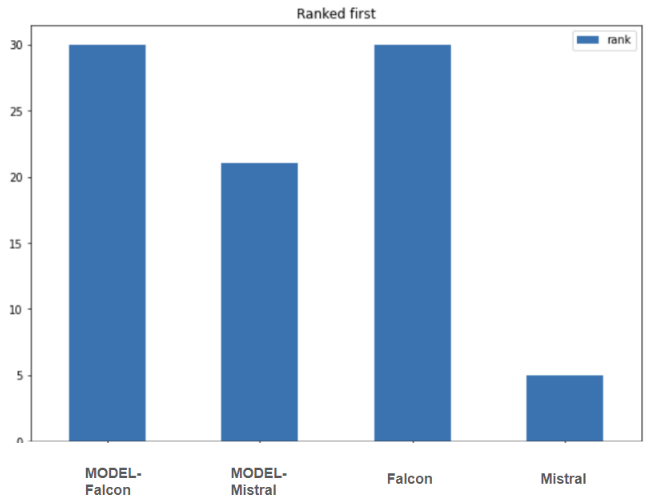


FIGURE 3 – Overall rankings of the four models in the second evaluation campaign

Figure 4 shows judgments regarding whether the model seemed human. These results were the most strongly correlated with overall preference. In this evaluation, Claire-Falcon (blue) was very slightly preferred over Falcon (green).

Finally, if we focus on “culture”-style prompts, as shown in Figure 5, we see that Claire-Falcon is clearly preferred over Falcon, a tendency that was observed in the first round of evaluations as well. Falcon takes the lead for “casual” prompts, however.

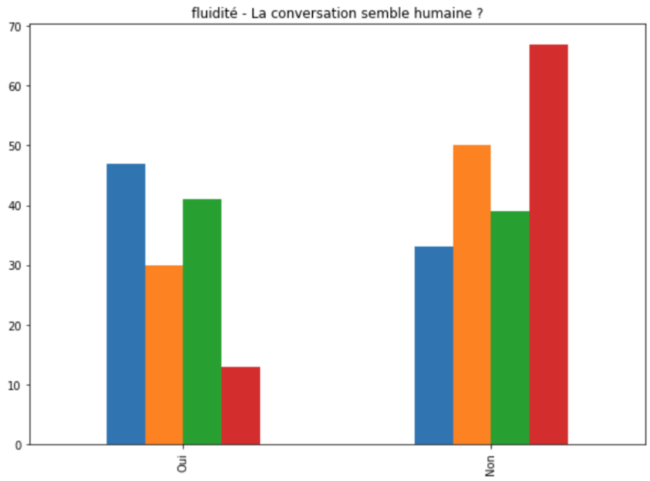


FIGURE 4 – Judgments concerning the number of times the response of a model was judged as human

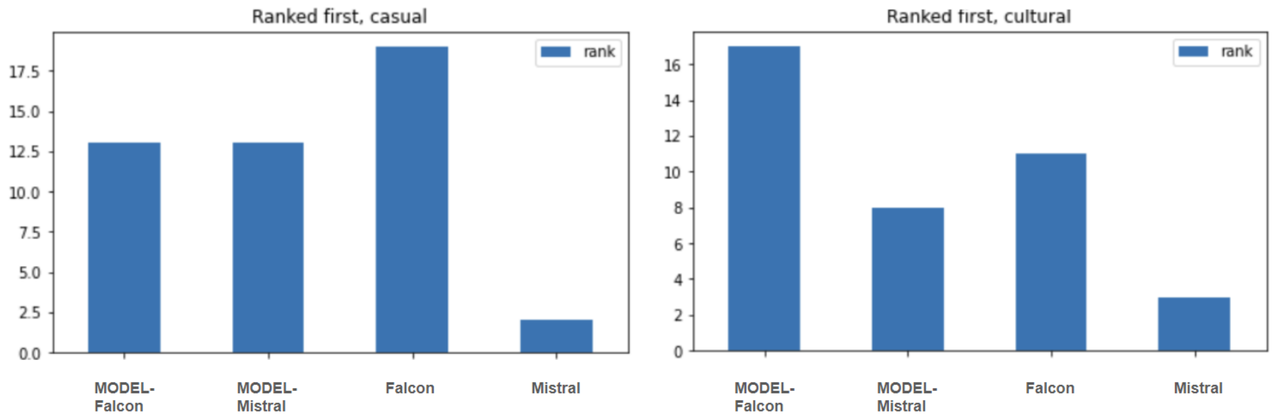


FIGURE 5 – Results showing that Claire-Falcon is clearly preferred for "culture" prompts. (A result reflected also in the first evaluation, which contained only culture prompts.)

C Sources for the original datasets used to train the principal Claire models

ACSYNT : Cognition, Langue, Langages, Ergonomie (CLLE) (2013)
Assemblée Nationale : Assemblée nationale (2023)
Orféo-CEFC : Debaisieux *et al.* (2016); Carruthers (2008); Tutin & Grossman (2014)
— **C-ORAL-ROM** : Cresti *et al.* (2004)
— **CRFP** : Équipe Delic (2004)
— **FLEURON** : no citation found
— **Valibel** : Dister *et al.* (2007)
Orféo : (other)
— **CFPB** : Dister & Labeau (2017)
— **Réunions De Travail** : (Work Meetings)
CFPP : Branca-Rosoff *et al.* (2013); CLESTHIA (2018)
CID : Bertrand *et al.* (2008); Blache *et al.* (2010); LPL (2021)

CLAPI : Baldauf-Quilliatre *et al.* (2016)
ESLO : Eshkol-taravella *et al.* (2011)
FREDSum : Rennard *et al.* (2023)
LinTO : Gravellier *et al.* (2021)
OFROM : Avanzi *et al.* (2012 2023)
Parole Publique :
— **Accueil UBS** : Nicolas *et al.* (2002); Antoine *et al.* (2009)
— **OTG** : Nicolas *et al.* (2002); Antoine *et al.* (2002)
Paris Stories : Kahane *et al.* (2021)
PFC : Durand *et al.* (2009); MoDyCo & RUG (2017)
Rhapsodie : Rhapsodie (2015); Branca-Rosoff *et al.* (2013); Avanzi (2012); Lacheret (2003); Mertens (1987); Avanzi *et al.* (2010); Eshkol-taravella *et al.* (2011); Durand *et al.* (2009)
SUMM-RE : Yamasaki *et al.* (2023)
TCOF : ATILF (2020)
Théâtre Classique : Théâtre classique (2022); Milling *et al.* (2021); Fischer *et al.* (2019)
Théâtre Gratuit : Théâtre Gratuit (2023)