GeneFRDebate: Generated French Debates from News Articles with Industrial-Expert Summaries

Rim Abrougui, Guillaume Lechien, Elisabeth Savatier, Benoît Laurent
Aday - Paris - France
innovations@aday.fr

Abstract

Summarizing domain-specific conversations, such as political debates, remains challenging despite advances in large language models (LLMs), and resources for French debates are particularly limited. We present GeneFRDebate, a new dataset of synthetic French political debates generated from real-world news articles using an LLM, while keeping expert-written summaries unchanged. Our pipeline combines prompt engineering, human curation, and quality evaluation using both automatic metrics and expert assessment. We also provide baseline experiments with small-scale LLMs ($\leq$8B parameters), demonstrating the dataset's usefulness for training and evaluation. This work shows that carefully generated synthetic data with human oversight can complement existing corpora, supporting research in multilingual and domain-specific dialogue summarization.

Keywords: French debate summarization, synthetic dataset, large language models (LLMs), prompt engineering, data curation

1. Introduction

In recent years, the task of dialogue summarization has received growing attention. With the progress of large language models (LLMs), this task has seen noticeable improvements. However, summarizing domain-specific conversations, such as political debates, remains challenging. The goal of summarization is to produce concise texts that preserve the original meaning, either by rephrasing the content (abstractive summarization) or by extracting key segments from the text (extractive summarization) (Rachabathuni, 2017).

To develop and benchmark robust summarization systems, datasets play a crucial role. Several corpora for written or formal dialogues have recently been introduced (Gliwa et al., 2019; Chen et al., 2021;Lin and Ng, 2019). Early work on spoken dialogue summarization dates back to the late 1990s and early 2000s (Zechner and Waibel, 2000), addressing multiple conversational domains including televised discussions, telephone interactions, and meeting transcripts (Tuggener et al., 2021). Despite these advances, available multilingual datasets are still mostly centered on document summarization, while dialogue-oriented and especially debate-specific datasets remain rare and underexplored (Zhao et al., 2024; Feng et al., 2022; Rennard et al., 2023; Roush and Balaji, 2020).

Building large, reliable dialogue resources continues to be time-consuming and expensive, requiring careful transcription, segmentation, and annotation. As a result, only a few corpora of multi-party conversations have been developed to support research on meeting and chat summarization. Notable examples include SAMSum (Gliwa et al., 2019) and ConvoSumm (Fabbri et al., 2021) for chat and dialogue summarization, as well as AMI (Kraaij et al., 2005), ICSI (Janin et al., 2003), ELITR (Nedoluzhko et al., 2022), and QMSum (Zhong et al., 2021) for meeting summarization (Shang, 2021).

For French, the Fredsum dataset (Rennard et al., 2023) is, to our knowledge, currently the only available corpus dedicated to dialogue summarization in political debates. It includes 144 manually transcribed debates with both extractive and abstractive summaries. The dataset covers major political themes such as immigration, healthcare, education, foreign policy, and security.

To extend existing resources, our study explores whether LLMs can be used to generate conversational datasets, particularly those involving political debates in French.

LLMs are increasingly adopted for text synthesis because they can produce human-like, contextually relevant, and linguistically diverse content (Long et al., 2024). However, hallucination remains a persistent issue, raising concerns about their reliability for creating training data without systematic validation. In addition, model size significantly influences generation quality. Larger models often yield more coherent outputs, but their use is limited by computational and resource constraints (Fu et al., 2024). Prior work, such as (Thulke et al., 2024), demonstrated that small-scale models can still achieve strong summarization performance when fine-tuned on synthetic data, as long as the data creation process is well-structured.

In this paper, we introduce GeneFRDebate, a Generated FRench political Debate dataset derived from written documents, using expert summaries originally produced for industrial purposes. Our ob-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1831–1841
11–16 May 2026.

jective is to evaluate whether medium-sized models, such as Mistral-24B-Instruct[1] , can generate coherent and diverse debates suitable for summarization research.

Our main contributions are:

1. A data generation pipeline in which an LLM transforms real-world French news articles into simulated political debates, while keeping the original expert-written summaries unchanged.
2. A framework for prompt engineering, data curation, and quality evaluation, combining automatic metrics and human validation.
3. Baseline experiments using our dataset for training small-scale LLMs ($\leq$8B parameters) to evaluate its practical utility.

Through this work, we aim to evaluate synthetic data generation and promote curation practices for dialogue summarization and encourage the development of multilingual and domain-specific resources for the NLP community.

2. GeneFRDebate: Generated French Debates from Industrial News Articles

The debates were generated from French news articles provided by industrial publishers. These articles are typically used in a professional setting to produce summaries written by domain experts. The summaries were crafted by reformulating the original articles and occasionally including key quotations or numerical data. As a result, this type of summary is hybrid, combining both abstractive and extractive characteristics.

In this work, we aimed to transform these written articles into oral-style debates to support research on automatic summarization of spoken interactions. For this purpose, we used the Mistral-24B-Instruct model, a resource-accessible LLM, and prompted it to convert the articles into realistic debates while keeping the expert-written summaries unchanged.

We first experimented with multiple prompting strategies to guide the model in producing coherent and natural debates. After an initial round of human evaluation, the most effective prompt was used to construct the final dataset. This section describes the prompt engineering process in detail.

2.1. Prompt Engineering

Several prompt formulations were tested to generate debates from French news articles. Each prompt was evaluated on a sample of 15 debates using three criteria: **(1)** readability, **(2)** coherence, and **(3)** thematic relevance. Each criterion was rated from 0 to 5. Debates were considered successful if they featured at least two speakers, presented clear opposing viewpoints, and followed a realistic conversational flow. For clarity, the prompts and examples are presented in English, translated from French.

2.1.1. Simple Prompting

In the first experiment, we used a minimal instruction asking the model to transform a written article into a political debate:

```
[INST] Transform the following
article into a political oral
debate: {article} [/INST]
```

This prompt produced coherent debates (average score 3.0) and moderate topic alignment (3.4). However, the dialogues tended to retain a written, formal tone, lacking spontaneity and oral cues (average style score 2.5).

2.1.2. Chain-of-Thought Prompting

To improve argument flow and orality, we introduced a more structured prompt encouraging the model to plan the debate composition step by step (CoT Prompt 1), we then tested two shorter variants to balance control and generation fluency (CoT Prompt 2 and CoT Prompt 3).

CoT Prompt 1

```
[INST] You are a political debate
expert. Analyze the article and
identify key
positions, arguments, and fig-
ures. Write a realistic debate
including:
- Introduction (1-2 sentences)
framing the topic
- Speaker 1 (Position A): main
arguments
- Speaker 2 (Position B): coun-
terarguments
- Optional moderator reactions or
interruptions
- Conclusion summarizing dis-
agreements or open question
Article to transform: {article}
[/INST]
```

CoT Prompt 2

```
[INST] You are a political debate
expert. Transform the following
article into
a realistic oral debate, clearly
showing opposing positions and
key arguments: {article} [/INST]
```

[1] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

CoT Prompt 3

```
[INST] You are a political debate
expert.  Transform the article
into a realistic,
spontaneous debate highlighting
opposing views and figures.  In-
clude natural
speech markers ("uh", "so") and
a few interruptions per speaker:
{article} [/INST]
```

As shown in Table 1, the simplified chain-of-thought prompt variants (CoT Prompt 2 and CoT Prompt 3) produced more structured and natural debates. In contrast, the initial chain-of-thought prompt led the model to focus mainly on extracting key elements from the article in a highly structured manner. However, the resulting texts lacked conversational dynamics, resembling more a synthesis than a real debate. On the other hand, for Prompt 2, spontaneity and the oral aspect were still missing. This led us to construct the final prompt (CoT Prompt 3), where we explicitly instructed the model to include speech markers.

Criterion	CoT1	CoT2	CoT3
Style	0.8	3.5	4.2
Coherence	2.8	3.9	4.1
Thematic relevance	3.9	4.0	4.0

Table 1: Human evaluation scores for different prompting strategies.

Table 2 provides an English illustration of the same debate for the three different prompts.

Using the best-performing prompt, we generated the final set of 420 synthetic debates. In the next section, we focused more on the data curation and quality evaluation of our dataset.

3. Dataset Curation and Quality Evaluation

3.1. Human Curation and Filtering

The curation was done semi-automatically . We started by removing automatically the empty and the non-valid outputs (like a repetiton of `[INST]` tokens). Then we manually reviewed to remove failed generations, such as those containing excessive repetitions of turns or incoherent exchanges. This filtering step ensured that only thematically consistent debates were retained in the final dataset.

With this step, we removed around 33% of the generated texts. Overall, we kept 283 debates covering 8 themes: *Transport*, *Social*, *Science*, *Health*, *Politics*, *Finance*, *Energy* and *Culture*.

The statistical characteristics of our dataset, compared to other public datasets, are highlighted in Table 3.

Prompt	Output
CoT1	Analysis of the article The different positions: **1. **Position A**:** Experts and analysts who fear a new surge in gas prices and supply risks for the European Union. **2. **Position B**:** Experts and analysts who believe that the situation should improve by 2026-2027...
CoT2	**Moderator:** Good evening everyone, and welcome to this debate on the energy situation in Europe. Perrine, let's start with you. Can you give us your analysis of the current situation? **Perrine Gauthier:** Of course. The situation is worrying...
CoT3	**Moderator:** Hello everyone, so, uh, we're going to talk about the energy situation in Europe. It's 2025 and it feels like 2022 is happening again. Should we fear another surge in gas prices? **Speaker 2 (S2):** Well, I think so, there's reason to be concerned...

Table 2: Examples of debates generated with different prompting strategies.

3.2. Human Evaluation

A human evaluation was conducted by native French speakers on a subset of 20 debates to assess the overall quality and coherence of the generated data. Each debate was independently analyzed by two annotators following a detailed annotation guide. Debates were assessed according to three criteria, each rated on a 0–5 scale:

- **Debate style**: measures whether the generated text resembles a realistic debate with multiple speakers, clear opposition, natural flow, and spontaneity. A score of 0 indicates no dialogue, while 5 represents a fluent and natural debate with distinct voices, clear argumentation, and spontaneous interactions.
- **Accuracy with summaries**: evaluates whether the key information from the expert-written summaries is present in the debates. A score of 0 indicates no information is preserved, while 5 indicates complete alignment with the summary and article, without omission or distortion.
- **Coherence**: assesses the logical flow and clarity of transitions within the debate. A score of 0 indicates incoherent sentences with no clear structure, while 5 reflects a fully coherent debate with smooth transitions, clear argument progression, and faithful representation of the main theme.

Dataset	Lang.	#Transcripts	#Words	#Turns	#Speakers	#Words/sum.
MediaSum	EN	463596	1553.7	30.0	6.5	14.4
MeetingBank	EN	6892	3800.3	146.9	3.2	87.2
ELITR	EN/CS	179	7549.9	884.5	6.5	327.9
FREDSum	FR	144	2595.5	49.5	4	238.9
GeneFRDebate	FR	283	726.4	15.3	3.5	152.9

Table 3: Dataset statistics

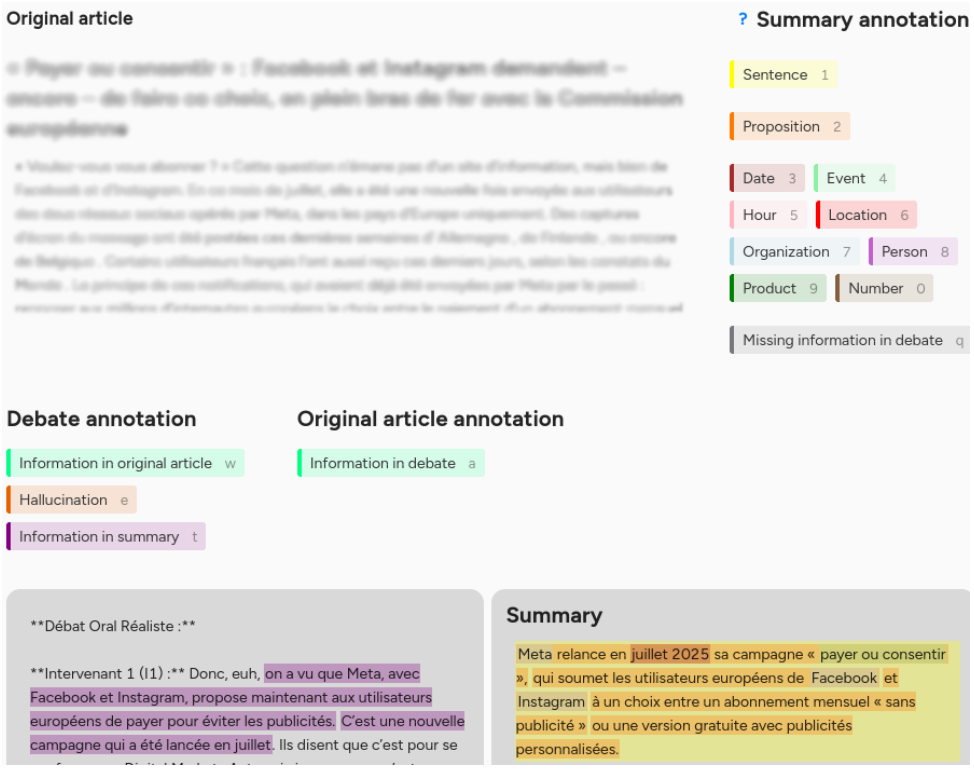

Figure 1: Human evaluation of the generated debates

For that, annotators analyzed the summaries by segmenting them into sentences, identifying discourse relations, and extracting key information and named entities. These elements were then checked in the debates to evaluate completeness and thematic alignment (c.f. Figure 1). The ultimate guide for evaluating debates is presented in Appendix A (cf. Appendix 7.1).

The inter-annotator agreement was calculated using the Cohen's Kappa coefficient (Kohen, 1960).For the *style* criterion, we obtained substantial agreement with a Kappa score of 0.67. For *coherence*, the agreement was moderate (0.56), while for *accuracy*, the agreement was fair (0.32). A third annotator reviewed the annotations and proposed compromises for cases of disagreement, which were subsequently discussed and resolved by consensus among the annotators. On average, the final scores were 3.5 for accuracy, and 3 for both style and coherence.

We then categorized the cases considered as errors. Four main types of errors were identified based on the evaluation criteria: hallucinations, omissions, logical incoherence, and stylistic errors, particularly regarding artificial or unnatural speech.

Hallucinations referred to cases where the model invented new speakers or entities that did not appear in the original article or summary (see *Pierre* and *Marie* in Table 4). However, these hallucinations did not have a negative impact. Indeed, some studies such as (Jiang et al., 2024) suggest that LLM hallucinations may sometimes foster creativity. Another type of error involved the addition of minor information not present in the source article, though such cases were rare. More frequently, we observed instances where the model distorted or reinterpreted information to make the debate more dynamic (see Table 4).

For omissions, this referred to debates that did not cover all the information included in the summaries

Regarding logical incoherence, we observed that the LLM sometimes repeated the same arguments or phrases across turns. Although such repetition

Information in Article	Generated Debate EN	Generated Debate FR
Will converting your car to bioethanol one day allow you to drive freely in low-emission zones? The driver advocacy association **40 Million Motorists (40MA) has launched a petition to this effect**, to push the public authorities to act. However, the initiative faces limitations: **vehicles originally running on bioethanol naturally benefit from a Crit'Air 1 sticker.**	**Pierre:** *So, uh*, **I think converting your car to bioethanol is a great solution for driving freely in low-emission zones.** **The 40 Million Motorists association launched a petition for this,** and I think it's a brilliant idea. **Marie:** *Wait*, *wait*, Pierre. You're forgetting that **vehicles converted to bioethanol don't benefit from the Crit'Air 1 sticker.** It's an injustice, but that's how it is.	**Pierre :** *Donc, euh,* **je pense que convertir sa voiture au bioéthanol est une super solution pour circuler librement dans les zones à faibles émissions.** **L'association 40 Millions d'Automobilistes a lancé une pétition pour ça,** et je trouve que c'est une idée géniale. **Marie :** *Attends*, *attends*, Pierre. Tu oublies que **les véhicules convertis au bioéthanol ne profitent pas de la vignette Crit'Air 1.** C'est une injustice, mais c'est comme ça.

Table 4: Example of how the model reformulates factual information from the article into a natural debate. Matching facts are color-coded. English translations are provided for clarity, alongside the original French version.

can occasionally occur in real debates, in this case it often appeared artificial and mechanical, affecting the overall naturalness of the discussion.

In addition, as shown in Table 4, the model frequently used interjections such as "uh" or expressions like "wait," which make the dialogue sound more spontaneous. However, the same "naturalness" template was often reused, with recurring patterns like "yes, but..." or "wait, wait..." appearing almost systematically in every opposition.

Another stylistic issue was the consistent use of informal address between speakers, while formal or polite forms, common in actual political debates, were rarely produced.

These stylistic and coherence issues could likely be mitigated with further refinement of the chain-of-thought prompting strategy introduced earlier. It is important to note that such stylistic issues were somewhat expected, since the news articles in our dataset cover a wide range of topics, and not all of them are suitable for real political debates.

Quantitative error statistics were then compiled (see Table 5). To calculate the statistics for logical incoherence, artificial style, and omission, we counted the debates among the 20 evaluated that received a score of 2 or lower for each corresponding criterion, respectively, coherence, style and accuracy.

Overall, the Mistral-24B Instruct model successfully transformed the articles into oral debates. In some cases, the LLM appears to adopt a two-stage approach, first producing a concise summary of the article and then generating the debate based on that summary.

Error Category	Frequency (%)
Hallucination	20
Omission	15
Logical Incoherence	50
Artificial Style	65

Table 5: Distribution of error types in the human evaluation.

3.3. Automatic Evaluation

We complemented the human analysis with automatic metrics comparing the generated debates and their associated summaries. We report results using standard similarity measures, including ROUGE (Lin, 2004), BERTScore (Zhang et al.), and cosine similarity (Li and Han, 2013). These automatic evaluations provide an objective estimation of textual similarity and semantic alignment between the debates and the expert-written summaries.

Metric	% Mean Value
ROUGE-1	36.1
ROUGE-2	16.8
ROUGE-L	19.1
BERTScore	69.5
Cosine Similarity	70.0

Table 6: Automatic evaluation of generated debates against expert-written summaries.

As we can observe in Table 6, generated debates have moderate lexical overlap but strong semantic

alignment with the summaries.

4. Benchmarking

To evaluate the quality and usefulness of the generated debates, we conducted a series of experiments comparing our dataset with the existing French political debate corpus, FREDSum. We used this dataset because, to the best of our knowledge, it is the only corpus that offers a comparable task with expert-written summaries in French. The goal was to assess whether synthetic debates could support abstractive summarization tasks and how well they align with real-world debates.

4.1. FREDSum Corpus

FREDSum provides three types of abstractive summaries that differ in their degree of abstraction and naturalness. The first version avoids coreference by systematically using proper names. The second is built from an extractive base, where key information is first selected and then rewritten into an abstractive summary. The third consists of fully natural, free-form summaries (Rennard et al., 2023). For our experiments, we used the second version as reference (abstractive summaries using an extractive base), since in our expertise it provides the best balance between faithfulness and readability. The dataset contains 115 debates for training and 29 debates for testing.

4.2. Lexical Comparison

To understand how our generated debates compare to real-world data, we computed lexical statistics for both GeneFRDebate and FREDSum. Lexical diversity was reported using two complementary measures: entropy and the MTLD (Measure of Textual Lexical Diversity) (McCarthy and Jarvis, 2010), which is independent of text length.

For GeneFRDebate, the mean MTLD score of 92.8 indicates a high level of lexical diversity, compared to 74.5 for FREDSum. Conversely, the average entropy for GeneFRDebate (7.2) is slightly lower than that of FREDSum (8.0), suggesting that the synthetic corpus maintains a lexical richness comparable to that of authentic political debates.

4.3. Experimental Setup

We evaluated how well the generated debates can be used to fine-tune small-scale language models for abstractive summarization. We tested three models: LLaMA-3B (Grattafiori et al., 2024; MetaAI, 2024), Mistral-small 7B Instruct (Chaplot, 2023), and Deepseek-R1-Llama-8B, a distilled version of Deepseek's model based on LLaMA 3.1–8B (DeepSeek-AI, 2025).

All models were fine-tuned using LoRA on both the FREDSum training set and our GeneFRDebate dataset. For GeneFRDebate, we experimented with two settings: the full dataset (283 debates) and a reduced subset of 115 debates, matching the size of the FREDSum training split, to study the impact of dataset size on performance.

The configurations were set as follows: rank 8, LoRA alpha 16, batch size 3 and 5 epochs of training. Evaluation was conducted on FREDSum test set using ROUGE-L and BERTScore metrics. The experimental prompts are reported in Appendix B (cf. Appendix 7.2).

4.4. Results and Discussion

Model		%RL	%Bertscore
ChatGPT		21.8	**72.8**
OpenAssistant		17.1	69.4
Deepseek-8B	*P*	16.0	66.3
	FF	20.2	67.5
	FG-all	21.1	69.2
	FG-train	21.5	69.1
	FFG	20.7	68.6
Mistral-7B	*P*	**23.2**	**72.7**
	FF	23.0	72.1
	FG-all	21.6	71.7
	FG-train	22.7	72.3
	FFG	16.7	64.2
LlaMa-3B	*P*	10.1	50.6
	FF	8.3	49.5
	FG-all	13.5	<u>59.9</u>
	FG-train	10.0	51.6
	FFG	11.9	55.8

Table 7: Evaluation Results: `P` for Prompt, `FF` for finetuning on FREDSum-train, `FG-all` for finetuning on full GeneFRDebate, `FG-train` for finetuning on 115 GeneFRDebate subset, `FFG` for finetuning on all GeneFRDebate and FREDSum.

In Table 7, we report the results, including both our models and the results reported in the FredSum paper, using ChatGPT (GPT-3.5) and OpenAssistant (Köpf et al.). Evaluation was done by computing the average of the scores.

As can be observed, the results of Mistral-7B and GPT-3.5 are close, although Mistral-7B outperforms ChatGPT in ROUGE-L metric. In this section,

we analyze in detail the impact of our dataset on various small-scale models, which are more accessible for experimentation. Our goal is not to identify the best model that surpasses larger LLMs, but rather to examine the effect of the data.

Deepseek-R1-Llama-8B Overall, fine-tuning improved performance compared to simple prompting. When trained on either 115 GeneFRDebate samples or the full dataset (283 debates), results were close to those obtained when fine-tuning on the original FREDSum data. However, training on the generated corpus led to an additional gain of about +1.5 in BERTScore and +1.0 in ROUGE-L. Combining both datasets did not produce significant qualitative differences.

Mistral-7B Instruct In general, we observe that Mistral-Instruct model is sensitive to overfitting during fine-tuning. In fact, fine-tuning did not improve results compared to zero-shot prompting. Varying the dataset size had no significant effect, and combining the two datasets degraded performance, likely due to overfitting issues.

LLaMA-3B This model achieved the lowest baseline scores and tended to produce more extractive outputs, often copying debate fragments instead of summarizing. However, fine-tuning on GeneFRDebate (particularly with the full 283-debate set) substantially improved both summary structure and informativeness, yielding +10 BERTScore and +3.5 ROUGE-L over simple prompting. This suggests that dataset size plays a stronger role for smaller models.

It is important to note that, to confirm the robustness of results, we re-split the GeneFRDebate subsets and repeated the fine-tuning; results were consistent, showing no statistically significant variation.

4.5. Style Analysis

Our analyses show that the GeneFRDebate data mainly influence the structure and style of the generated summaries rather than their factual content.

Deepseek-R1-Llama-8B The model produced coherent and well-organized summaries. When fine-tuned on GeneFRDebate, it generated more argumentative outputs, with clearer topic segmentation and discourse markers. For example:

> *"The debate between François Bayrou and Ségolène Royal covers several important topics, including (1) family reunification, (2) nuclear energy, and (3) national pride..."*

On FREDSum, however, it tended to summarize the debates in few sentences and reproduce fragments directly from the debates, resulting in less abstraction.

Mistral-7B Instruct For Mistral, the factual correctness and coverage were similar across datasets. However, similar to Deepseek, the summaries are slightly more detailed, often reformulating and structuring ideas. An example illustrates this difference:

> **GeneFRDebate-train:** *"The debate focuses on several issues: (1) family reunification for immigrants[...], (2) France's energy policy[...], (3) The French flag[...]"*
>
> **FredSum-Train:** *"The debate centers on two issues: (1) Family reunification in France for immigrants [...], (2) France's energy policy [...]. Finally, François Bayrou rejected Ségolène Royal's proposal to pass a law requiring every French person to display a French flag [...]"*

LLaMA-3B The smallest model benefited the most from the synthetic corpus. Fine-tuning improved the identification of speakers, reduced overcopying of dialogue fragments, and made the text flow more naturally. Below, an illustrative translated example from the model's output:

> *"François Bayrou responds to P2's question on family reunification... Ségolène Royal replies on nuclear power... Bayrou comments on the French flag."*

Increasing the size of the training set improved about 17% of the low-quality summaries that were not improved with either the 115 synthetic debate subset or the FREDSum dataset. This suggests that data quantity and stylistic consistency support better generalization for this model.

Overall, these results indicate that the GeneFRDebate corpus contributes primarily to stylistic improvements, such as making summaries more structured and discursive, while preserving factual accuracy. Therefore, it is indeed useful for studies in abstractive summarization.

5. Conclusion and Limitations

In this paper, we presented the first version of GeneFRDebate, a French synthetic dataset for abstractive political debate summarization, built from real-world news articles originally summarized for industrial purposes. Our approach combines the generation capacity of moderate-sized LLMs with human expertise for curation and validation. The

dataset was produced through controlled prompting and systematic filtering, aiming to create realistic, diverse, and coherent debates.

Despite its promising results, GeneFRDebate still has several limitations. First, the size of the dataset remains modest compared to similar corpora. We also observed some recurring issues in coherence and style, suggesting that a more refined prompting strategy could improve naturalness and reduce repetition. Future work will focus on a more detailed exploration of prompt engineering, including the integration of summaries alongside articles to help the model better identify and preserve key information during debate generation. Also, we observed some stylistic issues related to the diversity of article themes. In future work, we plan to filter the source articles and focus more specifically on political topics.

Our current annotation guide was based on information and sentence extraction, as well as named-entity identification to support human evaluation. In future releases, we plan to extend the dataset with these annotations and possibly include discourse and linguistic features, which could benefit the research community for fine-grained analysis and model training.

Last but not least, this work was limited to small and medium scale LLMs. It would be interesting to investigate how larger models perform on the GeneFRDebate corpus. Such experiments would also help assess how model scale affects the extraction of key information, reasoning quality, and potential biases in generated outputs.

6. Ethics Statement

All data used in this work come from internal industrial sources used for summarization purposes. All experiments were conducted locally using downloaded models, ensuring that no external API were used to avoid data leakage. We confirm that our data generation and evaluation processes comply with ethical research practices and respect intellectual property and privacy standards.

The GeneFRDebate dataset (debates and summaries) is available for research purposes only, under a license that permits academic use and prohibits any commercial exploitation. Interested researchers may contact us to request access to the dataset.

References

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Alexander R Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. *arXiv preprint arXiv:2106.00829*.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. Msamsum: Towards benchmarking multilingual dialogue summarization. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 1–12.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Tn. 2024. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 387–394, Mexico City, Mexico. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *CoRR*.

Jacob Kohen. 1960. A coefficient of agreement for nominal scale. *Educ Psychol Meas*, 20:37–46.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. corr, abs/2304.07327, 2023. doi: 10.48550. *arXiv preprint arXiv.2304.07327*.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Baoli Li and Liping Han. 2013. Distance weighted cosine similarity measure for text classification. In *International conference on intelligent data engineering and automated learning*, pages 611–618. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9815–9822.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

MetaAI. 2024. Llama 3 models. https://www.llama.com/models/llama-3/. Accessed: 2025-05-02.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182.

Pavan Kartheek Rachabathuni. 2017. A survey on abstractive summarization techniques. In *2017 International Conference on Inventive Computing and Informatics (ICICI)*, pages 762–765. IEEE.

Virgile Rennard, Guokan Shang, Damien Grari, Julie Hunter, and Michalis Vazirgiannis. 2023. Fredsum: A dialogue summarization corpus for french political debates. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4241–4253. Association for Computational Linguistics.

Allen Roush and Arvind Balaji. 2020. Debatesum: A large-scale argument mining and summarization dataset. *arXiv preprint arXiv:2011.07251*.

Guokan Shang. 2021. *Spoken Language Understanding for Abstractive Meeting Summarization*. Ph.D. thesis, Institut Polytechnique de Paris.

David Thulke, Yingbo Gao, Rricha Jalota, Christian Dugast, and Hermann Ney. 2024. Prompting and fine-tuning of small llms for length-controllable telephone call summarization. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 305–312. IEEE.

Don Tuggener, Margot Mieskes, Jan Milan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way?: a survey of dialogue summarization data sets. In *Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic (online), 7-11 November 2021*, pages 107–118. Association for Computational Linguistics.

Klaus Zechner and Alex Waibel. 2000. Minimizing word error rate in textual summaries of spoken language. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiutian Zhao, Ke Wang, and Wei Peng. 2024. Orchid: A chinese debate corpus for target-independent stance detection and argumentative dialogue summarization. *arXiv preprint arXiv:2410.13667*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

7. Appendices

7.1. Appendix A: Evaluation Guide

7.1.1. Objective

Evaluate the debates according to three criteria, each rated on a scale from 0 to 5.

7.1.2. Debate Style

Verify whether the generated debates correspond to the following definition:

> A debate is a discussion, often organized, on a specific topic. In its plural form (*debates*), it refers to the examination of a question within a parliamentary assembly or a large group of people (Larousse).

- **0** – Complete absence of dialogue (no exchange).
- **1** – Very poor dialogue: disorganized replies, for example.
- **2** – Dialogue format present, but no actual debate (no opposition, no argumentative progression, flat and artificial tone).
- **3** – Debate format respected: several distinct voices interact, minimal but perceptible opposition.
- **4** – Credible debate, with markers of opposition, reformulations, and more natural repetitions.
- **5** – Fluent, natural, realistic debate, with clear markers of opposition, follow-ups, and spontaneity.

7.1.3. Accuracy with the Summaries

The objective is to verify whether the information contained in the summaries appears in the debates.

If many elements in the debates do not appear in the summaries, consulting the full article may help clarify discrepancies.

- **0** – No correspondence: the debate does not cover any information from the summary.
- **1** – Very weak correspondence: only the same general theme as the summary.
- **2** – Some elements from the summary/article are included, but incompletely or inaccurately.
- **3** – The main themes of the summary/article are present, but some details are distorted or missing.
- **4** – The debate faithfully reflects the summary/article, with rare minor inaccuracies.
- **5** – The debate perfectly matches the summary/article, with no significant additions or omissions.

Procedure for Summary Analysis

Segmentation of the Summary. Divide the summary into sentences, whether simple or complex, according to their structure.

Analysis of Discourse Relations. If the summary contains complex sentences, examine the logical relations between their clauses. Also analyze the links between sentences in general: relations of cause, consequence, example, contrast, etc.

Identification of Key Elements. Identify major events and named entities, particularly people.

Search Within the Debates. Once this information has been extracted from the summary, verify its presence in the debates.

7.1.4. Coherence

Check discourse markers and logical relations, examine the fluidity of transitions, then extract the general theme of both the summary and the debate in order to compare them.

- **0** – Incoherent text: disconnected sentences, no guiding thread.
- **1** – Very weak coherence: missing transitions, unclear logic, difficult to follow.
- **2** – Minimal coherence: awkward transitions.
- **3** – Adequate coherence: general theme identifiable, transitions present but sometimes forced.
- **4** – Globally coherent debate: good transitions, clear logic, a few minor breaks.
- **5** – Fully coherent debate: smooth transitions, logical progression, theme perfectly maintained.

7.2. Appendix B: Experimental Prompts

Prompt

```
Summarize the following debate: {debate}
```

Figure 2: Main Prompt

Input For Finetuning

Mistral-7B (Instruct): [INST] Summarize the following debate: {debate} [/INST] {summary}

Other models: <source> Summarize the following debate: {debate} </source> <summary> {summary} </summary>

Figure 3: Finetuning input format for different models