Can Small-Scale LLMs Balance Content Accuracy and Speaker Faithfulness in Noisy French Dialogue Summarization?

Rim Abrougui* , **Guillaume Lechien*** , **Elisabeth Savatier, Benoît Laurent,**
Aday - Paris, France
innovations@aday.fr
* These authors contributed equally.

Abstract

Summarizing domain-specific and multi-speaker conversations, such as political debates, remains challenging under noisy ASR conditions. In industrial contexts, large language models (LLMs) are often impractical due to resource and confidentiality constraints. This work evaluates whether smaller LLMs (up to 8B parameters) can produce reliable summaries in such settings. Experiments on French debates show that noise significantly degrades accuracy and readability, while fine-tuning on clean, domain-related data improves robustness and reduces hallucinations. We also analyze person-name mentions as indicators of speaker faithfulness, finding that fine-tuning can help identify all speakers in far more debates than chain-of-thought prompting. However, evaluations on limited industrial data show that fine-tuning still struggles to generalize to unseen speakers and topics.

1 Introduction

Large Language Models (LLMs) have shown strong performance on dialogue tasks, but their growing size and computational cost limit real-world deployment, especially in industrial settings. Access to very large models is often restricted by hardware, cost, and confidentiality constraints. In our media and data-protection context, privacy requirements prevent the use of external APIs, so models must run locally, making large-scale deployment expensive. Exploring smaller and more accessible LLMs is therefore essential for cost efficiency and data sovereignty. Our task presents additional challenges. We focus on summarizing French political debates to extract speakers, discussion themes, and key arguments. Unlike casual dialogues, debates involve multiple speakers defending opposing viewpoints on complex topics.

In this work, we investigate how small LLMs (up to 8B parameters) perform abstractive summarization of these debates. Given the limited availability of French data and the noisy nature of ASR transcriptions, we simulate realistic noise on a public dataset. We compare three strategies: simple prompting, chain-of-thought prompting, and fine-tuning. Our contribution lies in evaluating the robustness of small-scale LLMs to noisy debate data and analyzing the faithfulness of Named Entity Person mentions, with a focus on speaker identification accuracy.

2 Related Work

LLMs have achieved strong results in dialogue summarization (Ramprasad et al., 2024), but their high computational demands limit their use in industrial contexts where speed, cost, and data privacy are critical. Moreover, ethical and security concerns remain central, as training and deploying LLMs often involve sensitive or proprietary data (Yao et al., 2024; Zhao and Song, 2024; Yan et al., 2024). These challenges have encouraged the adoption of smaller and more efficient models. Recent studies highlight that compact models, when properly tuned or guided, can achieve competitive performance on domain-specific tasks (Chen and Varoquaux, 2024; Wang et al., 2025).

In industrial applications, small LLMs have been adapted successfully for tasks such as telephone call summarization, where prompting and fine-tuning methods were used to control length and style (Thulke et al., 2024). However, real-world data often contain transcription errors, missing punctuation, and other noise. Previous work has explored robustness to ASR errors in low-resource domains, such as medical dialogues, by generating synthetic noisy data to improve summarizer stability (Binici et al., 2025). Earlier studies also tackled ASR issues using sub-word and phonetic representations (Li et al., 2018) or post-correction and restoration models to recover proper punctuation

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 153–157
February 26–March 1, 2026.

and casing (Dixit and Kirchhoff, 2020). Our work builds on these findings by evaluating how small LLMs perform in noisy, debate-style dialogues and by analyzing their faithfulness.

3 The Industrial Challenge: Noisy Transcripts and Missing Speakers

3.1 Dataset

Our ASR transcriptions are generated by a WFST based model using Kaldi implementation, and they lack both casing and punctuation. In our industrial setting, summarizing these debates is a key requirement. To study possible solutions, we relied on the FREDSum dataset (Rennard et al., 2023), which, to the best of our knowledge, is the only French dataset containing debates with corresponding abstractive summaries. To make FREDSum comparable to our internal data, we converted its manual transcripts into ASR-like transcripts that reproduce the characteristics of our industrial system. First, we concatenated each speech turn after removing speaker mentions at the beginning of each turn, and then applied typographical normalization. Next, we introduced different types of noise by randomly replacing some words with the out of vocabulary token, inserting interjections, and substituting certain words with their phonetic equivalents using our lexicon. Finally, to reproduce one of the specific artifacts of our production pipeline, we split the resulting text into sequences and merged them back to simulate interleaving errors observed in our real transcriptions. An example of the original and modified noisy transcript is shown in Table 1.

3.2 Experiments

For our experiments, we used four small-scale models; Flan-T5 large (Chung et al., 2024), the LLaMA-3B (Grattafiori et al., 2024; MetaAI, 2024), Mistral-7B-Instruct (Jiang et al., 2023) and a distilled version of Deepseek's model based on LLaMA 3.1–8B, namely DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025).

To illustrate the gap between clean and noisy data, we first evaluated all models using simple prompting, without adding any additional information. As shown in Table 2, performance dropped sharply on noisy debates, with Rouge-L (Lin, 2004) decreasing from 23.2 to 5.0 for Mistral and 16.0 to 9.5 for DeepSeek, and similar trends in BertScore (Zhang et al., 2020) (between -8 and -11 points). LLaMA-3B slightly improved due to repeated debate fragments, while Flan-T5 degraded further.

To evaluate our models on the noisy dataset, we conducted four experiments. All the prompts in this experiment were in French. To assess clarity the illustrations are in English. In experiment 1, we used simple prompting without providing any examples or additional context. The model received only the noisy debates and was asked to generate the summary based only on that input. This approach tested the model's ability to summarize information directly from raw text without additional guidance. In the second experiment, we incorporated specific instructions for the models, as illustrated in Figure 1. This setup encouraged the model to reason step by step before producing the final summary, allowing us to evaluate whether structured reasoning improves the quality and coherence of the generated summaries. The instructions in the chain-of-thought prompt are intentionally simple. After testing several variants, we found that when the instructions were too detailed, the models tended to focus on following each instruction literally and neglected the overall summary. Therefore, we simplified the instructions as much as possible to encourage more coherent reasoning.

Prompt

```
Summarize the following debate by following these steps:
identify the main ideas, select the essential passages,
then reformulate them in a clear, neutral, and fluid
style in a concise text of approximately 6 to 8
sentences: {debate}
```

Figure 1: Chain of Thought Prompting

In experiment 3, we fine-tuned the models using LoRA (Hu et al., 2022) on the Fredsum training set, specifically the cleaned version without the added noise described in Section 3. The LoRA configuration used a rank of 8 and an alpha value of 16 and we used a batch size of 3. The models were fine-tuned for up to 10 epochs maximum. The input structure provided to the models is illustrated in Figure 2. As we can observe, for Mistral-7B-Instruct, we used the instruction-style format with the `[INST]...[/INST]` tokens, following the model's native input convention. Other models were trained without these tokens, so we used a simpler `<source> / <summary>` structure. We also tested Mistral without the [INST] tokens, but this led to a noticeable performance drop, confirming the importance of using its instruction format.

Fredsum	Noisy
Laurence Ferrari : Nicolas Sarkozy, vos solutions ? *François Hollande* : Enfin, j'augmenterai de 25 % **l'**allocation **de** rentrée scolaire *Laurence Ferrari* : Vos solutions pour **le pouvoir d'achat** ? *Nicolas Sarkozy* : D'abord un mot sur les **syndicats** en Allemagne. **D'abord il** ne viendrait **à l'idée** de personne en Allemagne que les syndicats **appellent** à voter pour un candidat.	nicolas sarkozy vos solutions enfin j' augmenterai **euh** de 25 % **ll** allocations **d'** rentrer scolaire vos solutions pour **ls pouvoir d' achat pouvoir d' achat** d' abord un **<unk>** sur les **sindika** en allemagne d' **abords ille** ne viendrait **as** l' **heede** de personne en allemagne que les syndicats **apelle** a voter pour un candidat

Table 1: Example of normalization and noise insertion in FREDSum transcript

		R1	R2	RL	Bertscore
Deepseek-R1-R1-LLama-8B	*PC*	33.2	9.6	16.0	66.3
	PN	19.4	4.0	9.5	55.1
	CoT	20.6	4.2	10.4	57.2
	FC	**35.4**	**11.0**	**17.8**	**67.1**
	FN	33.2	10.3	17.6	65.7
Mistral-7B-Instruct-v0.3	*PC*	44.4	16.2	23.2	72.7
	PN	29.5	7.2	5.0	64.6
	CoT	32.0	8.4	15.9	65.4
	FC	18.3	4.5	11.6	58.2
	FN	9.1	8.2	7.2	51.3
Llama-3.2-3B	*PC*	21.0	4.4	10.1	50.6
	PN	23.7	3.9	11.5	59.4
	CoT	23.8	3.9	11.6	59.4
	FC	15.5	4.5	8.2	51.1
	FN	15.0	4.2	8.1	49.6
FlanT5-large	*PC*	15.1	4.2	9.7	53.5
	PN	15.1	4.0	7.7	44.9
	CoT	7.8	1.3	4.7	37.5
	FC	9.2	2.2	5.5	33.7
	FN	10.0	2.4	6.2	39.3

Table 2: Global Results: `PC` for Prompt on Clean dataset, `PN` for Prompt on Noisy dataset, `CoT` for Chain of Thought, `FC` for Finetuning on Clean dataset, `FN` for Finetuning on Noisy dataset

And last but not least, in experiment 4, we used the same configuration as in experiment 3, but with the noisy dataset.

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 2: Finetuning input format for different models

3.3 Results

For the evaluation, we report results using ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020), and we also conducted a human evaluation focused on readability. Since our inputs are noisy, we aimed to see whether the models could still generate readable summaries. As shown in Table 2, direct prompting produced acceptable results only for Mistral-7B. The chain-of-thought (CoT) strategy improved performance for both Mistral-7B and DeepSeek-R1-Llama-8B, while Llama-3B and especially Flan-T5 performed poorly on the noisy dataset. For human evaluation, we used a 1–5 scale to assess readability. Direct prompting resulted in an average score of 3.0 for Mistral and 2.5 for DeepSeek, meaning the outputs were readable with correct syntax, but some summaries were partly in English or contained fragments copied from the debates. Llama-3B and Flan-T5 both received an average of 1.5. The CoT approach did not lead to a significant improvement in readability. When fine-tuning on the clean dataset, the DeepSeek model achieved the best overall performance on both automatic metrics and human evaluation. Its outputs were more fluent and mostly in French, reaching an average readability score of 3.5, with only about 6% of summaries being partial extractions of debate fragments. For the other models, fine-tuning on the clean dataset did not help; instead, they tended to overfit and reproduce noisy input fragments. We also tested different hyperparameters (batch size, LoRA rank, and LoRA alpha), but these had no significant effect, especially on readability. Fine-tuning on the noisy dataset gave the worst performance for all models, except DeepSeek, which handled noise slightly better but still produced fragmentary outputs. Readability served as our main criterion for selecting the two most promising systems for industrial deployment. Therefore, in the next section, we focus on DeepSeek-8B (FN) fine-tuned on the clean dataset and Mistral-7B-Instruct (CoT) with chain-of-thought prompting for further analysis.

4 Focused Analysis of Top Systems

4.1 Impact of Noise Types

As discussed in section 3.1, our data contains two main types of noise: normalization noise and ASR-related errors typical of spoken transcription. We examine how these affect our top two systems. Figure 3 shows the histogram of BERTScore results.

For both systems, normalization noise is more challenging than ASR errors. The DeepSeek fine-tuned model is more robust overall, while Mistral with chain-of-thought prompting shows a larger drop under normalization noise. For ASR errors, both systems perform similarly, with very close scores, suggesting that they can capture the overall meaning even when some words or subwords from the input are missing.

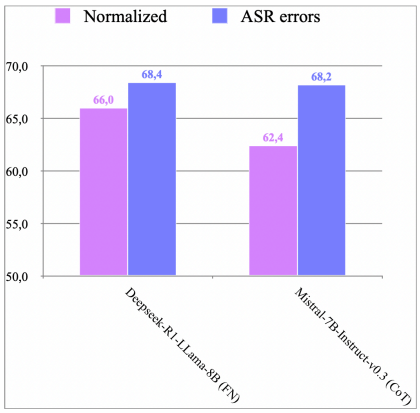

Figure 3: Impact of Noise Types - BertScore (%)

4.2 Person Name Recognition and Speaker Accuracy

In our industrial setting, providing accurate information is a key objective. We therefore evaluated the faithfulness of generated summaries by checking whether they preserved factual information from the input, focusing particularly on Named Entities of type Person. For this analysis, we used the multilingual version of GLiNER (Zaratiana et al., 2024), to identify person entities. Incorrect detections were manually removed, and equivalent mentions (e.g., *M. Mitterrand*, *François Mitterrand*) were grouped together using simple heuristics. We then measured the proportion of correct mentions and hallucinated entities in the generated summaries (Table 3). Overall, the fine-tuned DeepSeek model produced more accurate and faithful outputs, with fewer hallucinations than the Mistral model. We also focused on speaker identification. To evaluate this, we manually annotated the previous set of Named Entities to check whether the models could correctly identify the different participants. A summary was considered correct only if all speakers in the debate were detected. We observed that the fine-tuned DeepSeek model achieved better speaker accuracy, correctly identifying all speakers in 48.3% of the debates and one speaker in 10% of the cases. In contrast, the Mistral CoT system performed worse, fully detecting speakers in only 17.2% of the debates and focusing on a single participant in another 17%, which often led to biased summaries.

Model	Correct	Omission	Hallucination
DeepSeek-8B (FN)	62.9	37.1	5.6
Mistral-7B (CoT)	8.2	91.8	73.1

Table 3: Error Distribution of Named Entity *Person*

4.3 Preliminary evaluation on the industrial dataset

Due to confidentiality and time constraints, we only had access to five industry debates, and these have been summarized manually. We conducted a preliminary qualitative evaluation to observe model behavior in real conditions. Both models achieved similar automatic scores: ROUGE-L of 14 and BERTScore of 60.7 for DeepSeek-8B (FN), versus ROUGE-L of 13 and BERTScore of 59.5 for Mistral-7B (CoT). In human evaluation, both produced readable summaries, except for one unreadable case from Mistral. Regarding speaker faithfulness, performance was comparable: some speakers were omitted, while DeepSeek occasionally hallucinated interactions or attributed quotes to speakers seen during fine-tuning. Overall, Mistral captured discussion themes more clearly, but accurate speaker identification remained difficult for both models.

5 Conclusion

This study examined how small language models perform dialogue summarization under noisy, resource-constrained industrial conditions. Results show that transcription noise severely reduces accuracy and readability across all models. Fine-tuning on clean, domain-related data improves robustness and reduces hallucinations, especially for person entities. However, fine-tuned models still struggle to generalize to real industrial data. In such cases, chain-of-thought prompting can yield more balanced and general summaries. Future work includes mixed clean-noisy fine-tuning and parameter-efficient adaptation to improve faithfulness at low cost.

Acknowledgments

We sincerely thank the company members who took the time to provide access to the internal debates, allowing us to simulate our corpus and advance our work within the context of their activities. All experiments were conducted locally using downloaded models, ensuring that no external APIs were used and that no data was exposed.

References

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F Chen, and Stefan Winkler. 2025. MEDSAGE: Enhancing Robustness of Medical Dialogue Summarization to ASR Errors with LLM-generated Synthetic Dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23496–23504.

Lihu Chen and Gaël Varoquaux. 2024. What is the Role of Small Models in the LLM Era: A Survey. *Preprint*, arXiv:2409.06857.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

MSSRK Dixit and Sravan Bodapati Katrin Kirchhoff. 2020. Robust prediction of punctuation and truecasing for medical asr. *ACL 2020*, page 53.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Xiaopeng Zhang, Jin Li, Hongkai Xiong, and Qi Tian. 2023. From CLIP to DINO: Visual Encoders Shout in Multi-modal Large Language Models. *arXiv preprint arXiv:2310.06825*.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension. *Preprint*, arXiv:1804.00320.

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

MetaAI. 2024. Introducing LLaMA 3: Advancing Open Foundation Models for AI Everywhere. Accessed: 2025-05-02.

Sanjana Ramprasad, Elisa Ferracane, and Zachary C Lipton. 2024. Analyzing LLM Behavior in Dialogue Summarization: Unveiling Circumstantial Hallucination Trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12549–12561.

Virgile Rennard, Guokan Shang, Damien Grari, Julie Hunter, and Michalis Vazirgiannis. 2023. FRED-Sum: A Dialogue Summarization Corpus for French Political Debates. *arXiv preprint arXiv:2312.04843*.

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.

Linyong Wang, Lianwei Wu, Shaoqi Song, Yaxiong Wang, Cuiyun Gao, and Kang Wang. 2025. Distilling Structured Rationale from Large Language Models to Small Language Models for Abstractive Summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25389–25397.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (LLMs): A survey. *arXiv preprint arXiv:2403.05156*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Guoshenghui Zhao and Eric Song. 2024. Privacy-preserving large language models: Mechanisms, applications, and future directions. *arXiv preprint arXiv:2412.06113*.