

# Augmentation of a Clinical Note Corpus in French using LLMs

Nicolas Hiebel<sup>1</sup>   Priyanshu Priya<sup>2</sup>   Joël Mba Kouhoue<sup>1</sup>   Jesús Lovón<sup>2</sup>  
Gaël Dias<sup>2</sup>   Gaël Varoquaux<sup>1,3</sup>   Judith Abécassis<sup>1</sup>

(1) Soda, Inria Saclay, Palaiseau, France

(2) Normandie Caen Univ, UNICAEN, ENSICAEN, CNRS, GREYC, France

(3) Probabl, France

(1) `firstname.lastname@inria.fr` (2) `firstname.lastname@unicaen.fr`

## ABSTRACT

The PARHAF corpus was recently introduced to share clinical notes in French without privacy concerns. The purpose of this work is to extend the PARHAF corpus. We use the patient information included in PARHAF to generate new clinical notes with a large language model and we compare synthetic and real clinical notes through statistical analyses. Results show that the synthetic notes share many characteristics with real clinical notes, but feature a significantly less diverse vocabulary. With this limitation identified, we propose a methodology to further expand the corpus with new clinical scenarios and specialties using the UNESS database.

**KEYWORDS:** Natural Language Processing; Large Language Model; Text Generation; Synthetic Clinical Notes.

## 1 Introduction

Publicly available corpora in the clinical domain are scarce, particularly in languages other than English, for confidentiality considerations. This lack of available corpora is an obstacle to developing and sharing of Natural Language Processing (NLP) tools for clinical applications ([Šuster et al., 2017](#)).

The PARHAF corpus ([Tannier et al., 2026](#)) was created to support biomedical NLP research in French. Given a plausible clinical scenario, 104 clinicians produced notes as they would in routine practice. PARHAF contains 6,190 documents in total from 4,259 patient scenarios and 20 different specialties.

However, this number of documents is not sufficient to train or fine-tune state-of-the-art NLP models due to the vast amount of data required ([Yang et al., 2023](#)). Augmenting the PARHAF corpus to cover more clinical situations could help address this limitation, and the performance of current Large Language Models (LLMs) makes this a promising approach ([Huang et al., 2026](#)). Nevertheless, LLMs do not necessarily perform well in specialized domains ([Soroush et al., 2024](#)) and languages different than English ([Wiechetek et al., 2024](#); [Marco et al., 2024](#)).

In this work, we propose a text generation strategy to augment the PARHAF corpus. We use patient information and few-shot examples to generate new clinical notes that follow a patient and specialty distribution similar to that of PARHAF. We assess the similarity between the synthetic notes and the real notes with several statistics. We then propose an approach to further augment the PARHAF corpus, expanding it to new patients and medical specialties.

## 2 Validation of Augmentation Strategy

We first focus on generating documents from the same specialties as the PARHAF corpus in order to validate our generation strategy. We use a 31-billion-parameter version of the Gemma-4 model <sup>1</sup> to generate synthetic clinical notes.

Each clinical note in the PARHAF corpus is associated with a fictional scenario containing demographic information (sex and age), admission mode, discharge mode, and the primary diagnosis. We use this information to guide the model’s generation towards a specific (and fictional) patient.

In addition to the patient’s information, we give the model 10 examples of clinical notes with the same specialty following a few-shot strategy. We use BM25 to retrieve examples that are similar to the reference clinical note associated with the patient.

The usefulness of the PARHAF corpus for downstream NLP tasks has not yet been evaluated, making the assessment of the synthetic version’s quality even more difficult. Thus, we start with a statistical comparison of the real and the synthetic corpora using spaCy<sup>2</sup>. We selected different statistics based on traditional corpus comparison (number of words, vocabulary size...) and observations made during an early stage of generation (number of line breaks, number of punctuation marks...).

| Corpus       | Docs. | Tokens    | Vocab. | Sentence /doc. | Sentence length | Dependency length | Punctuation /sentence | Stopwords proportion |
|--------------|-------|-----------|--------|----------------|-----------------|-------------------|-----------------------|----------------------|
| PARHAF       | 4,259 | 2,881,520 | 64,964 | 56.46±40.41    | 12.00±13.89     | 3.96±10.30        | 1.74±2.45             | 0.26                 |
| PARHAF Synth | 4,259 | 2,297,261 | 31,847 | 46.85±23.79    | 11.56±11.48     | 3.37±5.50         | 1.70±2.06             | 0.26                 |

Table 1: Statistical comparison between the real corpus (PARHAF) and the synthetic corpus (PARHAF Synth). ±: standard deviation.

The results of the statistical analysis are shown in Table 1. For the same number of documents, the synthetic corpus contains fewer tokens than the real corpus (584,259 fewer tokens). The difference in vocabulary size stands out, and our current generation method shows lower vocabulary diversity than the real PARHAF corpus. Despite the synthetic corpus having around 20% fewer tokens than PARHAF, its vocabulary is 51% smaller.

Synthetic documents also exhibit lower structural diversity. While documents are shorter (there are on average 10 fewer sentences per document than in PARHAF), the difference in standard deviations shows that their lengths are more homogeneous. The same observation holds for dependency length.

Other metrics do not show a significant difference between the two corpora. Increasing vocabulary diversity appears to be the main point of focus for improving generation quality, further highlighting the importance of diversity.

## 3 Enhancing Diversity in the Synthetic Corpus

The strategy validated in Section 2 is limited to PARHAF’s 4,259 patient scenarios and 20 specialties. It can densify existing clinical settings but cannot extend coverage to new specialties, diagnoses, or

<sup>1</sup>[google/gemma-4-31B-it](https://google/gemma-4-31B-it)  
<sup>2</sup><https://spacy.io>

care pathways. Moreover, despite containing only 20% fewer tokens, the synthetic corpus has a 51% smaller vocabulary and lower structural variability than PARHAF, showing that patient metadata and same-specialty exemplars provide insufficient clinical and lexical diversity. We therefore propose a knowledge-enhanced retrieval-augmented generation (RAG) pipeline combining fictional patient cases, external biomedical knowledge, and PARHAF-based stylistic guidance.

**Proposed Pipeline.** The five-stage pipeline converts UNESS *progressive dossiers* (PDs) (temporally-aware fictional pedagogical cases for medical training and assessment) into fictitious French clinical notes (UNESS, 2023) as follows.

**Stage 1: Case extraction.** An LLM converts each PD into a structured scenario containing the patient profile, symptoms and history, examination and test findings, diagnosis, treatment, and follow-up, together with admission and discharge modes, type of care, and ICD-10 information when available. An LLM-based evaluator checks faithfulness, completeness, and consistency on this scenario; inadequate extractions are corrected, regenerated, or discarded.

**Stage 2: Knowledge augmentation.** The scenario is enriched with external knowledge bases, such as PubMed (Canese and Weis, 2013), UMLS (Bodenreider, 2004), and LiSA (UNESS, 2024), to add disease-specific context, terminology variants, and clinically coherent links among symptoms, investigations, diagnosis, and treatment.

**Stage 3: Exemplar (few-shot) retrieval.** PARHAF clinical notes are retrieved by specialty, document type, diagnosis, and clinical context to guide structure and style. For unseen specialties, exemplars from related specialties are used, while factual content remains grounded in the fictional scenario and retrieved knowledge.

**Stage 4: Clinical note generation.** The generative LLM receives the structured scenario, augmented knowledge, and retrieved exemplars and produces a single French clinical note in the target document format. The prompt enforces factual consistency, prohibits unsupported diagnoses and treatments, and explicitly labels the clinical note as fictitious.

**Stage 5: Evaluation and filtering.** Clinical notes failing factual-consistency, clinical-coherence, structural, privacy, or safety checks are regenerated or excluded. Evaluation covers diversity, quality, and downstream utility. Diversity is assessed by recomputing the statistics in Table 1. For quality, an LLM-as-a-judge evaluates format validity, clinical coherence, terminology, factual alignment, knowledge consistency, and stylistic authenticity. Borderline outputs and a representative sample of accepted notes are then reviewed by clinicians or UNESS pedagogical experts, as subtle errors such as implausible dosages or inappropriate management require expert assessment. For downstream utility, models are trained on human-authored fictitious data, generated data, and their combination to evaluate their performance on downstream tasks like named entity recognition.

## 4 Conclusion

We investigated LLM-based augmentation of the PARHAF corpus. Although the generated synthetic notes reproduce several statistical characteristics of PARHAF, they exhibit substantially lower lexical and structural diversity. To address this limitation, we proposed a knowledge-enhanced RAG pipeline combining UNESS clinical notes, biomedical knowledge, and PARHAF few-shot exemplars. Future work will implement and evaluate this pipeline for report quality, corpus diversity, and downstream clinical NLP utility.

# References

- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:267–270.
- Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1):2013.
- Bofeng Huang, Jacques Sun, Diane Bouchacourt, Nicolas Barascud, and Fajwel Fogel. 2026. [Where does the signal live? a web data recipe for medical encoder pretraining](#). *Preprint*, arXiv:2606.22079.
- Guillermo Marco, Julio Gonzalo, M.Teresa Mateo-Girona, and Ramón Del Castillo Santos. 2024. [Pron vs prompt: Can large language models already challenge a world-class fiction author at creative text writing?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 19654–19670, Miami, Floride, États-Unis. Association for Computational Linguistics.
- Ali Soroush, Benjamin S. Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W. Charney, Girish N Nadkarni, and Eyal Klang. 2024. [Large language models are poor medical coders — benchmarking of medical code querying](#). *NEJM AI*, 1(5).
- Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. [A short review of ethical challenges in clinical natural language processing](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87, Valencia, Spain. Association for Computational Linguistics.
- Xavier Tannier, Salam Abbara, Rémi Flicoteaux, Youness Khalil, Aurélie Névéol, Pierre Zweigenbaum, and Emmanuel Bacry. 2026. [Parhaf, a human-authored corpus of clinical reports for fictitious patients in french](#). *Preprint*, arXiv:2603.20494.
- UNESS. 2023. Webinaire : Atelier “dossiers progressifs et questions isolées” sur sides ng – jeudi 9 avril. <https://tinyurl.com/4fn4u42b>. Updated 7 April 2023; accessed 25 July 2026.
- UNESS. 2024. LiSA : de nouvelles fonctionnalités pour les étudiants et enseignants. <https://tinyurl.com/39xh9p3n>. Updated 29 February 2024; accessed 25 July 2026.
- Linda Wiecheteck, Flammie Pirinen, Maja Lisa Kappfjell, Trond Trosterud, Børre Gaup, and Sjur Nørstebø Moshagen. 2024. [The ethical question – use of indigenous corpora for large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 15922–15931, Torino, Italia.
- Eric Yang, Matthew D Li, Shruti Raghavan, Francis Deng, Min Lang, Marc D Succi, Ambrose J Huang, and Jayashree Kalpathy-Cramer. 2023. [Transformer versus traditional natural language processing: how much data is enough for automated radiology report classification?](#) *The British Journal of Radiology*, 96(1149).