Hallucination Detection in Automatically Generated Medical Reports: An Optimization Approach for Semantic Layers and Adaptive Thresholds

Souhir Khessiba[1,] Nadège Alavoine [1] , Damien Forest [1]

(1) PraxySante, Paris, France

Souhir.khessiba@praxysante.fr , nadege.alavoine@praxysante.fr , damien.forest@praxysante.fr

Résumé

Les Modèles de Langage (LLM) sont susceptibles aux hallucinations, générant parfois des informations inexactes d'ou un risque non négligeable, notamment dans le domaine médical où la fiabilité est essentielle. Cet article aborde deux objectifs : améliorer la qualité des dossiers médicaux et renforcer la fiabilité des cohortes de recherche. Nous présentons un système de détection des hallucinations dans les résumés médicaux générés par IA en optimisant les couches sémantiques de BERT. Notre méthodologie exploite BERT Score pour évaluer la similarité entre les phrases des rapports générés et des transcriptions originales. Notre contribution principale introduit un mécanisme à double seuil—critique et alerte—optimisé par l'algorithme Tree Parzen Estimator, contrairement aux approches traditionnelles à seuil unique. Les résultats démontrent des améliorations significatives dans la détection des hallucinations, avec une précision et un rappel supérieur aux méthodes de référence.Bien que notre étude soit limitée à la langue française,le système proposé assure améliore la fiabilité des informations médicales, répondant aux objectifs d'amélioration de la qualité documentaire et d'intégrité des données de recherche.

Abstract

Large Language Models (LLMs) are susceptible to hallucinations, generating inaccurate information—a critical concern in healthcare where precision impacts patient safety. This paper addresses two objectives: improving medical record quality and enhancing research cohort reliability. We present a system for detecting hallucinations in AI-generated medical summaries by optimizing BERT's semantic layers. Our methodology leverages BERTScore to evaluate similarity between sentences from generated reports and original transcriptions. Our main contribution introduces a dual-threshold mechanism—critical and alert—optimized using Tree Parzen Estimator, unlike traditional single-threshold approaches. Results demonstrate significant improvement in hallucination detection. The proposed system ensures accuracy of medical information, fulfilling objectives of enhancing documentation quality and research data integrity.

Mots-clés : Détection d'hallucinations, Comptes-rendus médicaux, Modèles BERT, Tree Parzen Estimator, Bert Score, Optimisation des couches

Keywords : Hallucination Detection, Medical Reports, BERT Models, Tree Parzen Estimator, Bert Score, Layer Optimization

1 Introduction

Hallucination in Large Language Models (LLMs) represents one of the major challenges in the development of artificial intelligence technologies today. This phenomenon, where systems generate incorrect information presented as factual, raises fundamental questions about their reliability across various application domains. In the medical context, this issue takes on particular importance. Medical summaries require absolute precision as they guide clinical decisions and patient monitoring. The integration of LLMs in this sector offers promising prospects for improving administrative efficiency and documentation, but the risks associated with hallucinations are amplified by the critical nature of healthcare. Erroneous information in a medical report can lead to serious consequences, from inappropriate diagnoses to inadequate treatments (Maynez et al., 2020).

Our work focuses on two primary objectives. First, we aim to improve the quality of medical records by developing a novel hallucination detection system that is used for French language which has less ressources than English This system leverages BERTScore, a metric that evaluates the semantic similarity between generated and reference texts using the contextual embeddings from BERT models. By optimizing the semantic layers of BERT models specifically for hallucination detection, we seek to enhance the accuracy of generated medical text and minimize the impact of hallucinations, ultimately ensuring the reliability of AI-generated content in critical healthcare settings. Second, we seek to enhance the reliability of automatically generated medical notes from consultation transcriptions by implementing a dual-threshold mechanism—critical and alert—that is algorithmically optimized rather than relying on traditional single empirically fixed thresholds. By leveraging BERTScore (Zhang et al., 2020) to evaluate semantic similarity between pairs of sentences from automated reports and original transcriptions, our approach offers a more nuanced and effective solution to the hallucination problem in medical text summarization.This system tackles the specific challenges posed by specialized medical terminology, intricate causal relationships, and varied care pathways, all of which require robust verification mechanisms for tasks such as generating medical documents, summaries, or reports using LLMs.

2 Related work

The development of artificial intelligence technologies has transformed natural language processing (NLP), particularly through the emergence of Transformer-based architectures (Vaswani et al., 2017). Models such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., s. d.) GPT-3 (Brown et al., 2020) and BART (Lewis et al., 2020), have revolutionized NLP across various domains, including the medical sector. These advances have enabled the automation of medical summary creation, significantly facilitating clinical documentation and administrative tasks in healthcare facilities. Alongside the advancement of LLM models, attention toward their limitations and potential risks has also increased. One of the most significant challenges with LLMs is the phenomenon of hallucination—where models generate content that is unfaithful to the source information or factually incorrect (Liu et al., 2023). (Carlini et al., 2021) demonstrated that LLM can be prompted to extract and generate private information from their training data, such as email addresses and phone numbers. This memorization behavior qualifies as hallucination since the model produces content unfaithful to the source input, generating private details absent from the immediate context, which also raises significant privacy concerns.

In the medical field, hallucinations present particularly many risks that extend beyond privacy concerns. Advanced models like Llama 3 (Touvron et al., 2023) and GPT-4o have demonstrated impressive capabilities in generating meaningful medical content and passing medical examinations (Kung et al., 2023), yet they remain susceptible to critical reliability issues. Researchers have identified two distinct categories of hallucinations affecting medical applications (Li et al., 2023): factual hallucinations, where generated information contradicts verifiable medical knowledge, and faithfulness hallucinations, where content deviates from the specific patient context provided. This distinction is especially critical given the highly contextualized and personalized nature of medical records.

To overcome these key challenges, researchers have developed various approaches for hallucination detection in LLM-generated medical content. These methods can be broadly categorized into reference-dependent and reference-free approaches. Reference-dependent metrics compare model outputs against verified knowledge sources, with specialized benchmarks like Med-HALT (Pal et al., 2023) specifically addressing hallucinations in medical contexts. These approaches evaluate the fidelity of generated content by measuring discrepancies against established medical knowledge databases or source documents. Reference-free approaches offer alternative detection methods that don't require external reference materials. Uncertainty-based methods as proposed by (Rebuffel et al., 2021) (Popat et al., 2018), analyze token probabilities, operating under the assumption that low-confidence predictions correlate with hallucinated content. More applicable to black-box scenarios, consistency-based detection methods generate multiple responses to the same prompt and measure their agreement. These techniques employ various similarity metrics including BLEU-based variation ratio (Huang et al., 2025), n-gram approximation (Manakul et al., 2023), and BERTScore (Zhang et al., 2020). BERTScore has proven particularly valuable for medical applications by leveraging contextual embeddings to assess semantic similarity between texts rather than relying on lexical matching—an important distinction in medical contexts where terminology variations are common but semantic integrity remains essential. Among the BERT-based models specialized for medical applications, several French language variants have been developed. ClinicalBERT (Huang et al., 2020) adaptations for French address challenges like gender agreement and terminology variations specific to the French healthcare system. CamemBERT-Bio (Touchent & de la Clergerie, 2024) extends the original CamemBERT (Antoun et al., 2024) with French biomedical texts, improving performance on medical entity recognition in French clinical documentation. Additionally, DrBERT was designed for French clinical applications, trained on medical reports from French hospitals (Labrak et al., 2023), while FlauBERT (Le et al., 2020) has been fine-tuned for healthcare applications. The original CamemBERT has also been applied to medical tasks through domain adaptation. BART-base-French leverages a denoising autoencoder architecture and is fine-tuned on domain-specific medical data to generate accurate French medical summaries while reducing hallucinations (Lewis et al., 2020). These models are valuable for French-speaking healthcare systems, where English-based models often fail to capture linguistic nuances and specialized French medical vocabulary. While these BERT-based models show promise for French medical NLP, research on optimization of BERT layers has shown important findings. (Zhang et al., 2020) found that using the middle layers of BERT rather than the final layer produces better correlation with human judgments when evaluating text similarity. This underscores the need for careful optimization of both the BERT layers and, consequently, the threshold settings to enhance performance in detection tasks. However, despite the recognized importance of threshold selection in classification and detection systems, research on optimizing thresholds specifically for hallucination detection remains limited. This gap is particularly significant in medical contexts, where the consequences of false positives and false negatives vary greatly in severity. The determination of appropriate thresholds represents a crucial task that has received insufficient attention in the current literature on medical hallucination detection.

3 Proposed methodology

In this section, we present a novel approach to hallucination detection in LLM through targeted optimization of semantic representations and detection thresholds which is illustrated in Figure 1. Current literature indicates that the most effective semantic representations in transformer-based models are not necessarily stored in the final layer, with multiple studies demonstrating that intermediate layers often contain more relevant information for specific semantic tasks. Building on this insight, our research implements a two-phase optimization process. Phase 1 focuses on optimizing the model layers by evaluating multiple BERT variants across all their respective layers, using semantic similarity corpora to determine which specific combination yields the most effective vector representations.

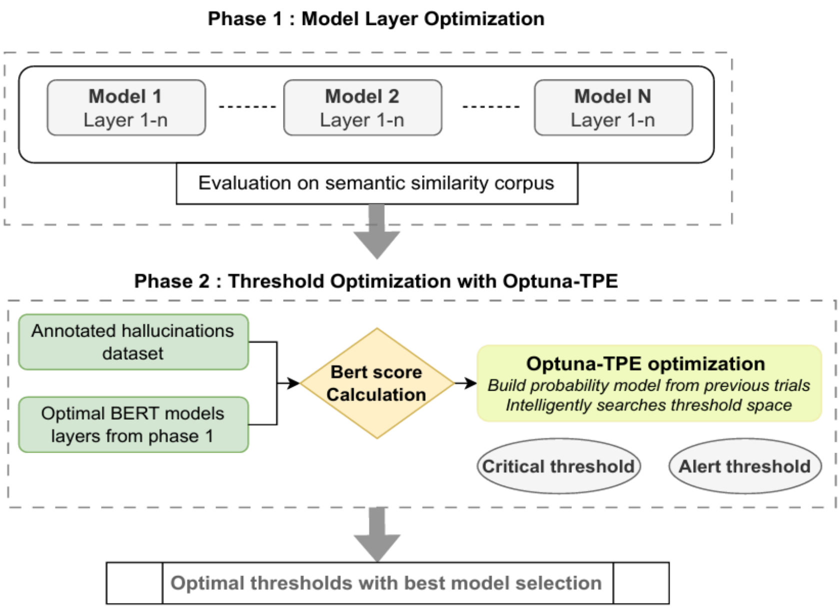

FIGURE 1: Anti-hallucination system process

Phase 2 leverages these optimal model-layer combinations to calculate BERTScore between reference and generated text, utilizing a specialized hallucination dataset containing medical transcripts, hallucinated medical reports, and their corresponding ground truth annotations of identified hallucinations in medical records. In the implementation, we perform pairwise comparisons between each sentence in the medical report and all sentences in the reference transcription. After preprocessing the text (including segmentation based on punctuation), we retain the highest similarity score for each report sentence, which represents the most semantically similar source-target pair. This approach operates on the principle that if a sentence in the medical report demonstrates high semantic proximity to any sentence in the transcription, it is unlikely to contain hallucinated content. Optuna's Tree-structured Parzen Estimator (TPE) algorithm is a hyperparameter optimization method designed to efficiently search the hyperparameter space by modeling the probability distribution of parameters (Akiba et al., 2019). Unlike traditional grid or random search methods, TPE focuses on areas of the search space that are more likely to yield better results, making it particularly effective for complex optimization tasks. In our approach, TPE is employed to intelligently optimizes two distinct thresholds: an Alert Threshold for potential hallucinations requiring review, and a Critical Threshold for severe hallucinations needing immediate intervention. This dual-threshold approach, combined with data-driven optimization rather than empirically set thresholds, enhances the system's ability to accurately distinguish between different severity levels of hallucinations while selecting the optimal model-layer combination for maximum detection performance.

3.1 Corpora for Semantic Similarity Evaluation

For the evaluation of semantic similarity in French medical texts, we employed two specialized corpora with different characteristics and annotation approaches, providing complementary perspectives on model performance (Table 1). CLISTER focuses specifically on clinical case reports with expert medical annotations, while DEFT 2020 offers a broader spectrum of medical and general content from varied french sources. Both datasets feature comparable train/test splits and utilize the same 0-5 similarity scale, enabling consistent evaluation across different textual domains. These two corpora served as benchmarks for evaluating LLM models on semantic textual similarity (STS) tasks.

Total sentence pairs	***Train set***	***Test set***	***Similarity scale***
CLISTER	600	400	0-5
DEFT 2020	600	410	0-5

TABLE 1: Details and specifications of DEFT2020 and CLISTER Corpora

3.2 Semantic Representation Selection and Optimization

3.2.1 BERT-based Models

Our study involved the evaluation of five BERT-type models, specifically selected for their relevance in processing medical texts in French. This selection reflects our main objective of working with models trained in the French language. We evaluated CamemBERT, a RoBERTa-based model trained on general French text from the OSCAR dataset (Antoun et al., 2024), CamemBERT-Bio, which extends the base model with fine-tuning on French biomedical corpora (Touchent & de la Clergerie, 2024). DrBERT (Labrak et al., 2023), developed specifically for French medical documentation with training on clinical documents from the French healthcare system; FlauBERT, an alternative French language model developed by (Le et al., 2020) to assess performance of general-purpose models on specialized content; and BioClinical BERT (Alsentzer et al., 2019), which was included to evaluate cross-lingual transfer capabilities from English to French medical contexts due to its training on both biomedical literature and clinical texts.

3.2.2 Optimization of BERT Layer Representations

To improve the quality of semantic representations, we developed a systematic approach for optimizing BERT layer embeddings illustrated in Figure 2. Inspired by BERTScore, we performed an extensive evaluation across every layer of the model to determine the most semantically meaningful representation space. For each BERT layer, we independently computed BERTScore, examining three critical factors: (1) correlation with human-rated semantic relevance, (2) precision in semantic similarity scoring, and (3) depth of contextual embedding analysis. This process aimed to isolate the optimal layer, the representation space that delivered the strongest semantic encoding for downstream tasks.

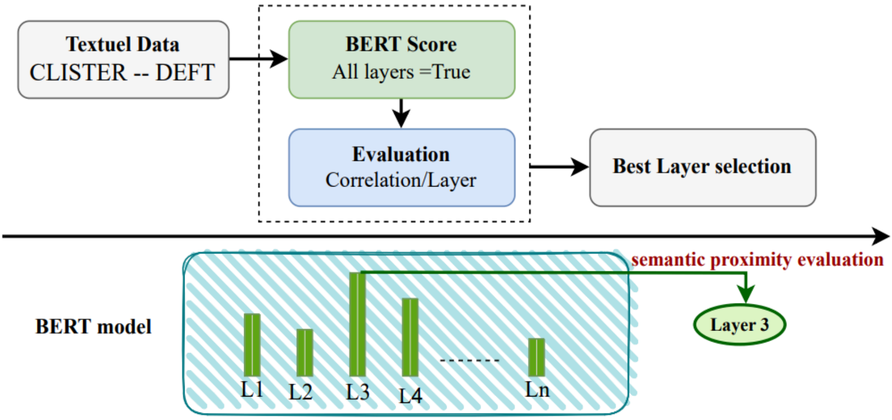

FIGURE 2: BERT layer optimization process for semantic representation

3.3 Threshold optimization for hallucination detection using Optuna

The optimization of detection thresholds represents a critical component in developing an effective hallucination detection system. Rather than relying on empirically determined threshold values, we implemented an optimization approach using Optuna (Akiba et al., 2019), which is an Hyperparameter Optimization (HPO) framework specifically designed for Machine Learning (ML) applications. Our process used the Optuna-Tree-Parzen-Estimator (TPE) algorithm, which offers significant advantages over traditional grid search or random search methods (Liashchynskyi & Liashchynskyi, 2019). The TPE algorithm functions by modeling the relationship between hyperparameters and their corresponding objective values as probability distributions. This approach enables more efficient exploration of the parameter space by building two models: one for hyperparameters yielding the best results and another for those yielding suboptimal results. As the optimization progresses, the algorithm increasingly samples from regions that have historically produced better performance, while still maintaining sufficient exploration of the parameter space. We implemented a sequential two-stage threshold optimization process. First, we optimized the Critical Threshold (CT) to identify severe hallucinations requiring immediate intervention.

Subsequently, we optimized the Alert Threshold (AT) while keeping the CT fixed at its optimal value. Importantly, we tailored our objective functions to the specific characteristics of each data type. For synthetic data, we prioritized maximizing precision and F1 score while minimizing false positives for the CT, and maximizing true positives for the AT. For authentic clinical data, we focused on maximizing precision and minimizing false positives for the CT, while optimizing both F1 Score and true positives for the AT. This differentiated approach reflects the distinct requirements of each environment: synthetic data allows for balanced metric optimization, while clinical contexts demand minimizing false alarms that could lead to unnecessary interventions. This adaptation ensures both theoretical robustness and practical reliability in medical applications.

Parameter	*Critical Threshold (CT)*	*Alert Threshold (AT)*
Range	{0.2,0.85}	{Optimal_CT +10%}
Optimization order	First	Second
Sample	TPE	
N Trails	30	

TABLE 2: Threshold optimization settings for hallucination detection system

During each trial, the optimization process calculated BERTScore between reference texts and potentially hallucinated texts using the optimal model-layer combination identified in Phase 1. These scores were then compared against the ground truth annotations from our specialized hallucination datasets containing medical transcripts and reports. The TPE algorithm systematically refined the threshold values based on performance feedback from each trial, ultimately converging toward optimal threshold values for our specific detection task. This two-stage optimization approach ensures that both thresholds are calibrated in a complementary manner, creating a dual-threshold system capable of distinguishing between different severity levels of hallucinations.

4 Experimentation

4.1 Hallucination Dataset

For this study, we used a synthetic corpus of 60 medical reports specifically constructed to optimize detection thresholds and alert parameters for our hallucination detection models. To create this synthetic dataset, we collected authentic medical transcriptions, generated alternative reports using language models, and manually injected hallucinations ourselves. This specialized synthetic dataset contains controlled hallucination instances paired with accurate transcriptions as ground truth, allowing systematic evaluation of different BERT variants in the medical domain. Additionally, we evaluated our models on a small but valuable real-world dataset consisting of 10 authentic anonymized clinical reports obtained from university hospital centers (CHUs), which contain genuine hallucinations detected through comparison with original physician-patient transcriptions. This proprietary clinical dataset remains confidential due to privacy requirements. Both datasets serve as evaluation benchmarks for our hallucination detection models.

4.2 Results and discussion

- ***Performance Evaluation of Optimized BERT Variants***

The analysis of optimal layer selection reveals distinct patterns across BERT variants (TABLE 3), with each model exhibiting unique preferences for semantic representation. CamemBERT-Large performs optimally at different layers depending on the corpus (layer 5 for CLISTER, layer 11 for DEFT), while FlauBERT consistently excels with its initial layer (0) across all evaluation scenarios. Dr-BERT-7G-cased demonstrates a marked difference between corpus-specific preferences (layer 3 for CLISTER, layer 12 for DEFT). These variations highlight how each model encodes semantic information at different architectural depths, with no consistent pattern across variants. Optuna efficiently identified these optimal layers with minimal computational overhead, providing a straightforward optimization approach that revealed the specific layer where each model achieves its best semantic representation for hallucination detection tasks.

Models	***Number of layers***	***Best layer on CLISTER***	***Best layer on DEFT***	***Best layer on mixed corpora***
CamemBERT-Large	24	5	11	7
CamemBERT-Bio	12	11	9	11
DrBERT-7G-cased	12	3	12	3
FlauBERT	12	0	4	0
Bio-Clinical-BERT	12	1	4	2

TABLE 3: Results of optimized best layer selection for BERT variants

- ***Hallucination Detection Threshold Optimization Results using synthetic data***

Table 4 presents the results of the optimized hallucination detection thresholds. We observe that the optimal critical thresholds vary significantly across models, demonstrating that each model requires a specific confidence threshold to achieve its best performance for hallucination detection. Similarly, the optimal alert thresholds follow a comparable pattern of model-specific variation. Optuna was effectively employed to determine these optimal thresholds, converging on the best solutions within relatively few trials as indicated in the "Best trials" column, demonstrating its efficiency for hyperparameter optimization. These optimized thresholds represent the best-performing configurations on our evaluation dataset, suggesting that threshold optimization should be considered an essential step when deploying BERT-based models for hallucination detection tasks.

Models	***Optimal critical threshold***	***Optimal alert threshold***	***Best trials***
CamemBERT-Large	0.703	0,7300	*3*
CamemBERT-Bio	0.721	0,790	*4*
DrBERT-7G-cased	0,683	0,735	*8*
FlauBERT	0.890	0.900	*5*
Bio-Clinical-BERT	0,806	0,845	*7*

TABLE 4: Results of optimized hallucination detection thresholds using synthetic dataset

Table 5 presents performance metrics for BERT variant models in hallucination detection across both critical and alert threshold optimization. For critical threshold settings, representing the best compromise between precision and recall, FlauBERT demonstrates superior overall performance with the highest F1-score (0.862) and precision (0.956). This indicates excellent capability in minimizing false positives while maintaining good recall, making it particularly valuable for balanced detection scenarios. CamemBERT-Bio achieves the highest recall (0.914) among all models with critical threshold, suggesting stronger sensitivity in capturing potential hallucinations, though at the cost of lower precision (0.780). This trade-off is important to consider for applications prioritizing comprehensive detection.

Models	***Performance metrics for optimal critical threshold***			***Performance metrics for optimal alert threshold***	
	Precision	***Recall***	***F1-Score***	***Precision***	***Recall***
CamemBERT-Large	0.896	0.800	0.845	0.555	0.689
CamemBERT-Bio	0.780	0.914	0.842	0.357	0.416
DrBERT-7G-cased	0.811	0.892	0.850	0.225	0.466
FlauBERT	**0.956**	0.785	**0.862**	**0.857**	0.380
Bio-Clinical-BERT	0.910	0.800	0.851	0.592	0.571

TABLE 5: Performance metrics for BERT variants

When examining optimal alert threshold results, we prioritized precision to minimize false alarms, with each model's alert threshold optimized based on its respective critical threshold. Each optimal alert threshold was identified by searching within an interval exceeding 10% of its specific critical threshold. FlauBERT maintained superior precision (0.857) across operational scenarios, while CamemBERT-Large demonstrated highest recall (0.689), making it effective at capturing potential hallucinations in high-precision settings. BioClinical BERT showed balanced performance. Notably, FlauBERT a general-purpose language model outperformed domain-specific medical models, possibly due to its superior handling of conversational language in medical transcriptions, larger parameter count, and diverse pre-training corpus allowing better contextualization across both medical and general language contexts.

- ***Clinical Validation: threshold optimization with authentic CHU healthcare data***

Table 6 presents the results of the optimized hallucination detection thresholds using real clinical data from CHU. We observe significant variations in optimal critical thresholds across models, with CamemBERT-Bio demonstrating a substantially lower critical threshold (0.6328) compared to FlauBERT (0.906) and Bio-Clinical-BERT (0.939). This suggests that CamemBERT-Bio requires a lower confidence score to effectively identify hallucinations in real clinical reports. The optimal alert thresholds follow a similar pattern, with CamemBERT-Bio requiring a lower confidence threshold (0.758) than its counterparts FlauBERT (0.943) and Bio-Clinical-BERT (0.975). The optimization process conducted through Optuna required a considerable number of trials (16-17 for two models), indicating the complexity of finding optimal configurations when working with real-world clinical data. These findings highlight the importance of model-specific threshold calibration when deploying BERT-based architectures for hallucination detection in authentic clinical settings.

Models	*Optimal critical threshold*	*Optimal alert threshold*	*Best trials*
CamemBERT-Bio	0.6328	0.758	*16*
FlauBERT	0.906	0.943	*11*
Bio-Clinical-BERT	0.939	0.975	*17*

TABLE 6: Results of optimized hallucination detection thresholds using authentic data

Table 7 presents performance metrics for three BERT variants evaluated on real clinical data from CHU. We selected these models based on their performance: three top-performing model from synthetic data experiments . For critical threshold optimization, we used Optuna to maximize precision while minimizing false positives, with Bio-Clinical-BERT achieving the highest F1-Score (0.604) and precision (0.928). For alert threshold optimization, we created a composite score prioritizing recall (70%) and F1-score (30%), where FlauBERT demonstrated superior performance with the highest recall (0.931) and FP-F1 score (0.851). Overall, FlauBERT emerges as the most balanced model across both thresholds, making it the optimal choice for clinical applications requiring different sensitivity levels.

Models	*Performance metrics for optimal critical threshold*			*Performance metrics for optimal alert threshold*	
	Precision	*Recall*	*F1-Score*	*Recall*	*Score FP-F1*
CamemBERT-Bio	0.909	0.344	0.50	0.896	0.830
FlauBERT	0.866	0.488	0.59	**0.931**	**0.851**
Bio-Clinical-BERT	0.928	0.448	0.604	0.7964	0.7964

TABLE 7: Performance metrics for BERT variants using real clinical data

5 System limitations

The dual-threshold optimization system we have developed for hallucination detection demonstrated promising results, while some limitations remain to be addressed. The current pairwise comparison method between sentences in the generated report and those in the original transcription may not perfectly capture cases where a single sentence in the medical report represents a condensed version of multiple sentences from the transcription. This situation is common in medical practice where clinicians often synthesize information from various parts of a consultation into a single concise conclusion.

6 Conclusion and perspectives

This study introduces a novel dual-threshold approach for detecting hallucinations in medical documentation generated by LLM models. By optimizing BERT semantic layers and implementing critical and alert thresholds using Tree Parzen Estimator, our system shows good results in hallucination detection compared to traditional methods. In clinical settings, this technology can enhance patient safety by ensuring accurate medical summaries. Future work should explore applications across diverse medical specialties and examine effectiveness in multiple languages beyond French. Testing our system on authentic medical consultations and records will provide more robust validation of its practical utility in clinical environments. Additionally, integrating explainable AI techniques could further enhance trust and adoption by healthcare professionals. The development of such systems represents an important step toward responsible AI deployment in healthcare, where information accuracy directly impacts patient outcomes and treatment decisions.

Références

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna : A Next-generation Hyperparameter Optimization Framework* (arXiv:1907.10902). arXiv. https://doi.org/10.48550/arXiv.1907.10902

Antoun, W., Kulumba, F., Touchent, R., Clergerie, É. de la, Sagot, B., & Seddah, D. (2024). *CamemBERT 2.0 : A Smarter French Language Model Aged to Perfection* (arXiv:2411.08868). arXiv. https://doi.org/10.48550/arXiv.2411.08868

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877‑1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Cardon, R., Grabar, N., Grouin, C., & Hamon, T. (s. d.). *Présentation de la campagne d'évaluation DEFT 2020 : Similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques*.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., & Raffel, C. (2021). *Extracting Training Data from Large Language Models* (arXiv:2012.07805). arXiv.
https://doi.org/10.48550/arXiv.2012.07805

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Éds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)* (p. 4171‑4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Hiebel, N., Ferret, O., Fort, K., & Névéol, A. (s. d.). *CLISTER : un corpus pour la similarité sémantique textuelle dans des cas cliniques en français*.

Huang, Y., Song, J., Wang, Z., Zhao, S., Chen, H., Juefei-Xu, F., & Ma, L. (2025). Look Before You Leap : An Exploratory Study of Uncertainty Measurement for Large Language Models. *IEEE Transactions on Software Engineering*, *51*(2), 413‑429. https://doi.org/10.1109/TSE.2024.3519464

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE : Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Labrak, Y., Bazoge, A., Dufour, R., Rouvier, M., Morin, E., Daille, B., & Gourraud, P.-A. (2023). *DrBERT : A Robust Pre-trained Model in French for Biomedical and Clinical domains* (arXiv:2304.00958). arXiv. https://doi.org/10.48550/arXiv.2304.00958

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). *FlauBERT : Unsupervised Language Model Pre-training for French* (arXiv:1912.05372). arXiv. https://doi.org/10.48550/arXiv.1912.05372

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Éds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 7871‑7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Li, J., Dada, A., Kleesiek, J., & Egger, J. (2023). *ChatGPT in Healthcare : A Taxonomy and Systematic Review* (p. 2023.03.30.23287899). medRxiv. https://doi.org/10.1101/2023.03.30.23287899

Liashchynskyi, P., & Liashchynskyi, P. (2019). *Grid Search, Random Search, Genetic Algorithm : A Big Comparison for NAS* (arXiv:1912.06059). arXiv. https://doi.org/10.48550/arXiv.1912.06059

Liu, J., Wang, C., & Liu, S. (2023). Utility of ChatGPT in Clinical Practice. *Journal of Medical Internet Research*, *25*(1), e48568. https://doi.org/10.2196/48568

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). *SelfCheckGPT : Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models* (arXiv:2303.08896). arXiv. https://doi.org/10.48550/arXiv.2303.08896

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Éds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 1906‑1919). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2023). *Med-HALT : Medical Domain Hallucination Test for Large Language Models* (arXiv:2307.15343). arXiv. https://doi.org/10.48550/arXiv.2307.15343

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). DeClarE : Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Éds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (p. 22‑32). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1003

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (s. d.). *Language Models are Unsupervised Multitask Learners*.

Rebuffel, C., Roberti, M., Soulier, L., Scoutheeten, G., Cancelliere, R., & Gallinari, P. (2021). *Controlling Hallucinations at Word Level in Data-to-Text Generation* (arXiv:2102.02810). arXiv. https://doi.org/10.48550/arXiv.2102.02810

Touchent, R., & de la Clergerie, É. (2024). CamemBERT-bio : Leveraging Continual Pre-training for Cost-Effective Models on French Biomedical Data. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Éds.), *Proceedings of the 2024 Joint International*

Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (p. 2692‑2701). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.241/

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna : A Next-generation Hyperparameter Optimization Framework* (No. arXiv:1907.10902). arXiv. https://doi.org/10.48550/arXiv.1907.10902

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). *Publicly Available Clinical BERT Embeddings* (No. arXiv:1904.03323). arXiv. https://doi.org/10.48550/arXiv.1904.03323

Huang, K., Altosaar, J., & Ranganath, R. (2020). *ClinicalBERT : Modeling Clinical Notes and Predicting Hospital Readmission* (No. arXiv:1904.05342). arXiv. https://doi.org/10.48550/arXiv.1904.05342

Labrak, Y., Bazoge, A., Dufour, R., Rouvier, M., Morin, E., Daille, B., & Gourraud, P.-A. (2023). *DrBERT : A Robust Pre-trained Model in French for Biomedical and Clinical domains* (No. arXiv:2304.00958). arXiv. https://doi.org/10.48550/arXiv.2304.00958

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Éds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 7871‑7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore : Evaluating Text Generation with BERT* (No. arXiv:1904.09675). arXiv. https://doi.org/10.48550/arXiv.1904.09675

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA : Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore : Evaluating Text Generation with BERT* (arXiv:1904.09675). arXiv. https://doi.org/10.48550/arXiv.1904.09675