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'où 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 the similarity between sentences from generated reports and original transcriptions. Unlike traditional single-threshold approaches, our main contribution introduces a dual-threshold mechanism—critical and alert—optimized using Tree Parzen Estimator. Results demonstrate significant improvement in hallucination detection. The proposed system ensures the accuracy of medical information, fulfilling the 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 developing 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 critical nature of healthcare amplifies the risks associated with hallucinations. Inaccurate 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 the French language, which has fewer resources 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.*, 2019) 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 works

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 (Schneider *et al.*, 2021) GPT-3 (Zong & Krishnamachari, 2022) and BART (Lewis *et al.*, 2019) 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 many risks that extend beyond privacy concerns (Tang *et al.*, 2024). 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.*, 2024) 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.*, 2022) (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.*, 2019). 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.*, 2019) adaptations for French address challenges like gender agreement and terminology variations specific to the French healthcare system. CamemBERT-Bio (Touchent *et al.*, 2023) 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.*, 2019) 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.*, 2019). 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 the optimization of BERT layers has shown important findings.(Zhang *et al.*, 2019) 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. Identifying appropriate thresholds represents a crucial task that has received insufficient attention in the current literature on medical hallucination detection. Moreover, recent advances such as SelfCheckGPT (Manakul *et al.*, 2023) have introduced zero-resource, black-box hallucination detection methods that leverage the generative model's outputs without requiring external references or access to training data. These approaches demonstrate promising directions toward scalable, model-agnostic detection frameworks that can complement threshold-based optimization and improve reliability in clinical

NLP applications. These developments underscore the importance of combining robust hallucination detection techniques with careful calibration of detection thresholds to ensure reliability, especially in sensitive medical contexts. Building on this foundation, our work focuses on optimizing threshold settings for hallucination detection in French clinical NLP models, leveraging both reference-free detection methods and layer-wise analysis of BERT representations to enhance performance and trustworthiness.

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.

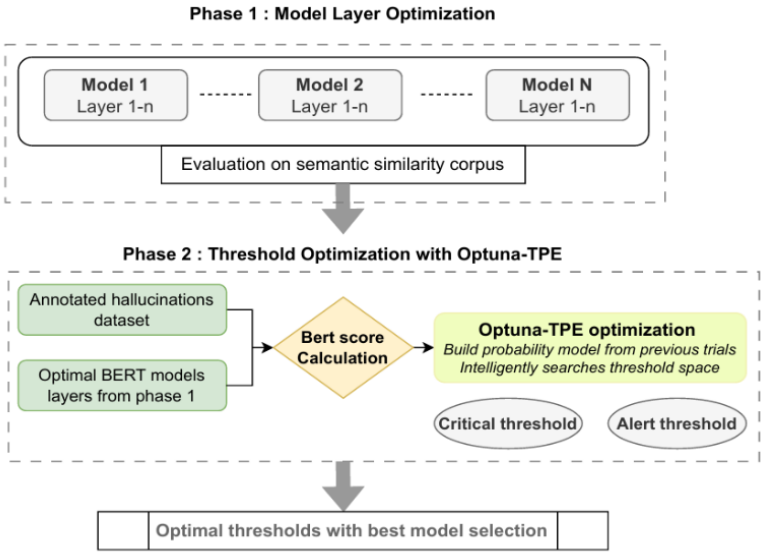

FIGURE 1 – Anti-hallucination system process

Phase 2 leverages these optimal model-layer combinations to calculate BERTScore between reference and generated text using 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, which involves data standardization to handle heterogeneous medical data formats, normalization of whitespace and special characters, systematic handling of null values, quality control filtering based on length ratios, and sentence-level segmentation based on punctuation patterns with attention

to medical abbreviations and terminology, 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 optimize 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

To evaluate 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 evaluated 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), and CamemBERT-Bio, which extends the base model with fine-tuning on French biomedical corpora (Touchent *et al.*, 2023). 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.*, 2019) 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 Algorithm 1. We conducted a detailed evaluation across every layer of each BERT model to determine the most semantically meaningful representation space. For each layer, we extracted embeddings and computed semantic similarity scores against a reference corpus. We then compared these scores with ground truth annotations to measure correlation. This process aimed to identify the optimal layer that delivers the strongest semantic encoding by selecting the layer with the highest correlation score for each model.

Algorithm 1 Model Layer Optimization

INPUT : BERT model variants $\{M_1, M_2, \ldots, M_n\}$, CLISTER-DEFT corpus with ground truth annotations
OUTPUT : Optimal layer for each BERT model
for each BERT model M_i **do**
  $best\_correlation \leftarrow 0$ {Initialize best correlation score}
  $best\_layer \leftarrow 0$ {Initialize best layer index}
  for each layer L_j in M_i **do**
    $embeddings \leftarrow extract\_embeddings(M_i, L_j)$
    $bert\_scores \leftarrow compute\_bert\_score(embeddings, corpus)$
    $correlation\_score \leftarrow calculate\_correlation(bert\_scores, ground\_truth\_labels)$
    if $correlation\_score > best\_correlation$ **then**
      $best\_correlation \leftarrow correlation\_score$
      $best\_layer \leftarrow L_j$
    end if
  end for
  $optimal\_models[M_i] \leftarrow best\_layer$
end for
RETURN $optimal\_models$

3.3 Threshold optimization for hallucination detection using Optuna

Threshold optimization is an essential step towards the design of an accurate hallucination detection system. Rather than relying on empirically determined threshold values, we implemented an optimization approach using Optuna (Akiba et al., 2019), which is a Hyperparameter Optimization (HPO) framework specifically designed for Machine Learning (ML) applications (Khessiba *et al.*, 2022) (Khessiba *et al.*, 2024). Our process used the TPE algorithm, which offers significant advantages over traditional Grid Search or Random Search methods (Liashchynskyi & Liashchynskyi, 2019). The TPE algorithm works 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 with customized objective functions for each threshold (Table 2).

For the Critical Threshold (CT), we developed a prioritized scoring system that heavily rewards configurations with zero false positives and high precision (≥ 0.999), assigning maximum scores

$(10.0 \times TP/N)$ to ideal cases with perfect precision and at least $N/3$ true positives, where N represents the total number of hallucinations in our ground truth dataset. Configurations with a single false positive were considered acceptable only if they maintained high precision (≥ 0.8) and sufficient true positives, while any setup with more than $N/6$ false positives was automatically rejected.

We've defined a multi-level scoring system for the alert threshold (AT), which balances the F1 score, precision, and true positive detection rates. The scoring system assigned premium values to near-perfect configurations (F1 ≥ 0.9, precision ≥ 0.9, TP $\geq 0.8N$) while incorporating progressive penalties for increasing false positive rates. This differentiated approach reflects the distinct requirements of each clinical context : the CT demands near-perfect precision to avoid unnecessary critical interventions, while the AT allows for more balanced metric optimization to enhance sensitivity while maintaining reasonable specificity.

Parameter	**Critical Threshold (CT)**	**Alert Threshold (AT)**
Range	{0.2, 0.98}	{0.2, 0.98}
Model Layer	Fixed from Phase 1	
Optimization Priority	Precision > Sensitivity	Balanced F1-Score
Optimization order	First	Second
Sample	TPE	
N Trials	30	
Max False Positives	N/6	Progressive Penalty
Min True Positives	N/3	0.8N

TABLE 2 – Threshold optimization settings for the hallucination detection system

During each trial,the optimization process calculated BERTScore between reference documents and potentially hallucinated ones 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. Among Bayesian optimization methods, TPE was selected for its computational efficiency with mixed parameter spaces and robust performance with limited evaluation budgets, as compared to GP-based methods, which can struggle with discrete parameters. 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 accordingly, creating a dual-threshold system capable of distinguishing between different severity levels of hallucinations.

4 Experimentation

4.1 Hallucination Dataset

Our study employed two datasets to evaluate hallucination detection models : a structured synthetic corpus and a real-world clinical dataset. These two datasets enable a complete evaluation of the models' performance within a set of systematic experimental conditions and authentic clinical scenarios (Table 3).

– Synthetic Dataset

We constructed a synthetic corpus of 54 medical reports with controlled hallucination instances and corresponding ground truth transcriptions. This dataset enables systematic evaluation and threshold optimization of detection models in the medical domain (described in 3).

– Authentic Clinical Dataset

To validate our findings in authentic clinical settings, we incorporated a smaller but highly valuable real-world dataset composed of 8 anonymized clinical reports obtained from university hospital centers (CHUs). This proprietary clinical dataset remains confidential following strict privacy requirements and ethical guidelines.

Dataset Type	Data Component	Files	Total Sentences	Avg. Sentences/File	Total Tokens	Avg. Tokens/File
Synthetic	Medical Reports	54	865	16.0 ± 5.6	19879	368.1 ± 110.1
Synthetic	Transcriptions	54	2646	49.0 ± 186.0	38834	719.1 ± 2020.9
Authentic	Medical Reports	8	133	16.6 ± 3.5	1804	225.5 ± 64.2
Authentic	Transcriptions	8	814	101.8 ± 45.5	12702	1587.8 ± 744.0

TABLE 3 – Medical Datasets Overview and Statistical Features

5 Results and discussion

– Performance Evaluation of Optimized BERT Variants

Analysis of optimal layer selection revealed that different BERT models exhibit distinct patterns and unique preferences for semantic representation (Table 4). 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).

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 4 – Results of optimized best layer selection for BERT variants

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.

– Hallucination detection threshold optimization results using synthetic data

Table 5 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-TPE 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 HPO. 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	**CT**	**AT**	**Best trials**	
			CT	**AT**
CamemBERT-Large	0.386	0.553	28	22
CamemBERT-Bio	0.704	0.479	11	8
DrBERT-7G-cased	0.317	0.390	14	25
FlauBERT	0.365	0.521	27	16
Bio-Clinical-BERT	0.440	0.491	23	27

TABLE 5 – Results of optimized hallucination detection thresholds using the synthetic dataset

Table 6 presents the performance metrics using synthetic data. The synthetic environment provides controlled evaluation conditions that enable comprehensive model assessment. Bio-Clinical-BERT demonstrates superior performance across both threshold configurations, achieving the highest F1-scores and maintaining excellent precision-recall balance. DrBERT-7G-cased shows competitive results with strong recall capabilities, while FlauBERT exhibits perfect precision in the critical threshold setting but with limited recall performance. The remaining models display moderate performance levels with varying precision-recall trade-offs. Overall, the synthetic data evaluation reveals that Bio-Clinical-BERT consistently outperforms other architectures, validating its effectiveness for hallucination detection tasks and confirming that the optimized layer selection contributes significantly to enhanced model performance.

Models	**Precision**	**Recall**	**F1-Score**
Critical Threshold Performance			
CamemBERT-Large	0.914	0.721	0.806
CamemBERT-Bio	0.251	0.990	0.406
DrBERT-7G-cased	0.894	0.639	0.745
FlauBERT	1.00	0.240	0.387
Bio-Clinical-BERT	0.957	0.684	0.798
Alert Threshold Performance			
CamemBERT-Large	0.518	0.939	0.668
CamemBERT-Bio	0.693	0.917	0.789
DrBERT-7G-cased	0.779	0.932	0.849
FlauBERT	0.600	0.924	0.727
Bio-Clinical-BERT	0.842	0.909	0.876

TABLE 6 – Results of optimized hallucination detection thresholds using synthetic data

– Clinical Validation : threshold optimization with authentic CHU data

Table 7 presents the results of optimized hallucination detection thresholds using authentic clinical data from CHU, revealing substantial differences in model behavior and calibration requirements. The optimal critical thresholds demonstrate significant model-specific variations, ranging from DrBERT-7G-cased's notably low threshold of 0.273 to Bio-Clinical-BERT's higher threshold of

0.470, indicating that each model requires different confidence levels to effectively identify hallucinations. Similarly, the alert thresholds vary considerably across models, from 0.411 for DrBERT to 0.491 for FlauBERT, suggesting distinct sensitivity patterns for each model. The optimization process itself required markedly different computational efforts, with trials ranging from just 14 for DrBERT-7G-cased to 28 for both CamemBERT-Large and Bio-Clinical-BERT, reflecting the varying complexity of finding optimal configurations for different architectures. CamemBERT-Bio demonstrated intermediate behavior with moderate thresholds and trial requirements.

Models	**CT**	**AT**	**Best trials**	
			CT	**AT**
CamemBERT-Large	0.339	0.463	28	9
CamemBERT-Bio	0.426	0.451	1	11
DrBERT-7G-cased	0.273	0.411	14	10
FlauBERT	0.334	0.491	27	15
Bio-Clinical-BERT	0.470	0.481	23	25

TABLE 7 – Results of optimized hallucination detection thresholds using authentic data

Table 8 presents the performance metrics for BERT variants using authentic clinical data, demonstrating the effectiveness of the optimized thresholds from Table 7. The results reveal significant performance differences across models and threshold configurations based on their specific objectives. For the optimal CT, the goal was to maximize true positives while minimizing false positives, which explains why all models achieve perfect precision (1.0) with no false positive hallucination detections. However, recall values vary substantially, with Bio-Clinical-BERT showing the highest recall (0.46), indicating superior ability to detect actual hallucinations, while other models demonstrate considerably lower recall rates. For the optimal alert threshold, the aim was to achieve a balanced performance between precision and recall. Bio-Clinical-BERT emerges as the superior model with the most balanced metrics. DrBERT shows competitive performance with strong recall but lower precision, while other models exhibit more moderate, balanced performance. Bio-Clinical-BERT outperforms all others, with the best layer selection step providing significant performance gains, which are essential for effective hallucination detection in clinical settings.

Models	**Critical threshold metrics**			**Alert threshold metrics**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
CamemBERT-Large	1.0	0.03	0.06	0.36	0.76	0.64
CamemBERT-Bio	1.0	0.26	0.42	0.75	0.50	0.60
DrBERT-7G-cased	1.0	0.06	0.12	0.6	0.80	0.68
FlauBERT	1.0	0.33	0.50	0.47	0.76	0.58
Bio-Clinical-BERT	1.0	0.46	0.63	0.68	0.66	0.75

TABLE 8 – Performance metrics for BERT variants using authentic data

Our evaluation on both synthetic and authentic clinical data confirms Bio-Clinical-BERT's superior performance with optimized layer selection across both contexts. However, our findings reveal that optimal detection thresholds vary significantly between synthetic and real-world clinical data, requiring dataset-specific calibration. Despite the presence of transcription errors and linguistic variations in authentic CHU data, Bio-Clinical-BERT demonstrates robust adaptability through appropriate threshold adjustments

6 Comparative Analysis with State-of-the-Art Approaches

We benchmarked our optimized BioClinical-BERT method against established baselines on a comprehensive dataset of authentic clinical consultations and reports. The evaluation dataset consists of real transcription-report pairs from medical practice, with all techniques assessed using identical ground truth annotations to ensure a fair comparison. In particular, we tested two variants of SelfCheckGPT : SelfCheckGPT-RoBERTa and SelfCheckGPT-GPT2, which represent recent state-of-the-art approaches for zero-resource, black-box hallucination detection. Table 9 summarizes the comparative performance across key metrics, illustrating the superiority of our approach in precision, processing speed, and overall detection quality. Notably, our method achieves a substantially lower average detection rate and processing time while maintaining higher precision and a competitive F1-score compared to these SelfCheckGPT variants.

Models	**Avg detection rate (%)**	**Avg Time (s)**	**Avg Precision**	**Avg F1-Score**
SelfCheckGPT-RoBERTa	18	69.08	0.24	0.32
SelfCheckGPT-GPT2	62.7	50.69	0.30	0.54
Our method (Optimized Bio-Clinical-BERT)	4.5	4.2	0.81	0.60

TABLE 9 – Comparative results of state-of-the-art hallucination detection techniques

7 System limitations

While the dual-threshold optimization system we developed for hallucination detection shows promising results, several limitations warrant further investigation. Our current approach relies on pairwise sentence-level comparisons between the generated medical report and the original transcription. However, this method may struggle to accurately handle cases where a single sentence in the report effectively condenses or synthesizes information from multiple sentences in the transcription. Such summarization is common in clinical practice, as healthcare professionals frequently integrate diverse pieces of information into concise, coherent conclusions. This inherent complexity poses challenges for the alignment and evaluation process, potentially leading to under-detection or misclassification of hallucinations in synthesized content.

8 Conclusion and perspectives

This study presents a novel dual-threshold approach for detecting hallucinations in medical documentation generated by large language models. By optimizing the semantic layers of BERT and implementing critical and alert thresholds through the Tree Parzen Estimator, our system achieves superior performance in hallucination detection compared to traditional methods and state-of-the-art approaches such as SelfCheckGPT variants. 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. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, p. 2623–2631.

ALSENTZER E., MURPHY J. R., BOAG W., WENG W.-H., JIN D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv :1904.03323*.

ANTOUN W., KULUMBA F., TOUCHENT R., DE LA CLERGERIE É., SAGOT B. & SEDDAH D. (2024). Camembert 2.0 : A smarter french language model aged to perfection. *arXiv preprint arXiv :2411.08868*.

CARLINI N., TRAMER F., WALLACE E., JAGIELSKI M., HERBERT-VOSS A., LEE K., ROBERTS A., BROWN T., SONG D., ERLINGSSON U. *et al.* (2021). Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, p. 2633–2650.

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, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

HUANG K., ALTOSAAR J. & RANGANATH R. (2019). Clinicalbert : Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv :1904.05342*.

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 analysis for large language models. *IEEE Transactions on Software Engineering*.

KHESSIBA S., BLAEICH A. G., ABDALLAH A. B., MANZANERA A., KHALIFA K. B. & BEDOUI M. H. (2024). A novel hybrid grid search and tree parzen estimator for deep learning hyperparameters optimization. In *2024 IEEE/ACS 21st International Conference on Computer Systems and Applications (AICCSA)*, p. 1–8. DOI : 10.1109/AICCSA63423.2024.10912622.

KHESSIBA S., BLAIECH A. G., MANZANERA A., BEN KHALIFA K., BEN ABDALLAH A. & BEDOUI M. H. (2022). Hyperparameter optimization of deep learning models for eeg-based vigilance detection. In *International Conference on Computational Collective Intelligence*, p. 200–210 : Springer.

KUNG T. H., CHEATHAM M., MEDENILLA A., SILLOS C., DE LEON L., ELEPAÑO C., MADRIAGA M., AGGABAO R., DIAZ-CANDIDO G., MANINGO J. *et al.* (2023). Performance of chatgpt on usmle : potential for ai-assisted medical education using large language models. *PLoS digital health*, **2**(2), e0000198.

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 preprint arXiv :2304.00958*.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. *arXiv preprint arXiv :1912.05372*.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2019). Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv :1910.13461*.

LI J., DADA A., PULADI B., KLEESIEK J. & EGGER J. (2024). Chatgpt in healthcare : a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, **245**, 108013.

LIASHCHYNSKYI P. & LIASHCHYNSKYI P. (2019). Grid search, random search, genetic algorithm : a big comparison for nas. *arXiv preprint arXiv :1912.06059*.

LIU J., WANG C. & LIU S. (2023). Utility of chatgpt in clinical practice. *Journal of medical Internet research*, **25**, e48568.

MANAKUL P., LIUSIE A. & GALES M. J. (2023). Selfcheckgpt : Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint 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, Online : Association for Computational Linguistics. DOI : 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 preprint arXiv :2307.15343*.

POPAT K., MUKHERJEE S., YATES A. & WEIKUM G. (2018). Declare : Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv :1809.06416*.

REBUFFEL C., ROBERTI M., SOULIER L., SCOUTHEETEN G., CANCELLIERE R. & GALLINARI P. (2022). Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, p. 1–37.

SCHNEIDER E. T. R., DE SOUZA J. V. A., GUMIEL Y. B., MORO C. & PARAISO E. C. (2021). A gpt-2 language model for biomedical texts in portuguese. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, p. 474–479. DOI : 10.1109/CBMS52027.2021.00056.

TANG L., SHALYMINOV I., WONG A. W.-M., BURNSKY J., VINCENT J. W., YANG Y., SINGH S., FENG S., SONG H., SU H. *et al.* (2024). Tofueval : Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv :2402.13249*.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. (2023). Camembert-bio : Leveraging continual pre-training for cost-effective models on french biomedical data. *arXiv preprint arXiv :2306.15550*.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F. *et al.* (2023). Llama : Open and efficient foundation language models. *arXiv preprint arXiv :2302.13971*.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*.

ZONG M. & KRISHNAMACHARI B. (2022). A survey on gpt-3. *arXiv preprint arXiv :2212.00857*.