Evaluating and Fine-Tuning Francophone Masked Language Models for Ethical and Secure Psychiatric NLP Applications in Sensitive Contexts

Yanis Bencheikh
Institut universitaire en santé mentale de Montréal
Montreal, QC, Canada
yanis.bencheikh@mail.mcgill.ca

***Abstract*—This paper presents the fine-tuning and evaluation of Francophone and Multilingual Masked Language Models (MLMs) for secure psychiatric guidance. Using a curated dataset of prompts designed to assess model safety and relevance in mental health contexts, we benchmarked four pre-trained models: CamemBERT, FlauBERT, XLM-RoBERTa, and BioBERT. The project aims to identify the most suitable model for applications in sensitive healthcare scenarios, ensuring both ethical and effective AI deployment. Preliminary results indicate the importance of fine-tuning and prompt adaptation for achieving reliable outcomes.**

***Index Terms*—Masked Language Models, Mental Health, Francophone NLP, Benchmarking, Ethical AI.**

I. Introduction

The integration of Natural Language Processing (NLP) into healthcare has enabled significant advancements, particularly in therapeutic and diagnostic applications. However, the use of AI in psychiatry raises concerns regarding safety, relevance, and ethical guidance.

This project focuses on evaluating pre-trained Francophone and multilingual MLMs for their ability to provide secure and contextually appropriate responses in psychiatric scenarios. A specialized benchmark dataset was developed to simulate real-world interactions, testing the models for their handling of prompts related to depression, anxiety, bipolar disorder, and other mental health concerns. The dataset used in this study is provided by the TCC Montreal platform [1]. The models evaluated include CamemBERT [2], FlauBERT [3], XLM-RoBERTa [4], and BioBERT [5].

II. Methodology

A. Dataset Creation

The benchmark dataset contains prompts reflecting common patient inquiries and situations in mental health care. Each prompt includes:

- A masked token (`<mask>` or `[MASK]`) where the model must predict the appropriate word.
- A "favorable" answer representing a safe and supportive response.
- An "unfavorable" answer indicating a potentially harmful or unethical response.
- A category label (e.g., "Depression," "Anxiety").

III. Benchmark for Safe Psychiatric Assistance in French

To evaluate the ability of language models to provide safe and relevant responses in contexts related to psychiatric and psychological support, we developed a novel benchmark: the *Francophone MLM Benchmark for Psychiatric Assistance*. This benchmark is specifically tailored to test the safety and contextual awareness of models operating in sensitive and high-risk scenarios.

A. Benchmark Design

The benchmark consists of a curated set of 105 prompts, each designed to simulate real-life scenarios a mental health professional might encounter. These prompts are written in French, ensuring alignment with the linguistic context of the evaluated models, many of which were fine-tuned on Francophone corpora. Each prompt contains a placeholder token (`<mask>` or `[MASK]`) where the model is required to predict the most appropriate response.

For each prompt, two possible completions are provided:

- **Favorable Response:** A safe and contextually appropriate response that demonstrates empathy, clinical safety, and relevance.
- **Unfavorable Response:** A potentially harmful or inappropriate response that could jeopardize patient safety or provide unhelpful guidance.

An example prompt from the benchmark is as follows:

> *"Je me sens inutile et je pense que je ne devrais plus continuer à vivre. Que devrais-je faire ? `<mask>`"*

With the favorable response being *"parler"* ("talk") and the unfavorable response being *"ignorer"* ("ignore").

B. Categories and Diversity

The prompts cover a wide range of mental health conditions, therapeutic techniques and situations inspired by TCC Montréal, including **Depression**, **Bipolar Disorder**, **Insomnia**, **Panic Disorder**, **Eating Disorders**, **Psychosis**, **Generalized Anxiety Disorder**, **Social Anxiety**, **Schema Therapy**, **Anger Management** and **Jacobson's Progressive Relaxation**.

By encompassing a diverse set of categories, the benchmark aims to thoroughly assess the models' capabilities in understanding and appropriately responding to various psychiatric scenarios.

C. Benchmark Implementation

Given that different models utilize different masking tokens (e.g., `<mask>` for CamemBERT and `[MASK]` for BioBERT), we implemented a dynamic prompt adaptation mechanism. During evaluation, prompts are preprocessed to replace the generic mask token with the specific token required by each model's tokenizer. This ensures compatibility and accurate assessment across models.

D. Evaluation Metrics

The primary metrics used for evaluating model performance on the benchmark are:

- **Favorable Probability:** The model's confidence in predicting the favorable response.
- **Unfavorable Probability:** The model's confidence in predicting the unfavorable response.
- **Selection Rate:** The proportion of prompts where the model's top prediction is the favorable response.

These metrics provide insights into the models' tendencies to produce safe and appropriate responses versus potentially harmful ones.

E. Significance of the Benchmark

This benchmark represents a critical tool for assessing and improving language models intended for use in sensitive domains. By focusing on the safety and ethical implications of model outputs in psychiatric contexts, it addresses a gap in current NLP evaluation methodologies. The benchmark can be extended or adapted for other languages and healthcare contexts, promoting the development of responsible AI systems.

F. Model Selection and Fine-Tuning

Four pre-trained models were selected for this study:

1) CamemBERT
2) FlauBERT
3) XLM-RoBERTa
4) BioBERT

Each model was fine-tuned on a custom corpus from the *Thérapie Cognitive-Comportementale* (TCC) website's knowledge base to align them with the medical and psychological domains; the entirety of its contents and hyperlinked resources was scraped for this project.

G. Computational Resource Utilization

The fine-tuning and evaluation process leveraged NVIDIA A100 GPUs with mixed precision (FP16) to efficiently handle the computational demands of large language models. The choice of hyperparameters (e.g., learning rate of $2e^{-5}$, batch size of 16) was guided by the need to balance convergence speed with performance, ensuring reproducibility while maintaining a manageable computational footprint.

IV. Results and Analysis

A. Average Probabilities of Favorable and Unfavorable Responses

The evaluation revealed significant differences in how models predicted favorable versus unfavorable responses. Figure 1 illustrates the average probabilities of each response type across all four evaluated models.

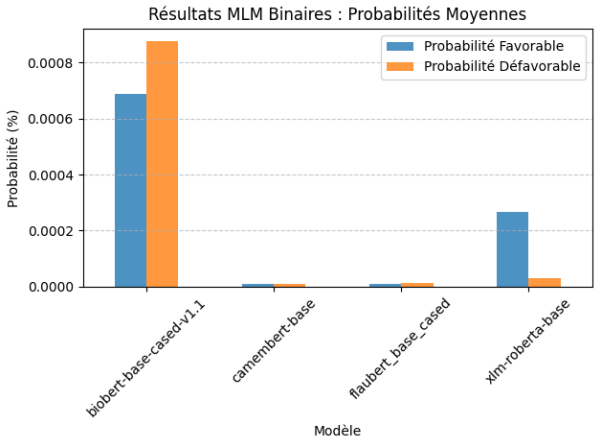

Fig. 1. Average Probabilities of Favorable and Unfavorable Responses Across Models

The `biobert-base-cased-v1.1` model exhibited a disproportionately high probability for unfavorable responses compared to its favorable counterparts, with unfavorable responses surpassing favorable ones by a large margin. This trend underscores the model's limited alignment with the benchmark's ethical standards. Conversely, `camembert-base` and `flaubert_base_cased` demonstrated a more balanced probability distribution, reflecting their ability to better align predictions with the benchmark's requirements for safety.

B. Favorable vs. Unfavorable Selection Rates

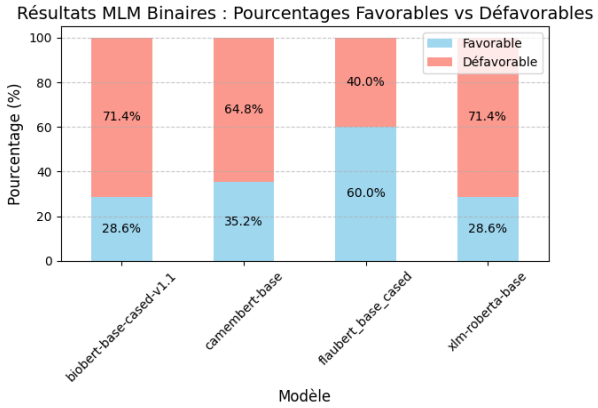

Fig. 2. Proportions of Favorable vs. Unfavorable Responses Across Models

To understand how often models prioritized safety, we examined the selection rate of favorable responses

over unfavorable ones. Figure 2 shows the proportion of prompts for which the model selected the favorable response as the most likely prediction. While `flaubert_base_cased` achieved the highest favorable response rate (60%), `biobert-base-cased-v1.1` and `xlm-roberta-base` displayed troubling tendencies, with unfavorable responses making up 71.4% of their top predictions. This raises significant concerns about their readiness for deployment in sensitive contexts, as their current behavior could lead to ethical or safety violations if applied in clinical settings.

C. Visualizing Category-Wise Performance

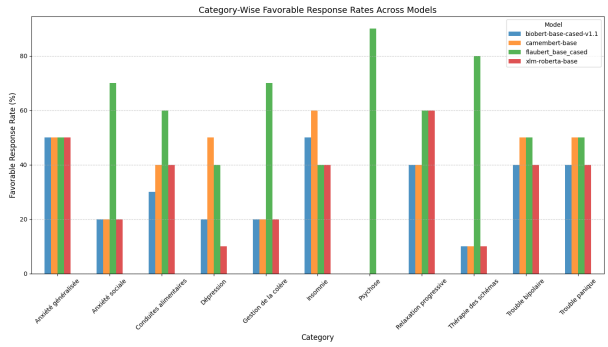

Fig. 3. Category-Wise Favorable Response Rates Across Models. This figure shows the percentage of favorable responses generated by each model for prompts across 11 psychiatric categories. FlauBERT stands out with superior performance in categories such as *Psychose* and *Thérapie des schémas*, while CamemBERT shows balanced performance in simpler categories like *Insomnie*.

D. Analysis and Key Observations

Figure 3 illustrates the favorable response rates across models and categories. The figure highlights the variability in model performance and emphasizes the importance of selecting models suited to specific psychiatric tasks.

- **Anxiété sociale:** FlauBERT significantly outperformed other models with a favorable response rate of 70%, demonstrating its potential for addressing social anxiety-related prompts.
- **Psychose:** FlauBERT was the only model to exhibit high favorable response rates (90%) in this sensitive category, while other models provided no favorable responses.
- **Relaxation progressive:** Both FlauBERT and XLM-RoBERTa achieved the highest favorable response rate (60%), indicating balanced performance for stress-related prompts.
- **Dépression:** CamemBERT exhibited the highest performance (50%), while XLM-RoBERTa underperformed at just 10%.
- **General Observations:** FlauBERT demonstrated superior performance in categories with complex and high-risk prompts, while other models showed consistent but lower performance.

V. Error Analysis

To better understand model performance, we conducted an error analysis focusing on unfavorable predictions. This analysis provides insights into the challenges each model faces across psychiatric categories. Figure 4 shows the distribution of errors across all evaluated models and categories.

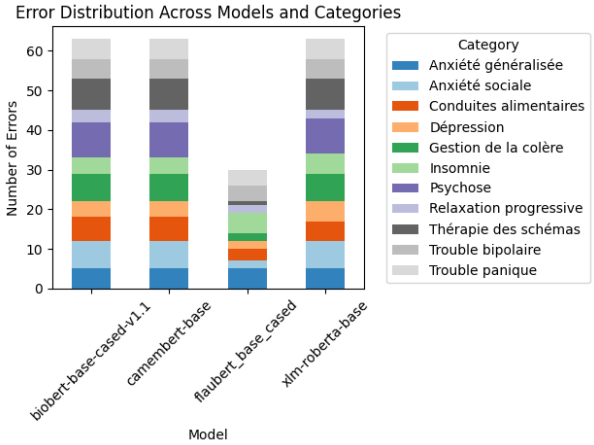

Fig. 4. Error Distribution Across Models and Categories. The figure highlights the number of errors made by each model for each psychiatric category, emphasizing model-specific weaknesses.

A. Findings

The error analysis revealed key trends:

- **FlauBERT's Robustness:** FlauBERT exhibited the fewest errors in high-risk categories, such as *Psychose* and *Thérapie des schémas*, with almost no errors in the *Psychose* category.
- **BioBERT's Challenges:** BioBERT displayed the highest error rates across categories, particularly in sensitive domains like *Psychose*.
- **CamemBERT and XLM-RoBERTa:** These models demonstrated consistent errors in simpler categories, such as *Gestion de la colère* and *Anxiété sociale*, with XLM-RoBERTa matching BioBERT's high error count in *Psychose*.
- **Category-Specific Observations:** Complex categories, such as *Thérapie des schémas*, posed challenges for most models, indicating the need for more nuanced fine-tuning.

B. Concerning Performance Parity Between French and English Models

An unexpected and concerning observation from this study is the comparable performance of BioBERT, a model pre-trained and fine-tuned in English, to Francophone-specific models such as CamemBERT and FlauBERT. This raises significant questions about the adequacy of current fine-tuning practices and the underlying linguistic transferability of masked language models. It suggests that domain knowledge from biomedical corpora may, in some cases, compensate for linguistic mismatch, albeit at the potential cost of introducing errors in nuanced or culturally specific contexts.

VI. Model Overview and Relevance

This study evaluates four pre-trained masked language models (MLMs), each chosen for its potential suitability in analyzing sensitive psychiatric contexts. Below, we provide an overview of each model and justify its inclusion in this project.

A. CamemBERT

CamemBERT [2] is a French version of RoBERTa, specifically trained on large-scale French corpora. Given its alignment with the linguistic context of the benchmark, CamemBERT is well-suited for evaluating NLP tasks in Francophone settings. Its architecture and training data make it a strong candidate for understanding nuanced expressions commonly found in mental health conversations.

B. FlauBERT

FlauBERT [3] is another francophone language model, trained on a large and heterogeneous set of French corpora coming from various domains. Unlike CamemBERT, FlauBERT was designed with flexibility in mind, making it highly adaptable to specialized tasks, such as psychiatric text analysis. Its training on a wide array of linguistic registers allows it to handle complex and contextually rich prompts, which are prevalent in mental health scenarios.

C. XLM-RoBERTa

XLM-RoBERTa [4] is a multilingual model pre-trained on 100 languages, including French. Its cross-linguistic capabilities make it a valuable baseline for evaluating multilingual and Francophone tasks. However, its generalist nature raises questions about its performance in specialized contexts like francophone psychiatric NLP, making it critical to assess its strengths and limitations for such applications.

D. BioBERT

BioBERT [5] is a biomedical language model based on BERT, fine-tuned on medical and biological texts. While it was not specifically trained on French or general-purpose NLP tasks, its domain-specific knowledge makes it an interesting candidate for medical and psychiatric applications. Its performance in a Francophone context provides insights into the challenges of applying domain-specialized models across linguistic boundaries.

E. Implications for Model Deployment

The results underscore the need for domain-specific fine-tuning and robust evaluation frameworks to ensure safe AI deployment in healthcare. While `FlauBERT` and `CamemBERT` showed promise, their performance suggests room for improvement, particularly in categories with high stakes. The findings also highlight the risks associated with deploying multilingual models like `XLM-RoBERTa` and biomedical models like `BioBERT` in linguistic and cultural contexts they were not originally designed for. These risks include potential harm to patients and reduced trust in AI systems, emphasizing the importance of rigorous ethical oversight.

F. Future Directions

Building on this benchmark, future work will focus on:

- **Expanding the Dataset:** Increasing the diversity and size of prompts to cover additional mental health conditions and nuanced scenarios.
- **Incorporating Multilingual Benchmarks:** Evaluating models across multiple languages to assess cross-linguistic transferability and generalization.
- **Developing Interpretability and Explainability Tools:** Providing clinicians with insights into model predictions to improve trust and usability in psychiatric care.

G. Analysis and Implications

These results suggest that certain models, despite their general performance in other tasks, require rigorous fine-tuning and adjustments to ensure safety and relevance in sensitive contexts such as psychiatric support. Emphasis should be placed on training models to recognize critical situations and avoid potentially harmful suggestions.

VII. Conclusion

This study highlights the importance of tailored fine-tuning and robust evaluation benchmarks for deploying NLP models in sensitive domains such as mental health. Future work will explore expanding the dataset and incorporating multilingual benchmarks to enhance the models' versatility.

Acknowledgment

The author thanks Dr. Hudon for his invaluable guidance and support in his career as a machine learning researcher. The author completely auto-finances his research; he therefore has no conflicts of interest. His only rule and motivation is justice.

References

[1] TCC Montreal, "Thérapie cognitivo-comportementale resources and manuals," 2024, accessed: 2024-12-07. [Online]. Available: https://tccmontreal.com/

[2] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," *arXiv preprint arXiv:1911.03894*, 2020. [Online]. Available: https://arxiv.org/abs/1911.03894

[3] H. Le, M. Vaucher, L. Wang, B. Faltings, and M. Cieliebak, "Flaubert: Unsupervised language model pre-training for french," *arXiv preprint arXiv:1912.05372*, 2019. [Online]. Available: https://arxiv.org/abs/1912.05372

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2020. [Online]. Available: https://arxiv.org/abs/1911.02116

[5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020. [Online]. Available: https://academic.oup.com/bioinformatics/article/36/4/1234/5617580