Evaluating CamemBERT for QCMs

Firas ABIDLI

Abstract
This report presents the evaluation of CamemBERT on a French multi-label QCM dataset in the pharmacy domain. Using exact match and macro F1-score as evaluation metrics, we compare zero-shot and fine-tuned performance and propose two strategies to further improve results: data augmentation and retrieval-based enhancement.

1. Dataset Analysis

The dataset includes 3100+ French-language QCMs focused on pharmacy:

- **Train:** 2171 questions
- **Dev:** 312 questions
- **Test:** 622 questions

Each question offers 5 choices (A–E) with one or more correct answers, labeled as `simple` or `multiple`.

2. Metrics

Two metrics were selected to evaluate performance:

- **Exact Match Accuracy:** A prediction is considered correct only if all selected answers exactly match the true labels.
- **Macro F1-score:** F1 is computed independently for each label and averaged, which treats all classes equally and is well-suited for multi-label classification with class imbalance [1].

This combination provides a strict measure (exact match) and a flexible one (F1) for a complete view of model performance.

3. Model Choice

Two French pretrained language models were considered:

- **CamemBERT** [2]
- **Flaubert-base-cased** [3]

We selected CamemBERT due to its wide adoption, strong support within the HuggingFace ecosystem, and training on the OSCAR corpus tailored for French NLP.

4. Zero-shot Evaluation

Without fine-tuning, CamemBERT yielded:

- **Exact Match:** 1.30%
- **Macro F1:** 39%

This shows its general understanding but highlights the need for domain-specific adaptation.

5. Fine-tuning Results

Phase 1: Supervised fine-tuning without including the `type` label. → *Macro F1:* 63.34%, *Exact Match:* 0.64%. Significant improvement in partial correctness.

Phase 2: Fine-tuning with `type` ("simple"/"multiple") added to input. → *Macro F1:* 63.72%, *Exact Match:* 1.30%. Slight gain, but type info alone did not drastically boost results.

6. Perspectives

To further enhance model performance, two strategies are worth exploring:

Retrieval-Augmented Generation (RAG) [4]: By retrieving external knowledge during inference, RAG can help the model base its predictions on relevant pharmacy-related content. This may improve *exact match accuracy* by providing more precise and factual context.

Data Augmentation [5]: Generating paraphrased or synthetic QCMs expands the training set and introduces more linguistic variety. This can help the model generalize better to unseen patterns and improve the *macro F1-score*, especially in cases with class imbalance or rare formulations.

Conclusion

Fine-tuning CamemBERT led to large gains in macro F1 over the base model. Adding `type` had limited impact, suggesting that future work should include external knowledge or advanced prompting. The final model is deployed in a web demo: `https://github.com/abidlifiras/llm-qcm-demo`

References

[1] M. Sokolova and G. Lapalme, *A Systematic Analysis of Performance Measures for Classification Tasks*, IPM, 2009.

[2] L. Martin et al., *CamemBERT: a Tasty French Language Model*, ACL 2020.

[3] A. Le et al., *FlauBERT: Unsupervised Language Model Pre-training for French*, ACL 2020.

[4] P. Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, NeurIPS 2020.

[5] S. Feng et al., *A Survey of Data Augmentation Approaches for NLP*, ACL 2021.