Une famille de modèles cliniques multilingues de type ColBERT diagnosticables grâce à un espace latent sémantique

François Remy[1]
(1) Parallia AI, 9000 Gand, Belgique
francois.remy@parallia.eu

RÉSUMÉ

Dans cet article, nous introduisons le concept de modèle ColBERT diagnosticable, un modèle de langage dont la compréhension est analysable directement à l'échelle du token. En effet, bien que les modèles ColBERT classiques permettent l'inspection de scores d'interaction token–token, cela ne suffit pas pour diagnostiquer les erreurs en contexte clinique : pour cela, il faut aussi un espace latent de référence, sémantiquement structuré, afin de rendre visibles les confusions de concept, de composition locale et de qualification contextuelle. Armé de cette capacité d'introspection dans la compréhension cognitive d'un modèle, il est possible d'entraîner des modèles plus solides en itérant sur les données de manière réactive. Cette perspective est appuyée par deux résultats récents : *ClinicalEncoder26AM* a obtenu le meilleur rappel multilingue en extraction d'entités cliniques à MultiClinNER avec une tête BIO légère, tandis que *ClinicalAligner26AM* s'est classé premier sur le transfert d'entités MultiClinCorpus avec un F1 supérieur à 0,95 dans presque tous les réglages.

ABSTRACT

A Family of Multilingual Clinical ColBERT Models Diagnosable Through Semantic Latents.

In this paper, we introduce the concept of a diagnosable ColBERT model, that is, a language model whose understanding can be analyzed directly at the token level. Although standard ColBERT models make token–token interaction scores inspectable, this is not sufficient to diagnose errors in clinical settings : to do so, one also needs a structured reference latent space that makes visible confusions of concept, local composition, and contextual qualification. With this capacity for introspection into a model's cognitive understanding, it becomes possible to train more robust models by iterating on the data in a reactive manner. This perspective is supported by two recent results : *ClinicalEncoder26AM* achieved the best multilingual recall for clinical entity extraction at MultiClinNER with a lightweight BIO head, while *ClinicalAligner26AM* ranked first on multilingual entity transfer in MultiClinCorpus, with an F1 above 0.95 in almost all settings.

MOTS-CLÉS : diagnosticable, colbert, clinique, biomédical, multilingue, espace latent.

KEYWORDS: diagnosable, colbert, clinical, biomedical, multilingual, latent space.

1 Introduction

En recherche clinique, expliquer qu'un modèle attribue un bon score de correspondance à deux passages ne suffit pas encore à diagnostiquer ce qu'il comprend réellement. Un score élevé peut masquer une confusion sur l'identité du concept, une mauvaise composition locale des termes, ou l'oubli d'une qualification contextuelle pourtant cliniquement décisive. Or, dans ce domaine, ces distinctions ne sont pas secondaires : il faut séparer une condition d'un test, un antécédent d'un problème actuel, un fait affirmé d'une négation, une hypothèse d'une certitude, ou encore le patient de l'un de ses apparentés (Harkema *et al.*, 2009). Les architectures de type ColBERT rendent déjà visibles des scores d'interaction token–token (Khattab & Zaharia, 2020), mais ces scores seuls disent mal pourquoi deux fragments se rapprochent, ni quelle partie de leur sens reste absente, ambiguë ou mal ancrée.

Notre thèse est donc la suivante : pour rendre un modèle clinique multilingue réellement diagnosticable, il faut pouvoir inspecter ses représentations tokenisées dans un espace latent de référence, structuré par la sémantique clinique. Un tel espace ne sert pas seulement à expliquer une décision après coup. Il sert aussi à repérer les concepts bien ancrés, les voisinages suspects et les distinctions contextuelles manquantes, notamment lorsque l'on cherche à relier les représentations à des terminologies cliniques comme SNOMED CT ou l'UMLS (Donnelly, 2006; Bodenreider, 2004). L'analyse peut alors se faire à plusieurs niveaux complémentaires : celui du terme isolé, du groupe de mots, de la phrase, puis du paragraphe ou du contexte discursif plus large. Cette lisibilité est précieuse pour l'audit, le débogage, la curation de données et, plus largement, pour construire une confiance plus informée dans des usages cliniques multilingues.

C'est cette perspective qui motive la famille de modèles étudiée ici. *ClinicalEncoder26AM* aligne des représentations multilingues de niveau token vers un espace clinique de référence afin de rendre leurs structures sémantiques plus interprétables. *ClinicalAligner26AM*, initialisé à partir de ce même encodeur, prolonge cette idée vers l'alignement interlingue au niveau token pour transférer plus précisément des entités d'une langue à l'autre. Autrement dit, le premier modèle aide à diagnostiquer la qualité des représentations cliniques elles-mêmes, tandis que le second montre que ces représentations peuvent aussi soutenir une localisation multilingue fine et actionnable.

2 Méthodologie

Notre famille de modèles repose d'abord sur *ClinicalEncoder26AM*, un encodeur multilingue diagnostiquable, post-entraîné à partir de BGE-M3 (Chen *et al.*, 2024) puis aligné à plusieurs niveaux sur un espace clinique de référence inspiré de travaux comme BioLORD (Remy *et al.*, 2024). Dans cette optique, le post-entraînement combine plusieurs objectifs hiérarchiques (niveau local ; niveau conceptuel, et niveau contextuel) ainsi que plusieurs sources complémentaires, notamment des notes cliniques synthétiques, des conversations médecin–patient, des ressources biomédicales annotées, une distillation multi-adaptateurs et un objectif final de type ColBERT. Ce montage permet de conserver la géométrie multilingue du modèle de base tout en l'ancrant davantage dans la sémantique clinique. Le contexte long de CE26AM joue aussi un rôle pratique, car il limite les erreurs de découpage et préserve mieux les indices locaux qui changent le sens clinique d'une mention. Pour participer à la tâche MultiClinNER (Gallego-Donoso *et al.*, 2026), nous réentrainons cet encodeur avec un classificateur BIO classique sur les données d'entrainement (Lima López *et al.*, 2026).

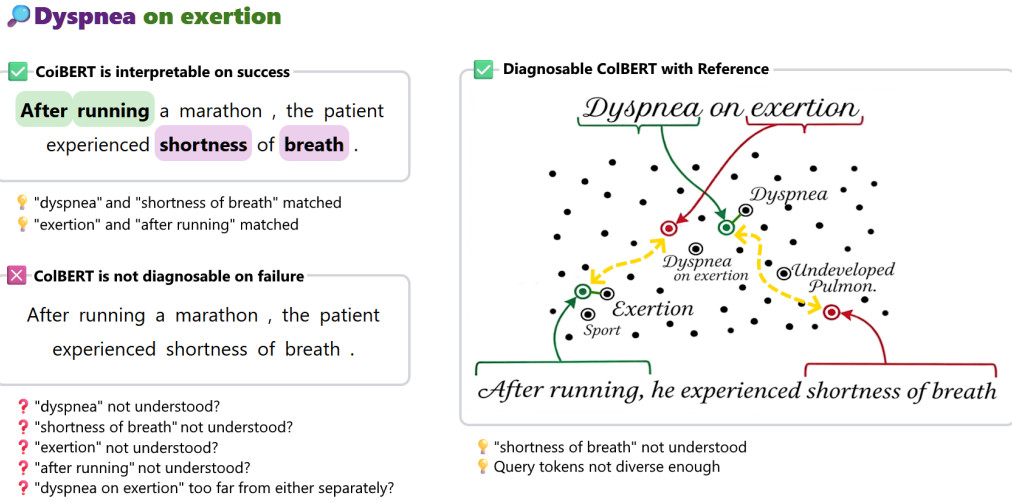

FIGURE 1 – Les modèles ColBERT classiques rendent les scores d'interaction requête–document inspectables, mais apportent peu d'aide lorsqu'aucune correspondance pertinente n'est trouvée pour un token. En ajoutant un espace latent de référence, un modèle de type Diagnosable ColBERT permet au contraire de distinguer plus clairement la sémantique latente et la compréhension en contexte, et donc de diagnostiquer les échecs de manière plus rapide et plus actionnable.

À partir de l'encodeur pré-entraîné, nous construisons aussi *ClinicalAligner26AM*, qui transpose la même philosophie vers l'alignement interlingue au niveau des tokens. Le modèle est initialisé à partir de CE26AM, puis entraîné pour rapprocher plus fidèlement les tokens ou groupes de tokens qui expriment le même contenu clinique dans des langues différentes. Autrement dit, là où CE26AM rend les représentations plus lisibles et plus cliniquement cohérentes, CA26AM exploite cette qualité pour transférer plus précisément des entités d'une langue source vers une langue cible.

3 Résultats empiriques

Les résultats empiriques soutiennent d'abord la qualité clinique des représentations apprises par CE26AM. Sur MultiClinNER, ce modèle atteint un niveau de rappel multilingue exceptionnel avec une simple tête BIO légère, jusqu'à obtenir les meilleures performances de rappel de la compétition sur la plupart des langues et types d'entités. Malgré cette simplicité, il reste compétitif sur les métriques globales de type F1 caractère, où il se maintient dans le Top 5 sur la quasi-totalité des configurations. Les courbes d'entraînement montrent en outre une meilleure efficacité en données que le modèle de base BGE-M3, ce qui suggère que le post-entraînement clinique n'améliore pas seulement l'interprétabilité, mais fournit aussi une initialisation supérieure.

Les résultats de CA26AM vont plus loin encore en montrant que ces représentations peuvent être exploitées pour un transfert d'entités extrêmement précis entre langues. Sur MultiClinCorpus, CA26AM se classe premier. Dans presque tous les réglages, le F1 au niveau caractère dépasse 0,95, et plusieurs combinaisons langue–entité approchent ou atteignent 0,97. Cette différence avec le scénario

MultiClinNER n'est pas surprenante : découvrir une entité dans un document reste plus difficile que la relocaliser dans une traduction lorsque sa présence est déjà connue. L'alignement au niveau token réduit fortement l'incertitude sur la zone pertinente, ce qui permet des frontières beaucoup plus précises et plus stables.

Pris ensemble, ces deux séries de résultats tendent à confirmer notre thèse. Le succès de CE26AM suggère qu'un espace latent cliniquement structuré capture déjà des régularités sémantiques multilingues robustes. Le succès de CA26AM montre que cette même structuration peut être convertie en un signal d'alignement assez fidèle pour un transfert exact et donc pour un diagnostic plus actionnable des erreurs. Autrement dit, la diagnosticabilité n'apparaît pas ici comme un luxe interprétatif ajouté après coup, mais comme une propriété étroitement liée à la performance pratique des modèles.

Modèle	Char R	Char P	Char F1	Strict R	Strict P	Strict F1
CA26AM+MCAI	0.96 ± 0.02	0.96 ± 0.02	0.96 ± 0.02	0.83 ± 0.03	0.82 ± 0.03	0.82 ± 0.03
CA26AM	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.81 ± 0.03	0.80 ± 0.03	0.81 ± 0.03
CE26AM	0.85 ± 0.02	0.81 ± 0.02	0.83 ± 0.02	0.74 ± 0.03	0.70 ± 0.03	0.72 ± 0.03

TABLE 1 – Résumé des résultats MultiClinAI (Gallego-Donoso *et al.*, 2026) de nos systèmes ClinicalAligner26AM et ClinicalEncoder26AM (rappel, précision et F1 pondérés au niveau des caractères, ainsi que leurs variantes strictes ; moyennes $\pm$ écart-type sur les 6 langues : fr,en,de,nl,sv,cz)

4 Discussion

L'intérêt principal de cette famille de modèles apparaît lorsqu'une erreur doit être expliquée plutôt que simplement mesurée. Supposons qu'une mention clinique soit mal retrouvée ou mal projetée dans une autre langue. L'inspection de CE26AM peut alors aider à distinguer plusieurs causes simples : abréviation mal comprise, concept voisin mais incorrect, ou qualification contextuelle insuffisamment marquée, par exemple une négation ou un statut historique.

Cette distinction est utile parce qu'elle oriente directement la curation. Selon le cas, il faudra enrichir les exemples synonymiques, renforcer les contrastes de négation ou de temporalité, ou ajouter des données parallèles multilingues pour renforcer leur cohérence.

Cette proposition reste toutefois un cadrage méthodologique plutôt qu'une théorie complète de la compréhension clinique. La construction optimale d'un espace latent de référence dépasse le périmètre d'un papier court, et un échafaudage utile pour le diagnostic peut encore rester incomplet vis-à-vis des besoins réels de recherche ou d'analyse de documents.

5 Conclusion

Déboguer une famille de modèles cliniques multilingues est donc plus aisé lorsque l'on peut inspecter des représentations contextualisées dans un espace latent de référence, et pas seulement des scores finaux. Cela permet d'améliorer plus rapidement la qualité des modèles grâce à l'analyse qualitative et à une meilleure sélection des données. Les résultats de CE26AM et de CA26AM sont déjà cohérents avec cette idée. L'intérêt principal de notre approche est de rendre les erreurs plus lisibles, plus localisables et donc plus corrigeables ; créant ainsi un cercle vertueux.

Références

BODENREIDER O. (2004). The unified medical language system (UMLS) : integrating biomedical terminology. *Nucleic Acids Research*, **32**(Database issue), D267–D270. DOI : 10.1093/nar/gkh061.

CHEN J., XIAO S., ZHANG P., LUO K., LIAN D. & LIU Z. (2024). M3-Embedding : Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv :2402.03216*. DOI : 10.48550/arXiv.2402.03216.

DONNELLY K. (2006). SNOMED-CT : The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, **121**, 279–290.

GALLEGO-DONOSO F., LIMA-LÓPEZ S., ROSELL J., FARRÉ-MADUELL E. & KRALLINGER M. (2026). The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction : Systems, Evaluation, and Datasets. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks* : Association for Computational Linguistics.

HARKEMA H., DOWLING J. N., THORNBLADE T. & CHAPMAN W. W. (2009). ConText : An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, **42**(5), 839–851. DOI : 10.1016/j.jbi.2009.05.002.

KHATTAB O. & ZAHARIA M. (2020). ColBERT : Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 39–48. DOI : 10.1145/3397271.3401075.

LIMA LÓPEZ S., ROSELL J., RODRÍGUEZ MIRET J., GALLEGO-DONOSO F. & KRALLINGER M. (2026). Multiclinai shared task training data. DOI : 10.5281/zenodo.18508039.

REMY F., DEMUYNCK K. & DEMEESTER T. (2024). BioLORD-2023 : Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, **31**(9), 1844–1855. DOI : 10.1093/jamia/ocae029.