

# **Développement d'un Modèle de Langage pour la Classification et l'Analyse de Contenus Pédagogiques en Odontologie**

## **Stage pour Ingénieurs/Étudiants de Master II : Développement d'un Modèle de Langage pour la Classification et l'Analyse de Contenus Pédagogiques en Odontologie**

### **Contexte**

En médecine bucco-dentaire, les supports pédagogiques (QCM, glossaires, chapitres de cours, fiches cliniques) sont souvent produits par différents enseignants, à des moments variés, et selon différentes « écoles ». Cette diversité, bien que riche, engendre fréquemment des incohérences terminologiques, des redondances, voire des contradictions entre les documents. Ces écarts peuvent créer une confusion importante chez les étudiants, notamment lorsqu'ils préparent des examens ou tentent de structurer leurs connaissances de manière cohérente.

En parallèle, les modèles de langage sont de plus en plus utilisés dans des applications éducatives, notamment en médecine et en odontologie (Lafourcade et al. 2025; Xu et al. 2025). Ces modèles offrent une opportunité unique pour développer des outils d'analyse pédagogique automatisée, capables de détecter les incohérences entre supports, d'extraire les notions clés, et de restructurer les contenus selon des logiques pédagogiques explicites.

### **Objectif du projet**

Détecter une méthode basée sur des modèles de langage pour détecter automatiquement les incohérences, redondances ou contradictions entre différents supports pédagogiques, sources fréquentes de confusion pour les étudiants.

### **Étapes du projet**

#### 1. Classification et clustering des contenus pédagogiques

- Réaliser une analyse exploratoire des données disponibles (QCM, glossaires, chapitres de cours, fiches cliniques).
- Appliquer des méthodes de classification supervisée (e.g., SVM, transformers) ou de clustering non supervisé (e.g., k-means, HDBSCAN, UMAP) pour regrouper les contenus selon : les thématiques abordées, les niveaux de compétences définis par les référentiels européens.
- Extraire les notions clés (termes, concepts, relations) à l'aide de techniques d'attention ou de pondération (TF-IDF, TextRank).
- Identifier les chapitres ou ressources les plus représentatifs pour chaque thématique ou niveau.

#### 2. Détection d'incohérences et de redondances entre supports

- Développer des méthodes pour détecter automatiquement les incohérences entre documents pédagogiques, en s'appuyant sur :
  - des incohérences terminologiques : usage de termes différents pour désigner un même concept, ou inversement ;
  - des incohérences logiques : contradictions dans les définitions, les recommandations cliniques ou les réponses aux QCM ;

- o des incohérences structurelles : désalignement entre les objectifs pédagogiques annoncés et les contenus réellement abordés ;
- o des redondances non justifiées : répétitions inutiles ou mal contextualisées d'un même contenu dans plusieurs supports.
- Utiliser des techniques de similarité sémantique (cosine similarity, embeddings) et des outils d'explicabilité (LIME, SHAP) pour justifier les regroupements ou les alertes d'incohérence.
- Approche par Natural Language Inference (NLI) ou Entailment
- Approche par question spécifique croisée à partir de la banque de questions

### **Compétences requises :**

- Connaissances en traitement du langage naturel (NLP), apprentissage automatique et clustering.
- Maîtrise de Python et des bibliothèques de deep learning (PyTorch, TensorFlow).
- Intérêt pour les questions d'impact environnemental et d'analyse de données éducatives.

### **Perspectives**

Ce stage permettra de développer des compétences en IA appliquée à l'enseignement de l'odontologie et de contribuer à l'innovation pédagogique. En intégrant une réflexion sur l'impact environnemental, il s'inscrit dans une démarche responsable et durable, en phase avec les exigences actuelles des industries technologiques et éducatives.

### **Contacts**

**Sébastien Valette (sebastien.valette@creatis.insa-lyon.fr)**

**Thomas Grenier (thomas.grenier@creatis.insa-lyon.fr)**

**Raphaël Richert (raphael.richert@insa-lyon.fr)**

### **Références**

Lafourcade C, Kérouédan O, Ballester B, Richert R. 2025. Accuracy, consistency, and contextual understanding of large language models in restorative dentistry and endodontics. *J Dent.* 157(September 2024):1–8.

Xu X, Liu S, Zhu L, Long Y, Zeng Y, Lu X, Li J, Dong Y. 2025. Development and evaluation of a retrieval-augmented large language model framework for enhancing endodontic education. *Int J Med Inform.* 203(May):106006. <https://doi.org/10.1016/j.ijmedinf.2025.106006>.