

Utilisation des LLMs pour l'Apprentissage Personnalisé

Mars 2025

Objectif du projet

Le projet vise à exploiter les grands modèles de langage (LLMs) pour améliorer l'apprentissage des élèves en générant des contenus pédagogiques personnalisés et adaptés à leurs besoins spécifiques. L'objectif est de combler les lacunes individuelles et de renforcer la compréhension des concepts à travers une approche interactive et adaptative.

Problématique

Disparité des élèves

Dans une même classe, les élèves possèdent des représentations mentales différentes pour un même concept. Par exemple, pour comprendre les fractions :

- Un élève peut les associer aux parts de pizza.
- Un autre aux morceaux de chocolat.
- Un troisième peut ne pas comprendre la notion de *couper en parts égales*.

Problème avec les LLMs : Les modèles de langage actuels ne tiennent pas compte de ces différences individuelles dans les représentations mentales.

Ignorance des prérequis

Les LLMs ne vérifient pas toujours les connaissances préalables des élèves avant de leur fournir des explications, ce qui peut entraîner des incompréhensions ou accentuer les lacunes existantes.

Solution proposée

Modélisation des connaissances des élèves

Objectif : Construire un graphe de connaissances qui capture :

- Les représentations mentales propres à chaque élève.
- Les connexions qu'ils établissent entre les concepts.

Ce graphe permettra d'identifier les forces et les faiblesses de chaque élève.

Alignement via des questions diagnostiques

Poser des questions spécifiques et ciblées pour évaluer le niveau de compréhension de chaque élève avant de générer du contenu pédagogique adapté.

Explications itératives et feedback adaptatif

Permettre aux LLMs de :

- Proposer des explications étape par étape.
- Vérifier la compréhension de l'élève à chaque étape.
- Ajuster les explications en fonction des réponses et des réactions de l'élève.

Objectif global

Créer un modèle LLM avancé capable de :

- Comprendre les réponses des élèves sous différentes formes (textes, images, PDF).
- Expliquer les concepts selon le niveau et les lacunes spécifiques de chaque élève.
- Corriger les erreurs de manière progressive et pédagogique.
- Adapter les explications selon les représentations mentales uniques de chaque élève.

Datasets Utilisés

Les datasets suivants sont utilisés pour entraîner et affiner les modèles de langage, afin de répondre aux besoins spécifiques de chaque élève dans l'apprentissage personnalisé.

1. SQuAD (Stanford Question Answering Dataset)

Le SQuAD est un dataset de questions-réponses qui repose sur des articles issus de Wikipédia. Chaque question est associée à un passage précis du texte, et la réponse est extraite de cet extrait.

- **Objectif :** Entraîner le modèle à rechercher des réponses spécifiques dans un contenu donné, essentiel pour des applications comme l'analyse de textes soumis par les élèves (ex. : PDFs, articles, etc.).
- **Application pédagogique :** Utile pour développer des systèmes qui comprennent les textes et fournissent des réponses précises aux questions posées, renforçant ainsi la compréhension des élèves dans différents domaines.

2. EdNet

EdNet est un dataset massif qui suit les interactions entre élèves et plateformes éducatives, incluant les questions posées, les réponses des élèves (correctes ou non), ainsi que le temps passé sur chaque question.

- **Objectif :** Analyser les comportements des élèves pour adapter les contenus pédagogiques à leurs besoins spécifiques. Le modèle peut ainsi identifier les erreurs fréquentes et proposer des corrections adaptées.

- **Application pédagogique** : Permet de personnaliser les explications en fonction des erreurs répétées et des temps de réponse des élèves, afin d'ajuster le niveau de difficulté ou reformuler les explications.

3. MathQA

MathQA est une ressource dédiée aux problèmes mathématiques complexes, incluant des solutions détaillées étape par étape pour chaque question.

- **Objectif** : Permet au modèle de comprendre et de générer des solutions détaillées pour des problèmes mathématiques, aidant ainsi les élèves à surmonter leurs difficultés.
- **Application pédagogique** : Idéal pour entraîner un modèle à fournir des explications progressives, permettant aux élèves de suivre et de comprendre chaque étape de la résolution d'un problème mathématique.

4. DocVQA

DocVQA est conçu pour entraîner des modèles à répondre à des questions basées sur des documents, incluant des PDFs, scans, ou images complexes.

- **Objectif** : Améliorer la capacité des modèles à lire et comprendre des documents structurés de manière complexe.
- **Application pédagogique** : Essentiel pour permettre aux modèles de traiter et analyser des documents envoyés par les élèves, notamment des copies manuscrites ou des PDF contenant des questions et des schémas.

5. Open-Assistant Conversations

Ce dataset regroupe des dialogues éducatifs dans lesquels un modèle joue le rôle de professeur et l'autre celui d'élève, avec des explications et corrections dynamiques tout au long de la conversation.

- **Objectif** : Entraîner un modèle à mener des dialogues interactifs, avec des ajustements en temps réel en fonction des réponses de l'élève.
- **Application pédagogique** : Idéal pour créer des systèmes de tutorat interactif, où le modèle s'adapte aux questions et besoins de l'élève, en fournissant un feedback immédiat et personnalisé.

6. les graphes

1 Recovering Concept Prerequisite Relations from University Course Dependencies

L'article traite du problème de récupération des relations de prérequis entre concepts à partir des dépendances entre les cours universitaires

les chercheurs ont commencé à utiliser Wikipédia pour identifier les relations de prérequis entre des concepts partagés universellement

Objectif : construire un graphe des concepts — une sorte de carte — qui montre quelles notions doivent être maîtrisées avant d’en apprendre d’autres.

2 Problème posé (CPR-Recover)

L’article introduit le problème CPR-Recover (Concept Prerequisite Relation Recovery), qui consiste à :

- **Entrées :**
 - Une liste de cours universitaires avec leurs dépendances (ex. : Cours 2 dépend de Cours 1).
 - Une représentation des concepts enseignés dans chaque cours (ex. : Matrices, Algèbre linéaire).
- **Sortie :**
 - Un graphe des concepts qui montre les relations de prérequis entre eux.

3 Hypothèses de base

La méthode repose sur deux hypothèses clés :

- **Causalité :**

Si un cours dépend d’un autre, c’est parce que les concepts enseignés dans le premier sont nécessaires pour comprendre le second.
- **Parcimonie (Sparsity) :**

Toutes les notions ne sont pas liées. Il existe peu de relations de prérequis dans le graphe des concepts par rapport au nombre total de paires possibles.

3.1 Pourquoi ces deux hypothèses sont importantes ?

- **Causalité** permet de construire un graphe basé sur des liens logiques entre les concepts.
- **Parcimonie** évite d’ajouter des liens inutiles et garde le graphe simple et lisible.

4 Méthodologie : Comment CPR-Recover fonctionne ?

- Chaque cours est représenté par un vecteur de concepts basé sur les articles Wikipédia mentionnés dans sa description.
- Le modèle cherche à minimiser une fonction mathématique pour trouver le graphe de concepts le plus simple.
- **Contraintes utilisées :**
 - Parcimonie
 - Causalité