

Corrigé Activité P1C3



Structure

Une note qui :

- décrit la tokenisation et la mise en nombres via embeddings ;
- explique que le modèle "pondère" les mots pertinents grâce au mécanisme d'attention ;
- résume le rôle du Transformer (encodeur/décodeur ou décodeur-seul) pour générer le texte ;
- précise que l'on adapte des foundation models par fine-tuning ;
- mentionne la nature statistique du procédé (prédiction du prochain token).

Exemple de prompt :

« Agissez comme pédagogue. En 200–250 mots, expliquez à des dirigeants non techniques comment fonctionne un modèle génératif de texte :

tokenisation/embeddings, **mécanisme d'**attention**, **architecture Transformer** (décodeur-seul vs encodeur-décodeur) et **fine-tuning**. Utilisez un exemple simple ("Le chat...") pour montrer comment l'attention relie les mots. Style clair, métaphores sobres, en français. »

Résultat ChatGPT

Comment fonctionne un modèle génératif de texte

Un modèle génératif de texte prédit mot après mot la suite la plus probable d'une phrase. Pour cela, il transforme d'abord chaque mot en **unités élémentaires** appelées *tokens*. Ces tokens deviennent des **vecteurs numériques** (*embeddings*) qui capturent le sens et les relations entre mots.

L'élément clé est le **mécanisme d'attention**. Imaginez un projecteur qui éclaire, à chaque instant, les mots les plus pertinents pour comprendre le mot en cours. Exemple : dans la phrase « Le chat dort sur le canapé », quand le modèle prédit « dort », l'attention se focalise sur « chat », et non sur « canapé ». Ce jeu de pondérations permet de garder la cohérence du sens.

Ces calculs sont organisés dans l'**architecture Transformer**, aujourd'hui dominante. Deux variantes existent :

- **Décodeur seul** (comme GPT) : se concentre sur la prédiction de la suite d'un texte.
- **Encodeur-décodeur** (comme T5) : comprend un texte d'entrée (encodeur), puis en génère une reformulation ou traduction (décodeur).

Enfin, les modèles sont souvent adaptés par **fine-tuning** : un entraînement complémentaire sur des données spécialisées (juridiques, médicales, sectorielles) ou via des retours humains, afin d'améliorer la pertinence dans un contexte donné.

En résumé, un modèle génératif n'« invente » pas au sens humain : il calcule, grâce à l'attention et à des milliards de paramètres, la suite statistiquement la plus plausible d'une phrase.