Fiche Technique : L'Architecture Transformer

1. L'Entree (Input)

Avant meme d'entrer dans les couches, le texte subit deux transformations :
- Embedding (Plongement) : Chaque mot est transforme en un vecteur numerique dense. C'est le 'sens' brut du mot.
- Positional Encoding : On ajoute un motif mathématique aux vecteurs pour donner l'ordre des mots (1er, 2eme...).

2. L'Encodeur (Le Lecteur)

Role : Analyser le contexte et comprendre les relations entre les mots.

A. Self-Attention (Auto-Attention)
Chaque mot se demande : 'Quelle importance les autres mots ont-ils pour moi ?'
Mecanisme Q, K, V (Query, Key, Value). Exemple : Dans 'La pomme est rouge', 'rouge' porte attention sur 'pomme'.

B. Add & Norm
Connexion Residuelle : On ajoute l'entree a la sortie pour que l'info circule bien.
Normalisation : On stabilise les chiffres.

C. Feed-Forward Network
Un petit reseau de neurones classique pour digerer l'information.

3. Le Decodeur (L'Ecrivain)

Role : Generer la sequence de sortie mot par mot.

A. Masked Self-Attention (Attention Masquee)
Interdiction de voir le futur. On applique un masque pour cacher les mots qui n'ont pas encore ete ecrits.

B. Cross-Attention (Attention Croisee)
Le pont entre l'Encodeur et le Decodeur. Le Decodeur regarde la phrase source complete (Encodeur) pour savoir quoi ecrire.

C. Sortie (Linear + Softmax)
Transforme le vecteur final en probabilites pour choisir le mot le plus probable dans le dictionnaire.

Resume des differences

- Encodeur seul (ex: BERT) : Pour comprendre, classer, extraire (SQuAD).
- Decodeur seul (ex: GPT) : Pour generer du texte, chatter.
- Encodeur-Decodeur (ex: T5) : Pour traduire, resumer.

Ton projet actuel : Tu as fait du 'From Scratch' (Encodeur) pour de l'Extraction de reponse.