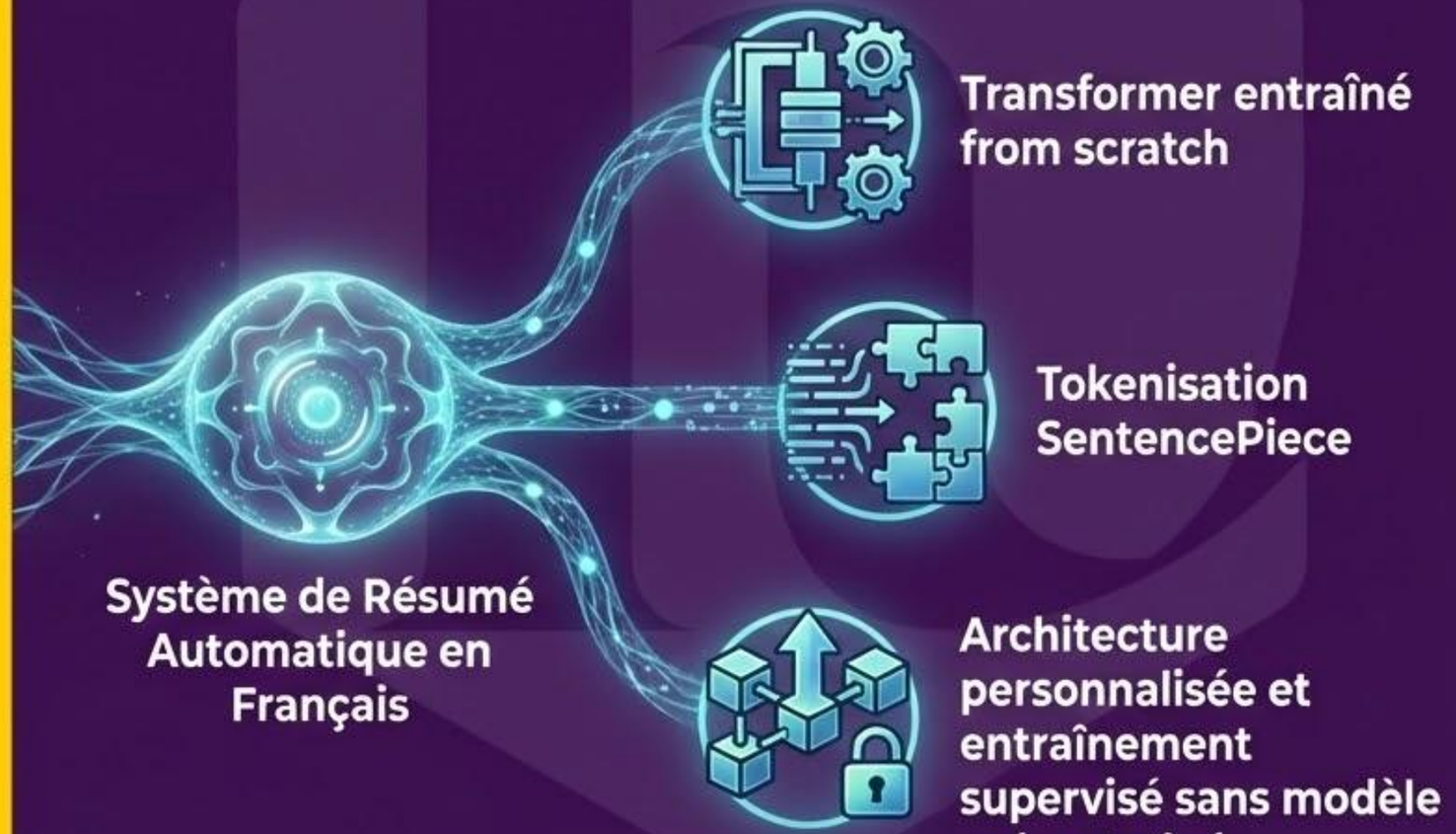
IA de Synthèse de Texte en Français



CONTEXTE

Problématique & Objectif du Projet

Pourquoi le résumé automatique ?

- Volumes importants de textes à traiter quotidiennement
- Besoin croissant de synthèse rapide et efficace
- Reformulation intelligente sans copier le contenu original
- Gain de temps pour les chercheurs et professionnels

Objectif du projet

Résumé **abstractif** : génération de nouvelles formulations

Modèle développé **from scratch** sans transfert

- Pipeline complet maîtrisé de bout en bout
- Compréhension approfondie des mécanismes internes

Le résumé automatique représente une tâche particulièrement complexe car elle nécessite à la fois une compréhension profonde du contenu source et une capacité de reformulation cohérente.

DONNÉES : Constitution du Dataset d'Entraînement

Source

Articles de presse en français provenant de sources journalistiques variées

Structure

text : article complet

summary : résumé cible

title : injecté comme signal

Séparation

Split strict en trois ensembles : train, validation et test sans chevauchement

Le titre est utilisé comme signal sémantique supplémentaire pour faciliter l'apprentissage from scratch et guider la génération du résumé.

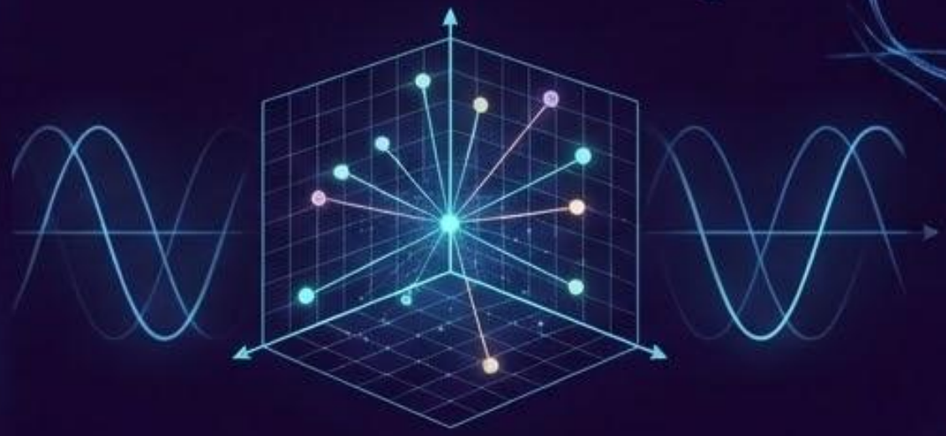
TOKENISATION & EMBEDDINGS

Tokenisation SentencePiece

16 000 tokens

Approche **Unigram** pour la segmentation du texte français avec un vocabulaire optimisé de **16 000 tokens**, permettant une couverture équilibrée des mots courants et des termes spécialisés.

Couche d'Embeddings



Embeddings appris **from scratch** durant l'entraînement :

- Dimension vectorielle : **256**
- Encodage positionnel sinusoïdal
- Initialisation aléatoire contrôlée
- Aucune représentation pré-entraînée

L'**absence d'embeddings pré-entraînés** signifie que toutes les représentations sémantiques sont construites uniquement à partir des données d'entraînement disponibles.

ARCHITECTURE

Architecture Transformer Encoder-Decoder

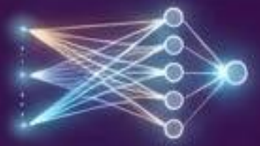
Couches

4 couches encodeur pour analyser le texte source

4 couches décodeur pour générer le résumé

Attention

4 têtes d'attention multi-têtes pour capturer différentes relations sémantiques



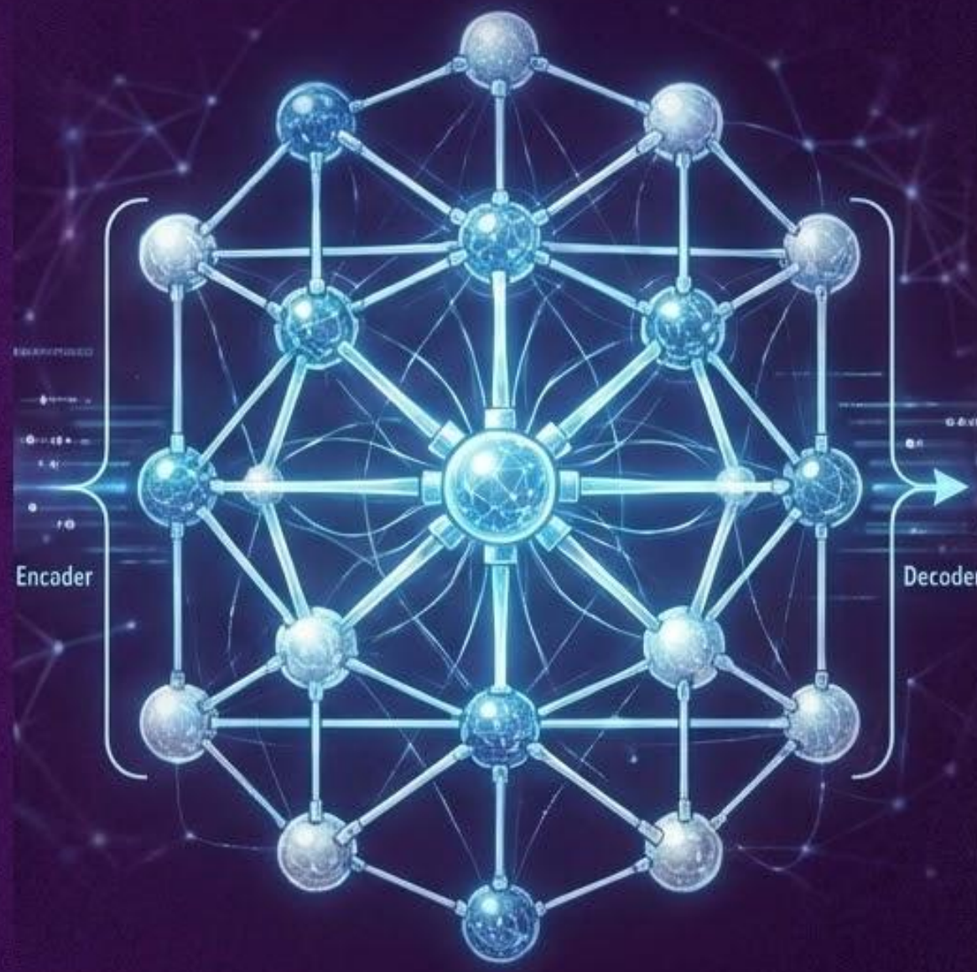
Feed-Forward

Dimension intermédiaire de 1024 dans les couches feed-forward

Régularisation

Dropout de 0.1 pour prévenir le surapprentissage

L'encodeur transforme le texte source en représentations contextuelles, tandis que le décodeur génère le résumé token par token en utilisant l'attention croisée. Architecture similaire aux standards modernes mais entièrement implémentée et entraînée from scratch.



CONFIGURATION

Stratégie d'Entraînement

Hyperparamètres & Technique

Optimiseur AdamW

Learning rate : 3×10^{-4}

Learning rate 3×10^{-4}

Fonction de perte

Cross-entropy + label smoothing (0.1)

Teacher Forcing

Utilisation des tokens de référence pendant l'entraînement

Gradient Clipping

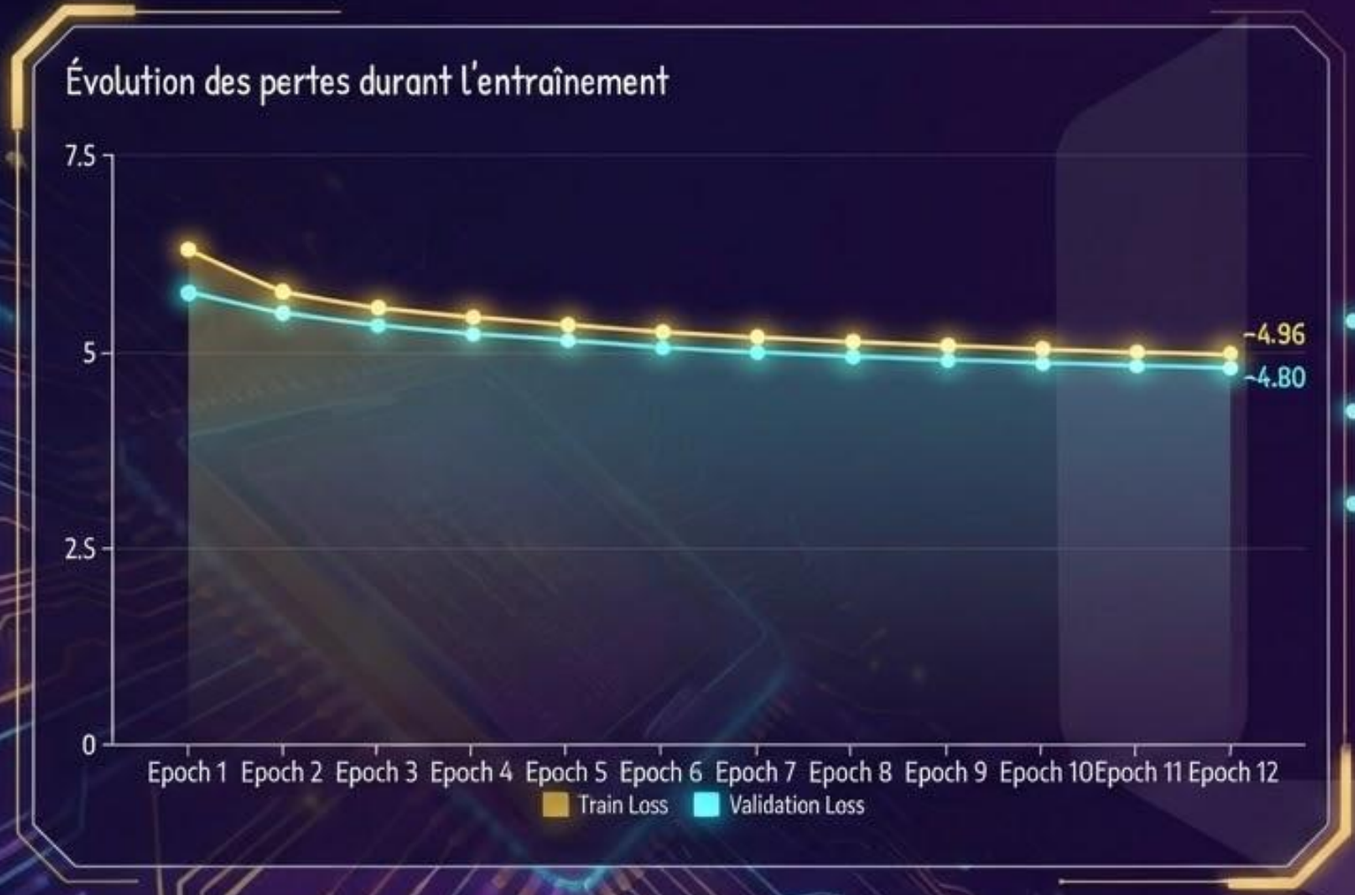
Stabilisation de l'entraînement

Infrastructure

Entraînement sur GPU RTX 4060 Laptop avec batches limités par epoch pour accélérer les itérations tout en conservant la stabilité d'apprentissage.

RÉSULTATS

Évaluation Quantitative des Performances



ROUGE-1

0.15

Chevauchement de unigrammes

ROUGE-2

0.02

Chevauchement de bigrammes

ROUGE-L

0.12

Plus longue sous-séquence commune

Les scores ROUGE restent modestes, ce qui est attendu et cohérent pour un modèle entraîné entièrement from scratch sans phase de pré-entrainement sur de larges corpus.

ANALYSE QUALITATIVE

Exemples de Résumés Générés

Observations positives

- ✅ Syntaxe globalement correcte
- ✅ Style de résumé appris avec succès
- ✅ Structure cohérente des phrases
- ✅ Formulations fluides et naturelles

Défis identifiés

- ⚠️ Contenu parfois incohérent
- ❓ Imprécisions factuelles
- ❓ Écart entre forme et fond
- ❓ Pertinence sémantique variable

Le modèle a appris la forme du résumé — sa structure et son style typique — avant d'acquérir une compréhension profonde du fond sémantique et de la précision factuelle.

⚠ LIMITES

Analyse Critique & Limites Observées

Problèmes identifiés

Hallucinations

Génération d'informations non présentes dans le texte source

Répétitions

Motifs répétitifs dans les résumés générés

Précision factoelle

Faible fidélité aux faits du document original

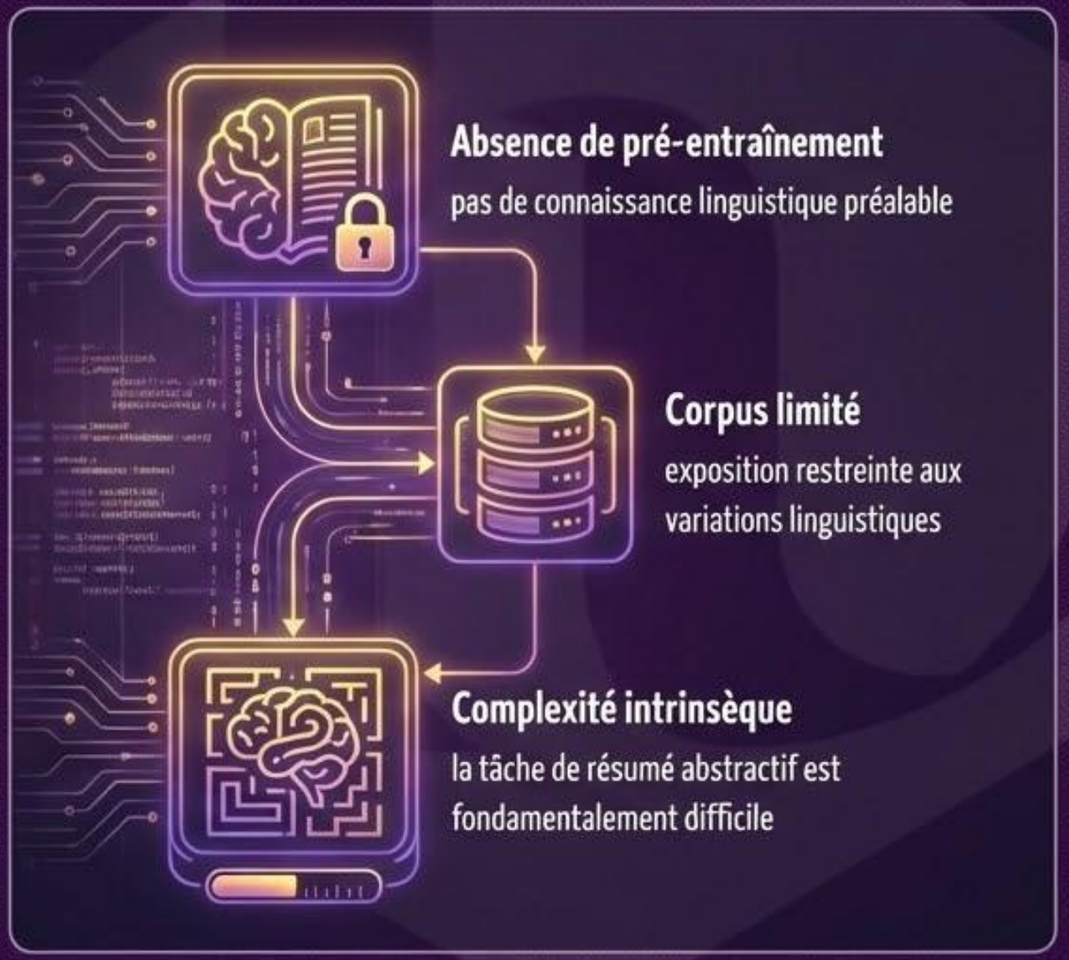
Précision factoelle

Faible fidélité aux faits du document original

Complexité intrinsèque

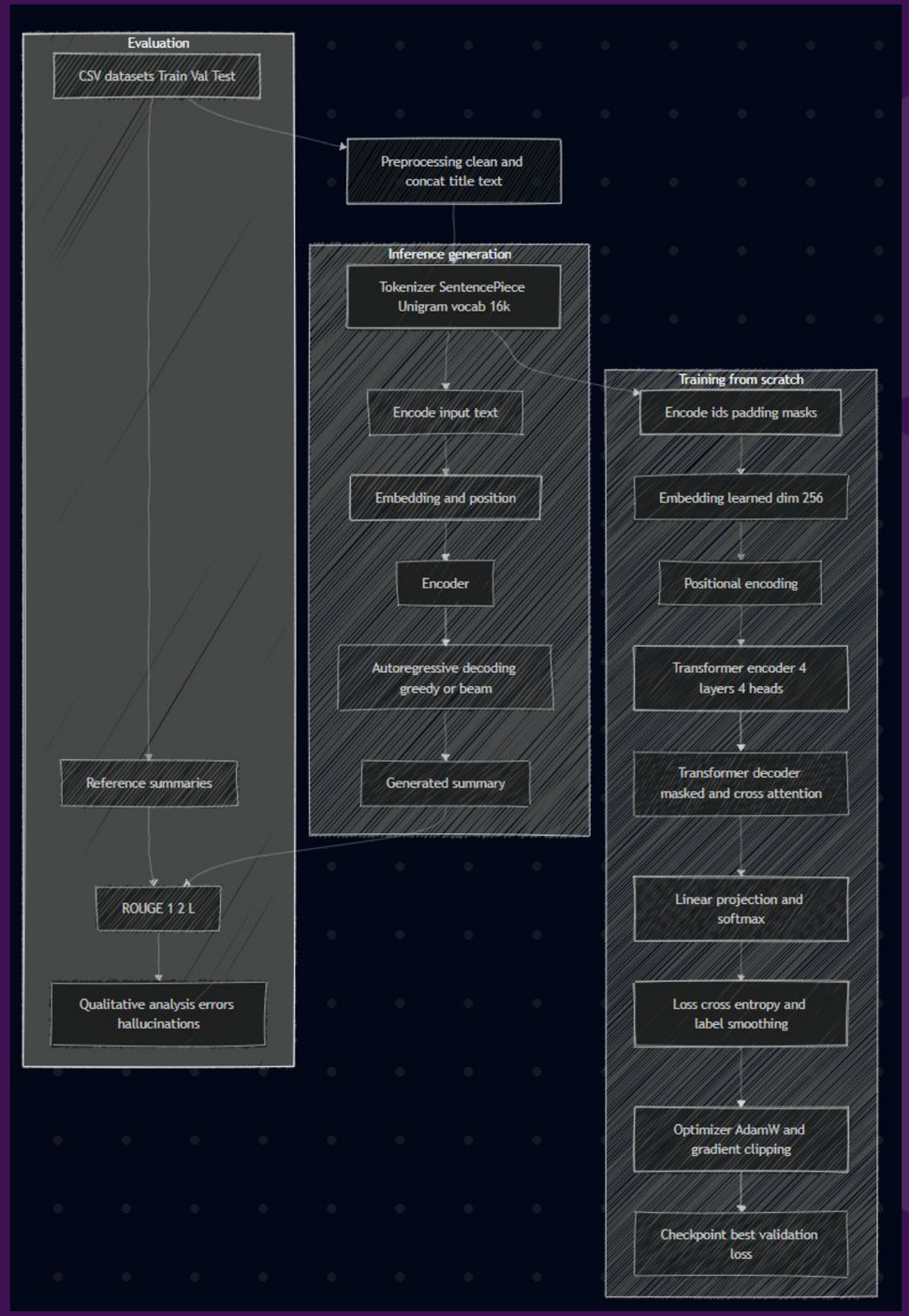
exposition restreinte aux modations restreinte aux variatons linguistiques

Causes sous-jacentes



Ces limitations sont typiques et attendues pour les modèles génératifs entraînés from scratch sans phase de pré-entraînement extensive.

PIPELINE :


Evaluation
CSV datasets Train Val Test
Preprocessing clean and concat title text
Inference generation
Tokenizer SentencePiece Unigram vocab 16k
Encode input text
Embedding and position
Encoder
Autoregressive decoding greedy or beam
Generated summary
Reference summaries
ROUGE 1 2 L
Qualitative analysis errors hallucinations
Training from scratch
Encode ids padding masks
Embedding learned dim 256
Positional encoding
Transformer encoder 4 layers 4 heads
Transformer decoder masked and cross attention
Linear projection and softmax
Loss cross entropy and label smoothing
Optimizer AdamW and gradient clipping
Checkpoint best validation loss

PERSPECTIVES

Conclusion & Perspectives d'Amélioration

Bilan

Pipeline complet fonctionnel, architecture moderne maîtrisée, résultats cohérents avec les contraintes

Perspectives

Pré-entrainement auto-supervisé, beam search avec contraintes, corpus plus large et diversifié

Ce projet met en évidence le rôle central du pré-entraînement dans les performances des modèles de langage modernes.

01

Pré-entrainement auto-supervisé

Pré-entrainement auto-supervisé Apprentnissage sur de vastes corpus non annotés

02

Beam search optimisé

Amélioration de la qualité de génération avec contraintes

03

Extension du corpus

Diversification et augmentation des données d'entrainement