# LLMs pour la Recherche en Littérature : Premières Expérimentations

Jean Barré

28 février 2024

Lattice : ENS-PSL-CNRS

# Analyse du sous-genre : Les romans d'aventures comme étude de cas

## Romans d'aventures vs Passages d'aventure

- « L'aventure est l'essence de la fiction » (Tadié, 1996) :
- Au-delà des étiquettes de sous-genre : Niveau du passage

## LLMs en tant que puissance d'annotation

- Les LLMs peuvent-ils détecter la stéréotypie dans les romans d'aventures ?
- Peut-on détecter le nombre de scènes d'aventures dans un grand corpus ?

# Stéréotypie dans les romans d'aventure

**Définition :**
« L'aventure est caractérisée par l'importance du changement de décor (historique/géographique/fantastique ou social) et de l'action violente mettant le héros en danger mortel ou en péril physique » (Letourneux, 2010)

**Détection de la stéréotypie**

- Travail au niveau du passage (4-5 pages)
- Quel type d'information est pertinent ? Suspense, Cadre spatiotemporel, Caractérisation ?

# Méthode

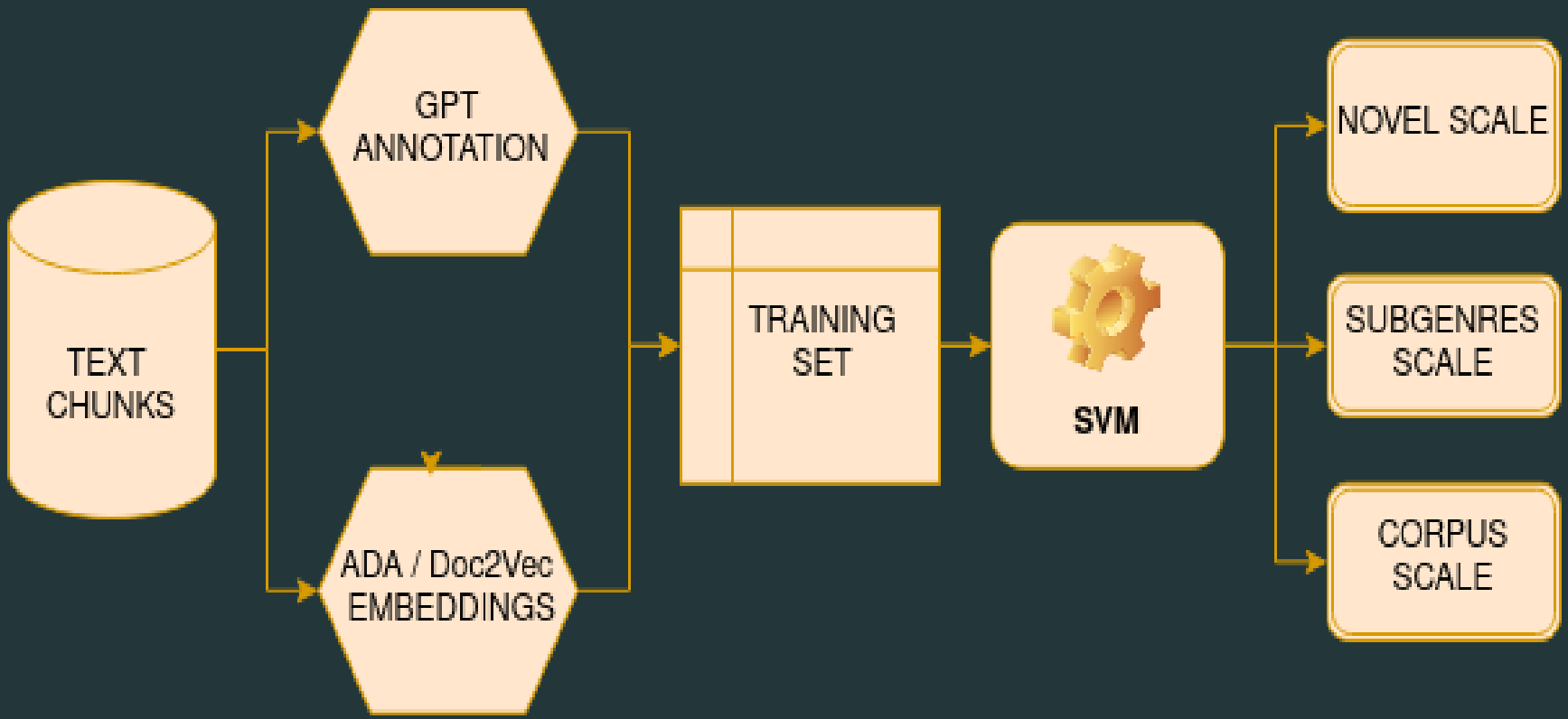


**Figure 1 :** Flux de travail

# Annotation : Ingénierie des prompts

**Prompt :** Donne-moi une sortie d'un mot : AVENTURE, si ce texte est typique du genre aventure, sinon écris NON_AVENTURE. Préfère la sortie NON_AVENTURE en cas de doute. AIDE : les romans d'aventures se caractérisent par l'importance du dépaysement (historique/géographique/fantastique ou social) et des actions violentes mettant le héros en danger mortel ou en péril physique. Une scène d'aventures typique consisterait en quelqu'un (décrit comme brave/héroïque) faisant quelque chose de dangereux de manière héroïque et dans un cadre sauvage.

**Annotation GPT 3.5 turbo comme vérité terrain**

- Annotation de 1000 exemples : aventures vs NON_aventures
- Évaluation de quelques exemples

# Pipeline d'entraînement

## Caractéristiques textuelles - 3 niveaux

- Tous les tokens des passages
- Tokens Fr-BookNLP (caractérisation + chronotope)
- Tokens aléatoires

## Embeddings

- Embeddings OpenAI
- Vecteurs de paragraphes

## Modélisation statistique

- SVM - État de l'art pour la classification de texte

# Résultats SVM

| | Tous les tokens | Tokens BookNLP | Tokens aléatoires |
|---|---|---|---|
| Embeddings ADA | **0.86** | 0.77 | 0.72 |
| DBoW | 0.78 | 0.69 | 0.63 |

**Table 1 :** Évaluation de référence

# Analyse Multiscale : Échelle du roman

## **Vingt mille lieues sous les mers** (Verne, 1869)

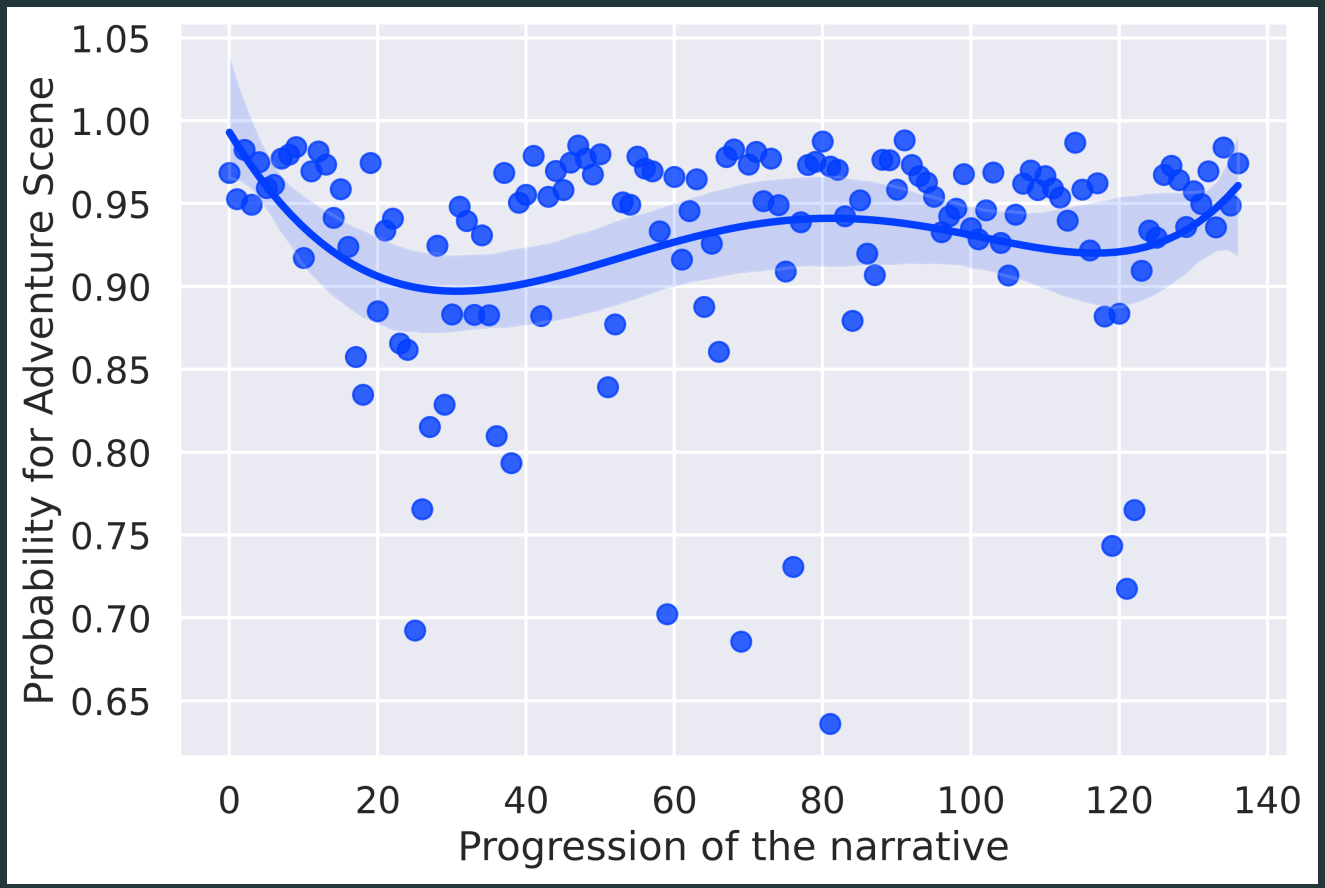


**Figure 2 :** Scènes d'aventures dans *Vingt mille lieues sous les mers*

# Analyse Multiscale : Échelle du roman

## l'Éducation Sentimentale (Flaubert, 1869)

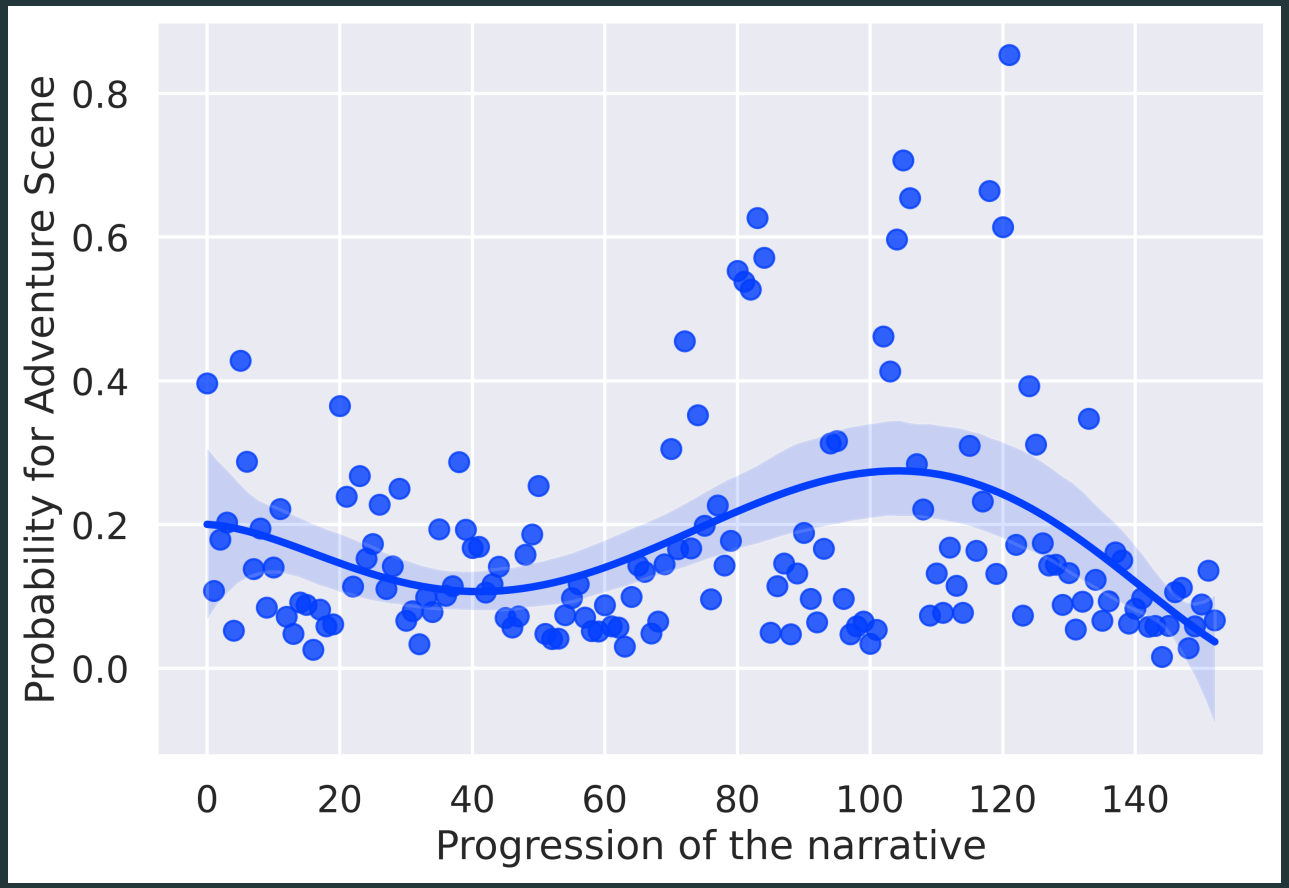


**Figure 3 :** Scènes d'aventures dans l'*Éducation Sentimentale*

# Analyse Multiscale : Échelle du sous-genre

## Étiquettes de sous-genre : Corpus des chapitres

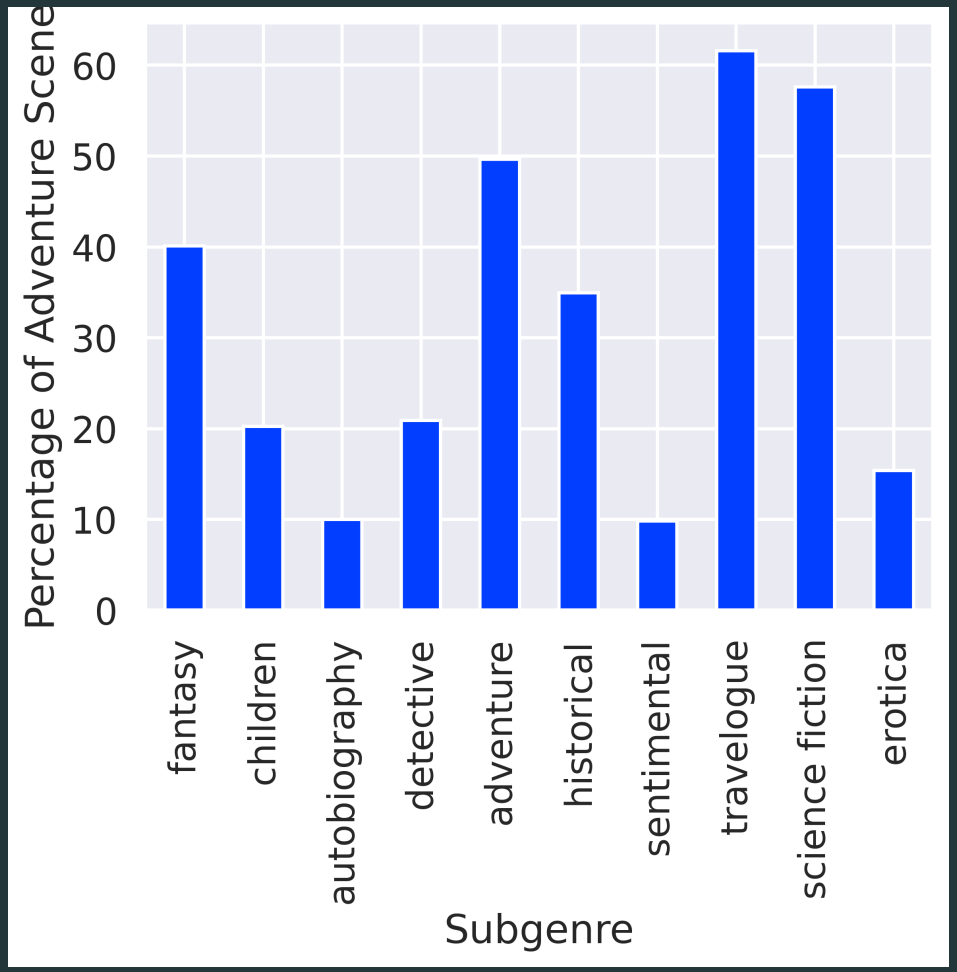


**Figure 4 :** Pourcentage de scènes d'aventurespar sous-genres

# Analyse Multiscale : Échelle du corpus

Corpus entier - 10 000 morceaux annotés à partir d'un corpus de 3000 romans

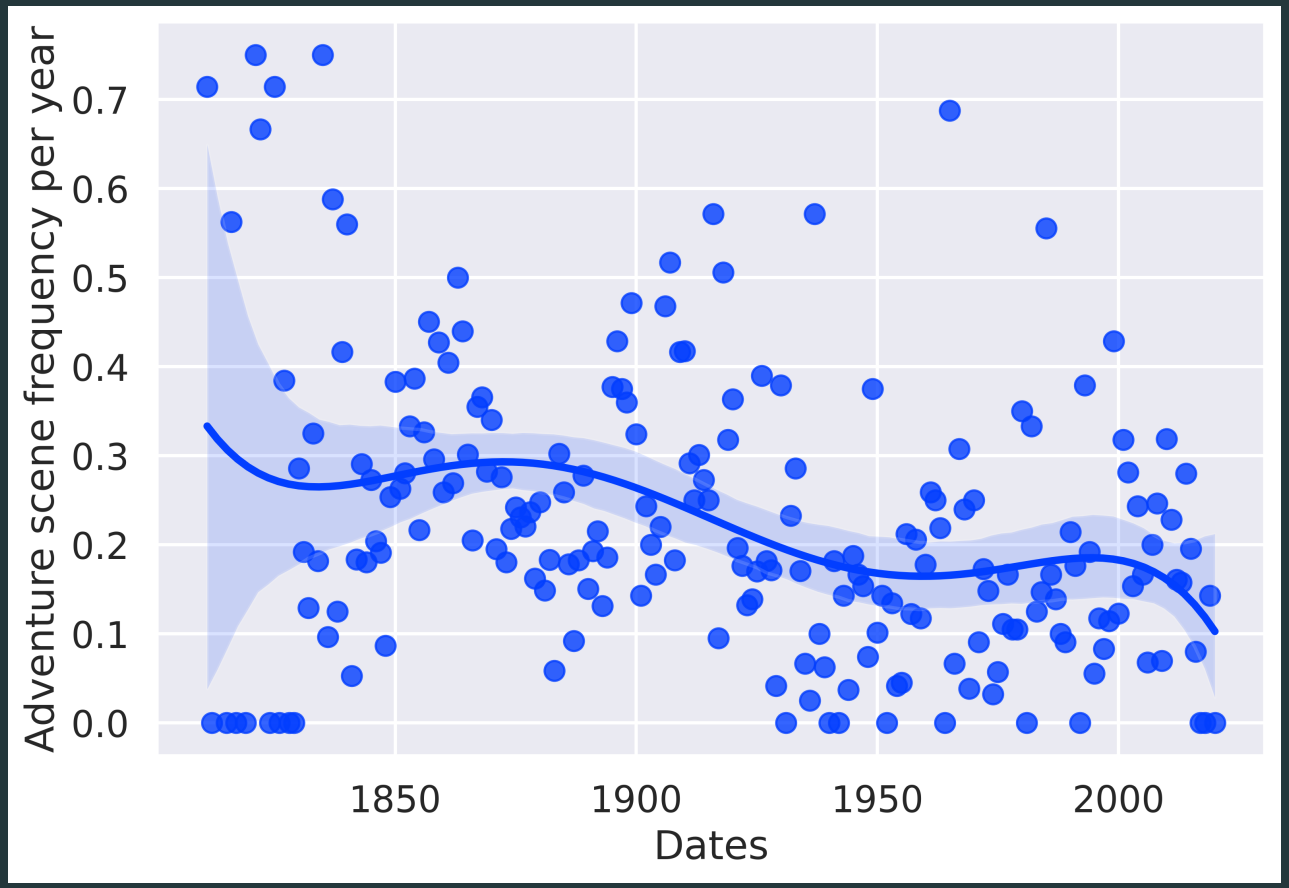


**Figure 5 :** Scènes d'aventures dans l'ensemble du corpus

# Conclusions et Perspectives

- Les LLMs comme puissance d'annotation : Lecture approfondie automatique ?
- Évaluation humaine de l'annotation synthétique
- Affiner un LLM "ouvert" (Llama2/3 ? Mistral/Hermes ?)
- Détecter la scène d'aventures stéréotypique pourrait être trop large : labels plus fins - Scène de combat, Scène d'amour, Scène de découverte, etc.
- LLMs : Nouvel outil pour la recherche en humanités numériques. Ils peuvent être utiles - mais peut-on leur faire confiance ? Et plus largement, avons-nous besoin d'eux ?

# Questions ?