Xiaohua CUI

Data Scientist NLP

Paris 75014 | +33 0766229976 | xiaohua.cui@outlook.com

xiaohua-cui Simounawind

COMPÉTENCES ET LANGUES

Expert en NLP/NLU
IA Générative: LLMOps, RAG, Agent, Prompt Engineering, Fine-Tuning, Langchain
R, SQL, Regex, XML, Python: Pandas, NumPy, Scikit-learn, SpaCy, NLTK, Gensim, Matplotlib, Seaborn
Machine Learning: Random Forest, XGBoost, SVM, PCA
Transformers & Pytorch
Azure, Databricks, Spark, Mlflow
Git, Shell, HTML5, CSS
Outils / IDE: Jupyter Notebook, VSCode SAS Viya, PowerBI, Excel

EXPÉRIENCE PROFESSIONNELLE

Data Scientist ML/GenAI — Paris, France

Groupe Covéa — 01/2025 - Présent

- Responsable du développement et de la mise en production des projets NLP et IA générative: anti-fraude, traitement et analyse des rapports / verbatims d'assurance, monitoring du modèle.

Stagiaire Data Scientist NLP — Paris, France

Groupe Covéa — 03/2024 - 09/2024

- PoC de solutions d'IA générative pour les tâches NLP classiques.
- Évaluation (coût et performance) des LLMs open-source et GPT pour la classification multi-label des messages clients sur Databricks.
- Expériences sur les paramètres (température, langue, top-p) avec prompt engineering, améliorant le F1-score de 0.6 à 0.79.
- Automatisation des expériences sur l'architecture LangChain et MLflow.
- Entraînement de modèles ML et fine-tuning de CamemBERT en comparaison. Résultats détaillés dans une série d'articles sur ***Medium***.

Stagiaire Data Science & Localisation — Pékin, Chine

Ourpalm — 04/2022 - 08/2022

- Annotation des messages client sur le système de boutique et développement d'un modèle RoBERTa en PyTorch pour l'analyse de sentiment en polarité, améliorant la précision jusqu'à 92%. Déploiement du modèle en collaboration avec l'équipe produit.

EXPÉRIENCE DE PROJET

Identification et classification des risques sur rapports d'expertise incendie

- Mise en place d'un pipeline d'extraction et de classification à partir de rapports d'expertise en PDF/JPG : OCR, indexation sémantique (FAISS), génération de prompts dynamiques et appel de LLM (LangChain + Pydantic).
- Analyse de plus de 50 000 sinistres via LLM pour identifier les causes liées aux batteries : typologie hiérarchique (type, cause, objet), consolidation temporelle, visualisation de tendances émergentes.

IA Frugale pour les LLMs français avec l'élagage et le PEFT

- Élagage au niveau des couches des modèles Transformer français (CamemBERT, FlauBERT), combiné avec LoRA pour réduire les coûts.
- Évaluation de la performance des modèles élagués sur des tâches sémantiques, de POS tagging et de NER : l'intensité d'élagage et la solution optimale peuvent être ajustées dynamiquement selon la nature des tâches.

FORMATION

Master Traitement automatique des langues — 09/2022 - 09/2024
Sorbonne Nouvelle — Paris

Master Linguistique Analyse des discours — 09/2021 - 06/2024
Université des langues et cultures de Pékin — Pékin, Chine

Licence Langue et littérature françaises — 09/2017 - 06/2021
Université Yanshan — Qinghuangdao, Chine