Antoine Chaffin

NohTow | Antoine Chaffin | Antoine Chaffin | antoine.chaffin.fr | antoine@chaffin.fr

Summary

I obtained an engineering degree at INSA Rennes, where I learned solid coding practices and spent six months at Polytechnique Montreal. I also completed a research-focused master's degree during my final year.
For my end-of-study internship at IRISA and throughout my Ph.D., I explored multimodality and semantic indexation, working to bridge the gap between textual and visual data for misinformation detection. My second main focus was on text generation, especially cooperative generation guided by external models. I also investigated reinforcement learning for training language models, integrating it into cooperative generation and studying multimodal rewards.
After my Ph.D., I joined LightOn, where I work on retrieval and explore encoders, late-interaction, and multimodal document retrieval.

Work Experience

R&D Machine Learning Engineer - LightOn February 2024 - Today

Research and development to improve the retrieval augmented generation product (**Paradigm**)

Visiting Researcher - Sorbonne University December 2021

One month internship in the **MLIA** team to collaborate on **cooperative generation**

Teaching Assistant - Rennes University 2020-2021

- Teaching **data analysis** to undergraduate students (**ISTIC**)
- Teaching **machine learning** to engineer students (**ESIR**)

Trainee - IRISA February - July 2020

End-of-studies research internship on the subject of **image repurposing detection using multimodal models**

Trainee - Them-is

- Realization of a web project for the **Basel-Mulhouse airport** June - July 2019
- Web development of the **firm's portfolio** July - August 2018

Education

IRISA, IMATAG — **Industrial Ph.D.** in Artificial Intelligence (Multimodal Misinformation Detection) – **ATALA 2024 Thesis Award** 2020-2023

UNIVERSITY OF RENNES 1 — **M.Sc. degree** in Research in Computer Science – **highest honors** 2019-2020

POLYTECHNIQUE MONTRÉAL — **International exchange** 2019

INSA RENNES — **Engineering degree** 2015-2020

Code

ModernBERT 1.5k⋆ GitHub

- Modernize the pre-training of encoder-only models using data scaling and architecture from recent Large Language Models

PyLate 589⋆ GitHub

– Built on top of Sentence Transformers, PyLate is designed to simplify and optimize training, inference, and retrieval with state-of-the-art late interaction models

PPL-MCTS & Which Discriminator 66⋆ GitHub

– Guiding language model to generate text matching an external constraint using MCTS

– Speed-up cooperative generation using unidirectional discriminator to leverage caching

Therapy 1⋆ GitHub

– Generate global explanations of a discriminator without requiring data using cooperative generation

WTF-RL 11⋆ GitHub

– Explore ground truth captions can be leveraged to train image captioning models using cross-modal rewards in a reinforcement learning training scheme

Models

ModernBERT Hugging Face

– Modernized encoders trained on 2T tokens achieving SOTA performance on classification and retrieval while being extremely efficient

ModernBERT-embed-large Hugging Face

– Fine-tuning of ModernBERT-large for dense retrieval

Reason-ModernColBERT Hugging Face

– Late interaction model trained for reasoning-intensive retrieval, outperforming model 45 times bigger

GTE-ModernColBERT Hugging Face

– State-of-the-art late interaction model trained using knowledge distillation

MonoQwen2-VL-v0.1 Hugging Face

– Fine-tuning of Qwen2-VL for visual document reranking using the MonoT5 approach

BioClinical-ModernBERT Hugging Face

– Continued pre-training of ModernBERT on biomedical and clinical data, achieving SOTA results in various tasks

Mambaoutai Hugging Face

– 1.6B mamba model trained on 300B tokens with exploration on enabling continued pre-training by the community

Publications

Antoine Chaffin, Vincent Claveau, and Ewa Kijak (2022). "PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* NAACL 2022.

Antoine Chaffin, Thomas Scialom, Sylvain Lamprier, Jacopo Staiano, Benjamin Piwowarski, Ewa Kijak, and Vincent Claveau (2022). "Which Discriminator for Cooperative Text Generation?" In: *Proceedings of the 45th International ACM Conference on Research and Development in Information Retrieval,* SIGIR 2022.

Vincent Claveau, **Antoine Chaffin**, and Ewa Kijak (2022). "Generating Artificial Texts as Substitution or Complement of Training Data". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* LREC 2022.

Sylvain Lamprier, Thomas Scialom, **Antoine Chaffin**, Vincent Claveau, Ewa Kijak, Jacopo Staiano, and Benjamin Piwowarski (2022). "Generative Cooperative Networks for Natural Language Generation". In: *Proceedings of the 2022 International Conference on Machine Learning,* ICML 2022.

Victor Sanh et al. (2022). "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *Proceedings of The Tenth International Conference on Learning Representations,* ICLR 2022.

Teven Le Scao et al. (2022). "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model". In: arXiv: 2211.05100.

Antoine Chaffin and Julien Delaunay (2023). ""Honey, Tell Me What's Wrong", Global Explanation of Textual Discriminative Models through Cooperative Generation". In: *Proceedings of the Sixth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP,* BlackboxNLP@EMNLP 2023.

Pierre Fernandez, **Antoine Chaffin**, Karim Tit, Vivien Chappelier, and Teddy Furon (2023). "Three Bricks to Consolidate Watermarks for Large Language Models". In: *Proceedings of the IEEE International Workshop on Information Forensics and Security,* WIFS 2023. **Best student paper award**.

Antoine Chaffin, Ewa Kijak, and Vincent Claveau (2024). "Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning". In: *Proceedings of The 2024 IEEE International Conference on Image Processing,* ICIP 2024.

Benjamin Clavié, **Antoine Chaffin**, and Griffin Adams (2024). "Reducing the Footprint of Multi-Vector Retrieval with Minimal Performance Impact via Token Pooling". In: arXiv: 2409.14683.

Benjamin Warner, **Antoine Chaffin**, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli (2024). "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference". In: arXiv: 2412.13663.

Antoine Chaffin and Raphaël Sourty (2025). "PyLate: Flexible Training and Retrieval for Late Interaction Models". In: *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. CIKM '25. Association for Computing Machinery.

Gautier Evennou, **Antoine Chaffin**, Vivien Chappelier, and Ewa Kijak (2025). "Reframing Image Difference Captioning with BLIP2IDC and Synthetic Augmentation". In: *Proceedings of the 2025 Winter Conference on Applications of Computer Vision,* WACV 2025.

Thomas Sounack, Joshua Davis, Brigitte N. Durieux, **Antoine Chaffin**, Tom J. Pollard, Eric Lehman, Alistair E. W. Johnson, Matthew B. A. McDermott, Tristan Naumann, and Charlotta Lindvall (2025). "BioClinical ModernBERT: A State-of-the-Art Long-Context Encoder for Biomedical and Clinical NLP". In: arXiv: 2506.10896.