4EU+ Doctoral research project:

Studying the efficiency of large language models in the reproduction of literary style to further characterize AI-written texts

The recent surge in popularity of generative AI models in the public sphere has brought to light several ethical and legal considerations regarding the epistemological status of the texts produced by such models, and the various circumstances in which one may exploit them (Bostrom, 2018). Unethical and unsupervised use of AI can lead to issues such as plagiarism, inaccurate and falsified information in official documents, as well as more general deontological negligence from private individuals and institutions alike. Therefore, it has become increasingly vital to provide the public with reliable tools that can detect the presence of AI tampering in documents and indicate the degree to which such transformations or so-called 'hallucinations' took place, so that the necessary precautions and fail-safes may be identified and implemented where needed. Several initiatives have already tackled the issue (Batane, 2010) – but so far with mixed results (AlAfnan et al., 2023).

This project proposes to take an active part in said effort by studying the literary style of text-generating AI, to try and identify features that may allow to differentiate between human-generated and AI-generated writings. This research avenue has already been somewhat explored in previous work (Canabac et al., 2021; AlAfnan et al., 2023) and shows promising results. However, on the one hand, these past efforts often limit their focus to empirical signs (Canabac et al., 2021) that are difficult to define from a formal perspective and prove equally difficult to apply systematically to AI-generated texts; on the other hand, past research efforts deal exclusively with non-literary writing formats, such as case studies, business correspondence, and academic writing. While not without their stylistic particularities, these formats are by design governed by objective considerations that restrain human expression through non-literary expression and communication. As such, they are especially vulnerable to reproduction from generative AIs: for example, OpenAI uses almost exclusively academic benchmarks to evaluate GPT-4's performance, boasting an efficacy rate of over 86% in six out of seven tests, and being state-of-the-art in the seventh test with a 67% rate (OpenAI, 2023).

Another challenge this project will tackle stems from the very nature of literary 'style' itself, the systematic analysis of which is often hindered by its elusive nature: although it is generally defined as an individual author's particularities in their use of language (Buffon, 1872), there are very few instances where selected stylistic markers are truly consistent (Genette, 1991) from one author to another, even less so from one reader to another, when interpreting stylistic expression. Computationally speaking, an author's 'style', or idiolect, can be detected using a stylometric approach: it has proven efficient in author-recognition tasks (Savoy, 2018; Kestemont et al., 2016) and rests upon a statistical analysis of the components of a text that is not unlike the process through which an AI learns to reproduce human writing styles. Although this approach does manage to perform a correct categorization on Tweet-sized texts that are AI-generated, past research shows that the efficiency of stylometry greatly decreases as the generator size – the number of parameters and thus the complexity of the AI model – increases (Kumarage et al., 2023). The stylometric approach to the automatic recognition task of AI-written texts struggles to keep up with the evolving industry of generative AI.

We believe that new innovative venues can be identified by studying the style of generative AI in a different corpus than those previously explored, i.e., through literary texts where personal style tends to be more visible and pronounced than in academic and business writing. More specifically, this project will aim to use the literary exercise of pastiche (Austin, 2013; Aron, 2013), a term which describes any work of art that aims to emulate the style of a specific artist, to try and understand better how an AI approach to the notion of style reproduction is compared to that of a human writer. By confronting the statistical/predictive approach of generative AI against the more sensitive and personal approach adopted by a human when asked to reproduce the style of a specific author, we hope to highlight the existence of undiscovered features that could help isolate AI-generated texts in a mixed corpus. By studying the style of AI in a literary exercise where it may underperform (Gunser et al., 2022), we anticipate we will be able to transfer our findings towards more academic exercises at which AI does excel.

These advances will be accomplished by using existing human pastiches whose literary quality has already been established (for example: the pastiches of La Fontaine, Proust and Reboux, among others), and then prompting generative AI models to create several pastiches imitating the same works as their human counterparts. This process will thus produce two corpora of a similar size that will then be explored and compared both empirically and stylometrically. Here, the challenge of scientific reproducibility will necessarily come to the fore, as AI models by their nature tend to generate different output even with identical prompting. We will thus aim to explore several different LLMs using a reproducible pipeline to replicate our analyses. But, given the accelerated nature of AI research, we are open to exploring new models and approaches should parallel research efforts come to light or the results achieved during the first phase of the project move us in that direction.

We are convinced that this project closely follows the guidelines of the third flagship of the 4EU+ Alliance, Digitisation – Modelling – Transformation, as it aims to contribute to the topical subject of AI ethics, epistemology, and digitised text analysis and exploitation through a cross-disciplinary study that tackles machine learning under the lens of literary creation and genetic criticism. As such, the profile of the future doctoral student should include a background in machine learning and an understanding of state-of-the-art large language models, as well as a background in literary studies, stylistics, and/or linguistics; ensuring that they may be able to grasp and model the underlying intertextual mechanisms behind style imitation and be prepared to apply them to computational language models.

Preliminary bibliography:

AlAfnan, Mohammad Awad et al. "Artificial Intelligence Chatbots Have a Writing Style? An Investigation into the Stylistic Features of ChatGPT-4", *Journal of Artificial Intelligence and Technology*, Vol. 3, 2023, p. 85-94.

Aron, Paul. "Le pastiche comme objet d'étude littéraire. Quelques réflexions sur l'histoire du genre", *Modèles linguistiques*, No. 60, 2009, p.11-27.

Austin, James F. *Proust, Pastiche, and the Postmodern or Why Style Matters*, Bucknell University Press, 2013.

Batane, Tshepo. "Turning to Turnitin to Fight Plagiarism among University Students", *Educational Technology & Society*, Vol. 13, No. 2, 2010, pp. 1-12.

Bostrom, Nick et al. "The Ethics of Artificial Intelligence", *Artificial Intelligence Safety and Security*, Chapman and Hall/CRC, p.13, 2018.

(de) Buffon, Georges-Louis Leclerc. *Discours sur le style*, 1753. Paris, Lecoffre fils, 1872.

Cabanac, Guillaume et al. *Tortured phrases: A dubious writing style emerging in science, Evidence of critical issues affecting established journals*, 2021.

Genette, Gérard. *Fiction et diction*, Paris, Seuil, collection "Poétique", 1991.

Gunser, Vivian Emily et al. "The Pure Poet: How Good is the Subjective Credibility and Stylistic Quality of Literary Short Texts Written with an Artificial Intelligence Tool as Compared to Texts Written by Human Authors?", *Proceedings of the Annual Meeting of the Cognitive Science Society*, No. 44, 2022, p.60-61.

Kestemont, Mike et al. "Authenticating the writings of Julius Caesar", *Expert Systems with Applications*, No. 63, 2016, p.86-96.

Kumarage, Tharindu et al. *Stylometric Detection of AI-Generated Text in Twitter Timelines*, 2023.

OpenAI. *GPT-4 Technical Report*, 2023.

Savoy, Jacques. *Machine learning methods for stylometry: authorship attribution and author profiling*, Springer, 2020.