Pseudollm: A multilingual LLM-enhanced tool for context-aware text pseudonymization

Arianna Bienati[1,2]

[1]University of Modena and Reggio Emilia (Italy); [2]Eurac Research, Bolzano (Italy)

Rationale

A small corpus of cover letters in English, Italian and German, annotated for quality ratings

The rationale that underpins this work was the need to carefully and fully pseudonymize a multilingual corpus of cover letters enhanced by **quality ratings**, collected through comparative judgments (CJ). To disclose authentic and **readable** cover letters with CJ judges, while maintaining participants' **confidentiality,** a full **pseudonymization** was needed.

The present of corpus linguistics: findable, accessible, interoperable, reusable resources

One important pillar of the project, was to FAIR-ify the corpus, allowing **data sharing** with other researchers and **reproducibility** of the results. A prerequisite of FAIR resources is to effectively anonymize metadata and pseudonymize the texts making up the corpus. This step is particularly necessary when dealing with spontaneous **speech** transcripts, **social media** data and **learner corpora**.

Theoretical framework

Personal data and the GDPR

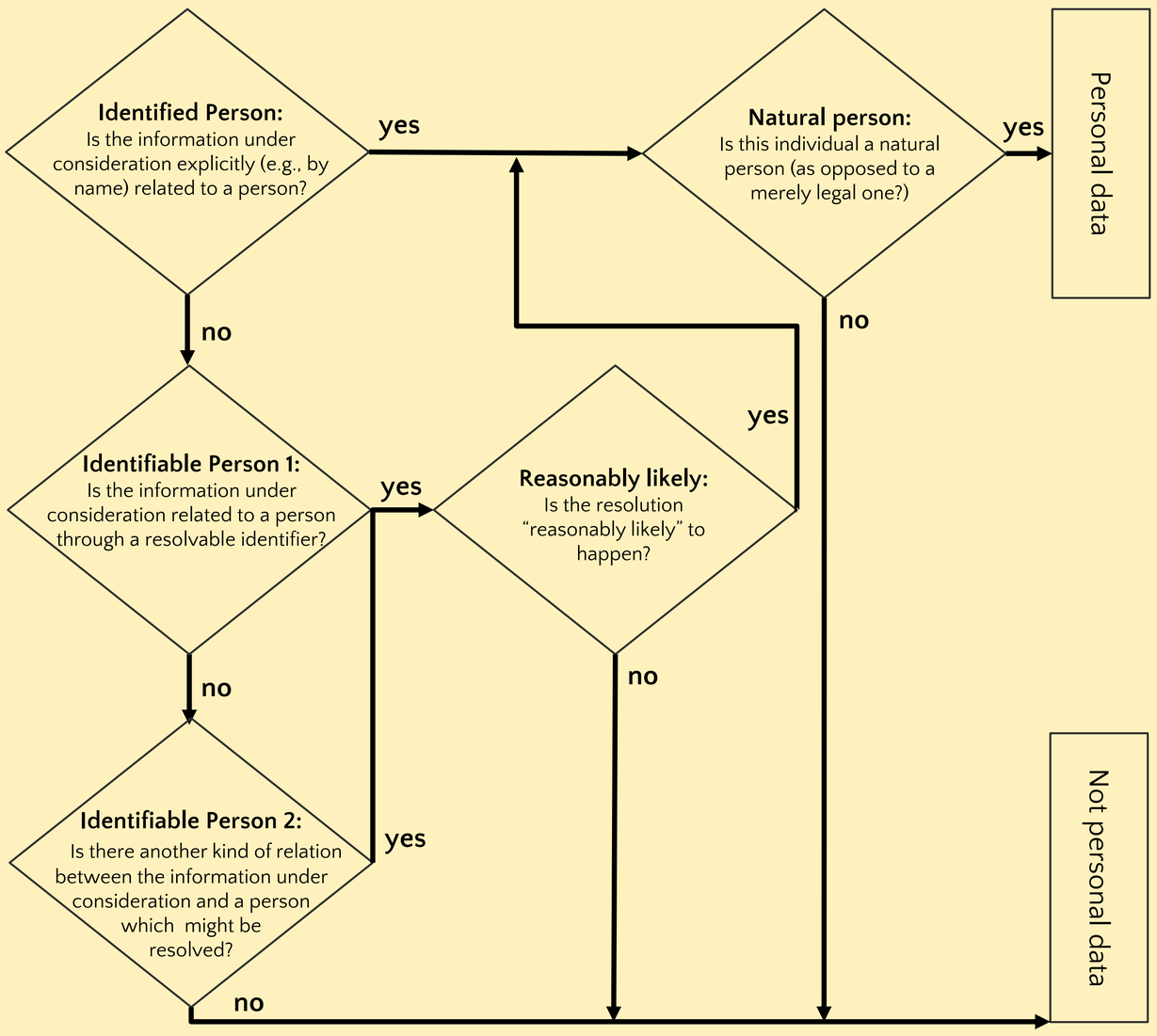
'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a **name**, an **identification number**, **location data**, an **online identifier** or to one or more **factors** specific to the **physical**, **physiological**, **genetic**, **mental**, **economic**, **cultural** or **social identity** of that natural person.

GDPR, Article 4

For example, in a cover letter you will find a narrative supporting the *curriculum vitae* of a person, thus academic institutions, workplaces, experiences abroad, specific research topics will be mentioned. This information, **taken together**, uniquely identify an individual. Therefore, they must be de-identified before disclosure. Even though sometimes this information is already public (e.g., published in bios or websites), we cannot leave the possibility of re-identification open through the sharing of the corpus.

Annotating personal data

Fink and Pallas (2020) systematize how to determine whether an information is personal data or not for each information in a dataset or unstructured data (like texts).



A concrete example (with author's consent)

*This working experience and an excursion to the **Vjosa** and **Vjosa** headwaters (**Albania** and **Greece**), led me to the decision to do my Master's thesis on the almost extinct and very special mayfly* **Prosopistoma pennigerum** *in the **Vjosa** and surrounding rivers.*

→ Candidate's identity is "one Google search away": the Master's thesis subject (***Prosopistoma pennigerum***) uniquely identifies who wrote the letter, but commercial industry-grade tools do not identify it!

Presidio (Microsoft)**:**

This working experience and an excursion to the **<LOCATION>** and **<LOCATION>** headwaters (**<LOCATION>** and **<LOCATION>**), led me to the decision to do my Master's thesis on the almost extinct and very special mayfly ***Prosopistoma pennigerum*** in the **<LOCATION>** and surrounding rivers.

Pseudollm:

This working experience and an excursion to the **Struma** and **Struma** headwaters (**Macedonia** and **Albania**), led me to the decision to do my Master's thesis on the almost extinct and very special mayfly ***Ephemera longicaudata*** in the **Struma** and surrounding rivers.

Related work

Learner corpora pseudonymization (Megyesi et al. 2018; Volodina et al. 2020): pseudonymization categories divided in **hard replacement** (e.g., dates, email), **placeholder** (geographical data, institutions) and **sensitive markup** (education, profession), depending on the degree to which these data points are structurally predictable.

Category	Replacement	Technique
Numbers, email, URLs	Standardized representation	Regex
Geo-data, institutions	Placeholder	Fixed word ontologies
Education, profession, ethical, political and religious views	Marked, but not replaced	Manual annotation

Staab et al. (2025): LLMs are very capable of inferring personal data from online texts. The framework proposed is **anonymization under adversarial inference**: "leverage the strong attribute inference capabilities of LLMs to inform a separate anonymizer language model". Extensive experimental evaluation on 13 LLMs and several baselines, including human evaluation.

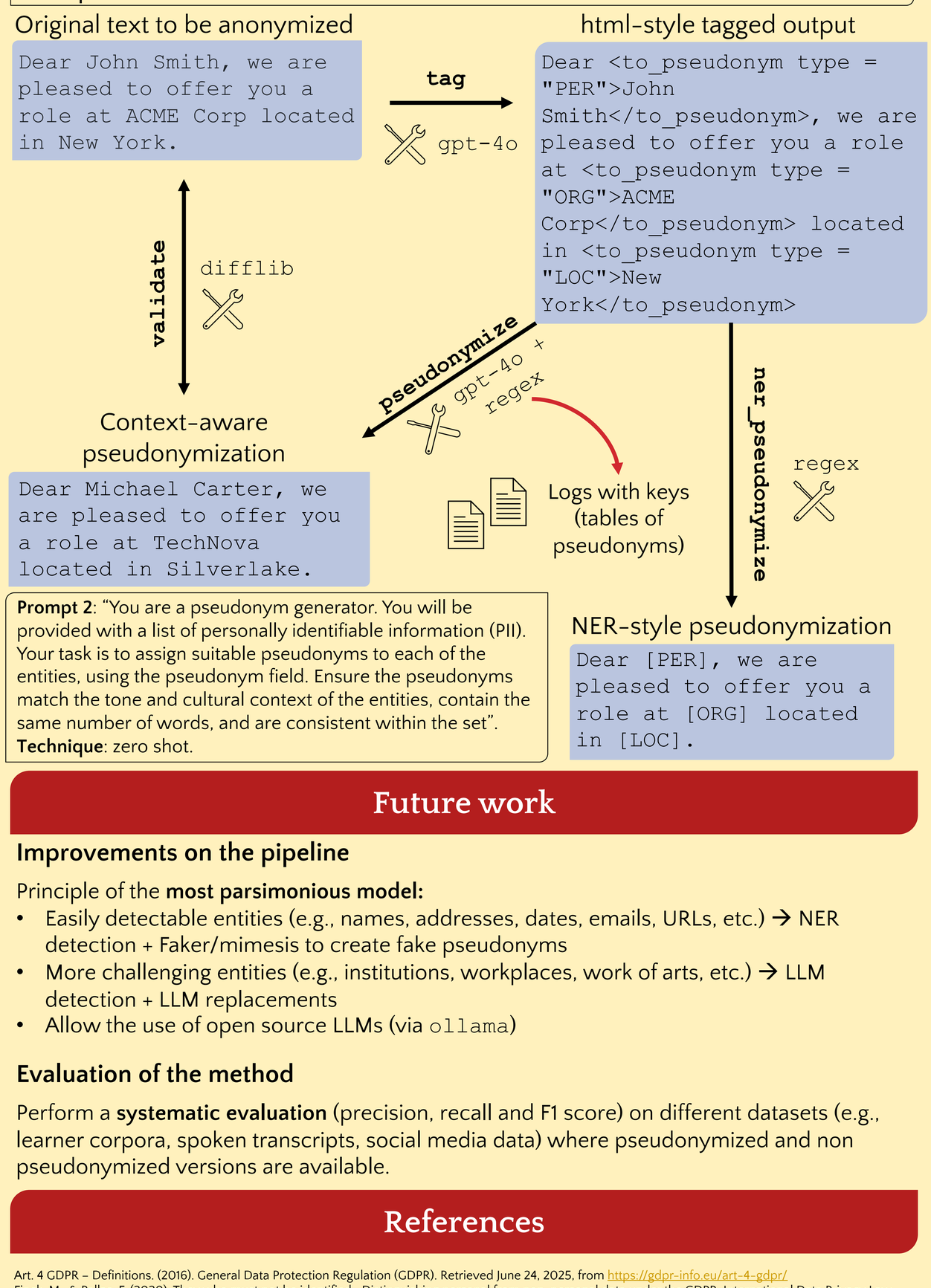
Related software / packages:

- Azure Language Studio (NER-style, proprietary)
- Presidio (NER-style, open source at https://github.com/microsoft/presidio)
- Faker (context-aware replacements, open source at https://github.com/joke2k/faker)
- Mimesis (context-aware replacements, https://github.com/lk-geimfari/mimesis)

Pseudollm pipeline*

* The code is released under the MIT License and is available at https://github.com/arianna-bienati/pseudollm

Prompt 1: "Annotate all Personally Identifiable Information in the following text (e.g., names, places, organizations, project names, etc.). Use the tags <to_pseudonym type = 'value'> </to_pseudonym> to tag them. Use the 'type' attribute to detail which kind of PII it is. You can choose between four types: PER for persons, LOC for locations, ORG for organizations and MISC for anything else. Just output the tagged text, without any further comment. Do not change the original text".
Technique: few shot.



Prompt 2: "You are a pseudonym generator. You will be provided with a list of personally identifiable information (PII). Your task is to assign suitable pseudonyms to each of the entities, using the pseudonym field. Ensure the pseudonyms match the tone and cultural context of the entities, contain the same number of words, and are consistent within the set".
Technique: zero shot.

Future work

Improvements on the pipeline

Principle of the **most parsimonious model:**

- Easily detectable entities (e.g., names, addresses, dates, emails, URLs, etc.) → NER detection + Faker/mimesis to create fake pseudonyms
- More challenging entities (e.g., institutions, workplaces, work of arts, etc.) → LLM detection + LLM replacements
- Allow the use of open source LLMs (via `ollama`)

Evaluation of the method

Perform a **systematic evaluation** (precision, recall and F1 score) on different datasets (e.g., learner corpora, spoken transcripts, social media data) where pseudonymized and non pseudonymized versions are available.

References

Art. 4 GDPR – Definitions. (2016). General Data Protection Regulation (GDPR). Retrieved June 24, 2025, from https://gdpr-info.eu/art-4-gdpr/

Finck, M., & Pallas, F. (2020). They who must not be identified—Distinguishing personal from non-personal data under the GDPR. International Data Privacy Law, 10(1), 11–36. https://doi.org/10.1093/idpl/ipz026

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M., & Volodina, E. (2018). Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. In I. Pilán, E. Volodina, D. Alfter, & L. Borin (Eds.), Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning (pp. 47–56). LiU Electronic Press. https://aclanthology.org/W18-7106

Staab, R., Vero, M., Balunović, M., & Vechev, M. (2025). Large Language Models are Advanced Anonymizers (arXiv:2402.13846). arXiv. https://doi.org/10.48550/arXiv.2402.13846

Volodina, E., Ali Mohammed, Y., Derbring, S., Matsson, A., & Megyesi, B. (2020). Towards Privacy by Design in Learner Corpora Research: A Case of On-the-fly Pseudonymization of Swedish Learner Essays. In D. Scott, N. Bel, & C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics (pp. 357–369). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.32