

SAFEWORDS: Language Anonymization with Ethical and Legal Safeguards using NLP within the HumanAlze framework

Rafael Muñoz^{1,*†}, Manuel Palomar^{1†}, Elena Lloret^{1†} and Iván Martínez-Murillo^{1†}

¹Centro de Inteligencia Digital, Universidad de Alicante

Abstract

This article describes the subproject **SAFEWORDS** (*Language Anonymization with Ethical and Legal Safeguards through NLP*), part of the coordinated project **HumanAlze** (*Humanizing AI with Language Technology*), funded by the Spanish Ministry of Science, Innovation and Universities under the 2025 call for projects in the field of Artificial Intelligence. HumanAlze brings together six Spanish research centers with the shared goal of humanizing Large Language Models (LLMs) by integrating linguistic, legal, and ethical knowledge into their development cycle. SAFEWORDS acts as a cross-cutting layer of the entire project, addressing two major challenges: (1) robust anonymization of natural language texts to ensure compliance with the GDPR and the EU AI Act, and (2) the definition of an ethical, legal, and data governance framework that underpins the rest of the subprojects. The proposed methodology combines multilingual detection of personally identifiable information (PII), differential privacy, synthetic substitute generation, and cryptographic traceability, along with evaluations of re-identification risk and preservation of linguistic utility for training LLMs. The experiments are validated on legal and administrative texts in Spain's co-official languages.

Keywords

text anonymization, AI regulations, ethical safeguards, HumanAlze

1. Introduction

The emergence of Large Language Models (LLMs) has radically transformed Natural Language Processing (NLP), but their widespread deployment has brought unprecedented challenges in terms of privacy, bias, traceability, and regulatory compliance [1]. The General Data Protection Regulation (GDPR) [2] and the EU Artificial Intelligence Act [3] now impose specific obligations on the processing of personal data, both in the data collection and training phases as well as during inference, with full applicability starting in August 2026.

Far from being purely technical issues, adapting LLMs to this regulatory framework requires close collaboration among computer scientists, linguists, and legal scholars. It is precisely this synergy that underpins the coordinated project **HumanAlze**, within which the subproject **SAFEWORDS** (*Language Anonymization with Ethical and Legal Safeguards through NLP*) is developed. Led by the GPLSI – CENID group at the University of Alicante (UA), SAFEWORDS pursues a dual objective: to design and implement a robust anonymization pipeline for multilingual textual corpora, and to develop the ethical, legal, and data governance framework that operates as a cross-cutting layer across the entire HumanAlze project.

The article is structured as follows: Section 2 describes the HumanAlze project and situates SAFEWORDS within its structure; Section 3 analyzes the regulatory context; Section 4 details the objectives and technical architecture; Section 5 develops the methodology; Section 6 explains the integration of

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

*Corresponding author.

†These authors contributed equally.

✉ rafael@dlsi.ua.es (R. Muñoz); mpalomar@dlsi.ua.es (M. Palomar); elloret@dlsi.ua.es (E. Lloret);
ivan.martinezmurillo@ua.es (I. Martínez-Murillo)

ORCID 0000-0001-8127-9012 (R. Muñoz); 0000-0002-1441-7865 (M. Palomar); 0000-0002-2926-294X (E. Lloret);
0009-0007-5684-0083 (I. Martínez-Murillo)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

SAFEWORDS with the other subprojects; Section 7 presents the main use case; and Section 8 provides the conclusions.

2. The HumanAIze project

This section presents the organization of the HumanAIze project, detailing its subprojects, work packages, and the role of SAFEWORDS in ensuring ethical, legal, and privacy-aware AI development.

2.1. Overview

HumanAIze (*Humanizing AI with Language Technology*) is a coordinated research project funded by the Spanish Ministry of Science, Innovation and Universities under the 2025 Call for Projects in the Field of Artificial Intelligence, with a duration of 36 months and a total budget of approximately 1,994 M [4].

Its overall objective is to *push the research frontier in LLMs*, delivering disruptive improvements in factuality, fairness, reliability, cultural awareness, multilingualism, sustainability, privacy, and evaluation, with the aim of developing a multilingual, open, made-in-Spain alternative to large commercial LLMs such as GPT or Gemini [5]. The project adopts the concept of *trustworthy-by-design*: rather than adding ethical and legal considerations as a post-processing step, these are integrated from the earliest stages of data acquisition and model design [6].

2.2. Subprojects and complementarity

HumanAIze brings together six Spanish research centers in a virtual laboratory with decades of prior collaboration. The project is organized into six specialized subprojects, each led by a different center and addressing a key aspect of trustworthy, multilingual, and multimodal AI:

SP1 – MultiReason (UPV/EHU): extension of LLMs to new languages, modalities, and reasoning capabilities; with special attention to Basque [7].

SP2 – AlignedLLM (USC): faithful and transparent alignment of LLMs through RLHF, DPO, and the generation of verifiable responses with citations [5].

SP3 – SAFEWORDS (UA): language anonymization with ethical and legal safeguards; cross-cutting ethical framework (this work).

SP4 – VERITAS-H (UJA): trust in LLMs for the biomedical domain [8].

SP5 – TRAIL-LLMs (UNED): robust and linguistically informed evaluation, from benchmarks to real-world applications.

SP6 – TRUSTMODAL (BSC): trustworthy large-scale multimodal models [9].

2.3. Work packages and SAFEWORDS contribution

The subprojects are organized into seven work packages (WPs). Table 1 summarizes the specific contribution of SAFEWORDS, and Figure 2.3 shows the relationships among the work packages.

3. Regulatory context and motivation

Since August 2024, the EU AI Act [3] complements the GDPR with graduated obligations for providers of “general-purpose AI models” and prohibits the indiscriminate collection of biometric or other sensitive data from the web. The first phase of implementation entered into force on February 2, 2025, while full applicability begins in August 2026.

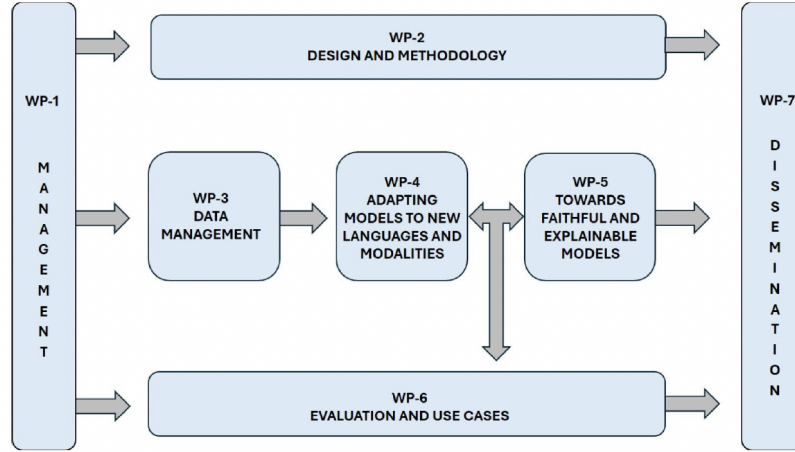


Figure 1: Architecture of the work packages

Table 1

Contribution of SAFEWORDS to the HumanAlze work packages.

WP	Task	SAFEWORDS Contribution
WP2	T2.3	Ethical, legal, and data protection framework
WP2	T2.4	Data governance framework
WP3	T3.1, T3.3	Multilingual repository of data and instructions
WP3	T3.4	Corpus anonymization pipeline
WP3	T3.5	Fairness and intersectionality suite
WP4	T4.3, T4.4	Anonymized corpora for extension to new languages
WP5	T5.2	Ethical and legal guardrails in aligned models
WP6	T6.4	Use case B: legal-administrative domain
WP7	T7.1–T7.5	Dissemination of methodology and open resources

State-of-the-art LLMs present specific vulnerabilities under this framework: it has been shown that they memorize fragments of their training data [10], which may lead to the unintended disclosure of personal information under adversarial queries. Tools such as *presidio* [11] or the OSCAR filters [12] are insufficient because they address entity detection in isolation, without simultaneously meeting the three requirements of the new framework: (i) effective protection against re-identification attacks [13]; (ii) cryptographic traceability of each transformation; and (iii) preservation of linguistic utility for training in low-resource languages. SAFEWORDS addresses this gap with a *privacy-by-design* methodology integrated into the full lifecycle of HumanAlze LLMs.

4. Objectives and technical architecture

The specific objectives of SAFEWORDS are:

1. **Ethical and data protection framework (T2.3):** to codify principles—privacy, non-discrimination, human oversight, sustainability—into operational checklists applicable to all phases of the project.
2. **Multilingual repository (T3.1, T3.3):** a versioned repository of corpora and instructions covering the official languages of Spain under FAIR principles [14].
3. **Anonymization (T3.4):** corpora free of PII that preserve utility for training LLMs.
4. **Fairness (T3.5):** a suite of metrics to audit and mitigate demographic and intersectional biases in data and models.

5. **Use case B (T6.4):** demonstrated anonymization in real legal and administrative documents.

Figure 4 shows the proposed anonymization pipeline, composed of four cascading modules.

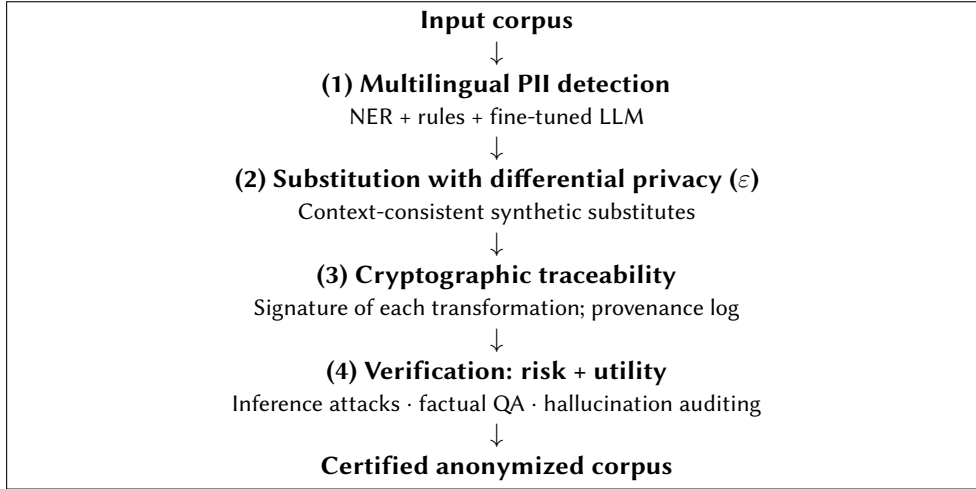


Figure 2: SAFEWORDS anonymization pipeline.

5. Methodology

The SAFEWORDS methodology follows an iterative and modular approach: each module in the pipeline can be improved independently without breaking the interface contracts with the rest. All experiments will be evaluated using the standard train/dev/test splits in NLP and reported following ACL reproducibility recommendations [15].

5.1. Multilingual PII detection

The detection of personally identifiable information (PII) constitutes the first link in the pipeline and largely determines the quality of the entire downstream process. We adopt a two-level entity taxonomy:

- **Direct identifiers:** first and last name, national ID (NIF/NIE/passport), Social Security number, email address, phone number, postal address, IBAN, and vehicle license plate.
- **Quasi-identifiers or contextual identifiers:** date of birth, municipality of residence, profession, medical diagnosis, legal status, or sexual orientation, which do not identify a person on their own but can do so when combined with other attributes [16].

For detection, we employ a three-layer hybrid approach:

1. **Rule-based and regular expression layer:** covers high-precision structured patterns (NIF, IBAN, postal codes, email addresses). This layer is deterministic, computationally inexpensive, and operates across all languages without requiring training.
2. **Supervised NER layer:** named entity recognition models fine-tuned on manually annotated data in Spanish, Catalan, Galician, and Basque, using multilingual variants of the Llama family and Salamandra models from BSC as base models [9]. For low-resource languages, multilingual transfer from Spanish is applied using domain-adaptive fine-tuning and synthetic data augmentation with automatically generated and expert-validated data [17]. A separate model is trained per domain (general, legal, biomedical) to capture domain-specific terminology and stylistic conventions.

3. **LLM-based ambiguity resolution layer:** for cases where previous layers produce uncertain or partially overlapping detections, a larger LLM is queried in a few-shot prompting mode. The LLM receives the text fragment, the list of candidate entities, and a description of the document context, and returns a binary decision (PII/non-PII) along with the corresponding GDPR category.

The three confidence levels produced by the hybrid pipeline feed into a probability calibration module that provides a risk score per entity, which is subsequently used to set the differential privacy budget.

5.2. Differential privacy and synthetic substitution

Once PII entities have been detected, the substitution module replaces them with synthetic values that preserve the semantic and referential coherence of the text. The process operates in three steps:

Step 1 — Allocation of the ϵ budget. The ϵ parameter of local differential privacy [18] is assigned adaptively: special GDPR categories (health data, sexual orientation, legal status) receive low ϵ values ($\epsilon \leq 1,0$), whereas less sensitive entities (municipality names, generic professions) allow higher values ($\epsilon \leq 5,0$). This differentiated budget scheme minimizes degradation of linguistic utility where regulatory risk is lower.

Step 2 — Substitute generation. The replacement value is generated using a lightweight language model (parameters in the 1B–3B range) fine-tuned on pairs (original entity, valid substitute) constructed from synthetic records. The model takes as input: (a) the entity category, (b) the sentence context (± 2 sentences), and (c) format constraints (e.g., “common male first name in Catalonia”, “date prior to the court ruling”). This ensures *context-plausible* substitutes, far superior to the [REDACTED] label, whose presence is easily detectable and destroys the linguistic signal for models trained on such data [19].

Step 3 — Referential coherence verification. Since the same individual may be mentioned multiple times in a document using different forms of reference (full name, appellation, pronoun, role), a per-document substitution dictionary is maintained to ensure that all mentions of the same entity receive the same substitute. A lightweight coreference resolution module identifies mention clusters before applying the replacement.

5.3. Cryptographic traceability

Each anonymization transformation—detection, substitution decision, and substitute generation—is recorded in a structured and immutable *provenance log*. The integrity of this log is ensured through SHA-256 hash chaining: each transformation record includes the hash of the previous record, forming a verifiable chain similar to a *ledger*.

Each log entry contains:

- Document identifier (hash of the original text).
- Entity position (start and end offsets in the text).
- GDPR category of the detected entity.
- Applied substitution mechanism and ϵ value.
- Timestamp and version of the detection model used.

This record complies with the transparency and traceability requirements imposed by Art. 53 of the AI Act [3] and facilitates regulatory audits without requiring access to the original data, since all information refers to the process rather than to sensitive content.

5.4. Risk evaluation and utility preservation

The evaluation follows a paired experimental design: for each corpus, both the original and anonymized versions are available, and the same evaluation procedures are applied to both.

Re-identification risk evaluation. Three canonical attacks of increasing sophistication are applied:

1. **Membership inference** [10, 13]: an adversarial classifier is trained that, given a text fragment and a generative model trained on the anonymized corpus, estimates whether that fragment was part of the training data. A highly accurate classifier indicates insufficient anonymization.
2. **Attribute inference**: given the anonymized text, an attacker attempts to predict the original values of sensitive attributes (e.g., the real name of the person or their diagnosis). Supervised classification models trained on auxiliary public-domain data are used.
3. **Record linkage**: the anonymized corpus is matched against external databases (e.g., public records from the National Statistics Institute, professional registries) using residual non-anonymized attributes. The successful linkage rate is the main risk indicator.

Evaluation of linguistic utility. Utility preservation is measured along two dimensions:

- **Direct corpus utility**: comparison of the perplexity of a reference LLM evaluated on the original corpus and on the anonymized corpus; a minimal difference indicates that the anonymized text preserves the same lexical and syntactic distribution.
- **Downstream training utility**: the same base model is trained on the original corpus and on the anonymized corpus, and their performance is compared on reading comprehension, information extraction, and factual question answering tasks. Any performance drop above a predefined threshold ($\Delta F_1 > 2,0$ points) triggers a re-sampling or fine-grained pseudonymization process, rather than proceeding with massive redaction that would destroy the linguistic signal.

Continuous improvement loop. The results of both evaluations feed back into the ε budget allocation module: if a given entity category exhibits high residual re-identification risk, its ε threshold is automatically reduced in the next anonymization iteration. This feedback loop turns the process into a continuous improvement system aligned with the *trustworthy-by-design* approach of HumanAIze.

5.5. Equity and Intersectionality Suite

Biases in LLMs can originate from the data, the training process, or the alignment phase [20]. SAFE-WORDS addresses the problem at the source by auditing corpora before they are consumed by modeling modules.

The fairness module operates in two phases:

1. **Diagnosis**: for each corpus, automated reports are generated quantifying: (a) demographic representation imbalance (gender, age, geographic origin); (b) conditional demographic parity in classification tasks; (c) consistency under counterfactual substitutions (e.g., replacing male names with female names and checking whether model predictions change); and (d) intersectional harm scores, which detect whether bias is amplified at the intersection of multiple demographic attributes [21].
2. **Mitigation**: when metrics exceed thresholds defined in the ethical framework (T2.3), the following strategies are automatically applied in order of increasing computational cost: reweighting of training instances, counterfactual augmentation of underrepresented groups, and, as a last resort, post-training calibration of the model’s classification layers.

The suite is executed at every training *checkpoint* of HumanAIze models, and its results are directly consumed by the TRAIL-LLMs evaluation module (T6.1, T6.2) and by AlignedLLM’s legal guardrails protocol (T5.2).

5.6. Ethical framework and data governance

The SAFEWORDS ethical framework is not a static document but a set of operational tools—checklists, audit templates, and escalation procedures—integrated into the consortium’s daily workflow. It is structured in four layers:

1. **Data protection and ethics:** establishes thresholds exceeding GDPR minima regarding data minimization, processing of special categories, and necessity principle. Includes a privacy incident notification protocol with defined timelines and responsible parties.
2. **FAIR governance [14]:** each corpus and model produced by HumanAIze is published with a structured datasheet documenting provenance, license, distribution statistics, and known biases, ensuring findability, accessibility, interoperability, and reusability.
3. **Human feedback:** defines the protocol for recruitment, training, instructions, and compensation of human annotators who generate preference data for RLHF in SP2 (AlignedLLM). The goal is to ensure that collected preferences are representative of Spain’s cultural and linguistic diversity and not biased by annotators’ working conditions.
4. **Environmental sustainability [22]:** for each training experiment, electricity consumption, GPU time, and estimated CO₂ equivalent footprint are automatically recorded using codecarbon. Research teams receive periodic reports and can monitor their cumulative consumption in real time against the subproject’s carbon budget. Efficient training techniques such as *LoRA*, knowledge distillation, and recycling of preexisting models are prioritized to minimize the footprint.

6. Integration of SAFEWORDS within HumanAIze

SAFEWORDS acts as the ethical and legal backbone of the project, extending its influence across all subprojects through well-defined integration mechanisms.

SP1 – MultiReason (UPV/EHU). The anonymized corpora (T3.4) serve as the primary training material for extending base models to low-resource languages (T4.3). Governance guidelines determine which web sources may be used in the continuous pretraining of coofficial languages in Spain [7], preventing un-anonymized personal data from contaminating the models.

SP2 – AlignedLLM (USC). The ethical framework (T2.3) and legal guardrails (T5.2) establish the limits that aligned models must respect via RLHF and DPO [23]. SAFEWORDS provides specific evaluation sets to detect whether an aligned model inadvertently reveals personal information memorized during training [10].

SP4 – VERITAS-H (UJA). The biomedical domain contains the most sensitive personal data categories: clinical histories, diagnoses, and treatments. SAFEWORDS designs anonymization techniques adapted to medical terminology, which VERITAS-H will apply in Use Case C [8].

SP5 – TRAIL-LLMs (UNED). The SAFEWORDS fairness suite provides bias and privacy benchmarks that TRAIL-LLMs incorporates into its evaluation framework (T6.1, T6.2), ensuring that models meet fairness thresholds in addition to standard performance metrics.

SP6 – TRUSTMODAL (BSC). Anonymization techniques are extended to non-textual modalities: voice —removal of identifying paralinguistic data— and image —pixelation of faces and text— in the multimodal corpora of T4.5 and T4.6 [9].

7. Use Case: Legal and Administrative Domain

Use Case B of HumanAIze applies SAFEWORDS techniques to the anonymization of legal and administrative documents in Spanish. This domain is particularly critical because: (i) documents often contain special categories of data (health status, judicial situation, sexual orientation) subject to strict GDPR restrictions; and (ii) the Spanish Public Administration must demonstrate compliance with the AI Act before August 2026.

The evaluation corpus includes judicial rulings from the General Council of the Judiciary, administrative resolutions from various Ministries, healthcare correspondence, and data protection files provided by the Spanish Data Protection Agency (AEPD) for research purposes. All documents are labeled with the GDPR category of each detected sensitive entity.

Evaluation combines automatic and manual metrics:

- *Precision, recall, and F_1* for IIP entity detection, validated against annotations by legal experts.
- *Re-identification rate* estimated via membership inference attacks and record linkage against auxiliary databases.
- *Linguistic utility* measured as the difference in performance on reading comprehension and information retrieval in the legal domain between an LLM trained on the original corpus and one trained on the anonymized corpus.
- *GDPR compliance verification* via a semi-automatic protocol derived from the SAFEWORDS ethical framework [24].

Results are compared with state-of-the-art systems—*presidio* [11] and *spaCy*-NER fine-tuned on Spanish legal data. Preliminary results show that the combination of differential privacy and synthetic substitution outperforms simple redaction in linguistic utility while maintaining an equivalent level of protection. All resources (anonymized corpora, models, and source code) will be released under open licenses (CC-BY for data, Apache 2.0 for code) via HuggingFace and Zenodo.

8. Conclusions and Future Work

SAFEWORDS addresses one of the most urgent challenges in the ethical deployment of LLMs in Europe: ensuring that training corpora and the resulting systems do not violate citizens’ privacy rights or breach the GDPR or the AI Act. Its technical proposal covers the entire cycle from data acquisition to evaluation, and its role as a transversal layer within HumanAIze ensures that all subprojects operate under a uniform ethical and legal standard, enriching the corpora used by MultiReason and TRUSTMODAL, the guardrails adopted by AlignedLLM, the domain-specific techniques leveraged by VERITAS-H, and the benchmarks employed by TRAIL-LLMs.

Future work includes: (i) extending IIP recognizers to the biomedical domain with VERITAS-H; (ii) multimodal anonymization of voice and image in collaboration with TRUSTMODAL; (iii) exploring *federated learning* to train IIP recognizers without centralizing sensitive data; and (iv) open publication of all produced resources as assets for the Spanish-speaking NLP community.

Acknowledgments

This work has been funded by the project “SAFEWORDS: Language Anonymization with Ethical and Legal Safeguards through NLP” (AIA2025-163322-C63) (coordinated project HumanAIze) funded by MICIU/AEI/10.13039/501100011033. The authors acknowledge the collaboration of the IXA group (UPV/EHU), LComp (USC), CEATIC (UJA), NLP & IR (UNED), and the Language Technology Laboratory at BSC.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for grammar and spelling checks, proofreading, and text polishing. After using this tool, the authors reviewed and edited the content as necessary and take full responsibility for the final content of the publication.

References

- [1] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, et al., Ethical and social risks of harm from language models, in: arXiv preprint arXiv:2112.04359, 2021. <https://arxiv.org/abs/2112.04359>.
- [2] Parlamento Europeo y Consejo de la UE, Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (RGPD), Technical Report L 119, Diario Oficial de la Unión Europea, 2016. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
- [3] Parlamento Europeo y Consejo de la UE, Reglamento (UE) 2024/1689 por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de IA), Technical Report L 2024/1689, Diario Oficial de la Unión Europea, 2024. <https://www.euaiact.com/>.
- [4] E. Agirre, A. Soroa, P. Gamallo, M. Garcia, R. Muñoz, N. Fernández, L. A. Ureña, E. Martínez, J. Gonzalo, I. da Cunha, M. Villegas, Humanizando la IA con Tecnología del Lenguaje (HumanAIze): Memoria Científico-Técnica, Technical Report, Ministerio de Ciencia, Innovación y Universidades, España. Convocatoria 2025 de Proyectos de Investigación en el Ámbito de la Inteligencia Artificial, 2025.
- [5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., Training language models to follow instructions with human feedback, in: Advances in Neural Information Processing Systems (NeurIPS 2022), volume 35, 2022, pp. 27730–27744. <https://arxiv.org/abs/2203.02155>.
- [6] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, et al., Constitutional AI: Harmlessness from AI Feedback, Technical Report, Anthropic, 2022. <https://arxiv.org/abs/2212.08073>.
- [7] J. Etxaniz, G. Ormazabal, A. Salaberria, A. Urizar, I. Palomino, O. L. de Lacalle, et al., Latxa: An open language model and evaluation suite for basque, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Association for Computational Linguistics, 2024, pp. 1–18. Best Resource Paper Award.
- [8] H. S. C. Alonso, O. Sainz, J. Etxaniz, O. L. de Lacalle, I. Gonzalez-Dios, E. Agirre, et al., MedExpQA: Multilingual benchmarking of large language models for medical question answering, Artificial Intelligence in Medicine 152 (2024) 102893. doi:10.1016/j.artmed.2024.102893.
- [9] M. Villegas, et al., Salamandra: A family of open-weight language models for spanish and co-official languages, <https://huggingface.co/BSC-LT/salamandra-7b>, 2024. Barcelona Supercomputing Center (BSC).
- [10] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, et al., Extracting training data from large language models, in: Proceedings of the 30th USENIX Security Symposium, USENIX Association, 2021, pp. 2633–2650.
- [11] Microsoft Corporation, Presidio: Data protection and de-identification SDK, <https://github.com/microsoft/presidio>, 2020. Accedido: marzo 2026.
- [12] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, et al., Quality at a glance: An audit of web-crawled multilingual datasets, in: Transactions of the Association for Computational Linguistics, volume 10, 2022, pp. 50–72. doi:10.1162/tac1_a_00447.
- [13] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P 2017), 2017, pp. 3–18. doi:10.1109/SP.2017.41.
- [14] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., The FAIR guiding principles for scientific

data management and stewardship, *Scientific Data* 3 (2016) 160018. doi:10.1038/sdata.2016.18.

- [15] J. Dodge, S. Gururangan, D. Card, R. Schwartz, N. A. Smith, Show your work: Improved reporting of experimental results, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2185–2194. URL: <https://aclanthology.org/D19-1224/>. doi:10.18653/v1/D19-1224.
- [16] L. Sweeney, Simple demographics often identify people uniquely, in: *Carnegie Mellon University, Data Privacy Working Paper 3*, Pittsburgh, PA, 2000. <https://dataprivacylab.org/projects/identifiability/>.
- [17] S. Ruder, M. E. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing, *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Tutorials* (2019) 15–18. doi:10.18653/v1/N19-5004.
- [18] C. Dwork, A. Roth, *The Algorithmic Foundations of Differential Privacy*, volume 9, *Foundations and Trends in Theoretical Computer Science*, 2014. doi:10.1561/04000000042.
- [19] P. Lison, I. Pilán, D. Sanchez, M. Wieling, L. Øvrelid, Anonymisation models for clinical narratives: a comparative study, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021)*, Association for Computational Linguistics, Online, 2021, pp. 3476–3488. doi:10.18653/v1/2021.acl-long.270.
- [20] S. Barocas, M. Hardt, A. Narayanan, *Fairness and machine learning: Limitations and opportunities*, in: *NeurIPS 2017 Tutorial*, ampliado como libro, 2019. <https://fairmlbook.org>.
- [21] J. Foulds, R. Islam, K. N. Keya, S. Pan, An intersectional definition of fairness, in: *Proceedings of the 36th IEEE International Conference on Data Engineering (ICDE 2020)*, 2020, pp. 1918–1921. doi:10.1109/ICDE48307.2020.00203.
- [22] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, 2019, pp. 3645–3650. doi:10.18653/v1/P19-1355.
- [23] T. Liu, Y. Zhao, R. Joshi, M. Khalman, M. Saleh, P. J. Liu, et al., Understanding reference policies in direct preference optimization, in: *Proceedings of the 2025 Annual Conference of the North American Chapter of the ACL (NAACL 2025)*, Association for Computational Linguistics, 2025.
- [24] T. Dalenius, Towards a methodology for statistical disclosure control, *Statistik Tidskrift* 15 (1977) 429–444.