SAFEWORDs: un marco reproducible para anonimización conforme al RGPD y evaluación de generación en lenguas cooficiales

Rafael Muñoz[1], **Manuel Palomar**[1], **Elena Lloret**[1], **Nuria Fernández**[1]
[1] CENID-Universidad de Alicante (UA)
rafael@dlsi.ua.es, mpalomar@dlsi.ua.es
elloret@dlsi.ua.es, nuria.fernandez@ua.es

Abstract

Los Grandes Modelos de Lenguaje (LLMs) abren oportunidades para el Procesamiento del Lenguaje Natural (PLN) en contextos institucionales, si bien plantean riesgos críticos en entornos regulados y multilingües, especialmente en lo relativo a protección de datos personales, trazabilidad de decisiones y equidad entre lenguas con distinta disponibilidad de recursos. Presentamos SAFEWORDs, proyecto que acaba de iniciarse en el marco del proyecto coordinado *"HumanAIze"* (Plan Nacional de Inteligencia Artificial 2025, España), que propone un marco reproducible de *privacy-by-design* y *ethics-by-design* para la evaluación y alineación de LLMs en las lenguas oficiales de la Península Ibérica (español, catalán, valenciano, gallego y euskera). El marco integra: (i) anonimización automática conforme al RGPD, con protocolos explícitos de detección de fuga residual y verificación adversarial; (ii) transformación orientada a la accesibilidad textual y al lenguaje claro; y (iii) evaluación en el dominio biomédico, donde la sensibilidad de los datos y la precisión terminológica exigen mecanismos adicionales de control generativo. Desde el punto de vista metodológico, se comparan configuraciones *zero-shot* y *few-shot*, y se documentan prompts, hiperparámetros y recursos para facilitar la replicabilidad y la gobernanza de recursos. Además de sintetizar resultados de referencia de la literatura para contextualizar métricas y órdenes de magnitud esperables, el trabajo discute implicaciones éticas y limitaciones del enfoque propuesto. La propuesta se alinea con las líneas de trabajo de SEPLN y con los objetivos de LANLP, al establecer protocolos transferibles para el desarrollo de tecnologías lingüísticas confiables en ecosistemas caracterizados por variación dialectal y lenguas infrarepresentadas.

Keywords: modelos de lenguaje; anonimización; RGPD; lenguaje claro; dominio biomédico; lenguas ibéricas; evaluación multilingüe; IA confiable

1. Introducción

El desarrollo reciente de los Grandes Modelos de Lenguaje (LLMs) ha ampliado significativamente el alcance del Procesamiento del Lenguaje Natural (PLN), posibilitando aplicaciones avanzadas en contextos institucionales, administrativos y científicos. Sin embargo, su integración en entornos europeos regulados exige garantizar principios de robustez, legalidad, transparencia y respeto a los derechos fundamentales, especialmente en lo que concierne a la protección de datos personales y la equidad lingüística.

El proyecto HumanAIze (Plan Nacional de Inteligencia Artificial 2025) aborda estos retos desde la perspectiva de una inteligencia artificial centrada en el ser humano, con especial atención a las lenguas de la Península Ibérica: español, catalán, valenciano, gallego y euskera. Este enfoque reconoce la diversidad lingüística del territorio y la necesidad de evaluar el comportamiento de los modelos de lenguaje en escenarios multilingües con distintos grados de disponibilidad de recursos.

En este marco, SAFEWORDs constituye la contribución específica del equipo de la Universidad de Alicante dentro de HumanAIze. Su objetivo es desarrollar y validar un marco metodológico para la evaluación y alineación de LLMs en contextos regulados y multilingües, tomando como referencia tres casos de uso representativos:

1. **Anonimización automática conforme al RGPD**, aplicada a textos administrativos, jurídicos y sanitarios en lenguas peninsulares.

2. **Transformación orientada a la mejora de la accesibilidad textual**, con especial atención a escenarios de lenguaje claro en documentos institucionales.

3. **Aplicación en el dominio biomédico**, donde la sensibilidad de los datos y la precisión terminológica requieren mecanismos robustos de generación y control de salida.

Estos tres casos de uso comparten un núcleo metodológico común: la necesidad de evaluar sistemáticamente el comportamiento de los LLMs bajo restricciones normativas y lingüísticas explícitas, comparando distintas configuraciones experimentales y analizando riesgos como la fuga de información, la inconsistencia interlingüística o la degradación semántica.

SAFEWORDs se concibe, por tanto, no como una aplicación aislada, sino como un marco experimental orientado a la evaluación reproducible y la gobernanza de recursos lingüísticos en entornos de alto impacto social.

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 46–54

2. Trabajo Relacionado

El marco propuesto en SAFEWORDs se sitúa en la intersección de cuatro líneas de investigación activas: (i) alineación normativa de modelos generativos, (ii) anonimización automática y reconocimiento de entidades nombradas, (iii) generación controlada y accesibilidad textual en dominios especializados, y (iv) desarrollo reciente de modelos de lenguaje para lenguas ibéricas.

2.1. LLMs y alineación normativa

La literatura reciente ha puesto de relieve los riesgos estructurales asociados al despliegue de modelos fundacionales en contextos regulados. Bender et al. (2021) identifican problemas sistémicos relacionados con sesgos y falta de control en modelos generativos de gran escala. Desde la perspectiva de la seguridad, Carlini et al. (2021) demostraron empíricamente que estos modelos pueden memorizar y revelar datos sensibles presentes en los datos de entrenamiento, con implicaciones directas para el cumplimiento del Reglamento General de Protección de Datos (RGPD) (Parlamento Europeo and Consejo de la Unión Europea, 2016) y de la Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales (Jefatura del Estado (España), 2018).

Estas preocupaciones resultan especialmente relevantes en entornos institucionales europeos, donde la generación automática debe evaluarse bajo criterios explícitos de robustez, legalidad y minimización del riesgo.

2.2. Anonimización y reconocimiento de entidades

La relación entre los Grandes Modelos de Lenguaje (LLMs) y la anonimización de datos atraviesa una fase de transformación dual entre 2024 y 2026. Por un lado, los LLMs han demostrado ser herramientas de anonimización avanzadas capaces de superar las limitaciones de los métodos tradicionales basados en reglas o Reconocimiento de Entidades Nombradas (NER) simple, gracias a su capacidad para comprender el contexto y realizar una redacción adaptativa de información personal identificable (PII) (Staab et al., 2024; Ponomarenko et al., 2026). Investigaciones recientes destacan el auge de la "anonimización adversarial", donde se utilizan modelos para identificar inferencias indirectas que podrían conducir a la reidentificación, permitiendo así proteger datos que parecen anónimos pero que son vulnerables ante la capacidad deductiva de modelos avanzados (Miranda et al., 2024; Zamroz and Morozov, 2024).

Sin embargo, el despliegue de estas tecnologías enfrenta desafíos críticos, principalmente el equilibrio entre privacidad y utilidad. Mientras que técnicas como la Privacidad Diferencial (DP) y el uso de Modelos de Lenguaje Pequeños (SLMs) locales ofrecen garantías de seguridad superiores y cumplimiento regulatorio (como GDPR o HIPAA en entornos clínicos), suelen conllevar una degradación en la calidad de las respuestas o un aumento en los costos operativos (Garza et al., 2025; Yang et al., 2024). Además, persiste una preocupación significativa sobre la "capacidad inferencial" de los modelos propietarios cerrados, lo que ha impulsado el desarrollo de marcos de trabajo de código abierto y arquitecturas *self-hosted* que permiten procesar datos sensibles sin exponerlos a nubes de terceros, garantizando así una soberanía de datos real en sectores altamente regulados (Jonnagaddala and Wong, 2025; Mancera et al., 2025).

La desidentificación automática en textos clínicos ha sido ampliamente estudiada mediante modelos neuronales supervisados (Dernoncourt et al., 2017). No obstante, la incorporación de LLMs generativos introduce nuevos desafíos, como la generación inconsistente de sustitutos léxicos o la persistencia de información residual que permita la reidentificación indirecta.

2.3. Generación controlada y accesibilidad en el dominio biomédico

La simplificación textual automática ha sido extensamente estudiada en inglés (Alva-Manchego et al., 2020; Shardlow, 2014). En español, trabajos recientes publicados en *Procesamiento del Lenguaje Natural* han abordado la simplificación en contextos de salud mediante modelos basados en transformers. En particular, el ajuste fino de modelos BART para la simplificación de textos sanitarios (Alarcón et al., 2023) y la construcción de corpus comparables para la evaluación en salud (Campillos-Llanos et al., 2022) evidencian avances significativos en generación controlada en dominio especializado.

SAFEWORDs se sitúa en esta línea de investigación, integrando accesibilidad textual y generación en el dominio biomédico dentro de un marco unificado de evaluación normativa y lingüística.

2.4. Lenguas ibéricas y evaluación multilingüe

El desarrollo de infraestructuras multilingües para lenguas peninsulares ha sido impulsado, entre otras iniciativas, por El proyecto ILENIA[1] ha impulsado el desarrollo de una familia de modelos de lenguaje entrenados sobre grandes volúmenes

[1] https://proyectoilenia.es

de datos: SALAMANDRA (34 idiomas) (Gonzalez-Agirre et al., 2025), AITANA (valenciano) (GPLSI – Language and Information Systems Group, University of Alicante, 2026), LATXA (euskera) (Etxaniz et al., 2024), CARVALHO (gallego) (Gamallo et al., 2024) y ANIA (catalán) (González-Agirre et al., 2024). Estos modelos sientan una base sólida para el desarrollo de sistemas instruidos en tareas concretas con resultados excelentes en tareas como generación, traducción y su aplicación en casos de usos reales.

3. Marco metodológico

SAFEWORDs (UA) contribuye a HumanAIze con un marco metodológico de *ethics-by-design* y *privacy-by-design* para el desarrollo, control y evaluación de LLMs en las lenguas oficiales de España. Este marco operacionaliza directrices y umbrales ético-legales en forma de *checklists* y puntos de control aplicables a datos, modelos y procedimientos de evaluación (WP3–WP4), tal como se ilustra en la Figura 1.

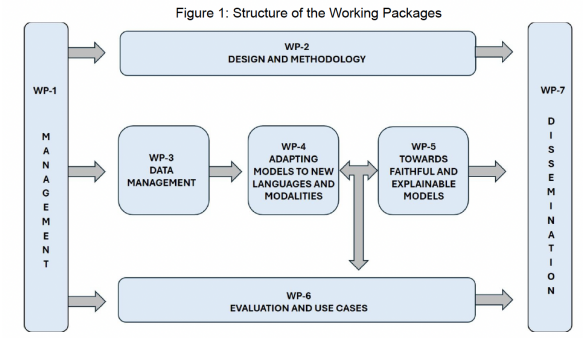

Figure 1: Paquetes de trabajo del proyecto HumanAIze.

El marco se implementa como un **pipeline trazable** que conecta cuatro etapas: (i) adquisición y curación de datos, (ii) anonimización y mitigación de riesgos de reidentificación, (iii) alineación mediante instrucciones y evaluación comparativa, y (iv) verificación previa a la liberación de recursos y modelos, bajo umbrales estrictos de privacidad y equidad.

En coherencia con el planteamiento del proyecto, la evaluación se diseña explícitamente en escenarios **zero-shot** y **few-shot**, comparando distintas configuraciones de prompts y verificando el cumplimiento del RGPD y el EU AI Act, dado que la verificación automática completa sigue siendo un problema abierto en el estado del arte.

3.1. Diseño transversal: configuraciones y reproducibilidad

Para cada caso de uso se comparan, como mínimo, las siguientes configuraciones:

- **Zero-shot**: instrucciones directas con restricciones normativas explícitas.
- **Few-shot**: incorporación de ejemplos representativos, multilingües cuando sea posible.
- **Ablaciones de guardrails**: con y sin módulos de control (p. ej., validadores de formato, listas de exclusión léxica, detectores de entidades nombradas).

La reproducibilidad se garantiza mediante:

- Versionado de datos, prompts y configuraciones experimentales.
- Registro sistemático de hiperparámetros (modelo, temperatura, longitud máxima de contexto, etc.).
- Firma y trazabilidad de recursos (datos, prompts, salidas) como parte del control de procedencia y cadena de custodia.

3.2. Pipeline de privacidad y verificación RGPD/EU AI Act

Adoptamos un enfoque **privacy-by-design** en el que cada corpus de entrada es analizado mediante reconocedores multilingües para detectar:

- Identificadores directos (correos electrónicos, identificadores fiscales, números de teléfono, direcciones postales, etc.).
- *Contextual cues* que pueden facilitar ataques de reidentificación por enlace (p. ej., la combinación de cargo profesional, localidad y fecha de evento).

Cuando resulta técnicamente viable, las referencias a datos personales son:

- **eliminadas** mediante purga directa, o
- **reemplazadas** por **sustitutos sintéticos** (*surrogates*) generados bajo presupuestos de privacidad diferencial, preservando la utilidad textual (p. ej., coherencia morfológica de género y número) y registrando el proceso para garantizar la trazabilidad.

El pipeline se registra de forma que permita cumplir con las obligaciones de trazabilidad y transparencia establecidas en el marco regulatorio vigente.

3.2.1. Protocolo de detección de *fuga residual* y verificación de cumplimiento

Definimos **fuga residual** (*residual leakage*) como la presencia en las salidas del modelo de información que permita: (i) identificar directa o indirectamente a una persona física, o (ii) recuperar datos personales mediante inferencias o ataques de enlace. La evaluación combina tres niveles de verificación:

(1) Verificación automática

- Detección post-generación mediante herramientas de NER y expresiones regulares multilingües orientadas a información de identificación personal (PII).
- Validadores de formato: patrones de DNI/NIE, números de teléfono, correos electrónicos y códigos postales.
- Heurísticas de enlace: detección de co-ocurrencias infrecuentes que puedan resultar identificadoras (p. ej., combinación única de profesión, micro-localización y fecha).

(2) Verificación adversarial Se generan *prompts de ataque* diseñados para inducir la revelación de PII:

- Consultas de recuperación directa ("¿quién es la persona X?", "proporciona el nombre completo").
- Reformulaciones orientadas a forzar la generación de detalles identificadores ("amplía la información anterior").
- Ataques por consistencia: generación múltiple de la misma entrada para detectar divergencias reveladoras.

(3) Auditoría humana Muestreo estratificado de salidas (por lengua, dominio y tipo de PII) con los objetivos de:

- Identificar falsos negativos de los detectores automáticos.
- Evaluar el riesgo de enlace y reidentificación contextual en casos límite.

Métricas de evaluación Se reportan: (i) tasa de fuga residual, (ii) tasa de falsos negativos del detector de PII, (iii) severidad de la fuga (PII directa frente a inferencial), y (iv) degradación de utilidad textual (comparación entre muestras anonimizadas y originales mediante métricas automáticas y valoración humana).

3.3. Caso de uso A: Lenguaje claro en el dominio administrativo-jurídico

HumanAIze define explícitamente que la producción de lenguaje claro se abordará mediante la transformación de rasgos lingüísticos en **tres niveles**: discursivo, morfosintáctico y léxico, y que los desarrollos resultantes se incorporarán a la herramienta **arText** (da Cunha, 2022) como vía de transferencia tecnológica.

Diseño y evaluación

- Corpus anotados manualmente para medir la adecuación de las salidas generadas.
- Evaluación automática basada en precisión y recall sobre transformaciones etiquetadas.
- Estudios con usuarios reales para contrastar la claridad percibida y la comprensión efectiva de los textos transformados.

3.4. Caso de uso B: Anonimización para la preservación de la privacidad

Este escenario evalúa técnicas de anonimización asistidas por LLMs sobre textos legales y administrativos en lenguas peninsulares. Se comparan distintas configuraciones de prompt en modalidades zero-shot y few-shot, con verificación explícita del cumplimiento de los requisitos de privacidad establecidos por el RGPD y el EU AI Act.

El diseño experimental contempla:

- Comparativa de técnicas (enmascarado, sustitución léxica y generación de sustitutos sintéticos) y sus posibles combinaciones.
- Evaluación del cumplimiento normativo, verificando que las salidas no contengan datos personales no anonimizados.
- Introducción de documentos con alta densidad de información sensible como casos de estrés del sistema (*adversarial inputs*).

3.5. Caso de uso C: Dominio biomédico

La evaluación en el dominio biomédico se realiza sobre corpus anotados y validados por expertos clínicos. Se incorporan métricas específicas del dominio, protocolos de *test* rigurosos y análisis comparativo con *benchmarks* establecidos en la literatura, incluyendo *shared tasks* relevantes cuando aplique.

4. Resultados de referencia en la literatura

Dado que SAFEWORDs se presenta en este trabajo como una contribución metodológica y de diseño evaluativo en el marco de HumanAIze, incluimos resultados de referencia procedentes de la literatura reciente (principalmente para español y lenguas cooficiales en España) con el fin de contextualizar el tipo de métricas y los órdenes de magnitud esperables en las tareas relacionadas con nuestros casos de uso. Esta sección no reporta resultados propios del proyecto, sino que sintetiza *baselines* y recursos publicados disponibles en el momento de la escritura de este trabajo.

En español, el desarrollo de modelos de lenguaje específicos ha avanzado de forma notable con iniciativas como ILENIA[2] en el que se pueden encontrar una gran diversidad d edatasets, modelos texto y voz, y demostradores. En tareas de accesibilidad en salud, se han publicado enfoques basados en ajuste fino de BART/mBART con resultados expresados en términos de SARI, junto con evidencias de mejora en escalas de legibilidad estándar. La construcción de corpus comparables de simplificación médica en español ha permitido, asimismo, disponer de *benchmarks* de referencia a escala considerable.

4.1. Implicaciones para SAFEWORDs

Estos resultados respaldan dos decisiones metodológicas centrales del equipo UA en SAFEWORDs: (i) la necesidad de protocolos de evaluación que combinen métricas automáticas y verificación cualitativa en tareas de accesibilidad y dominio biomédico, y (ii) la viabilidad de enfoques adaptados a bajo recurso en lenguas cooficiales mediante generación de datos sintéticos y evaluación con métricas estandarizadas.

En fases posteriores del proyecto, SAFEWORDs utilizará estos resultados y órdenes de magnitud como referencias para diseñar comparativas y análisis de robustez bajo restricciones normativas (anonimización conforme al RGPD) y lingüísticas (español, catalán/valenciano, gallego y euskera).

5. Recursos, gestión de datos y liberación de recurso

En coherencia con los principios de HumanAIze, SAFEWORDs (UA) no se limita a evaluar modelos, sino que contribuye a un ciclo completo de *data governance* y liberación controlada de activos, con el objetivo de facilitar la adopción científica e institucional de tecnologías lingüísticas en las lenguas oficiales de España. Este planteamiento se articula en torno a dos pilares: (i) un repositorio multilingüe y versionado de datos de tarea e instrucción, y (ii) un conjunto de umbrales y listas de verificación en materia de privacidad y equidad que actúan como compuertas de calidad (*quality gates*) antes de la reutilización o difusión de cualquier activo.

5.1. Repositorio multilingüe y versionado (datos de tarea e instrucción)

HumanAIze establece WP3 como el núcleo de gestión de datos del proyecto, con un portfolio balanceado de corpus (desde *crawls* web filtrados hasta colecciones específicas de dominio) y un pipeline que registra procedencia, licencia y propiedades estadísticas, garantizando la trazabilidad y la comparabilidad longitudinal entre versiones.

Antes de liberar cualquier activo del proyecto (datos, instrucciones, modelos o herramientas), se verifica que cumple umbrales estrictos de privacidad y equidad, como parte del proceso de gobernanza versionada definido en WP3 y alineado con las directrices del WP2.

Dentro de este marco, SAFEWORDs contribuye de forma explícita a:

- Establecer y curar un **repositorio multilingüe, versionado y auditado** de datos de tarea e instrucción que cubra dominios prioritarios y las lenguas oficiales de España (T3.1, T3.3).
- Verificar que **cada activo** cumple umbrales estrictos de privacidad y equidad antes de su uso en entrenamiento o evaluación (T3.4, T3.5 y tareas relacionadas en WP5).

5.2. Privacidad, trazabilidad y control de riesgo

La memoria del proyecto define un enfoque **trustworthy-by-design** y **privacy-by-design** en el que cada corpus de entrada es auditado automáticamente y complementado con revisiones humanas en casos límite, traduciendo los requisitos del RGPD y del EU AI Act a *checklists* operativas para la adquisición, retención, archivo y acceso controlado a datos.

Este marco incorpora mecanismos explícitos para:

- Mitigar riesgos de reidentificación mediante pruebas de estrés (p. ej., *membership inference*, *attribute inference* y *record linkage*) y, cuando sea necesario, aplicar técnicas de pseudonimización fina o re-muestreo dirigido.
- Garantizar la trazabilidad y la transparencia mediante el registro completo del pipeline de

[2] https://proyectoilenia.es/recursos-modelos-datasets/

Trabajo (tarea)	Recurso / escenario	Resultados reportados
ILENIA (modelos de lenguaje)	Familia de modelos de lenguaje en español y lenguas cooficiales	Publicación de modelos y recursos de referencia para en la web del proyecto y hugginface
Alarcón et al. (simplificación en salud)	Ajuste fino de BART/mBART para simplificación de textos sanitarios en español	SARI: 59.7 (mBART fine-tuned); 29.74 (mBART preentrenado en generación de resúmenes); mejora de legibilidad según escala Inflesz
Campillos-Llanos et al. (corpus biomédico)	CLARA-MeD: corpus comparable de simplificación médica en español	24.298 pares de textos (profesional vs. simplificado); $>$96M tokens

Table 1: Resultados de referencia de la literatura relacionados con los casos de uso de SAFEWORDs. Esta tabla no reporta resultados propios del proyecto.

procesamiento, facilitando el cumplimiento de las obligaciones regulatorias aplicables a sistemas de IA de propósito general.

HumanAIze establece como principio que los recursos del proyecto (datos, modelos y software) sean **reutilizables, reproducibles y debidamente documentados**, y que se distribuyan mediante plataformas de recursos digitales relevantes para la comunidad investigadora, facilitando tanto el escrutinio científico como la adopción institucional o industrial.

En el caso de SAFEWORDs, la liberación de recursos está condicionada a la superación de controles previos de privacidad y equidad, alineados con las directrices del WP2 y el régimen de trazabilidad del WP3.

6. Consideraciones éticas y limitaciones

En HumanAIze, los aspectos éticos, legales y sociales no se tratan como una capa adicional posterior, sino como un componente *trustworthy-by-design* que se traduce en guías, recomendaciones y listas de verificación operativas que condicionan la adquisición de datos, el entrenamiento, la evaluación y la diseminación de resultados.

SAFEWORDs (UA) se alinea con este enfoque codificando principios y umbrales en *checklists* aplicables a todo el ciclo de vida de los activos del proyecto —privacidad, no discriminación, supervisión humana y sostenibilidad computacional— y verificando su cumplimiento antes de cualquier uso o liberación.

6.1. Privacidad, RGPD y EU AI Act

HumanAIze explicita que, desde agosto de 2024, el EU AI Act (Reglamento 2024/1689) complementa el RGPD con obligaciones graduadas en función del nivel de riesgo del sistema, y adopta una postura *privacy-by-design* a lo largo de todo el ciclo de vida. En la práctica, todo corpus entrante se escanea con reconocedores multilingües para identificar identificadores directos (correos electrónicos, identificadores fiscales) y señales contextuales susceptibles de facilitar ataques por enlace.

Cuando resulta técnicamente viable, las referencias personales se eliminan o se reemplazan mediante sustitutos sintéticos generados bajo presupuestos de privacidad diferencial; el proceso se registra íntegramente para satisfacer los requerimientos de trazabilidad y transparencia del EU AI Act.

Como medida adicional de control del riesgo, el proyecto contempla pruebas de estrés de reidentificación (incluyendo *membership inference*, *attribute inference* y *record linkage*) y comparativas emparejadas entre muestras anonimizadas y originales para cuantificar la pérdida de utilidad textual. Este conjunto de prácticas guía el caso de uso de anonimización de SAFEWORDs sobre textos legales y administrativos, con el objetivo explícito de que ningún identificador personal persista tras la ingesta y de que los activos resultantes sean defendibles bajo el marco regulatorio vigente.

6.2. Equidad, sesgos y evaluación culturalmente informada

HumanAIze establece un programa sistemático de auditoría y mitigación de sesgos en paralelo al desarrollo de modelos. Se prevé la construcción de un *benchmark* multilingüe (lenguas oficiales de España e inglés) para evaluar estereotipos, desequilibrios de representación y diferencias de rendimiento entre grupos demográficos, con anotación por especialistas y revisión comunitaria orientada a capturar señales culturales que típicamente no aparecen en conjuntos de datos de orientación anglocéntrica.

La evaluación se operacionaliza mediante métricas diagnósticas más allá de la exactitud agregada (p. ej., paridad demográfica condicional, consistencia bajo sustituciones contrafactuales y medidas de daño interseccional), y se integra como compuerta de calidad: regresiones significativas

en estas métricas bloquean la promoción de versiones de modelo. Cuando se detectan sesgos, se contemplan estrategias de mitigación mediante reponderación de datos, debiasing basado en modelo y calibración *post-hoc*, documentando los compromisos entre equidad, utilidad y coste computacional.

6.3. Supervisión humana y recogida de retroalimentación

El proyecto incorpora directrices explícitas sobre la interacción humana y la recogida de retroalimentación (*feedback*) como parte del marco de recomendaciones de WP2, con el objetivo de garantizar un enfoque centrado en el ser humano a lo largo de todo el proyecto.

En SAFEWORDs, estas directrices condicionan tanto la generación de datos de alineación e instrucción como la evaluación humana en escenarios sensibles (administrativo-jurídico y biomédico), evitando mecanismos de recogida de retroalimentación que sean coercitivos, sesgados o no representativos de la diversidad de usuarios.

6.4. Sostenibilidad y control de huella computacional

HumanAIze incorpora como requisito la minimización de la huella de carbono y el diseño de prácticas de *green AI* (WP2), promoviendo la eficiencia computacional y la reutilización de modelos existentes siempre que no se comprometan los objetivos científicos del proyecto.

SAFEWORDs adopta estas restricciones como parte de su diseño experimental, priorizando configuraciones y protocolos reproducibles sobre iteraciones de entrenamiento que no aporten evidencia metodológica adicional.

6.5. Limitaciones

El marco presentado tiene varias limitaciones inherentes que deben tenerse en cuenta al interpretar sus resultados:

- **Verificación incompleta de privacidad.** Aunque se aplican auditorías automáticas, pruebas adversariales y ataques de reidentificación, no existe garantía absoluta de ausencia de riesgo residual en contextos abiertos y multifuente; el riesgo se minimiza pero no se elimina.
- **Cobertura desigual por lengua y dominio.** La disponibilidad de datos y recursos varía considerablemente entre el español y las lenguas cooficiales, lo que puede comprometer la comparabilidad entre configuraciones y obliga a diseños de evaluación cuidadosamente estratificados.
- ***Trade-off* entre utilidad y cumplimiento normativo.** Determinados escenarios pueden requerir técnicas de pseudonimización fina o re-muestreo dirigido para preservar la fidelidad factual tras la anonimización; la pérdida de utilidad asociada debe monitorizarse sistemáticamente.
- **Dependencia de la evaluación humana.** La auditoría culturalmente informada y la supervisión humana son componentes esenciales del marco, pero introducen variabilidad interanotador y costes de escalado que deben gestionarse mediante protocolos de anotación robustos y muestreo estratificado.

Finalmente, en línea con los principios de HumanAIze, todos los recursos del proyecto (datasets, modelos y software) serán abiertos, reutilizables y debidamente documentados; en SAFEWORDs, cualquier liberación estará condicionada a superar los umbrales estrictos de privacidad y equidad definidos por el propio marco del proyecto.

7. Conclusiones

Este trabajo ha presentado SAFEWORDs como contribución al proyecto HumanAIze, orientada a operacionalizar un marco *trustworthy-by-design* para Grandes Modelos de Lenguaje en las lenguas oficiales de España (español, catalán, valenciano, gallego y euskera). SAFEWORDs articula un enfoque metodológico integrado que combina: (i) alineación normativa y verificación de cumplimiento del RGPD y el EU AI Act, (ii) control generativo y evaluación comparativa en configuraciones *zero-shot* y *few-shot*, y (iii) un esquema reproducible de auditoría de fuga residual, detección de sesgos y trazabilidad a lo largo del ciclo de vida de los activos.

La propuesta se valida conceptualmente en tres casos de uso de alto impacto social: (1) anonimización para la preservación de privacidad en textos administrativos, jurídicos y sanitarios, (2) mejora de la accesibilidad textual, incluyendo lenguaje claro en documentación institucional, y (3) aplicación en el dominio biomédico, donde la sensibilidad de los datos y la precisión terminológica exigen protocolos de evaluación más estrictos. Frente a aproximaciones centradas exclusivamente en métricas de rendimiento, el marco propuesto prioriza criterios operativos de legalidad, robustez y equidad multilingüe como condiciones necesarias para el despliegue responsable de sistemas basados en LLMs.

Como trabajo futuro, SAFEWORDs avanzará en: (i) la consolidación del repositorio multilingüe

versionado de datos de tarea e instrucción, (ii) la ampliación de los protocolos de auditoría adversarial y las pruebas de reidentificación, (iii) la evaluación sistemática por lengua y dominio con participación de expertos, y (iv) la preparación de activos (datasets, prompts, herramientas y, cuando proceda, modelos) para su liberación condicionada a umbrales estrictos de privacidad y equidad, reforzando la transferibilidad y reutilización científica en línea con los objetivos de HumanAIze.

Declaración ética

El proyecto garantiza el cumplimiento del RGPD en todas las fases de tratamiento de datos, minimiza los riesgos de reidentificación mediante auditorías automáticas y humanas, y evita prácticas extractivas en comunidades lingüísticas minoritarias. Los datos experimentales utilizados en el desarrollo del marco han sido previamente anonimizados conforme a los protocolos descritos en este trabajo.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia, Innovación y Universidades (España) en el marco de la convocatoria 2025 de Proyectos de Investigación en el Ámbito de la Inteligencia Artificial (AIA2025), referencia **AIA2025-163322-C63**.

8. References

Rodrigo Alarcón, Paloma Martínez, and Lourdes Moreno. 2023. Tuning bart models to simplify spanish health-related content. *Procesamiento del Lenguaje Natural*, (70):111–122.

Fernando Alva-Manchego et al. 2020. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *ACL*, pages 4668–4679.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623.

Leonardo Campillos-Llanos, Ana R. Terroba Reinares, Sofía Zakhir Puig, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for spanish medical text simplification. *Procesamiento del Lenguaje Natural*, (69):189–196.

Nicholas Carlini et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Iria da Cunha. 2022. Un redactor asistido para adaptar textos administrativos a lenguaje claro. *Procesamiento del Lenguaje Natural*, 69:39 – 49.

Franck Dernoncourt et al. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Pablo Gamallo, Pablo Rodríguez, Diana Santos, Salvador Sotelo, N. Miquelina, S. Paniagua, D. Schmidt, I. de Dios-Flores, P. Quaresma, D. Bardanca, J. R. Pichel, V. Nogueira, and Senén Barro. 2024. A galician-portuguese generative model. In *EPIA 2024 – Portuguese Conference on Artificial Intelligence (Conference Proceedings)*.

A. Garza et al. 2025. Prvl: Quantifying the capabilities and risks of large language models for pii redaction. *arXiv preprint arXiv:2508.05545*.

Aitor González-Agirre, Montserrat Marimon, Carlos Rodriguez-Penagos, Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller, Jorge Palomar-Giner, Baybars Kulebi, and Marta Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of catalan. In *Proceedings of the LREC-COLING 2024 Conference*, pages 2556–2566.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruíz-Fernández, and Marta Villegas. 2025. Salamandra technical report. arXiv:2502.08489.

GPLSI – Language and Information Systems Group, University of Alicante. 2026. Aitana-6.3b (model card). https://huggingface.co/gplsi/Aitana-6.3B. Accessed: 2026-02-13.

Jefatura del Estado (España). 2018. Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. BOE núm. 294, de 06/12/2018. Referencia: BOE-A-2018-16673. Entrada en vigor: 07/12/2018. ELI: https://www.boe.es/eli/es/lo/2018/12/05/3/con.

J. Jonnagaddala and M. C. Wong. 2025. Strategies for privacy-preserving ehr analysis using llms.

J. Mancera et al. 2025. Pba-llm: Privacy- and bias-aware nlp using named-entity recognition (ner). *arXiv preprint arXiv:2507.02966*.

Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, et al. 2024. Preserving privacy in large language models: A survey on current threats and solutions. *arXiv preprint arXiv:2408.05212*.

Parlamento Europeo and Consejo de la Unión Europea. 2016. Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, 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 (Reglamento general de protección de datos). DO L 119, 4.5.2016, pp. 1–88. CELEX: 32016R0679. Aplicación desde 25/05/2018.

V. Ponomarenko et al. 2026. Capid: Context-aware pii detection for question-answering systems. *arXiv preprint arXiv:2602.10074*.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1).

Robin Staab, Mark Vero, Mislav Balunović, et al. 2024. Large language models are advanced anonymizers. *arXiv preprint arXiv:2402.13846*.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2024. Robust utility-preserving text anonymization based on large language models. *arXiv preprint arXiv:2407.11770*.

P. I. Zamroz and Yu. V. Morozov. 2024. Large language models and personal information: security challenges and solutions through anonymization. *Komp'ûternì sistemi ta merežì*.