

# Proyecto Espacio de Observación de Inteligencia Artificial en Español

## Ámbito 1.1 Leaderboard

### Informe Técnico Año 2

**Enrique Amigó, Alejandro Benito, Jorge Carrillo-de-Albornoz, Andrés Fernández,  
Víctor Fresno, Julio Gonzalo, Guillermo Marco, Roser Morante, Jacobo Pedrosa,  
Laura Plaza, Eva Sánchez**

Natural Language Processing and Information Retrieval Group, UNED

Autor de contacto: Julio Gonzalo - [julio@lsi.uned.es](mailto:julio@lsi.uned.es)

## Resumen

En este informe se presenta el trabajo del segundo año de proyecto en torno al **Leaderboard ODESIA**, que se ha desarrollado en el marco del convenio UNED - ONTSI/Red.es para crear un *Espacio de Observación de Inteligencia Artificial en Español*. El objetivo del Leaderboard ODESIA es permitir la medición de la brecha de efectividad entre los modelos del lenguaje para el español y el inglés en la IA. En el segundo año se han realizado los siguientes avances: (i) se ha ampliado el conjunto de tareas de evaluación con cuatro tareas nuevas del Leaderboard ODESIA CORE (es decir, tareas con datos anotados dentro del proyecto y con un test privado que no se distribuye para evitar problemas de contaminación de los modelos); (ii) se ha medido la brecha para modelos discriminativos utilizando las 15 tareas de la versión 2 del leaderboard; (iii) se han evaluado 10 modelos de lenguaje para cada idioma, estableciendo los modelos con mejor rendimiento para el español; (iv) se ha desarrollado una nueva versión de la aplicación web del leaderboard, con más funcionalidades; (v) ampliando los compromisos establecidos en el convenio, se ha trabajado en la construcción de dos datasets para evaluación de modelos generativos (uno de exámenes de acceso a la universidad y otro de resúmenes de dominio legal con texto claro), se ha trabajado en una experimentación adicional en tareas de escritura creativa, y se ha comenzado la evaluación de modelos generativos en modo few-shot sobre las tareas del Leaderboard ODESIA.

El Leaderboard ODESIA proporciona una infraestructura de evaluación para modelos de lenguaje preentrenados en inglés y español que permite una comparación directa entre el rendimiento de modelos en uno y otro idioma, y por tanto posibilita **medir la brecha de efectividad inglés-español** de los sistemas de Procesamiento del Lenguaje Natural (PLN). También permite comparar modelos preentrenados para el español con una combinación de datasets públicos y privados.

**Datasets.** Después del trabajo del segundo año, el leaderboard consta en la actualidad de diez tareas en el ODESIA CORE (las que tienen al menos un juego de pruebas privado desarrollado en el proyecto) y cinco tareas más en el ODESIA EXTENDED (tareas preexistentes de dominio público incorporadas al leaderboard). Además se están desarrollando datasets para evaluación de modelos generativos en modo few-shot (una de ellas se ha terminado este segundo año). En conjunto, en el leaderboard se cubren las siguientes áreas de aplicación: detección y caracterización de contenidos tóxicos en redes sociales, desinformación, biomedicina, administración y lenguaje claro, conocimiento general de modelos generativos, escritura creativa, similitud textual y sistemas de pregunta/respuesta. En términos de tareas abstractas, se cubren las de clasificación binaria, multiclase, jerárquica, multilabel, clasificación en modo Learning with Disagreement, Sequence labeling, Generación de texto (resúmenes, creativo), Regresión, y dos tipos de Question Answering (Machine Reading, Multiple Choice Questions).

**Aplicación web Leaderboard.** El Leaderboard ODESIA utiliza las infraestructuras del toolkit de evaluación y el servicio online de evaluación EvALL 2.0 desarrollados dentro del convenio. Proporciona tres espacios de comparación de modelos para cada una de las versiones (v1 y v2 por ahora): uno para modelos en español, otro para modelos en inglés, y otro para comparar cuantitativamente los mejores modelos en ambos idiomas. La interfaz del Leaderboard ODESIA se ha mejorado respecto al año anterior de la siguiente manera: se ha integrado en un único ecosistema Leaderboard/EvALL/ODESIA portal, se ha creado el espacio Leaderboard v2, y se han mejorado la navegabilidad y las funcionalidades.

**Experimentación.** Se ha experimentado con diez modelos de lenguaje discriminativos de dominio público en cada idioma, para un total de 3120 procesos de entrenamiento y evaluación. En español, el modelo más efectivo es el modelo multilingüe XLM-Roberta-large. En español, es seguido de cerca por PlanTL-GOB-ES-roberta-large-bne (eficacia promedio de 0,55 frente a 0,57 del XLM-Roberta-large). En inglés, el mejor modelo es Roberta-large, seguido de cerca por XLM-Roberta-large (0,59 y 0,57). **La brecha porcentual es del  $(20 \pm 6) \%$  a favor del inglés.**

Adicionalmente a lo previsto inicialmente en el convenio, se ha experimentado con modelos generativos en varios ámbitos: (i) se ha evaluado GPT-4 en modo zero-shot sobre tres de las tareas del leaderboard en inglés y español, midiéndose una brecha promedio del 18,39 %, muy similar a la de los modelos discriminativos; (ii) se ha desarrollado un dataset (UNED-ACCESO) de más de 1,000 preguntas de exámenes de acceso a la universidad de la UNED correspondientes a 11 asignaturas diferentes, traducidas manualmente al inglés. Sobre este dataset, los modelos abiertos (Llama-2, Mistral y Gemma) dan una brecha promedio del 12,20 % a favor del inglés, mientras que los dos mejores modelos disponibles (GPT-4 y Claude 3 Opus) son ligeramente mejores en español, lo que apunta a posibles problemas de contaminación; (iii) se ha finalizado un dataset de resúmenes con texto claro de documentación legal (CURIA-2024), y se está trabajando en la evaluación de modelos generativos sobre este dataset, y (iv) se está trabajando en una evaluación de GPT-4 en una tarea de escritura creativa.

## Índice

|   |           |
|---|-----------|
| <b>1. Introducción</b>  | <b>3</b>  |
| <b>2. Datasets y tareas de PLN usados para la experimentación</b>           | <b>6</b>  |
| 2.1. Criterios de selección de datasets y tareas                            | 7         |
| 2.2. Dataset 1: EXIST 2022  | 9         |
| 2.3. Dataset 2: DIPROMATS 2023  | 12        |
| 2.4. Dataset 3: DIANN 2023  | 14        |
| 2.5. Dataset 4: EXIST-2023  | 14        |
| 2.6. Dataset 5: SQUAD/SQAC-2024   | 15        |
| 2.7. Dataset 6: CURIA-2024  | 16        |
| 2.8. Dataset 7: UNED ACCESO 2024  | 17        |
| 2.9. Dataset 8: PRON VS PROMPT  | 18        |
| 2.10. Calibración de los datasets para medir la brecha en efectividad EN-ES | 18        |
| <b>3. Aplicación web Leaderboard ODESIA v2</b>                              | <b>20</b> |
| 3.1. Perfiles de usuario  | 21        |
| 3.2. Casos de uso   | 21        |
| 3.3. Requisitos y funcionalidades   | 22        |
| 3.4. Plataforma web   | 22        |
| 3.5. Arquitectura   | 23        |
| 3.5.1. Capa Interfaz de usuario   | 24        |
| 3.5.2. Capa Servidora   | 24        |
| 3.6. Roles de usuario   | 24        |
| 3.7. Flujos del Leaderboard ODESIA  | 25        |
| 3.7.1. Registro de usuario  | 25        |
| 3.7.2. Inicio de sesión   | 25        |
| 3.7.3. Evaluación comparativa español/inglés                                | 26        |
| 3.7.4. Visualización de resultados comparativos español/inglés              | 27        |
| 3.8. Formatos de entrada  | 30        |
| 3.9. Navegación   | 30        |
| 3.10. Aspectos técnicos de desarrollo de proyectos de software              | 31        |

|  |           |
|--|-----------|
| <b>4. Experimentos de evaluación con datasets con test privado (Core Tasks)</b>              | <b>32</b> |
| 4.1. Modelos de lenguaje seleccionados . . . . .   | 32        |
| 4.2. Entrenamiento de modelos e hiperparámetros . . . . .                                    | 34        |
| 4.3. Evaluación: Métricas y baselines . . . . .  | 35        |
| 4.3.1. EXIST-2022 . . . . .  | 35        |
| 4.3.2. DIANN 2023 . . . . .  | 36        |
| 4.3.3. DIPROMATS 2023 . . . . .  | 36        |
| 4.3.4. EXIST-2023 . . . . .  | 36        |
| 4.3.5. SQUAD/SQAC-2024 . . . . .   | 37        |
| 4.4. Evaluación: resultados experimentales Core Tasks . . . . .                              | 37        |
| <b>5. Medición del gap: experimentación extendida con datasets públicos (Extended Tasks)</b> | <b>39</b> |
| 5.1. Medición de la brecha de efectividad español - inglés . . . . .                         | 39        |
| 5.2. Métricas y baselines . . . . .  | 39        |
| 5.2.1. MLDoc . . . . .   | 39        |
| 5.2.2. MultiCONER 2022 . . . . .   | 40        |
| 5.2.3. STS 2017 . . . . .  | 40        |
| 5.2.4. SQAC/SQUAD . . . . .  | 40        |
| 5.2.5. DIANN . . . . .   | 40        |
| 5.3. Evaluación: resultados experimentales . . . . .   | 40        |
| <b>6. Medición del gap: experimentación extendida con modelos generativos</b>                | <b>42</b> |
| 6.1. UNED ACCESO: Tests de conocimiento general de acceso a la universidad . . . . .         | 42        |
| 6.1.1. Métrica . . . . .   | 42        |
| 6.1.2. Baseline . . . . .  | 43        |
| 6.1.3. Descripción de experimentos y parametrización . . . . .                               | 43        |
| 6.1.4. Resultados . . . . .  | 45        |
| 6.2. GPT-4 y DIPROMATS . . . . .   | 49        |
| 6.3. CURIA: Generación de microresúmenes legales . . . . .                                   | 50        |
| 6.3.1. Métrica . . . . .   | 50        |
| 6.3.2. Descripción de experimentos futuros . . . . .   | 52        |
| 6.4. PRON vs PROMPT: Generación de sinopsis . . . . .  | 52        |
| 6.4.1. Métrica . . . . .   | 52        |
| <b>7. Conclusiones y trabajo futuro</b>  | <b>52</b> |
| <b>A. Apéndice: Aspectos técnicos relevantes en el desarrollo de proyectos software</b>      | <b>58</b> |
| A.1. Mantenibilidad . . . . .  | 58        |
| A.2. ENS: Esquema Nacional de Seguridad . . . . .  | 58        |
| A.3. ENI: Esquema Nacional de Interoperabilidad . . . . .                                    | 58        |
| A.4. Reglamento General de Protección de Datos . . . . .                                     | 58        |
| A.5. Informe de técnicas de Search Engine Optimization . . . . .                             | 59        |
| A.6. Diseño de la navegabilidad . . . . .  | 59        |

## 1. Introducción

En este informe se presenta la Versión 2 del **Leaderboard ODESIA**, que se ha desarrollado en el marco del convenio UNED - ONTSI/Red.es para crear un *Espacio de Observación de Inteligencia Artificial en Español*. El objetivo general del proyecto es la comparación de la presencia y efectividad del español y el inglés en la Inteligencia Artificial (IA), con objeto de cuantificar la brecha entre ambas lenguas. Uno de los ámbitos de esa comparación es el del estado del arte y, dentro de este, es crucial conocer el rendimiento comparado de los sistemas en aplicaciones de Procesamiento del Lenguaje Natural sobre el inglés y

el español. Con este fin se ha creado el Lederboard ODESIA<sup>1</sup>, que proporciona una infraestructura de evaluación para modelos de lenguaje preentrenados en inglés y español que permite una comparación directa entre el rendimiento de modelos, y por tanto **medir la brecha de efectividad inglés-español** de los sistemas de Procesamiento del Lenguaje Natural (PLN).

Desde 2017/2018, los sistemas de PLN discriminativo se han estandarizado mucho, y en su inmensa mayoría consisten en la aplicación de un modelo de lenguaje preentrenado (LLM o Large Language Model) a cada problema concreto mediante un proceso conocido como fine-tuning. Los LLMs son redes neuronales profundas (casi todas ellas de una clase concreta, el transformer) que han aprendido uno o varios idiomas de forma auto-supervisada. El fine-tuning consiste en entrenar de forma supervisada una última capa de neuronas, sobre la representación que genera el modelo preentrenado de cada palabra del texto (así como del texto completo) en las capas anteriores.

Un testbed o benchmark para evaluar comparativamente modelos de lenguaje suele consistir en una colección diversa de tareas de Procesamiento del Lenguaje Natural, susceptibles de ser abordadas por los modelos de lenguaje mediante fine-tuning.<sup>2</sup> Para cada tarea se elige una métrica de evaluación pertinente, y se reportan tanto los resultados para cada tarea como el resultado agregado sobre todas ellas, que suele ser alguna forma de promedio. Un leaderboard de modelos de lenguaje no es más que una aplicación que permite a los desarrolladores enviar el output de los modelos para cada una de esas tareas y visualizar comparativamente sus resultados con respecto al resto de modelos. Si el conjunto de tareas del leaderboard tiene la calidad y diversidad suficientes, la agregación de los resultados de un modelo preentrenado para todas las tareas es un indicador de su calidad intrínseca, es decir, de lo bien que ha aprendido y generalizado las características del idioma (o idiomas) en la fase de entrenamiento auto-supervisado (self-supervised).

Desde la explosión de los modelos generativos a finales de 2022 (con la aparición de ChatGPT), se han continuado desarrollando conjuntos de tareas que permitan evaluar a los modelos de lenguaje, con dos diferencias notables: la primera, que son datasets destinados a evaluar las capacidades de los sistemas (en lugar de las tareas de PLN que pueden resolver): conocimiento general, razonamiento, conocimiento matemático, generación de código, etc. La segunda, que la evaluación de grandes modelos generativos se realiza en modo zero-shot o few-shot: ya no se proporciona un conjunto de entrenamiento para tarea, sino que se evalúa la capacidad de los modelos generativos para resolverlas a partir de la descripción de la tarea y/o unos pocos ejemplos.

Existen muchos leaderboards para evaluar LLMs en inglés: entre los más conocidos se encuentran GLUE (Wang et al., 2018a), SUPERGLUE (Wang et al., 2019a), HELM (Liang et al., 2022), Big-Bench (Srivastava, 2022), Big-Bench Hard (Suzgun et al., 2022), MMLU (Hendrycks et al., 2021), AGIEval (Zhong et al., 2023), Chatbot Arena (Chiang et al., 2024), AlpacaEval<sup>3</sup> o el leaderboard de HuggingFace para comparar modelos open-source<sup>4</sup>. Es difícil, sin embargo, encontrar leaderboards equivalentes en otros idiomas. Aunque hay leaderboards conocidos para el ruso<sup>5</sup> y para el chino<sup>6</sup> (Zeng, 2023; Gu et al., 2024; Huang et al., 2023), es más difícil encontrar leaderboards para idiomas del ámbito de la Unión Europea, aunque se están empezando a crear, como para el alemán<sup>7</sup>, el neerlandés<sup>8</sup>, el italiano (Basile et al., 2023) o el francés. En este último caso, por ejemplo, existe la iniciativa FLUE benchmark<sup>9</sup>, pero no permite enviar outputs para su evaluación, sino artículos científicos de los que los autores de FLUE extraen las métricas relevantes y las incluyen en su tabla de resultados.

Existen algunos leaderboards multilingües, que solo recientemente se han empezado a diseñar específicamente para medir la distancia de rendimiento entre lenguas. El benchmark XTREME (Cross-lingual

<sup>1</sup><http://leaderboard.odesia.uned.es/>

<sup>2</sup><https://www.ruder.io/nlp-benchmarking/>

<sup>3</sup>[https://github.com/tatsu-lab/alpaca\\_eval](https://github.com/tatsu-lab/alpaca_eval)

<sup>4</sup>[https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)

<sup>5</sup><https://russiansuperglue.com>

<sup>6</sup><https://www.cluebenchmarks.com>

<sup>7</sup><https://scandeval.com/german-nlu/>

<sup>8</sup><https://scandeval.com/dutch-nlu/>

<sup>9</sup><http://fluebenchmark.com>

TRansfer Evaluation of Multilingual Encoders) cubre 40 idiomas (tipológicamente diversos) de 12 familias distintas (Hu et al., 2020). Está diseñado para evaluar las capacidades de transferencia translingüe del aprendizaje de los modelos de lenguaje, es decir, la capacidad para resolver una tarea en un idioma de destino habiendo realizado el aprendizaje supervisado en otro idioma diferente. XTREME se centra en el aprendizaje en inglés y la evaluación en otros idiomas, así que no dispone de datos de entrenamiento equiparables entre idiomas, y por tanto no es adecuado para realizar evaluaciones de los modelos de cada idioma. Además, los datos de test son en su mayoría traducciones directas del inglés. Recientemente algunos trabajos se han enfocado en la comparación de modelos generativos entre idiomas, como en (Ahuja et al., 2023) donde se comparan 70 idiomas incluido el español, en (Bang et al., 2023) y (Lai et al., 2023) donde se evalúa en concreto ChatGPT en varios idiomas, o en (Zhang et al., 2023) donde se evalúan los LLMs en un contexto multilingüe, multimodal y multinivel con preguntas de exámenes reales. En el ámbito de las lenguas europeas destaca ScandEval (Nielsen, 2023)<sup>10</sup>, un leaderboard muy reciente para modelos preentrenados del noruego, el sueco y el danés; sin embargo, en el momento de escribir este informe no permite enviar outputs para su evaluación directa, sino que proporciona un paquete de software<sup>11</sup> para que los desarrolladores puedan evaluar sus propios modelos. Tampoco parece diseñado, en principio, para establecer una comparación directa entre el estado del arte en cada uno de los idiomas que contempla.

Con respecto al español, todavía no existe ningún leaderboard propiamente dicho para evaluar modelos de lenguaje. En general, cada nuevo modelo preentrenado es evaluado por sus autores de forma independiente, y la evaluación se suele incluir en el informe técnico o artículo científico que acompaña la publicación del modelo. En cada caso, los conjuntos de datasets de referencia, aunque con grado variable de solapamiento, son diferentes. El primer modelo para el español de adopción masiva, BETO (Cañete et al., 2020), fue evaluado sobre un conjunto de siete tareas conocido desde entonces como GLUES (GLUE español), y ese conjunto ha sido reutilizado y expandido en las evaluaciones asociadas a modelos posteriores como BERTin (De la Rosa et al., 2022), MarIA (Gutiérrez-Fandiño et al., 2021), RigoBERTa (Serrano et al., 2022), etc. Actualmente los grandes modelos generativos entrenados en cantidades masivas de datos demuestran habilidades en español, al contener varios idiomas en su entrenamiento (como GPT-3.5 y GPT-4 de OpenAI, Llama-2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024), Gemma y Gemini de Google (Akter et al., 2023) o Claude 3 de Anthropic<sup>12</sup>). Sin embargo es necesario contar con infraestructuras para poder evaluar este rendimiento en nuestro idioma, y poder compararlo respecto al inglés.

Nuestro objetivo principal es crear un leaderboard que permita evaluar comparativamente modelos de lenguaje en inglés y en español, de modo que se pueda estimar la distancia de rendimiento entre los modelos preentrenados de ambos idiomas en condiciones de igualdad respecto a la cantidad, calidad y comparabilidad de los datos de entrenamiento. Como objetivo secundario, y teniendo en cuenta la ausencia de leaderboards para el español, pretendemos que el leaderboard sea también una forma de medir el rendimiento de los modelos de lenguaje para el español con independencia de su comparabilidad con el inglés. En una evaluación reciente (Agerri and Agirre, 2022) se observó que no parece haber todavía modelos discriminativos para el español superiores a los modelos multilingües (en el caso de ese trabajo, a XLM-Roberta-Large), en parte por una falta de evaluación sistemática que incluya tanto a los modelos del español como a los multilingües.

En este documento describimos el trabajo realizado en el segundo año para el Leaderboard ODESIA, que se ha desarrollado a partir de la Versión 1 que se presentó en el informe del Año 1. Resumimos en la Tabla 1 los datasets y tareas que se han añadido en la Versión 2, así como los cambios en las funcionalidades y la interfaz.

Este informe se estructura de la siguiente manera: en la Sección 2 se describen los datasets desarrollados dentro del proyecto y sus tareas correspondientes, y el proceso de calibración de los datasets para poder

<sup>10</sup><https://scandeval.github.io/>

<sup>11</sup><https://github.com/saattrupdan/ScandEval>

<sup>12</sup>[https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bb618857627/Model\\_Card\\_Claude\\_3.pdf](https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bb618857627/Model_Card_Claude_3.pdf)

|   | Versión 1   | Añadido en Versión 2  |
|---|---|---|
| <b>ODESIA CORE</b><br>Tareas con partición de training y partición test privada                   | DIANN 2023 T1: Detección de discapacidades<br>DIPROMATS 2023 T1: Propaganda identification<br>DIPROMATS 2023 T2: Coarse propaganda characterization<br>DIPROMATS 2023 T3: Fine propaganda characterization<br>EXIST 2022 T1: Sexism detection<br>EXIST 2022 T2: Sexism categorisation | EXIST-2023 T1: Sexism Identification LeWiDi<br>EXIST-2023 T2: Source Intention LeWiDi<br>EXIST-2023 T3: Sexism Categorization LeWiDi<br>SQAC-SQUAD - 2024: Question answering           |
| <b>ODESIA EXTENDED</b><br>Tareas preexistentes con partición de training y partición test pública | MLDoc: clasificación de noticias<br>SQUAD/SQAC Question Answering (Machine Reading)<br>MultiCoNER 2022: entidades nombradas<br>STS 2017: similitud textual<br>DIANN T2: negación en ámbito biomédico  |   |
| <b>Tareas zero-shot o few-shot</b>  |   | UNED-ACCESO: exámenes de acceso a la universidad de respuesta múltiple<br>CURIA: resúmenes en texto claro de documentos legales (en curso)  |
| <b>Aplicación web</b>   | Browsing del leaderboard v1   | Remodelación completa de la interfaz y la navegación. Login de usuario y envío de sistemas. Integración en ecosistema ODESIA EvALL-Leaderboard-Portal. Incorporación del leaderboard v2 |

Tabla 1: Resumen de las extensiones de la Versión 2 del Leaderboard ODESIA respecto a la Versión 1.

realizar una medición directa de la brecha de efectividad entre los dos idiomas. A continuación, en la Sección 3 se describe la aplicación web Leaderboard ODESIA en su versión actual. En la Sección 4 se describe la experimentación con modelos de lenguaje discriminativos sobre los datasets ODESIA CORE, y en la Sección 5 la experimentación con el conjunto completo de datasets ODESIA EXTENDED. En la Sección 6 se describe el trabajo adicional hecho para evaluar modelos generativos, y se reportan resultados para el dataset UNED ACCESO en los que se comparan GPT-4 y Claude 3 con varios modelos abiertos. En la Sección 7 se esbozan las conclusiones principales del trabajo realizado hasta ahora.

## 2. Datasets y tareas de PLN usados para la experimentación

En esta sección se describen los criterios que se han seguido para seleccionar los datasets que se han usado para la experimentación y se explican los detalles de cada uno de los datasets, incluyendo las tareas de PLN que se evalúan con el dataset. Los datasets con tareas discriminativas y particiones de test privadas se agrupan en el banchmark ODESIA Core Tasks y se usan para medir la brecha de efectividad, mientras que los datasets con tareas discriminativas y particiones de test públicas se agrupan en el bechmark ODESIA Extended Tasks y se usan como una extensión para medir la brecha de efectividad. Este año se

han creado también datasets con tareas generativas que se usarán para medir la brecha de efectividad, y que de momento no se han incorporado al leaderboard.

## 2.1. Criterios de selección de datasets y tareas

- En cada dataset, los datos en inglés y en español se deben haber recolectado y anotado siguiendo la misma metodología, y el volumen de datos debe ser comparable entre ambos idiomas. No consideramos que cumplan este requisito los datasets que son traducción el uno del otro, ya sea automática o manual. En ambos casos pueden producirse sesgos en la evaluación (por supuesto, el supuesto de traducción automática introduce sesgos más acusados).
- En cada dataset, el subconjunto de test (sobre el que se evalúan los sistemas) no debe haber sido distribuido públicamente. De esta manera se evita el sobreajuste de datos y se minimiza la posibilidad de contaminación en los modelos (es decir, que el modelo haya visto las anotaciones manuales en la fase de preentrenamiento). Como [Sainz et al. \(2023\)](#) indican, la contaminación provoca una sobreestimación del rendimiento de un modelo contaminado con respecto a sus homólogos no contaminados. Desde un punto de vista científico las consecuencias son muy perjudiciales, ya que se publican conclusiones científicas erróneas.
- Para los datasets existentes, debe ser posible crear un subconjunto adicional de test privado utilizando la misma metodología con la que se construyó el dataset original. En este sentido, es preferible adaptar datasets en los que podamos contar con el equipo original que los desarrolló y los anotó.

Otros aspectos que hemos considerado son los siguientes:

- La selección de tareas debe estar más orientada hacia las aplicaciones de la tecnología que hacia evaluaciones puramente lingüísticas. Las tareas más lingüísticas son adecuadas para evaluar el grado de conocimiento que tienen los modelos sobre la lengua, pero tienen una relación menos directa con el comportamiento de las aplicaciones de IA usadas por ciudadanos y organizaciones. Nuestro interés es realizar mediciones que tengan algún tipo de correspondencia con el uso práctico de estas tecnologías, así que queremos priorizar las tareas más cercanas al mercado de aplicaciones.
- Las tareas deben tener un grado de dificultad realista. Por un lado, hay datasets que pueden resolverse con un grado muy alto de acierto, pero no porque las tareas sean realmente sencillas, sino porque los sistemas aprenden los sesgos del dataset. En esos casos, el comportamiento de los sistemas fuera del laboratorio (es decir, sobre conjuntos de datos que no tienen los sesgos del conjunto de entrenamiento) es mucho peor que en las evaluaciones de laboratorio. En el otro extremo se encuentran los "diagnostic datasets", en los que se escogen ejemplos para el conjunto de test que requieran un conocimiento profundo del lenguaje para poder resolverse. Este tipo de datasets son muy útiles para identificar los puntos débiles de los modelos de lenguaje y para mejorarlos, pero son más difíciles que las situaciones promedio que nos encontramos en entornos de aplicación, de forma que tampoco ofrecen una imagen realista del rendimiento de las aplicaciones de Procesamiento del Lenguaje Natural. Nos interesa utilizar datasets en los que el rendimiento de los modelos de lenguaje se acerque al que encontraríamos en entornos reales de aplicación.
- Dificultad similar entre idiomas. Hemos establecido mecanismos para calibrar el leaderboard en función de la dificultad relativa intrínseca de los datasets en inglés y español, de manera que si para una tarea determinada hay una dificultad intrínseca (independiente del conocimiento del lenguaje que tengan los sistemas) diferente entre los dos idiomas, somos capaces de medirla y recalibrar la evaluación para que no contamine las diferencias observadas en el conjunto de test entre los dos idiomas. Sin embargo, conviene descartar los datasets en los que la dificultad intrínseca entre los dos idiomas es muy grande, porque esa es una señal de que la metodología con la que se han seleccionado y/o anotado en ambos idiomas seguramente no es equivalente.

- Los datasets y las tareas deben tener diversidad para ser representativos de las distintas aplicaciones del Procesamiento del Lenguaje Natural. Nos centraremos en tareas discriminativas que sean susceptibles de ser abordadas directamente por los modelos de lenguaje: en particular, de clasificación y de etiquetado. En el segundo año hemos empezado a trabajar también en tareas generativas por ser especialmente relevantes en el mercado actual de la IA, aunque su evaluación es mucho más compleja: o se invierte mucho esfuerzo en evaluaciones no automatizables, o se evalúan de forma poco precisa con medidas automatizadas de similitud textual con modelos realizados por humanos. También es deseable cierta diversidad en los dominios y en el tipo de textos.
- Accesibilidad. Los datos de entrenamiento deben ser fáciles de conseguir para los desarrolladores, idealmente mediante un sólo acuerdo centralizado con la UNED como proveedora.

Este conjunto de requisitos no se cumple en la mayoría de datasets disponibles, especialmente dada la necesidad que tenemos de expandirlos para disponer de un juego de pruebas privado que no haya sido publicado.

Los siguientes datasets se crearon para el leaderboard en el primer año de trabajo, y constituyen el leaderboard ODESIA CORE. En todos ellos se ha realizado al menos una anotación adicional para disponer de un juego de pruebas privado que no pueda ser leído por los modelos de lenguaje en su proceso de entrenamiento, garantizando así que la evaluación está libre de problemas de contaminación:

- DIPROMATS 2023. Este dataset fue creado desde cero para incorporarlo al Leaderboard ODESIA en la Versión 1. Se trata de un conjunto de tuits emitidos por diplomáticos de cuatro potencias mundiales (la Unión Europea, Rusia, China y Estados Unidos), anotados en función de las técnicas de propaganda que utilizan para transmitir una imagen determinada de sus países o de sus competidores a nivel global. Hay tres tareas asociadas con este dataset: identificación de propaganda, caracterización a grano grueso (cuatro técnicas) y caracterización a grano fino (15 técnicas subsumidas en las anteriores). Se trata de un problema de clasificación multiclase y multietiqueta. Se enmarca dentro de los problemas relacionados con la desinformación.
- EXIST 2022. Este dataset fue extendido para su incorporación al leaderboard en la Versión 1, creando un subconjunto adicional de datos anotados como conjunto de test privado. Se trata de un conjunto de tuits anotado en función de si contienen mensajes sexistas o no, y de qué tipo de sexismo se trata. Se enmarca dentro del problema de la toxicidad en redes sociales.
- DIANN 2023. Este dataset fue adaptado y extendido para su incorporación al Leaderboard. Se creó una partición de evaluación para la Versión 1 del Leaderboard. Los textos son resúmenes de artículos sobre biomedicina, y la tarea consiste en identificar menciones de discapacidades. Se trata por tanto de una tarea de etiquetación de secuencias.

En la Versión 2 del Leaderboard ODESIA CORE se han incorporado los siguientes datasets:

- EXIST 2023. Se trata de un dataset creado en su integridad para la Versión 2 del Leaderboard. Se compone de tuits etiquetados en función del tipo de sexismo expresado o descrito en ellos. Se trata, además, de un dataset desarrollado siguiendo el paradigma de “aprendizaje con desacuerdo” (Learning with Disagreement, LeWiDi) (Uma et al., 2021a), lo que lo convierte en el primer dataset para el entrenamiento y prueba de sistemas de detección de sexismo en textos construido conforme a este paradigma. Consta de tres particiones (entrenamiento, desarrollo, evaluación) y anotaciones para tres tareas: Detección de sexismo, categorización e identificación del emisor de sexismo. Se enmarca dentro del problema de la toxicidad en redes sociales.
- SQUAD/SQAC 2024. Este dataset contiene una partición de evaluación creada para la Versión 2 del Leaderboard ODESIA. Contiene artículos de divulgación científica del CSIC para el español y de Cambridge University para el inglés. La tarea que este dataset permite evaluar es la de comprensión de texto extractiva en sistemas de pregunta-respuesta. La tarea consiste en responder a preguntas sobre un

texto, de tal manera que la respuesta sea un fragmento extraído directamente del texto. Se trata de una tarea de etiquetación de secuencias. El hecho de que los documentos anotados en los SQUAD/SQAC originales sean de fuentes distintas a los de nuestra anotación hace que, desde el punto de vista de los sistemas supervisados, este dataset sea particularmente complejo, ya que implícitamente se está midiendo la capacidad de transferencia del aprendizaje entre dominios. Además, es un dataset particularmente apropiado para evaluar modelos generativos en modo zero-shot (sin ejemplos) o few-shot (unos pocos ejemplos); de hecho, una de las motivaciones para incluirlo este año ha sido su proximidad a las aplicaciones más comunes a nivel empresarial de los modelos generativos: la capacidad de los modelos para extraer y sintetizar información de información corporativa en formato texto o semiestructurado, en una aproximación RAG (Retrieval-Augmented Generation) no supervisada.

Adicionalmente, se ha desarrollado otros datasets pensados exclusivamente para evaluar modelos generativos en modo zero-shot o few shot. De momento no se han añadido al leaderboard, ya que los modelos discriminativos no pueden abordarlos.

- **UNED ACCESO 2024.** El dataset contiene 1003 preguntas de opciones múltiples de once asignaturas del Curso de acceso para Mayores de 25 años de la UNED. Las preguntas y sus respuestas se han traducido manualmente al inglés (sin intervención de ningún sistema de traducción automática, para evitar sesgos). Este dataset permite evaluar el conocimiento general de los modelos generativos, de forma similar a otros datasets como MMLU. Las diferencias con MMLU son: la disponibilidad de las preguntas en dos idiomas mediante traducciones manuales, y que el dataset no se hace público, de forma que se limitan los problemas de contaminación. Sobre este dataset se ha finalizado una experimentación exhaustiva sobre los mejores modelos generativos actuales (Claude 3 Opus y GPT-4) y sobre varios modelos abiertos (Llama-2, Mistral y Gemma).
- **CURIA 2024.** Este dataset contiene particiones de entrenamiento, desarrollo y evaluación. Los textos que lo conforman son sentencias de tribunales de justicia de la Unión Europea, que están acompañados de micro-resúmenes en lenguaje claro. En este caso, la tarea que permite evaluar el dataset CURIA-2024 es el resumen simplificado de textos legales, por tanto, se trata de una tarea de generación de texto. El dataset está finalizado y la experimentación se realizará en el tercer año de proyecto.
- **Pron vs Prompt.** El dataset consiste en 120 sinopsis para 60 títulos de películas imaginarias, 30 propuestos por GPT4 y 30 por un novelista de prestigio (Patricio Pron, premio Alfaguara de novela). Se solicitó a ambos, al escritor y GPT-4, que escribieran sinopsis de aproximadamente 600 palabras para cada título, incluyendo tanto los propuestos por ellos mismos como por su contraparte. En este caso, se está realizando una evaluación sistemática por parte de críticos y académicos, con la que se espera medir de forma precisa y fiable las capacidades de escritura creativa de los modelos de forma puntual.

En conjunto, los 5 datasets del Leaderboard que tienen tests privados (ODESIA CORE) aportan 10 tareas distintas que abarcan problemas de clasificación (binaria, multiclase, multietiqueta, jerárquica, clasificación con disagreement) y etiquetado de secuencias. Hay varios tipos de textos: tuits en redes sociales, mensajes de autoridades y diplomáticos, resúmenes científicos y noticias de divulgación científica. Y en cuanto a dominios, se abarcan el biomédico, el de política y relaciones internacionales, las redes sociales, y dominios científicos variados.

Los detalles de cada dataset se especifican en sus informes técnicos correspondientes. A continuación se presenta un resumen de cada uno de ellos.

## 2.2. Dataset 1: EXIST 2022

EXIST (sEXism Identification in Social neTworks) 2022 es un dataset desarrollado para facilitar la investigación en detección automática de sexismo en redes sociales. Se compone de textos cortos

| Dataset         | Tareas                                 | Tarea abstracta                            | Dominio        | Área de aplicación  |
|-----------------|--|--|----------------|---------------------|
| DIANN 2023      | detección de discapacidades            | etiquetado                                 | Biomedicina    | Entidades nombradas |
| DIPROMATS 2023  | Identificación de propaganda           | Clasificación binaria                      | Geopolítica    | Desinformación      |
|                 | caracterización de propaganda (gruesa) | Clasificación jerárquica multilabel        | Geopolítica    | Desinformación      |
|                 | Caracterización de propaganda (fina)   | Clasificación jerárquica multilabel        | Geopolítica    | Desinformación      |
| EXIST 2022      | Detección de sexismo                   | Clasificación binaria                      | Redes sociales | Toxicidad           |
|                 | Categorización de sexismo              | Clasificación jerárquica multiclase        | Redes Sociales | Toxicidad           |
| EXIST-2023      | Identificación de sexismo              | Clasificación binaria LeWiDi               | Redes sociales | Toxicidad           |
|                 | Intención de la fuente                 | Clasificación jerárquica LeWiDi            | Redes Sociales | Toxicidad           |
|                 | Categorización de sexismo              | Clasificación jerárquica multilabel LeWiDi | Redes sociales | Toxicidad           |
| SQUAD-SQAC 2024 | Machine Reading                        | etiquetado                                 | Ciencia        | Pregunta-respuesta  |

Tabla 2: Resumen de los datasets usados en el Leaderboard ODESIA CORE v2.

procedentes de redes sociales etiquetados en función del tipo de sexismo expresado o descrito en ellos. Contiene datos de dos redes sociales diferentes: Twitter<sup>13</sup> y Gab<sup>14</sup>. Se trata, por tanto, de mensajes cortos intercambiados en alguna de las dos redes anteriores y de dominio general (es decir, no versan, a priori, sobre ninguna temática en particular).

En total, el dataset se compone de 12,390 textos etiquetados: 6,226 en español y 6,164 en inglés. La distribución de los textos por partición (entrenamiento/test), fuente (Twitter/Gab) e idioma (inglés/español) se muestra en la Tabla 7. Sobre ellos se definen dos tareas:

- Tarea 1: detección de sexismo (clasificación binaria). Los sistemas deben decidir, para cada tweet, si contiene mensajes sexistas o no.
- Tarea 2: detección y caracterización de sexismo (clasificación multiclase). Los sistemas deben decidir, para cada tweet, si es o no sexista, y en caso afirmativo qué tipo de sexismo, entre las siguientes categorías: ideological and inequality, stereotyping and dominance, objectification, sexual violence, misogyny and non-sexual violence.

Cada texto del dataset tiene asignadas 1 o 2 etiquetas, dependiendo de si se trata de un texto sexista o no. La primera etiqueta responde a la pregunta: *¿Es el texto sexista, en cualquiera de sus formas, o describe conductas o situaciones en las que se produce discriminación sexista (es decir, es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista)?* En función de la respuesta a esta pregunta, la etiqueta puede tomar uno de los siguientes dos valores:

- **SEXIST**: el tuit es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista. Ejemplos de textos del dataset anotados con esta etiqueta son:

<sup>13</sup><https://twitter.com/>

<sup>14</sup><https://gab.com/>

| Entrenamiento |        |         |        | Test    |        |
|---------------|--------|---------|--------|---------|--------|
| Twitter       |        | Gab     |        | Twitter |        |
| Español       | Inglés | Español | Inglés | Español | Inglés |
| 5,211         | 5,152  | 490     | 492    | 522     | 513    |
| 10,363        |        | 982     |        | 1,035   |        |

Tabla 3: Distribución de datos de EXIST 2022 por partición, fuente e idioma.

- (1) *Que materialistas se han vuelto las mujeres de hoy en día, aún recuerdo cuando las podíamos enamorar con puras mentiras.*
  - (2) *Las mujeres no deberían ni maquillarse. Ya sabemos que son rompe bolas arregladas o desarregladas. Joda las amamos. JAJAJA.*
  - (3) *Te tacharán de machista y misógino. Las mujeres pueden pegar los hombres no.*
- NON-SEXIST: el tuit no perjudica, discrimina o menosprecia a las mujeres ni se refiere a contenidos o situaciones en las que sí se haga. Ejemplos de textos del dataset anotados con esta etiqueta son:
    - (4) *Mi mamá: Si te llegas a ir del país, vas a saber lo dura que es la vida, porque ni siquiera te dignas a fregar las cosas que usas.*
    - (5) *Alguien por favor me regala un bate que yo entre los fachas los señores de las nuevas masculinidades y los que mucho feminismo pero el porno no me lo toqueis quiero reventar una cabeza.*
    - (6) *El impactante video de una mujer al volante que sale milagrosamente ilesa de un intento de asesinato...*

La segunda etiqueta responde a la pregunta: *De acuerdo con la faceta de la mujer que está siendo atacada, ¿qué tipo de sexismo encontramos en el tuit?*. El sexismo puede afectar a las mujeres en muchas facetas de sus vidas, incluidas las funciones domésticas y de crianza, las oportunidades profesionales, la imagen sexual y las expectativas de vida, por nombrar algunas. En función de la respuesta, la etiqueta puede tomar uno de los siguientes valores:

- IDEOLOGICAL-INEQUALITY: Descrédito ideológico, negación de la desigualdad y narrativa invertida. El tuit desacredita el movimiento feminista, rechaza la existencia de desigualdad entre hombres y mujeres, o presenta a los hombres como víctimas de la opresión de género. Ejemplos de tuits en esta categoría son:
  - (7) *Mucho feminismo pero poco interés con casos de pedofilia, que hipócrita de mierda.*
  - (8) *Mucho feminismo mucho feminismo pero andamos tirando historias burlándose de una piba solo porque el novio le metió los cuernos y volvió. Gente hipócrita en esta vida existen y muchas.*
  - (9) *Mi hermana y mi madre se burlan de mí por defender todo el tiempo los derechos de todos y me acaban de decir feminazi, la completaron.*
- STEREOTYPING-DOMINANCE: Estereotipos y dominancia. El tuit expresa ideas falsas sobre las mujeres que sugieren que son más adecuadas o inapropiadas para ciertas tareas o afirma que los hombres son superiores a las mujeres. Ejemplos de textos del datasets con esta etiqueta son:
  - (10) *Un taxista enojado con una mujer que pasaba en bicicleta le gritó “mujer tenias que ser” creo que la quiso insultar pero no se alto orgullo serlo bobo.*
  - (11) *@RuloTanki @PrinncessRoyal Uy andas en tus días? Jajajja.*
  - (12) *Paula2R faber\_acuria A las mujeres hay que amarlas...solo eso... Nunca las entenderás.*

- **OBJECTIFICATION:** Objectificación. El tuit presenta a las mujeres como objetos al margen de su dignidad y personalidad, o asume o describe ciertas cualidades físicas que las mujeres deben tener para cumplir con los roles tradicionales de género. Ejemplos de textos anotados con esta etiqueta son:
  - (13) *@EveryWhereTraps Tienes buen tamaño para ser chica, más tamaño ya no encuentras casi ropa y menos zapatos bonitos.*
  - (14) *¿Cómo iba vestida?”, “seguramente estaba provocando”, “mmm ese short esta muy corto”, “no te pongas eso que pareces una puta”, “bueno iba sola por la calle a esa hora”, “se lo busco”, “bueno, para que toma entonces”, “le gustó”, “¿porque no gritó?*
  - (15) *Pareces una puta con ese pantalón - Mi hermano de 13 cuando me vio con un pantalón de cuero.*
- **SEXUAL-VIOLENCE:** Violencia sexual. Se realizan sugerencias sexuales, solicitudes de favores sexuales o acoso de carácter sexual (violación o agresión sexual). Ejemplos de textos anotados con esta etiqueta son:
  - (16) *@giulianarob @Gattarok @DemenArt ¿Misógino? Misoginia es odio a las mujeres y, en el porno los hombres las hacen disfrutar mucho.*
  - (17) *#MeToo Estas 4 no han conseguido su objetivo. El juez estima que se abrieron de patas <https://t.co/GSHiiwqY6A> tanta lagartona hay en este <https://t.co/8t5VmFIUFn>.*
- **MISOGYNY-NON-SEXUAL-VIOLENCE:** Misoginia y violencia no sexual. El tuit expresa odio y violencia hacia la mujer. Ejemplos de textos anotados con esta etiqueta son:
  - (18) *Las mujeres de hoy en día te enseñan a querer... estar soltero.*
  - (19) *Cualidades = odio a las mujeres, extranjeros, trabajadores, catalanes, vascos...*
  - (20) *Odio la misoginia más de lo que odio a las mujeres.*

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset EXIST 2022", que fue entregado en el Año 1.

### 2.3. Dataset 2: DIPROMATS 2023

DIPROMATS consiste en un conjunto de tweets en español y en inglés emitidos por diplomáticos y autoridades de cuatro potencias mundiales: EEUU, Rusia, China y la UE. Sobre ellos se ha realizado una anotación (de expertos) en la que se identifican los tweets que contienen algún tipo de propaganda, y una caracterización de las técnicas de propaganda usadas, tanto en grano grueso (cuatro categorías) como en grano fino (quince subcategorías). El dataset contiene 24,248 tweets anotados para las tres tareas, y se utiliza dentro del proyecto de dos formas: como parte del leaderboard bilingüe para evaluación comparada de modelos del lenguaje en español e inglés, y como parte de los datasets utilizados para la medición de la brecha de la IA entre ambos idiomas.

La Tabla 4 muestra la distribución de tuits y autoridades por áreas geopolíticas para el inglés, y la Tabla 5 para el español.

|             | China | Rusia | UE    | EEUU  | Total  |
|-------------|-------|-------|-------|-------|--------|
| Tuits       | 3,647 | 3,591 | 3,553 | 3,956 | 14,747 |
| Autoridades | 106   | 111   | 186   | 216   | 619    |

Tabla 4: Distribución del dataset DIPROMATS 2023 en inglés por áreas geopolíticas

Utilizando las anotaciones manuales, se han definido tres tareas:

- **Tarea 1: Identificación de propaganda**, que se plantea como un problema de clasificación binaria que consiste en determinar si un tuit contiene o no técnicas de propaganda.

|             | China | Rusia | UE    | EEUU  | Total |
|-------------|-------|-------|-------|-------|-------|
| Tuits       | 2,997 | 1,391 | 2,465 | 2,738 | 9,501 |
| Autoridades | 25    | 22    | 48    | 40    | 135   |

Tabla 5: Distribución del dataset DIPROMATS 2023 en español por áreas geopolíticas

- **Tarea 2: Caracterización de la propaganda (grano grueso)**, que tiene como objetivo categorizar los mensajes propagandísticos según el tipo de propaganda que contienen. La categorización propuesta considera múltiples técnicas identificadas mediante revisión de la literatura existente, que se agrupan según sus características retóricas. Proponemos una tarea de clasificación multiclase y multietiqueta, en la que los sistemas tienen que asignar los tuits a una o más de las categorías definidas. Hay dos tipos de categorías:
  - Una categorización de **grano grueso** con cuatro clases de propaganda (más una clase negativa):
    - Grupo 0: Not propaganda
    - Group 1: Appeal to Commonality
    - Grupo 2: Discrediting the opponent
    - Group 3: Loaded Language
    - Grupo 4: Appeal to authority
- **Tarea 3: Caracterización de la propaganda (grano fino)**, idéntica a la anterior pero con un conjunto de subcategorías que refinan la clasificación anterior. En este caso se distinguen 15 subtipos de propaganda: Flag Waving, Ad Populum / Ad antiquitatem, Name Calling, Undiplomatic Assertiveness / Whataboutism, Scapegoating, Propaganda Slinging, Appeal to Fear, Demonization, Personal Attacks, Doubt, Reductio Ad Hitlerum, Loaded Language, Appeal to False Authority y Bandwagoning.

Como parte del leaderboard, el dataset tiene varias características interesantes:

- Se trata de un problema complejo (especialmente al nivel de caracterización de grano fino), difícil de resolver por cualquier sistema de PLN. Frente a otros datasets, éste tiene la ventaja de que no saturará fácilmente (momento en el cual no sirve para medir diferencias entre sistemas).
- Se trata de un problema de clasificación jerárquico, multiclase y multietiqueta. Aunque este tipo de problemas es muy habitual en aplicaciones prácticas de la IA, en entornos de laboratorio suelen simplificarse ignorando las características jerárquicas de las clases, o reduciendo el problema artificialmente a una sola etiqueta por ítem.
- Por tratarse de fuentes diplomáticas, abarca una gran variedad de registros dialectales, desde los propios del país donde trabaja cada diplomático hasta su grado de bilingüismo (hay diplomáticos que son nativos en el idioma del país de destino, otros lo han adquirido como segunda lengua).
- Se trata de una anotación de expertos, más costosa que el crowdsourcing pero que permite el uso de tipologías de anotación más sofisticadas, como es el caso de DIPROMATS.

Los detalles del dataset DIPROMATS 2023 están recogidos en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español: Informe del dataset DIPROMATS", entregado en el Año 1.

## 2.4. Dataset 3: DIANN 2023

El dataset DIANN-2023 se ha compilado y anotado en la UNED en el marco del proyecto. Consiste en un dataset bilingüe de resúmenes de artículos científicos relacionados con enfermedades raras anotados manualmente con discapacidades. El dataset ha sido concebido con el objetivo de entrenar sistemas de reconocimiento de entidades nombradas especializados en la detección de discapacidades. Una parte del corpus se creó en 2018 para la competición de IberLEF “Disability annotation on documents from the biomedical domain (DIANN)”<sup>15</sup> (Fabregat et al., 2018), que proponía una tarea de reconocimiento de entidades centrada en la identificación de discapacidades. Otra parte se ha creado en 2023 con el fin de incorporarla como partición privada de test al leaderboard del proyecto ODESIA. La anotación de esta última parte es la que ha sido financiada por el proyecto.

El corpus se proporciona en dos particiones, una de entrenamiento y otra de evaluación. La partición de entrenamiento contiene 500 textos en cada lengua. Estos textos se corresponden con las particiones de entrenamiento y evaluación hechas públicas para la competición DIANN en Iberlef 2018, donde se proporcionaban 400 archivos de entrenamiento y 100 de evaluación por lengua. Además se dispone de una partición privada de test que contiene 100 textos para cada lengua. Puesto que esta es la partición que se usa para evaluar sistemas en el leaderboard, esta partición no se hará pública y no se proporciona información sobre sus contenidos más allá de la información referente al tamaño y la metodología de anotación.

En la Tabla 6 se muestran los detalles de tamaño para ambas particiones y el número anotaciones en la partición de entrenamiento. Los tokens se han calculado contando unidades separadas por espacios en blanco.

| Partición | Entrenamiento |          |          |                  | Evaluación |          |          |
|-----------|---------------|----------|----------|------------------|------------|----------|----------|
| Idioma    | # docs        | # tokens | # líneas | # discapacidades | # docs     | # tokens | # líneas |
| Español   | 500           | 98948    | 5923     | 1555             | 100        | 21103    | 1203     |
| Inglés    | 500           | 89325    | 6901     | 1656             | 100        | 19087    | 1234     |

Tabla 6: Información sobre el tamaño del corpus DIANN-2023.

En el corpus se han anotado todas las discapacidades mencionadas en los textos. Por tanto, la única etiqueta usada es la de ‘discapacidad’. La distribución por lengua en la partición de entrenamiento es la siguiente:

- **Español:** 1555 menciones de discapacidades anotadas.
- **Inglés:** 1656 menciones de discapacidades anotadas.

Dentro del leaderboard, este dataset es representativo de uno de los dominios de aplicación más relevantes del PLN, el dominio biomédico; y de las tareas de etiquetado, que son las más prototípicas del PLN discriminativo junto con las de clasificación.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español, Informe del dataset DIANN-2023", entregado en el Año 1.

## 2.5. Dataset 4: EXIST-2023

En la línea de EXIST 2022, EXIST (sEXism Identification in Social neTworks) 2023 es un dataset desarrollado para facilitar la investigación en detección automática de sexismo en redes sociales. Se compone de textos cortos, tuits, procedentes de redes sociales etiquetados en función del tipo de sexismo expresado o descrito en ellos. A diferencia de EXIST 2022, se trata de un dataset desarrollado siguiendo el paradigma de “aprendizaje con desacuerdo” (Learning with Disagreements, LeWiDi) (Uma et al., 2021a), lo que lo convierte en el primer dataset para el entrenamiento y prueba de sistemas de detección de

<sup>15</sup>Página web de la competición DIANN 2018: <http://nlp.uned.es/diann/>

sexismo en textos construido conforme a este paradigma, y el primer dataset de este tipo que se incorpora al Leaderboard ODESIA.

En este paradigma, en lugar de depender de una única etiqueta “correcta” para cada ejemplo o instancia del dataset, el modelo se entrena para aprender de anotaciones conflictivas o diversas. De esta manera, las perspectivas, sesgos o interpretaciones diferentes de los anotadores pueden ser tenidas en cuenta por los sistemas, permitiendo un aprendizaje más justo y equitativo. Esto se deriva del hecho de que el dataset ha sido anotado por diferentes anotadores, con distintas características sociodemográficas, de manera que todas las notaciones (aun siendo en algunos casos contradictorias) forman parte del dataset final.

El dataset se compone de 10,034 textos etiquetados, 5,307 en español y 7,727 en inglés. Los textos proceden de conversaciones de Twitter. Cada texto está anotado con diferentes tipos de etiquetas. Como en EXIST 2022, en EXIST 2023 la primera etiqueta responde a la pregunta: *¿Es el texto sexista, en cualquiera de sus formas, o describe conductas o situaciones en las que se produce discriminación sexista (es decir, es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista)?* y la clasificación es binaria. A diferencia de EXIST 2022, en EXIST 2023 la segunda etiqueta responde a la pregunta: *¿Cuál crees que es la intención de la persona que escribió el tweet?* Las categorías anotadas son: DIRECT, REPORTED, JUDGEMENTAL. La tercera etiqueta responde a la pregunta: *De acuerdo con la faceta de la mujer que está siendo atacada, ¿qué tipo de sexismo encontramos en el tweet?* y las categorías son las mismas que en EXIST 2022.

El dataset presenta tres particiones para cada lengua: entrenamiento, desarrollo y test. La distribución de los textos por partición e idioma se muestra en la Tabla 7.

|                | Entrenamiento | Desarrollo | Test  | Total  |
|----------------|---------------|------------|-------|--------|
| <b>Español</b> | 3,660         | 549        | 1,098 | 5,307  |
| <b>Inglés</b>  | 3,260         | 489        | 978   | 7,727  |
| <b>Total</b>   | 6,920         | 1,038      | 2,076 | 10,034 |

Tabla 7: Distribución de EXIST 2023 por partición e idioma.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Año 2 - Informe del dataset EXIST 2023", entregado con este informe.

## 2.6. Dataset 5: SQUAD/SQAC-2024

En este caso, la tarea que este dataset permite evaluar es la de comprensión de texto en sistemas de pregunta-respuesta con respuestas extractivas. La tarea consiste en responder a preguntas sobre un texto, de tal manera que la respuesta sea un fragmento extraído directamente del texto. Los textos son noticias del CSIC (para español) y de Cambridge University (para inglés). En todos los casos, las respuestas son fragmentos del texto y no se incluyen preguntas que no se puedan contestar a partir del texto. Esta tarea es interesante por su dificultad, ya que requiere tanto entender el lenguaje como tener una representación del conocimiento del mundo en general y del mundo representado en cada texto.

SQUAD/SQAC-2024 es una extensión de los datasets SQUAD/SQAC. SQAC (Spanish Question Answering Corpus) (Gutiérrez-Fandiño et al., 2021) es un dataset de pregunta-respuesta, con respuestas extractivas en español. En la tarea de PLN asociada, dada una pregunta y un párrafo, el sistema debe localizar el span (fragmento) más pequeño que contiene la respuesta. La metodología para crearlo está basada en la de SQuAD (Stanford Question Answering Dataset) v1.1 (Rajpurkar et al., 2016), un dataset de pregunta-respuesta extractivo en inglés. Si bien el dataset SQAC se creó por la necesidad de tener un corpus de pregunta-respuesta en español que no fuera una traducción del inglés, el dataset SQUAD/SQAC-2024 se ha creado para disponer de una partición de evaluación privada que permita evaluar Large Language Models (LLMs) sin riesgo de contaminación de datos, tanto en inglés como en español. Por tratarse de textos obtenidos de fuentes distintas, este dataset es particularmente complejo, ya que se requiere cierto grado de transferencia entre lo aprendido con unas fuentes y lo aplicado a otras.

El dataset contiene noticias académicas del CSIC (Centro Superior de Investigaciones Científicas)

para el español<sup>16</sup> y de Cambridge University para el inglés.<sup>17</sup> Las noticias son de dominios científicos variados y suelen ser cortas, entre 712 y 2,760 palabras en inglés, y entre 514 y 2,818 palabras en Español. Además, están dirigidas al público general, por lo que no se usa lenguaje especializado.

En la Tabla 8 se proporciona información sobre el tamaño del dataset por idioma.

|                | # textos | # tokens  | $\mu$ tokens/texto | # pares pregunta-respuesta | $\mu$ preguntas/texto |
|----------------|----------|-----------|--------------------|----------------------------|-----------------------|
| <b>Español</b> | 110      | 962,502   | 840                | 1,144                      | 10,4                  |
| <b>Inglés</b>  | 110      | 1,235,638 | 1,045              | 1,182                      | 10,7                  |
| <b>Total</b>   | 220      | 2,198,140 | —                  | 2,379                      | —                     |

Tabla 8: Información sobre el tamaño del dataset por idioma.

Las unidades que conforman el dataset son pares de pregunta-respuesta. Hay 1,144 pares en español y 1,182 pares en inglés, realizados sobre 110 textos en cada lengua. Los textos en español son un poco más cortos, con una media de 840 palabras, en comparación con los textos en inglés, con una media de 1,045 palabras. Se ha realizado una media de aproximadamente 10 preguntas por texto tanto en inglés como en español.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Año 2 - Informe del dataset SQUAD/SQAC-2024", entregado con este informe.

## 2.7. Dataset 6: CURIA-2024

CURIA-2024 es un dataset compuesto por sentencias de tribunales en inglés y en español, con sus correspondientes resúmenes simplificados en lenguaje claro. Se enmarca, por tanto, en el dominio jurídico. El dataset se ha elaborado con textos descargados de la página web del Tribunal de Justicia de la Unión Europea. Este organismo está formado por el Tribunal General (de primera instancia) y el Tribunal de Justicia. Todas las sentencias de ambos tribunales son publicadas en sus páginas web.<sup>18</sup> Aparte de las sentencias, CURIA-2024 contiene los micro-resúmenes de las sentencias, que han sido creados por profesionales expertos en lenguaje legal.

Puesto que el dataset se compone de sentencias con sus resúmenes, la unidad de referencia es el par texto/resumen. Como se muestra en la Tabla 9, el dataset en inglés contiene 2,230 pares texto/resumen sumando un total de 18,153,858 tokens, mientras que el dataset en español contiene 1,961 pares sumando un total de 17,842,388 tokens. El dataset está dividido en tres particiones por lengua: entrenamiento, desarrollo y evaluación. Las particiones se han creado siguiendo el mismo procedimiento para las dos lenguas: se ha tomado un 80 % para entrenamiento, un 10 % para desarrollo y un 10 % para evaluación.

| Partición | Entrenamiento |            | Desarrollo |           | Evaluación |           | Total  |            |
|-----------|---------------|------------|------------|-----------|------------|-----------|--------|------------|
| Idioma    | # docs        | # tokens   | # docs     | # tokens  | # docs     | # tokens  | # docs | # tokens   |
| Español   | 1,568         | 14,173,589 | 196        | 1,820,298 | 197        | 1,848,501 | 1,961  | 17,842,388 |
| Inglés    | 1,784         | 14,461,442 | 223        | 1,724,975 | 223        | 1,967,441 | 2,230  | 18,153,858 |

Tabla 9: Número de pares sentencia-resumen en el dataset CURIA-2024.

Mediante este dataset se pueden evaluar sistemas de resumen simplificado de textos legales, por lo que se trata de una tarea de generación de texto.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset CURIA-2024".

<sup>16</sup><https://www.csic.es/es/actualidad-del-csic/noticias>

<sup>17</sup><https://www.cam.ac.uk/news>

<sup>18</sup>[https://curia.europa.eu/jcms/jcms/j\\_6/es/](https://curia.europa.eu/jcms/jcms/j_6/es/)

## 2.8. Dataset 7: UNED ACCESO 2024

Muchos benchmarks se han propuesto como evaluaciones de una sola tarea, pero con el surgimiento de modelos de lenguaje generales como BERT (Devlin et al., 2018a), se ha popularizado el desarrollo de benchmarks más completos para poder medir las capacidades generales de estos modelos. GLUE (Wang et al., 2018b) y SuperGLUE (Wang et al., 2019b) son benchmarks populares que evalúan el rendimiento de los modelos de lenguaje con distintas tareas de PLN. Más recientemente se han introducido benchmarks que contienen una amplia gama de tareas de PLN para la evaluación, como Big-Bench (Srivastava, 2022), Big-Bench Hard (Suzgun et al., 2022), MMLU (Hendrycks et al., 2021) y HELM (et al, 2023). La mayoría de los datasets presentes en estos benchmarks proponen evaluaciones que se realizan con conjuntos de datos creados artificialmente para tareas concretas de PLN, en vez de proponer escenarios reales de evaluación como los exámenes con los que se evalúa a los humanos. El reconocimiento de esta limitación ha llevado a que en los últimos tiempos se haya puesto énfasis en la importancia de las evaluaciones centradas en evaluar capacidades humanas. Se han introducido así benchmarks como AGIEval (Zhong et al., 2023), que se centran en un tipo de evaluación que pone el énfasis en tareas cognitivas a nivel humano, en escenarios reales. Este benchmark incluye exámenes de acceso a la universidad, pruebas de admisión a la facultad de derecho, concursos de matemáticas y pruebas de cualificación de abogados. Por otro lado, se han desarrollado diversos datasets a partir de exámenes, que abarcan distintas tareas (no sólo de respuesta múltiple), como en RACE (Lai et al., 2017). Las preguntas de respuesta múltiple se han erigido como uno de los métodos preferidos para evaluar los nuevos modelos generativos, debido a la dificultad que presenta su evaluación y la ausencia de una métrica estándar.

Así, en la intersección entre las evaluaciones con preguntas de múltiple respuesta, y las evaluaciones centradas en capacidades humanas y basadas en exámenes, se enmarca el dataset UNED ACCESO 2024. El dataset se compone de preguntas tipo test con 3 o 4 respuestas extraídas de exámenes de los Cursos de Acceso para Mayores de 25 años de la UNED de los siguientes grados: Administración y Dirección de Empresas, Biología, Bioquímica, Economía, Fundamentos de Informática, Lengua Castellana, Literatura, Matemáticas, Matemáticas Aplicadas a las Ciencias Sociales, Matemáticas Avanzadas y Psicología. El dataset se presenta en español y en inglés. La versión en español se ha obtenido directamente de la transcripción de los exámenes de Acceso para mayores de 25 años, mientras que la parte en inglés se ha construido mediante la traducción (manual) de estos exámenes.

En la Tabla 10 se incluye el número de preguntas, el número de exámenes y el número de palabras por asignatura, así como el número de opciones en la respuesta que tienen los exámenes de cada asignatura.

| Asignatura                                    | # Preguntas | # Exámenes | # Respuestas<br>por pregunta | # Palabras |
|---|-------------|------------|------------------------------|------------|
| Administración y Dirección de Empresas        | 87          | 6          | 3                            | 3936       |
| Biología                                      | 119         | 6          | 3                            | 2872       |
| Bioquímica                                    | 59          | 4          | 3                            | 1466       |
| Economía                                      | 51          | 3          | 4                            | 1726       |
| Fundamentos de Informática                    | 63          | 6          | 4                            | 1987       |
| Lengua Castellana                             | 94          | 4          | 4                            | 2816       |
| Literatura                                    | 91          | 6          | 4                            | 5130       |
| Matemáticas                                   | 73          | 11         | 3                            | 1538       |
| Matemáticas Aplicadas a las Ciencias Sociales | 94          | 10         | 3                            | 2941       |
| Matemáticas Avanzadas                         | 24          | 5          | 3                            | 470        |
| Psicología                                    | 248         | 14         | 4                            | 5669       |
| Total   | 1003        | 75         | —                            | 30551      |

Tabla 10: Dataset UNED ACCESO: distribución del número de preguntas y exámenes por asignatura, y número de opciones de respuesta por pregunta que tienen los exámenes de cada asignatura.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset UNED Acceso 2024".

## 2.9. Dataset 8: PRON VS PROMPT

**Motivación** Históricamente, hitos en el desarrollo de la Inteligencia Artificial como las victorias de Deep Blue y AlphaGo en ajedrez y Go, respectivamente, han marcado el avance tecnológico en diversas área de investigación. En el caso del Go, el sistema de IA desarrolló estrategias de juego novedosas que han sido imitadas, desde entonces, por todos los maestros humanos: se demostró que la IA podía ser creativa. Pero los juegos de mesa tiene características muy particulares que los hacen muy adecuados para los sistemas de IA. Desde la aparición de ChatGPT, la atención se ha desplazado hacia tareas como la escritura creativa, mucho más complejas. Estos modelos de lenguaje desafían la concepción de la inteligencia humana y las fronteras de la creación artística tradicionalmente considerada exclusiva de los humanos. Tanto es así que la colaboración humano-máquina en la industrias creativas es cada vez mayor (Adelani et al., 2023). De hecho, tal es su relevancia que los propios desarrolladores de OpenAI, en la interfaz de ChatGPT, ofrecen como primera sugerencia de uso de su modelo de lenguaje la opción "crea un historia".

**Objetivo** A pesar de todo, no existen aproximaciones de evaluación objetiva y rigurosa en esta tarea. Este dataset tiene el propósito general medir comparativamente la calidad en la creación de textos creativos de GPT4, el modelo con mejor rendimiento hasta la fecha, y un escritor profesional. Así, el objetivo de este dataset es doble: por un lado, se quiere realizar un estudio comparativo entre las capacidades creativas de generación de texto de una Inteligencia Artificial avanzada (GPT-4) y un novelista consagrado, Patricio Pron (Premio Alfaguara de novela). Y, por otro lado, se diseñará una metodología rigurosa de evaluación de textos creativos que servirá para puntuar la salida de cualquier modelo. A medio plazo, esperamos que esta experimentación sea el punto de partida para el desarrollo de métricas de evaluación automática que permitan conocer de la manera más objetiva posible la calidad de las producciones literarias de los modelos de lenguaje.

**Descripción** El dataset consiste en 120 sinopsis para 60 títulos de películas imaginarias, 30 propuestos por GPT4 y 30 por Patricio Pron. Se solicitó a ambos, al escritor y a GPT-4, que escribieran sinopsis de aproximadamente 600 palabras para cada título, incluyendo tanto los propuestos por ellos mismos como por su contrincante.

**Evaluación** Los textos generados están (a fecha de finalización del año 2 del proyecto) siendo sometidos a una evaluación a ciegas por un panel de expertos, compuesto por críticos y académicos, para garantizar una valoración objetiva de la calidad, creatividad y coherencia narrativa de las sinopsis. Los resultados estarán disponibles a lo largo del tercer año del proyecto.

## 2.10. Calibración de los datasets para medir la brecha en efectividad EN-ES

Los datasets descritos deben servir para medir la distancia de rendimiento entre los modelos del lenguaje en español e inglés. Definir un indicador de brecha en efectividad entre lenguas requiere tener en cuenta muchas variables, que en muchos casos dependen también del problema y de los datos disponibles para la evaluación. El primer problema a resolver es que no todas las métricas de efectividad tienen las mismas propiedades de escala. La mayoría de las métricas, como por ejemplo la tasa de aciertos, están acotadas entre cero y uno. Otras métricas no tienen una cota superior. Por tanto, no se pueden equiparar las diferencias obtenidas para varios problemas en los que se emplean diferentes métricas. Es decir, un intervalo en una métrica puede tener una relevancia completamente distinta del mismo intervalo en otra métrica. Por tanto, es necesario establecer un intervalo-unidad o intervalo de referencia.

El segundo gran problema es que la efectividad obtenida puede ser sensible a la dificultad intrínseca de los datos de evaluación. Por ejemplo, sobre un tamaño de datos de entrenamiento menor, es normal obtener valores de efectividad menores. Esto no significa que exista una brecha intrínseca en cuanto a

efectividad de los sistemas en cada uno de los idiomas. Para controlar este aspecto, será necesario tomar como referencia un sistema base que cumpla ciertas características. El sistema base no debe incluir ningún tipo de tecnología de la lengua específica de idioma. Es decir, debe de ser un sistema no pre-entrenado para ningún idioma, y que además no emplee herramientas de procesamiento lingüístico. Por ejemplo, emplear un clasificador SVM sobre conjuntos de *tokens* para clasificar textos no supone pre-entrenamiento ni pre-procesamiento lingüístico. Por tanto, las diferencias de efectividad entre idiomas vendrán determinadas únicamente por la dificultad del conjunto de datos de evaluación.

Teniendo en cuenta estos dos factores, tomaremos como intervalo de referencia en cada idioma la distancia entre la efectividad del sistema base que denotaremos como  $b$ , y un punto de referencia en la escala de la métrica que denotaremos como  $r$ . Es decir, tomaremos como intervalo unidad  $|b - r|$ .

El punto de referencia  $r$  también requiere un análisis previo. En casos en los que la métrica no esté acotada superiormente o en casos en los que la efectividad de los sistemas sea muy baja, este punto debería ser la cota inferior. Por otro lado, en casos en los que exista una cota superior y la efectividad sea alta debería tomarse como punto de referencia el valor superior en la escala de la métrica. Por ejemplo, dado un sistema base con efectividad cercana al uno en una métrica acotada superiormente, por ejemplo, 82 % de acierto, y tomando como punto de referencia el 100 % de acierto, el intervalo unidad debería ser del 18 %.

Una vez definido este intervalo unidad  $|b - r|$  la aportación lingüística efectiva en un idioma se calculará como el ratio de la diferencia entre efectividad del sistema evaluado y el sistema base respecto al intervalo unidad:

$$\Delta = \frac{s - b}{|b - r|} \cdot 100$$

La aportación lingüística en cada idioma cumple las siguientes propiedades. En primer lugar, la aportación es nula cuando el mejor sistema se comportan igual que el sistema base ausente de tecnología lingüística:

$$s = b \implies \Delta = 0$$

En segundo lugar, dada una efectividad fija por parte del sistema base, la aportación es proporcional a la diferencia entre efectividad del sistema evaluado y la del sistema base:

$$b = k \implies \Delta \propto s - b$$

En tercer lugar, dado una diferencia fija entre el sistema y el sistema base, la contribución será proporcional a la inversa del intervalo unidad:

$$s - b = k \implies \Delta \propto \frac{1}{|b - r|}$$

Esto quiere decir que, tomando como referencia la máxima puntuación ( $r = 1$ ) a medida que la efectividad del sistema y del sistema base se aproximen al punto máximo, la contribución será mayor. Por ejemplo, una mejora de 0.97 a 0.98 es más importante que una mejora de 0.67 a 0.68. De la misma forma, a valores bajos de efectividad, tomando como punto de referencia ( $r = 1$ ), una mejora de 0.1 a 0.2 será más significativa que una mejora de 0.3 a 0.4.

El indicador de la brecha de efectividad entre idiomas inglés (I) y español (E) se calculará mediante el indicador:

$$Ind(I, E) = \Delta_I - \Delta_E = \frac{s_I - b_I}{|b_I - r|} - \frac{s_E - b_E}{|b_E - r|}$$

Este indicador cumple las siguientes propiedades. En primer lugar, es simétrico respecto a los idiomas:

$$Ind(I, E) = -Ind(E, I)$$

En segundo lugar, un comportamiento idéntico en ambos idiomas se corresponde con una brecha cero:

$$Ind(I, I) = Ind(E, E) = 0$$

Tomando como punto de referencia  $r = 0$ , se obtiene una brecha nula cuando ambos presentan la misma diferencia porcentual respecto al baseline.

$$\left. \begin{array}{l} r = 0 \\ \frac{s_I - b_I}{b_I} = \frac{s_E - b_E}{b_E} \end{array} \right\} \Rightarrow Ind(I, E) = 0$$

que es lo mismo que decir que ambos tienen la misma proporción de efectividad respecto a sus sistemas base.

$$\left. \begin{array}{l} r = 0 \\ \frac{s_I}{b_I} = \frac{s_E}{b_E} \end{array} \right\} \Rightarrow Ind(I, E) = 0$$

Tomando como punto de referencia  $r = 0$ , se obtiene una brecha del 100 % cuando la efectividad del sistema en inglés consigue la diferencia porcentual conseguida por el sistema en español (brecha cero) más la efectividad del sistema base en inglés.

$$\left. \begin{array}{l} r = 0 \\ s_I = b_I \cdot \frac{s_E}{b_E} + b_I \end{array} \right\} \Rightarrow Ind(I, E) = 100 \%$$

Tomando como referencia  $r = 1$  (cota máxima), la brecha es nula cuando la efectividad del sistema en ambas lenguas es proporcional la diferencia entre los sistemas base y la cota superior:

$$\left. \begin{array}{l} r = 1 \\ \frac{s_I}{1 - b_I} = \frac{s_E}{1 - b_E} \end{array} \right\} \Rightarrow Ind(I, E) = 0$$

Tomando como referencia  $r = 1$ , hay una brecha del 100 % cuando el sistema en español no supera al sistema base, mientras que el sistema en inglés obtiene la máxima puntuación.

$$\left. \begin{array}{l} r = 1 \\ s_I = 1 \\ s_E = b_E \end{array} \right\} \Rightarrow Ind(I, E) = 100 \%$$

Este es el indicador que usamos en los experimentos descritos en la sección 5 sobre medición de la brecha de efectividad.

### Medición de brecha inglés-español en efectividad

Representa la diferencia entre idiomas entre mejoras porcentuales sobre un sistema base no lingüístico. Sea  $\mathcal{D}$  el conjunto de dominios,  $P_d$  el peso asignado a cada dominio, y  $\mathcal{H}_d$  el conjunto de aplicaciones seleccionadas para dicho dominio, se calcula:

$$E, 1.a = \sum_{d \in \mathcal{D}} P_d \sum_{h \in \mathcal{H}_d} \frac{s_I^h - b_I^h}{|b_I^h - r^h|} - \frac{s_E^h - b_E^h}{|b_E^h - r_0^h|}$$

donde  $s_I^h$ ,  $b_I^h$  y  $r^h$  representan la efectividad de la herramienta  $h$ , del sistema base y el punto de referencia en inglés. La notación para el español es análoga.

### 3. Aplicación web Leaderboard ODESIA v2

En las secciones anteriores se han descrito tanto la motivación y objetivos, como los criterios de selección de los dataset incluidos en el Leaderboard ODESIA v2. Como se ha expuesto, el objetivo principal del Leaderboard ODESIA es proporcionar una infraestructura de evaluación para modelos del lenguaje preentrenados en inglés y español, que permita, además, una comparación directa de rendimiento en ambos idiomas.

A lo largo de esta sección se describe en detalle el diseño y arquitectura de la aplicación. En primer lugar, se describen los perfiles de usuario y los casos de uso. Una vez definidos los casos de uso, se indican los requisitos necesarios para alcanzar las funcionalidades deseadas y la arquitectura.

### 3.1. Perfiles de usuario

Se pretende que el Leaderboard ODESIA se convierta en un marco de referencia para seguir el avance del estado del arte en PLN en español e inglés, especialmente para usuarios especialistas como investigadores o desarrolladores. No obstante, la popularidad de la IA obtenida en los últimos años gracias a los agentes conversacionales como ChatGPT, puede hacer que otros usuarios, menos asiduos a este tipo de infraestructuras, vean en el Leaderboard ODESIA un referente. En concreto, se han identificado los siguientes perfiles:

- **Investigadores:** es el perfil principal a quien va dirigido esta plataforma, dado que permite evaluar un modelo de lenguaje o sistema en varias tareas. Mediante el Leaderboard ODESIA, un investigador puede: (i) encontrar cual es el estado del arte para ciertas tareas, (ii) enviar resultados para evaluar sus sistemas, (iii) comparar los resultados de sus sistemas con los resultados existentes en español e inglés; (iv) encontrar información sobre la brecha de efectividad de los sistemas del español en relación al inglés.
- **Desarrolladores:** tanto a investigadores, como freelance, desarrolladores de empresas o administraciones, el Leaderboard ODESIA ofrece un marco comparativo entre los idiomas inglés y español de la eficiencia de modelos de lenguaje preentrenados. Esta infraestructura dotará a la comunidad de desarrollo de un mecanismo de evaluación preciso y fácil de usar, garantizando la comparación entre los distintos modelos evaluados al generarse todas las evaluaciones bajo los mismos parámetros.
- **Administraciones y empresas:** aunque este tipo de usuarios son menos frecuentes en este tipo de herramientas, como ya se ha comentado, la popularidad obtenida por los modelos del lenguaje en los últimos años está haciendo que nuevos actores entren en juego. Es por ello, que administraciones y empresas puedan encontrar en el Leaderboard ODESIA un marco donde validar la eficiencia de los posibles sistemas de información que deseen adquirir o subvencionar. Para las entidades que financian investigación en PLN e IA, el Leaderboard proporciona información muy relevante sobre los puntos débiles del estado del arte.

Puesto que el Leaderboard se ofrece en inglés y en español, será accesible a un espectro amplio de usuarios.

### 3.2. Casos de uso

Aunque, a priori, el objetivo propuesto para el Leaderboard ODESIA pueda parecer sencillo, es imprescindible tener en cuenta muchos factores a la hora de desarrollar una aplicación de dichas características. Por ejemplo, es imprescindible tener en cuenta que los procesos de evaluación son procesos costosos en tiempo de ejecución y recursos, pudiendo llegar a colapsar el sistema si no se tiene un correcto control sobre ellos. Así mismo, se hace necesario tener presente los distintos tipos de licencia disponibles a la hora de compartir datos, así como la mejor manera de compartir esos datos.

Por todo ello, y como requisito necesario para tener un correcto control de la aplicación y los distintos flujos de ejecución, se hace necesario que el Leaderboard ODESIA sea una aplicación accesible únicamente con registro de usuario. Atendiendo a esto, se proponen los siguientes flujos de ejecución:

- **Registro de usuario:** mediante este flujo un usuario no registrado previamente podrá crear una cuenta en la infraestructura Leaderboard ODESIA. Para ello, deberá proporcionar cierta información, como por ejemplo, datos de identificación y correo electrónico, así como aceptar los términos de uso de la aplicación. El Leaderboard tiene una página con un formulario para este fin.
- **Inicio de sesión:** mediante este flujo de ejecución un usuario previamente registrado podrá iniciar sesión en el Leaderboard ODESIA y acceder así a las diferentes opciones de evaluación o a la gestión de su perfil.

- **Gestión de perfil:** dado que el Leaderboard ODESIA es una aplicación en la que el registro de usuario es necesario, se debe dotar a los usuarios de una interfaz en la que gestionar su perfil y la información asociada al mismo.
- **Evaluación comparativa español/inglés:** mediante este flujo de evaluación un usuario registrado podrá realizar la evaluación de sus modelos del lenguaje preentrenado. Es importante destacar que los modelos deben abordar todas y cada una de las tareas en ambos idiomas, dado que de otro modo la comparativa sería imposible. Para ello, el Leaderboard ODESIA proporcionará una interfaz amigable mediante la cual, con solo cuatro pasos, cualquier usuario, independientemente de su perfil y experiencia, pueda evaluar sus modelos. En concreto, el usuario solo tendrá que: (i) descargar los datos; (ii) generar las predicciones con su modelo; (iii) subir los archivos con las predicciones de su modelo; (iv) pulsar el botón *enviar*.
- **Visualización resultados comparativos español/inglés:** por ultimo, es importante que los resultados de las evaluaciones comparativas entre inglés y español de los distintos modelos estén disponibles para la comunidad. Para ello, el flujo de visualización permite analizar los resultados desde la comparativa por idiomas, así como individualmente para cada idioma.

### 3.3. Requisitos y funcionalidades

El objetivo principal del Leaderboard ODESIA es el de establecer un marco competitivo para la comparación de modelos de lenguaje en diferentes tareas de PLN, con el fin último de comparar el desempeño de los modelos en español y en inglés. Para ello, en esta plataforma se pone a disposición del usuario una infraestructura que permita evaluar predicciones generadas para cada tarea contra Gold Standard privado, es decir, que garantizando que no ha sido usado en el entrenamiento de los modelos de lenguaje.

Además, estas tareas se deben revisar, cambiar y actualizar constantemente para que se adapten al estado del arte del momento en un ámbito actualmente en gran expansión donde la aparición de nuevos modelos se mide en días e incluso en horas. Es por ello que se hace imprescindible que la gestión de estas tareas sea sencilla y escalable. Para ello se definen diferentes conjuntos de tareas en forma de versiones. En este informe presentamos las tareas de la Versión 2 del Leaderboard.

Dado que los resultados obtenidos deben ser persistentes para ser comparados en el tiempo con nuevos modelos, los resultados obtenidos se deben registrar en una base de datos, y para poder asociarlos a un usuario, es imprescindible que la plataforma cuente con un sistema fiable de gestión de usuarios, en el que se contemple tanto el registro como la autenticación. Además debe disponer de una administración superior que permita prevenir comportamientos improcedentes, como uso de la plataforma con otros fines que no sean para los que está diseñada.

Para la evaluación de las predicciones aportadas por el usuario, es necesario contar con una herramienta de evaluación que recoja las buenas prácticas en el área de evaluación de sistemas de información, es por ello que se utiliza PyEvall. PyEvALL es una herramienta de evaluación para sistemas de información que permite evaluar un conjunto extenso de métricas que abarcan multitud de contextos, como clasificación o clustering. PyEvALL está implementada en Python y está basada en la teoría de la medida, recogiendo así las buenas prácticas en el área de evaluación de sistemas de información. El objetivo de PyEvALL es dotar a la comunidad científica, así como a otros actores en el área del desarrollo de sistemas de información, con una herramienta de evaluación de referencia que cubra multitud de contextos de evaluación y proporcione un amplio abanico de métricas.

### 3.4. Plataforma web

Para el desarrollo de este proyecto, se deben contemplar tres aspectos principales: (i) la gestión de contenidos mediante una herramienta que permita crear, editar y mantener diferentes tipos de contenido (páginas, tareas, resultados) interrelacionados entre ellos; (ii) administración de funcionalidades, con la que un usuario de rol administrador no especializado en desarrollo web pueda gestionar fácilmente el funcionamiento de la plataforma en sus funcionalidades más básicas como la gestión de usuarios, archivos,

análisis de registros de error, configuración, etc; (iii) uso de la herramienta PyEvALL para la evaluación de los archivos de predicciones.

Para contemplar estos tres aspectos en este proyecto, se valoraron inicialmente tres plataformas de desarrollo de webs: Django, Drupal y Flask, todas ellas basadas en formatos de código abierto y que cumplen en mayor o menor medida las especificaciones básicas especificadas en la Sección 3.3. En el informe del Año 1 se describieron las tres opciones con sus ventajas e inconvenientes. Finalmente, se optó por una estructura híbrida entre el gestor de contenidos Drupal que abarca la mayor parte de las necesidades del proyecto en general y la implementación de una API Rest independiente con Flask/Python que permite suplir las necesidades que el gestor no alcance. Todo ello en una estructura interna basada en contenedores Docker que permite gestionar y escalar el sistema fácilmente en caso necesario.

### 3.5. Arquitectura

El diseño de la arquitectura del Leaderboard ODESIA sigue el patrón general para este tipo de aplicaciones donde se identifican claramente la capa de interfaz de usuario y la capa de servidor. En este caso, tal y como se puede ver en la Figura 1, la capa de servidor está compuesta a su vez por tres componentes a modo de contenedores docker donde cada uno ofrece un servicio principal a la aplicación: el contenedor PHP, el contenedor de la base de datos y el contenedor EvALL.

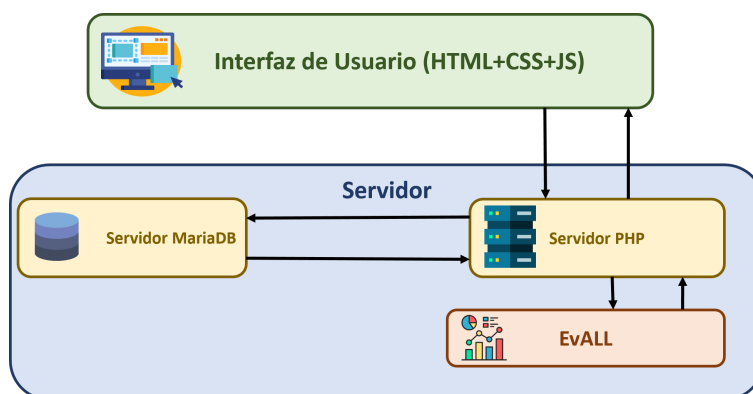


Figura 1: Arquitectura Leaderboard

La arquitectura de este proyecto, y de acuerdo con el requisito presentado para una gestión y mantenimiento sencillos de la infraestructura, trabaja en contenedores aislados implementados en Docker. De esta forma, cada parte del proyecto puede trabajar independientemente para evitar problemas de concurrencia de software, pero al mismo tiempo estos contenedores trabajan de forma unificada en un entorno de red virtualizada por el propio sistema y manteniendo comunicación entre los contenedores mediante diferentes protocolos.

Para ello, dentro de este sistema, podemos encontrar diferentes redes virtualizadas de las que destacan principalmente una privada, a la que solo tienen acceso los contenedores que lo requieran, y una red pública conectada con el entorno físico donde se ejecuta el sistema. De este modo, se garantiza una comunicación eficiente entre los diferentes contenedores y se maximiza la seguridad en aspectos como la base de datos y/o EvALL. Atendiendo a esto, los contenedores incluidos en la arquitectura son:

- **Contenedor Apache/PHP/Drupal:** que contiene el sistema de archivos con la lógica de la programación del sistema. Es la parte central que gestiona los otros nodos del sistema. Este contenedor tiene doble conexión, una dentro del propio entorno virtualizado (privada) y otra a la red pública, dado que por un lado tiene que trabajar con los diferentes contenedores y por otro tiene que ser accesible a los usuarios.
- **Contenedor MariaDB:** que contiene el SGBD MariaDB y la base de datos del proyecto. Este contenedor únicamente es accesible dentro de la red privada.

- **EvALL:** mediante este contenedor la aplicación accede a la herramienta de evaluación PyEvALL, que es la que realmente realiza la evaluación final. Así mismo, mediante este módulo se realizan los controles necesarios para que los procesos no colapsen el servidor y los recursos del mismo. Todo este proceso se realiza mediante un conjunto de protocolos establecidos entre el contenedor PHP y este contenedor. Este contenedor únicamente es accesible dentro de la red privada.

En las siguientes secciones se detallan las funcionalidades de cada capa, centrándose en la interacción entre ellas.

### 3.5.1. Capa Interfaz de usuario

La capa de interfaz de usuario del Leaderboard ODESIA se desarrolla en un marco de trabajo con Drupal como base. En este entorno, se ha planteado una interfaz de usuario general, distribuida en dos componentes principales, la parte pública y la parte privada.

La parte pública será accesible por cualquier usuario, tanto registrado como no, y recogerá fundamentalmente la parte informativa de la infraestructura, como puede ser la portada con las diferentes tablas de resultados o las preguntas frecuentes, así como el acceso a los flujos de registro e inicio de sesión. Por otro lado, dentro del área privada, se pueden encontrar diferentes niveles separados por roles. Entre ellos cabe destacar los usuarios registrados que podrán acceder al flujo de evaluación comparativo completo, así como a la descarga de datos.

### 3.5.2. Capa Servidora

Tal y como se ha comentado anteriormente, para la arquitectura de este proyecto se parte de la base de trabajar en contenedores aislados implementados en Docker, de tal manera que cada parte del proyecto pueda trabajar de manera independientemente para evitar problemas de concurrencia de software, aunque al mismo tiempo estos contenedores trabajan de forma unificada en un entorno de red virtualizada por el propio sistema.

Dentro de este sistema, el contenedor servidor PHP/Drupal es el núcleo de interconexión entre los distintos contenedores y la capa de interfaz de usuario. Mediante este modulo se centralizan todas las peticiones de la parte publica, y se redirigen a los distintos contenedores privados: servidor MariaDB y EvALL.

Como ya se ha expuesto, el servidor MariaDB, contiene el SGBD MariaDB y la base de datos del proyecto. Este contenedor únicamente es accesible dentro de la red privada. Por otro lado, el contenedor EvALL forma parte de la parte privada de EvALL 2.0, mediante la cual se accede a la librería de evaluación PyEvALL. EvALL 2.0 proporciona una herramientas imprescindible para este tipo de infraestructuras al disponer de múltiples métricas que cubren diferentes contextos de evaluación. En particular, para este proyecto, mediante una sola llamada se puede abordar contextos de evaluación tan dispares como clasificación mono-label o clasificación multi-label con jerarquía. Para un mayor detalle sobre su funcionamiento se recomienda consultar el informe "Proyecto Espacio de Observación de Inteligencia Artificial en Español - Ámbito 1.3 Aplicación Web EvALL 2.0 - Informe Técnico Año 2", sobre la aplicación EvALL 2.0 desarrollada en el marco de este mismo convenio.

## 3.6. Roles de usuario

A tenor del diseño y la arquitectura propuesta, los casos de uso, y las especificaciones de la aplicación, se identifica el siguiente conjunto inicial de roles de usuario:

- **Usuario anónimo:** en este caso el usuario no está registrado y solo puede acceder a la parte pública de la aplicación. Cualquier usuario puede registrarse con un correo electrónico. Su solicitud debe ser aprobada por un administrador.
- **Usuario registrado:** puede acceder tanto a la parte pública como a la parte privada de la aplicación, donde podrá enviar los resultados de su sistema a evaluar. Para ello, el usuario registrado debe iniciar sesión con su usuario y contraseña.

- **Usuario administrador:** el rol de administración solo está reservado para el personal de la UNED. Mediante este rol, se podrán aceptar las solicitudes de registro de los usuarios, o editar o borrar sus perfiles en casos necesarios. Además, se podrán gestionar los resultados publicados o las tareas de los distintos benchmarks.

### 3.7. Flujos del Leaderboard ODESIA

De acuerdo a los casos de uso analizados, y presentados en la sección 3.2 se han desarrollado los siguientes flujos en la aplicación actual.

#### 3.7.1. Registro de usuario

The screenshot shows the 'CREAR NUEVA CUENTA' (Create New Account) page of the ODESIA Leaderboard application. At the top, there is a navigation bar with the application logo and several menu items: Inicio, Participa, Resultados v1, Tareas, Metodología, FAQ, Login/Registro, and Idioma. Below the navigation bar, the main heading 'CREAR NUEVA CUENTA' is centered. Underneath, there are three tabs: 'Iniciar sesión', 'Crear nueva cuenta' (which is active), and 'Reinicializar su contraseña'. The 'Crear nueva cuenta' tab contains a form with the following fields: 'Dirección de correo electrónico' (Email address), 'Nombre de usuario' (Username), 'País' (Country), and 'Afiliación' (Affiliation). Each field has a small red asterisk indicating it is required. Below the email field, there is a note: 'The email address is not made public. It will only be used if you need to be contacted about your account or for opted-in notifications.' Below the username field, there is a note: 'Varios caracteres están permitidos, incluyendo los espacios, puntos (.), guiones (-), comillas (') y el signo @.' At the bottom of the form, there is a green button labeled 'Crear nueva cuenta'.

Figura 2: Registro

El Leaderboard ODESIA es una aplicación orientada a usuarios registrados, como ya se expuso en la sección 3.2, debido fundamentalmente a la necesidad de supervisar los procesos de evaluación en el servidor y a la necesidad de controlar el acceso a los datos. Por todo ello, es preciso dotar al usuario con un flujo que permita a cualquier usuario anónimo registrarse en la aplicación, pudiendo así acceder a la parte privada de la misma. Para ello, el usuario solo debe rellenar un pequeño formulario con unos datos básicos y un correo electrónico (Figura 2). El proceso se realiza en dos pasos, en primer lugar el usuario debe cumplimentar el formulario de registro y enviarlo, y en un segundo paso dicha solicitud debe ser aceptada por un administrador.

Debido a la similitud con el flujo desarrollado en la aplicación EvALL 2.0 desarrollada en el marco de este mismo convenio, se remite al informe "Proyecto Espacio de Observación de Inteligencia Artificial en Español - Ámbito 1.3 Aplicación Web EvALL 2.0 - Informe Técnico Año 2" para un mayor detalle del mismo.

#### 3.7.2. Inicio de sesión

Al igual que en flujo anterior, Leaderboard ODESIA debe proporcionar al usuario con un flujo que permita iniciar sesión a usuarios ya registrados y poder así acceder a la parte privada. Para ello, el usuario solo debe rellenar un pequeño formulario con los datos de acceso, nombre de usuario y contraseña, y enviar la información para su comprobación (Figura 3).

Debido a la similitud con el flujo desarrollado en la aplicación EvALL 2.0 desarrollada en el marco de este mismo convenio, se remite al informe "Proyecto Espacio de Observación de Inteligencia Artificial en Español - Ámbito 1.3 Aplicación Web EvALL 2.0 - Informe Técnico Año 2" para un mayor detalle del mismo.

## INICIAR SESIÓN

[Iniciar sesión](#) [Crear nueva cuenta](#) [Reinicializar su contraseña](#)

**Nombre de usuario \***  
  
Escriba su nombre de usuario en Leaderboard.

**Contraseña \***  
  
Escriba la contraseña asignada a su nombre de usuario.

[Iniciar sesión](#)

Figura 3: Inicio de sesión

### 3.7.3. Evaluación comparativa español/inglés

Esta sección describe el flujo de participación en el Leaderboard ODESIA que proporciona un entorno preparado para que investigadores y/o desarrolladores puedan evaluar sus sistemas contra diferentes tareas relacionadas con el PLN y al mismo tiempo comparar estos resultados con los obtenidos por otros usuarios. A continuación, se detallan en profundidad las funcionalidades del flujo, los casos de uso, así como su lógica, mostrando en cada paso las pantallas implementadas para ello.

#### Funcionalidad

Con este flujo se pretende ofrecer al usuario un marco sobre el que validar el funcionamiento de un modelo creado enfrentándolo a diferentes tareas tanto en inglés como en español. Para ello, este, tiene que enviar los resultados obtenidos en cada tarea y tras una evaluación contra un Gold Standard privado se puede comparar la precisión del sistema frente a los intentos de otros usuarios tanto por tarea como en una tabla de resultados general donde se muestra el mejor sistema para cada tarea/idioma.

Para ello se ha valorado el estilo de una de las herramientas más populares en este ámbito SuperGLUE Benchmark,<sup>19</sup> en el que por un lado el usuario encuentra un sencillo formulario de envío de archivos de resultados y por otro, en una tabla de "leaderboard", puede ver donde se encuentra su sistema en una clasificación general.

#### Lógica del flujo

Tal como se ha comentado anteriormente, el flujo principal de esta aplicación está basado en otras más populares como SuperGLUE Benchmark, en la que un usuario debe enfrentar un modelo a diferentes tareas simultáneamente y valorar los resultados obtenidos en cada una de ellas. A esto se le añaden principalmente dos cualidades, que las tareas a las que se tiene que enfrentar son bilingües, y que los resultados de cada tarea se valoran con diferentes métricas.

Para ello el proceso de participación está planteado en básicamente 3 pasos, sin contar el flujo de registro y login:

1. **Acceso a la descripción de las tareas:** En la página de tareas<sup>20</sup> de la aplicación, el usuario puede visitar cada una de las tareas seleccionadas para ver el objetivo a alcanzar en esta.
2. **Acceso a los conjuntos de datos:** En la página Participar se proporcionan instrucciones sobre cómo participar (Figura 4a). Una vez registrado el usuario puede descargar un paquete .zip con los conjuntos de datos de cada tarea y un archivo README donde se detalla el formato de entrega de los archivos de salida generados por su sistema para ser evaluados.

<sup>19</sup><https://super.gluebenchmark.com/>

<sup>20</sup><http://leaderboard.odesia.uned.es/leaderboard/tareas>

## How can I participate?

1. To participate you must be registered and logged in. You can do it here
2. Once you are logged in, you can download the datasets from a link below.
3. Now you can start training your model for the different tasks.
4. When you consider that your model is ready... ¡submit your results!
5. In the FAQ In the FAQ you will find detailed information on the submission format for each of the tasks..
6. Remember, the file name must be the same for each task and language as the one shown in the 'Output file name' column displayed on the Tasks page. Tareas.
7. Create a zip file with only the files containing the predictions, no folders.
8. Use the form on the left and submit the zip file.

Download dataset

(a) Participación y descarga de datasets.

### Send results

Name \*

Correo electrónico \*

Affiliation

System name \*

Model url

System description

Description paper URL

Github URL

Submission languages \*

☐ English

☐ Español

☐ Ambos

ZIP File \*

No file selected.

☐ Accept los términos y condiciones de uso \*

Enviar

(b) Envío de resultados.

Figura 4: Participación en el Leaderboard.

3. **Entrega de los resultados:** Una vez generados los resultados para cada tarea, el usuario debe cumplimentar un sencillo formulario donde los puede enviar dentro de un archivo .zip (Figura 4b).

Una vez enviado este formulario, el sistema comprueba que se encuentre un archivo para cada tarea y para cada idioma de las definidas, en caso contrario retorna un error. Una vez realizada esta comprobación, se evalúa cada uno de los output del sistema del usuario contra un Gold Standard interno (que es un archivo que no solamente es privado sino que se han tomado diferentes medidas de seguridad para que no sean accesibles desde el exterior) y la respuesta a cada evaluación es almacenada como resultado. Con esto finaliza el flujo y el usuario puede ver si su sistema aparece o no en el leaderboard.

### 3.7.4. Visualización de resultados comparativos español/inglés

El usuario tiene diferentes opciones en cuanto a la visualización de los resultados obtenidos para un sistema que haya publicado. En la pantalla principal de la aplicación, el usuario puede encontrar tres tablas de resultados:

**Mejor sistema general para cada tarea/idioma** En esta tabla de resultados, se puede visualizar directamente cual es el mejor resultado obtenido para una tarea en un idioma, de modo que se puede apreciar si existe algún gap destacable entre los sistemas participantes por idioma (5).

| Odesia Core Tasks ?                                     |                     |                     |      | Odesia Extended Tasks ?                            |                     |                     |       |
|---|---------------------|---------------------|------|--|---------------------|---------------------|-------|
| Tasksit   | Best result Spanish | Best result English | ?    | Tasks  | Best result Spanish | Best result English | ?     |
| Media total   | 0.57                | 0.59                | 42%  | Total mean   | 0.81                | 0.87                | 21.5% |
| EXIST 2022: Sexism detection (ES)                       | 0.77                | 0.81                | 17%  | MLDOC - Document classification                    | 0.96                | 0.98                | 40%   |
| EXIST 2022: Sexism categorisation (ES)                  | 0.57                | 0.58                | 10%  | Multilingual Complex Named Entity Recognition 2022 | 0.71                | 0.75                | 5%    |
| DIPROMATS 2023 Propaganda identification (ES)           | 0.82                | 0.82                | 11%  | SQAC-SQUAD 2016                                    | 0.77                | 0.88                | 25%   |
| DIPROMATS 2023: Coarse propaganda characterization (ES) | 0.47                | 0.55                | 48%  | Semantic Textual Similarity 2017                   | 0.80                | 0.86                | 16%   |
| DIPROMATS 2023 Fine propaganda characterization (ES)    | 0.26                | 0.47                | 299% |  |                     |                     |       |
| DIANN 2023: Detección de discapacidades (ES)            | 0.84                | 0.79                | 1%   |  |                     |                     |       |
| EXIST-2023: Sexism Identification (ES)                  | 0.64                | 0.64                | 10%  |  |                     |                     |       |
| EXIST-2023: Source Intention (ES)                       | 0.42                | 0.36                | -4%  |  |                     |                     |       |
| EXIST-2023: Sexism Categorization (ES)                  | 0.40                | 0.40                | 12%  |  |                     |                     |       |
| SQAC-SQUAD 2024: Question Answering (ES)                | 0.46                | 0.46                | 19%  |  |                     |                     |       |

Figura 5: Mejor sistema por lengua y tarea.

Además, situando el ratón por encima del resultado, se puede obtener más información sobre el sistema que ha generado dicho resultado, principalmente el nombre del sistema y la métrica utilizada para la evaluación.

**Mejores sistemas en español** En la segunda tabla que se muestra en la pantalla principal del Leaderboard ODESIA, se pueden ver los 10 sistemas que mejores resultados obtienen en todas las tareas en español y el resultado que han obtenido para cada una (Figura 6). Hay una tabla para las tareas Core y otra las tareas Extended.

**Mejores sistemas en inglés** En la segunda tabla que se muestra en la pantalla principal del Leaderboard ODESIA, se pueden ver los 10 sistemas que mejores resultados obtienen en todas las tareas en inglés y el resultado que han obtenido para cada una (Figura 7).

## Odesia Core Tasks ?

| Gap English - Spanish |                                    |                 | Results for Spanish               |  |   | Results for English                                     |   |  |
|-----------------------|------------------------------------|-----------------|-----------------------------------|--|---|---|---|--|
| #                     | Sistema                            | Arithmetic mean | EXIST 2022: Sexism detection (ES) | EXIST 2022: Sexism categorisation (ES) | DIPROMATS 2023 Propaganda identification (ES) | DIPROMATS 2023: Coarse propaganda characterization (ES) | DIPROMATS 2023: Fine propaganda characterization (ES) | DIANN 2023: Detección de discapacidades (ES) |
| 1                     | distilbert-base-multilingual-cased | 0.459           | 0.72                              | 0.47                                   | 0.75  | 0.34  | 0.09  | 0.78   |
| 2                     | distilbert-base-spanish-uncased    | 0.473           | 0.72                              | 0.51                                   | 0.77  | 0.34  | 0.07  | 0.75   |
| 3                     | xlm-roberta-base                   | 0.515           | 0.74                              | 0.50                                   | 0.79  | 0.47  | 0.10  | 0.84   |
| 4                     | ixambert-base-cased                | 0.488           | 0.73                              | 0.50                                   | 0.77  | 0.32  | 0.06  | 0.83   |
| 5                     | bert-base-multilingual-cased       | 0.488           | 0.72                              | 0.47                                   | 0.78  | 0.35  | 0.10  | 0.84   |

Figura 6: Leaderboard del español (Core Tasks).

## Odesia Core Tasks ?

| Gap English - Spanish |                                    |                 | Results for Spanish               |  |  | Results for English                           |   |   |
|-----------------------|------------------------------------|-----------------|-----------------------------------|--|--|---|---|---|
| #                     | Sistema                            | Arithmetic mean | EXIST 2022: Sexism detection (EN) | EXIST 2022: Sexism categorisation (EN) | DIANN 2023: Detección de discapacidades (EN) | DIPROMATS 2023 Propaganda identification (EN) | DIPROMATS 2023: Coarse propaganda characterization (EN) | DIPROMATS 2023: Fine-grained propaganda characterization (EN) |
| 1                     | distilbert-base-multilingual-cased | 0.472           | 0.74                              | 0.53                                   | 0.68   | 0.77  | 0.45  | 0.16  |
| 2                     | distilbert-base-uncased            | 0.497           | 0.77                              | 0.55                                   | 0.66   | 0.78  | 0.47  | 0.14  |
| 3                     | bert-base-cased                    | 0.513           | 0.76                              | 0.53                                   | 0.72   | 0.81  | 0.50  | 0.21  |
| 4                     | ixambert-base-cased                | 0.505           | 0.77                              | 0.53                                   | 0.73   | 0.78  | 0.49  | 0.14  |
| 5                     | xlm-roberta-base                   | 0.517           | 0.76                              | 0.53                                   | 0.76   | 0.80  | 0.54  | 0.16  |

Figura 7: Leaderboard del inglés (Core Tasks).

### 3.8. Formatos de entrada

Como ya se ha expuesto, el Leaderboard ODESIA hace uso de la herramienta de evaluación EvALL 2.0 para realizar las evaluaciones de las distintas tareas sobre las distintas métricas y contextos de evaluación. Es por ello, que el envío de resultados se hace en el formato json propuesto en EvALL 2.0. Para una correcta comprensión del mismo, se expone aquí de nuevo.

```
[
  {
    "test_case": "1",
    "id": "1",
    "value": "TRUE"
  },
  {
    "test_case": "1",
    "id": "2",
    "value": "TRUE"
  }
]
```

```
[
  {
    "test_case": "1",
    "id": "1",
    "value": ["TRUE"]
  },
  {
    "test_case": "1",
    "id": "2",
    "value": ["TRUE", "FALSE"]
  }
]
```

a. Formato json de la librería PyEvALL.

b. Formato json multi-label de la librería PyEvALL.

Figura 8: Formatos json de entrada.

El formato de entrada para PyEvALL es único para todos los contextos de evaluación, y está encapsulado en un formato json. El formato json está compuesto, tal y como se ve en la Figura 8.a, por una lista de objetos json en la que cada uno representa una instancia de evaluación.

En concreto, cada atributo representa:

- **test\_case:** un experimento o caso de uso determinado, pudiendo ser, por ejemplo, diferentes ejecuciones de un algoritmo de clasificación, o diferentes queries en un contexto de ranking.
- **id:** el identificador unívoco de la instancia en el dataset.
- **value:** la clase objetivo o valor que toma la instancia.

Mediante este formato se pueden realizar evaluaciones en el contexto de clasificación mono-label, tanto con o sin jerarquía. Para realizar evaluaciones multi-label, tanto con o sin jerarquía, se debe utilizar el formato indicado en la Figura 8.b, donde, como se puede ver, el atributo **value** es una lista de elementos que representan las clases objetivos o posibles valores de la instancia.

Así mismo, y por similitud, los formatos de los distintos datasets se han convertido a formato json, siguiendo el mismo esquema pero preservando los atributos en cada caso.

### 3.9. Navegación

El Leaderboard ODESIA consta de ocho páginas principales como se observa en la cabecera de la portada del Leaderboard en la Figura 9: Inicio, Participa, Resultados, Tareas, Metodología, FAQ, Perfil, Login/Registro, Idioma.

- En la portada se presenta el nombre de la aplicación, un resumen de los contenidos en cajas, tablas de resultados para Core Tasks y Extended Tasks por separado, una tabla con los resultados que se usan para medir la brecha español-inglés, y al final una llamada a la participación.
- La página Participa contiene instrucciones sobre cómo participar y el formulario de registro para usuarios no registrados, y en el caso de usuarios registrados contiene el formulario para el envío de datos.

# ODESIA Leaderboard

## Evaluación de modelos de lenguaje en inglés y español

**Objetivos:** establecer una comparación directa entre el rendimiento de modelos en inglés y español para medir la brecha de efectividad.

**Método:** evaluación sobre el Benchmark ODESIA, una colección de tareas de Procesamiento del Lenguaje Natural con conjuntos de datos comparables en inglés y español.

| Objetivos  | Resultados   | Tareas  | Metodología   |
|--|--|---|---|
| <p>El <b>Leaderboard ODESIA</b> permite (I) medir la brecha de efectividad de los modelos de lenguaje en español respecto al inglés; (II) evaluar de forma comparada modelos de lenguaje en español. Si has desarrollado un modelo de lenguaje en español, ¡envía tus resultados!</p> <p><a href="#">Ver más detalles aquí</a></p> | <p>La <b>brecha de efectividad promedio</b> entre Español e Inglés es del <b>20%</b>, con un error estándar de <math>\pm 6\%</math>. Hay que destacar que la brecha es más acusada en las tareas más difíciles (hasta superar el 200% en la tarea con mayor dificultad intrínseca), y por tanto el valor promedio tiene una representatividad relativa.</p> <p><a href="#">Ver más detalles aquí</a></p> | <p>Se utilizan dos conjuntos de tareas: (I) <b>ODESIA CORE</b>, , bilingual tasks with private test data (this avoids contamination, that the models have seen the evaluation keys in the pre-training phase); and (II) <b>ODESIA EXTENDED</b>, que añade un conjunto de cinco tareas bilingües estándar y disponibles de forma pública.</p> <p><a href="#">Ver más detalles aquí</a></p> | <p><b>ODESIA Leaderboard</b> utiliza un conjunto de 14 tareas bilingües para comparar el estado del arte en inglés y español. Sobre cada tarea (I) se estima la dificultad intrínseca aplicando varios algoritmos no lingüísticos y (II) se calibran los mejores resultados en cada idioma usando esa dificultad intrínseca.</p> <p><a href="#">Ver más detalles aquí</a></p> |

Figura 9: Parte superior de la página principal del Leaderboard ODESIA

- En la página Resultados se muestran las tablas de resultados para las Core Tasks y Extended Tasks en español y en inglés, así como la tabla donde se muestran los resultados usados para medir la brecha, con la medición de la brecha.
- En la página de Tareas (Figura 10) se muestran todas las tareas del Benchmark ODESIA, tanto las Core como las Extended. Haciendo clic en cada tarea se accede a información sobre la tarea.
- En la página de Metodología se ofrece información sobre la metodología usada para evaluar los sistemas y para medir la brecha, mientras que en la página de FAQ se contestan preguntas que pueden ser de interés para el usuario.
- En la página Perfil el usuario registrado accede a sus datos. En la página Login/Registro el usuario se puede registrar si no lo estaba o puede hacer el login o logout si ya está registrado.

### 3.10. Aspectos técnicos de desarrollo de proyectos de software

En el Apéndice A se recogen los aspectos técnicos y de documentación relevantes en el proceso de desarrollo de código que afectan a esta aplicación. Aunque este aspecto queda, en cierto modo, fuera de los ámbitos del convenio, se ha intentado abordar, hasta donde llegan nuestras posibilidades como universidad, todos los puntos sugeridos por la SEDIA.

## TAREAS LEADERBOARD

| Título   | Test    | Idioma | Métrica Leaderboard | Nombre archivo Output     |   |
|--|---------|--------|---------------------|---------------------------|---|
| DIANN 2023: Detección de discapacidades (ES)                 | PRIVADO | es     | F1                  | DIANN_2023_T1_es.json     | 1 |
| DIPROMATS 2023 Propaganda identification (ES)                | PRIVADO | es     | F1                  | DIPROMATS_2023_T1_es.json | 1 |
| DIPROMATS 2023: Coarse propaganda characterization (ES)      | PRIVADO | es     | F1                  | DIPROMATS_2023_T2_es.json | 1 |
| DIPROMATS 2023 Fine propaganda characterization (ES)         | PRIVADO | es     | F1                  | DIPROMATS_2023_T3_es.json | 1 |
| EXIST 2022: Sexism detection (ES)                            | PRIVADO | es     | Accuracy            | EXIST_2022_T1_es.json     | 1 |
| EXIST 2022: Sexism categorisation (ES)                       | PRIVADO | es     | Macro F1            | EXIST_2022_T2_es.json     | 1 |
| EXIST-2023: Sexism Identification (ES)                       | PRIVADO | es     | ICM                 | EXIST_2023_T1_es.json     | 1 |
| EXIST-2023: Source Intention (ES)                            | PRIVADO | es     | ICM                 | EXIST_2023_T2_es.json     | 1 |
| EXIST-2023: Sexism Categorization (ES)                       | PRIVADO | es     | ICM                 | EXIST_2023_T3_es.json     | 1 |
| MLDOC - Document classification                              | PÚBLICO | es     | F1                  | MLDOC_2018_es.json        | 1 |
| Multilingual Complex Named Entity Recognition 2022           | PÚBLICO | es     | F1                  | MULTICONER_2022_es.json   | 1 |
| SQAC-SQUAD 2016  | PÚBLICO | es     | F1                  | SQAC_SQUAD_2016_es.json   | 1 |
| SQAC-SQUAD 2024: Question Answering (ES)                     | PRIVADO | es     | F1                  | SQAC_SQUAD_2024_es.json   | 1 |
| Semantic Textual Similarity 2017                             | PÚBLICO | es     | Pearson correlation | STS_2017_es.json          | 1 |
| DIANN 2023: Detección de discapacidades (EN)                 | PRIVADO | en     | F1                  | DIANN_2023_T1_en.json     | 1 |
| DIPROMATS 2023 Propaganda identification (EN)                | PRIVADO | en     | F1                  | DIPROMATS_2023_T1_en.json | 1 |
| DIPROMATS 2023: Coarse propaganda characterization (EN)      | PRIVADO | en     | F1                  | DIPROMATS_2023_T2_en.json | 1 |
| DIPROMATS 2023 Fine-grained propaganda characterization (EN) | PRIVADO | en     | F1                  | DIPROMATS_2023_T3_en.json | 1 |

Figura 10: Página Tareas del Leaderboard ODESIA

## 4. Experimentos de evaluación con datasets con test privado (Core Tasks)

El leaderboard puede ser utilizado por cualquier equipo de investigación o desarrollo interesado en evaluar modelos de lenguaje preentrenados en español, en inglés, o multilingües. Para ello se deben descargar (previa firma de los acuerdos de distribución pertinentes) los datos anotados de entrenamiento para cada una de las tareas en uno o los dos idiomas, y los datos de test sin anotar. Una vez aplicado el proceso de fine-tuning del modelo sobre cada una de las tareas (en cada idioma pertinente), se suben las salidas del sistema mediante la aplicación web del leaderboard y se obtienen los resultados para cada tarea y los resultados agregados, que pasan a formar parte (de forma optativa) del leaderboard público.

Además del servicio que proporciona el leaderboard a los desarrolladores de modelos, también nos interesa disponer de datos de rendimiento para los modelos de lenguaje más populares, especialmente para el español. Por ello hemos realizado una selección de modelos de lenguaje disponibles en Hugging Face, los hemos aplicado a cada una de las tareas para inglés y español y hemos añadido los datos a la versión 2 del Leaderboard.

### 4.1. Modelos de lenguaje seleccionados

Para establecer una comparación inicial entre modelos del lenguaje en inglés y español, hemos hecho una selección de modelos preentrenados en español, en inglés, y multilingües, escogiendo entre los más utilizados en la literatura y en Hugging Face. En conjunto se trata de cinco modelos en español, cuatro en inglés, y cinco modelos multilingües, lo que nos permite considerar diez modelos en el leaderboard español y nueve en inglés.

El leaderboard incluye tareas tanto discriminativas (EXIST, DIPROMATS, etc.) como generativas (CURIA, UNED Acceso). De manera natural, las tareas discriminativas suelen resolverse con modelos encoders. Y, las tareas generativas, con modelos decoder, también llamados modelos generativos; el objetivo de entrenamiento de los modelos decoder es generar texto fluido. Sin embargo, para resolver esta tarea, han demostrado adquirir una gran capacidad de comprensión de lenguaje tal que parecen ser capaces de resolver tareas discriminativas, aunque no estén originalmente diseñados para eso.

De momento en el leaderboard sólo se incluyen los modelos discriminativos. Respecto a modelos generativos, en este año se ha realizado una primera experimentación con el dataset UNED-ACCESO,

utilizando los modelos decoder o generativos que ahora mismo son el estado del arte: GPT4 (OpenAI et al., 2024), ChatGPT<sup>21</sup>, Claude 3<sup>22</sup>, Mistral (Jiang et al., 2023), Llama-2 (Touvron et al., 2023) y Gemma<sup>23</sup>. Además, está en proceso la evaluación de estos modelos con el dataset CURIA de resúmenes en texto claro. También se ha evaluado GPT-4 en modo zero-shot (proporcionando al sistema sólo la guía de anotación) para las tres tareas de DIPROMATS 2023.

Los modelos discriminativos incluidos en el leaderboard son los siguientes:

| Español                         | Inglés                  | Multilingües                       |
|---------------------------------|-------------------------|------------------------------------|
| roberta-large-bne               | roberta-large           | xlm-roberta-large                  |
| roberta-base-bne                | roberta-base            | xlm-roberta-base                   |
| bert-base-spanish-wwm-cased     | bert-base-cased         | bert-base-multilingual-cased       |
| distilbert-base-spanish-uncased | distilbert-base-uncased | distilbert-base-multilingual-cased |
| bertin-roberta-base-spanish     |                         | ixambert-base-cased                |

Tabla 11: Modelos de lenguaje usados en la evaluación.

## Modelos en inglés

- **Bert-base-cased** BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018a) fue el primer modelo que utilizó la arquitectura transformer, que después ha sido la predominante en toda la investigación posterior. El modelo base tiene 12 capas transformer blocks, 12 attention heads, y 110 millones de parámetros. El preentrenamiento se realizó de forma autosupervisada en un gran corpus en inglés optimizando dos objetivos simultáneamente:
  - Masked Language Model (MLM). En cada frase el modelo oculta de forma aleatoria el 15 % de las palabras, e intenta predecir las palabras ocultas.
  - Next Sentence Prediction (NSP). En el preentrenamiento, dado un par de oraciones, el modelo debe predecir si las dos frases aparecían consecutivamente en el corpus de entrenamiento o no.
- **RoBERTa-base** RoBERTa (Liu et al., 2019) un modelo que introduce algunas variaciones respecto a BERT. La principal diferencia es que RoBERTa usa un masking dinámico, en el que en cada epoch (iteración de aprendizaje sobre el corpus de entrenamiento) se enmascaran distintas partes de la oración. RoBERTa-base tiene 123M de parámetros.
- **Roberta-large** La única diferencia entre ROBERTA-BASE y ROBERTA-LARGE es el número de parámetros utilizado para su entrenamiento. Si bien ROBERTA-BASE fue entrenado con 123 millones de parámetros, ROBERTA-LARGE lo hizo con 354 millones de parámetros.
- **Distilbert-base-uncased** es un modelo preentrenado con el mismo corpus que BERT, pero se trata de un modelo más pequeño y más rápido (Sanh et al., 2019). El modelo fue preentrenado con tres objetivos:
  - Que tuviera la capacidad de devolver las mismas probabilidades que el modelo BERT
  - Que fuera capaz de aprender como su predecesor al ocultar el 15 % de las palabras en la entrada.
  - Que fuera capaz de generar estados ocultos como su predecesor.

Con esto se pretendía conseguir un modelo igual de potente que BERT, pero a su vez más rápido y pequeño.

<sup>21</sup><https://chat.openai.com>

<sup>22</sup><https://www.anthropic.com/news/claude-3-family>

<sup>23</sup><https://ai.google.dev/gemma?hl=es-419>

## Modelos en español

- **Roberta-base-bne** Este es un modelo basado en RoBERTa base que ha sido preentrenado con el mayor conjunto de datos disponible en español hasta la fecha de su entrenamiento. Forma parte del conjunto de modelos MarIA (Gutiérrez-Fandiño et al., 2021) desarrollados por el Barcelona Supercomputing Center y la Biblioteca Nacional de España dentro del Plan Nacional de Tecnologías del Lenguaje español. El corpus de entrenamiento está compuesto por un total de 570 GB de texto limpio, realizado por la Biblioteca Nacional de España a partir del rastreo de páginas web entre 2009 y 2019. Sin embargo, la cantidad de datos para entrenar roberta-base-bne fue inferior, para obtener un modelo más ligero.
- **Roberta-large-bne** Modelo RoBERTa large entrenado de forma similar al anterior.
- **Bertin-roberta-base-spanish** Este es un modelo entrenado desde cero sobre la porción en español de mC4 (Xue et al., 2020), que contiene unos 416M de documentos. Se trata de un modelo RoBERTa-base con 12 capas, 12 attention heads cada una, y 125M de parámetros.
- **Bert-base-spanish-wwm-cased**, conocido como *BETO* (Cañete et al., 2020), es un modelo entrenado en un gran corpus español y que tiene un tamaño similar a BERT-Base. Al igual que en las versiones anglosajonas, esta versión distingue entre mayúsculas y minúsculas.
- **Distilbert-base-spanish-uncased** Se trata de la versión "destilada" de BETO. Fue entrenada por sus mismos autores y tiene el propósito de ofrecer un modelo de un tamaño más reducido que BETO, pero con un rendimiento similar.

## Modelos bilingües

- **xlm-roberta-base** (Conneau et al., 2019) Se trata de la versión multilingüe de RoBERTa, abarca más de 100 idiomas y, a pesar de que en la página de Huggingface asegura que está preentrenada con 2,5TB de datos CommonCrawl, está entrenado con menos cantidad de datos que la versión *large*.
- **xlm-roberta-large** Esta es la versión que sí utiliza toda la cantidad de datos de los 2,5TB. Al igual que el modelo inferior, su entrenamiento está realizado para más de 100 idiomas diferentes.
- **bert-base-multilingual-cased** Modelo BERT preentrenado con los 104 idiomas mejor representados en la Wikipedia.
- **distilbert-base-multilingual-cased** Modelo DistilBERT preentrenado con un corpus similar al modelo anterior.
- **ixa-ehu/ixambert-base-cased** (Otegi et al., 2020) Modelo preentrenado multilingüe para inglés, español y euskera. El corpus de entrenamiento está compuesto por las Wikipedias en inglés, español y euskera, junto con artículos de noticias en euskera extraídos de periódicos online.

### 4.2. Entrenamiento de modelos e hiperparámetros

El entrenamiento eficaz de modelos de lenguaje avanzados es fundamental para que den su mejor rendimiento. Un aspecto fundamental de este proceso es la selección óptima de hiperparámetros, los cuales pueden influir significativamente en la capacidad del modelo para aprender de los datos. En este contexto, hemos implementado una estrategia de búsqueda exhaustiva, conocida como Grid Search, para optimizar los hiperparámetros de nuestros modelos de lenguaje.

Grid Search es una técnica de optimización de hiperparámetros que sistematiza el proceso de experimentación al construir y evaluar un modelo para cada combinación de parámetros especificada en una cuadrícula predefinida. Esto permite identificar la configuración de hiperparámetros que resulta en el mejor rendimiento del modelo.

La aplicación de Grid Search asegura una exploración exhaustiva del espacio de hiperparámetros, proporcionando una base sólida para la toma de decisiones informadas sobre la configuración óptima para el entrenamiento de modelos.

Para la optimización de los modelos de lenguaje, hemos definido la siguiente cuadrícula (grid) de hiperparámetros para explorar:

- **Tamaño de *batch*:** Especifica el tamaño del lote (batch size) durante el entrenamiento. Se consideraron tamaños de 32 y 16, ajustando el equilibrio entre la memoria consumida y la precisión del gradiente.
- **Tasa de aprendizaje:** Se utiliza para el ajuste de los pesos del modelo en cada iteración. Se exploraron valores de 0.00001, 0.00003 y 0.00005, buscando optimizar la velocidad y estabilidad del aprendizaje.
- **Weight decay:** Controla la regularización sobre los pesos del modelo. Se usaron los valores de 0.1 y 0.01. La regularización ayuda a prevenir el sobreajuste al penalizar pesos grandes.

Para cada combinación de hiperparámetros especificada en la grid, se entrenó un modelo de lenguaje desde cero, utilizando un conjunto de datos estandarizado. La evaluación del rendimiento de cada modelo se realizó a través de las métricas específicas que se detallan en la siguiente Sección, tratando de identificar la configuración de hiperparámetros que maximiza el rendimiento del modelo sobre el conjunto de desarrollo. Cada modelo se entrenó un total de 12 veces en diferentes configuraciones sobre cada tarea e idioma.

La aplicación del Grid Search ha permitido una exploración sistemática y completa del espacio de hiperparámetros, identificando la configuración que conduce al mejor rendimiento de los modelos de lenguaje en el contexto de nuestro proyecto. La metodología adoptada asegura que la selección de hiperparámetros se base en evidencia empírica, mejorando la confiabilidad y eficacia de los modelos entrenados, así como la fiabilidad del gap reportado.

Hemos realizado una configuración base para todos los modelos, con alguna alteración para los problemas de clasificación binaria y multiclase, ya que los modelos DistillBert no soportan el '*gradient\_checkpoint*'. Sin embargo, algunos modelos requieren tener activada esta opción debido a la carga de memoria que reciben a la hora de procesar la información en la capa de *Attention*, sobre todo en modelos grandes con muchos parámetros y en modelos que no convergían por el número de pasos o epochs. Según el conjunto de datos y sus tareas, hemos ajustado diferentes configuraciones para comprobar la consistencia de los modelos. En cualquier caso, la configuración final para cada tarea se estableció como única para todos los modelos.

### 4.3. Evaluación: Métricas y baselines

En esta sección resumimos las métricas y los algoritmos baseline, sin información lingüística, utilizados como referencia para calibrar la efectividad de los modelos del lenguaje entre inglés y español.

Para poder comparar resultados en datasets de dos idiomas es necesario estimar primero la dificultad intrínseca de cada dataset, de forma que puedan calibrarse los resultados en un idioma y otro para compararlos directamente. Para ello, en cada dataset hemos estimado la dificultad intrínseca mediante algoritmos de Machine Learning que no usan ningún tipo de información lingüística.

Respecto a las métricas, salvo que se indique lo contrario se ha utilizado la implementación de las métricas en la librería PyEval desarrollada dentro del proyecto.

#### 4.3.1. EXIST-2022

##### Métrica

Para evaluar las dos tareas de EXIST-2022, se ha utilizado F1 Macro (ver informe correspondiente).

##### Baseline

Para generar los baselines de las tareas de EXIST, vectorizamos el conjunto de datos y test sin ningún tipo de preprocesamiento, para evitar utilizar información lingüística. A partir de los conjuntos resultantes, entrenamos modelos de regresión logística, xgboosts y Support Vector Machines (SVM). Tomamos como baseline la media de los resultados obtenidos. El proceso es el mismo para las dos tareas.

#### 4.3.2. DIANN 2023

##### Métrica

La métrica que hemos usado para evaluar la tarea de DIANN 2023 es F1, para ello hemos usado Segeval,<sup>24</sup> un framework en python que evalúa el etiquetado de secuencias.

##### Baseline

Para realizar la tarea de NER de DIANN, utilizamos Conditional Random Fields (CRF), una clase de métodos de modelado estadístico que se aplican a menudo en el reconocimiento de patrones. Lo hemos usado como fórmula de predicción estructurada. No se ha usado ningún tipo de información lingüística.

#### 4.3.3. DIPROMATS 2023

##### Métrica

La evaluación se ha llevado a cabo considerando las tareas como de clasificación, abarcando tanto clasificación binaria, para la Tarea 1, como multietiqueta, para las Tareas 2 y 3. Es importante destacar que la tarea de clasificación multietiqueta de estas dos últimas presenta una complejidad no trivial desde el punto de vista de las métricas de evaluación. Esto se debe a que las clases involucradas mantienen una relación jerárquica entre sí. Por ejemplo, en la Tarea 2, un error de clasificación entre el grupo 2 y el grupo 3 se considera menos grave que un error entre el grupo 2 y el grupo 0.

Por ello, además de las métricas estándar, se reporta la métrica ICM (Amigo and Delgado, 2022), la cual se adapta de manera óptima a las particularidades de nuestra tarea de clasificación. La métrica ICM está diseñada para considerar con severidad variable los errores de clasificación entre diferentes grupos, basándose en su relación jerárquica. Esto la hace especialmente adecuada para tareas en las que los errores entre ciertas clases son inherentemente menos críticos que entre otras, como en DIPROMATS.

Aparte, ya que son métricas estándar, se reportan la Precisión (P), Recall (R) y la puntuación F1 (F1) de cada clase, para proporcionar una evaluación básica del rendimiento de los modelos en las tareas de clasificación mencionadas.

##### Baseline

Para la primera tarea de DIPROMATS, usamos los mismos algoritmos que hemos usado en la clasificación binaria de EXIST 2022. Para las tareas 2 y 3 de DIPROMATS, y continuando con la idea de evitar usar información semántica en los modelos, usamos un clasificador multietiqueta, basado en el algoritmo de KNN (K-Nearest Neighbors).

#### 4.3.4. EXIST-2023

##### Métrica

La métrica empleada fue la oficial de la competición en el modo de evaluación soft-soft (Plaza et al., 2023), ICM-soft. ICM-soft se ha desarrollado dentro del proyecto ODESIA, y es una extensión de la métrica ICM (Amigo and Delgado, 2022) que permite evaluar un sistema bajo el paradigma "learning-with-disagreements" (LeWiDi) (Uma et al., 2021b) comparando sus salidas (dadas como probabilidades de pertenencia a una o varias clases) con un "soft gold standard" especificado de esta misma forma. Además, ICM-soft permite evaluar distintos tipos de problemas de clasificación: binaria (Tarea 1), jerárquica multiclase (Tarea 2) y jerárquica multiclase multietiqueta (Tarea 3). Es la única métrica existente que permite evaluar tareas complejas de clasificación (multilabel, jerárquica o ambas) en modo learning with disagreement.

**Baseline para cada tarea** Para establecer un baseline para cada una de las tareas de EXIST-2023 se entrenó una red neuronal simple con una única capa oculta para la clasificación de los textos. La red se entrenó a 20 epochs con un learning rate de 0.001. El algoritmo de optimización escogido fue Adam. La función de pérdida usada por la red neuronal fue la entropía binaria para las tres tareas. La arquitectura de la red neuronal constó de tres componentes principales: dos capas totalmente conectadas y una función de activación unitaria lineal rectificadora (ReLU). Los textos de entrada suministrados a la red se convirtieron primeramente a vectores de 10.000 dimensiones mediante el método TF-IDF. La última capa de la red

<sup>24</sup><https://huggingface.co/spaces/evaluate-metric/segeval>

neuronal produce salidas que corresponden al número de capas objetivo de cada tarea específica: 2 para la tarea 1, 4 para la tarea 2 y 6 para la tarea 3. Para obtener un baseline robusto en cada tarea, se promediaron los resultados de 10 ejecuciones distintas para cada tarea.

#### 4.3.5. SQUAD/SQAC-2024

##### Métrica

Para evaluar la tarea de SQAC-SQUAD-2024 hemos usado la misma métrica que se utilizó para evaluar tanto SQAC como SQUAD v1.1. Para calcular el F1 score, primero se hace un preprocesamiento de las predicciones y gold standard, luego cada par de respuestas (predicción-gold standard) se tokenizan y se cuenta cuantas palabras coinciden. Con este dato, se calcula el F1.

##### Baseline

El algoritmo baseline consiste en cotejar, mediante distancia coseno cada frase del texto con la pregunta, tomando la más semejante como respuesta candidata.

#### 4.4. Evaluación: resultados experimentales Core Tasks

Los resultados experimentales para la evaluación de modelos en Core Tasks pueden verse en la Tabla 12. Algunos aspectos destacables de la evaluación de estos modelos iniciales son:

- En el leaderboard para español, los mejores resultados los obtienen dos modelos roberta-large: el modelo multilingüe xlm-roberta-large (0,564 de media) y el modelo de María PlanTL-GOB-ES-roberta-large-bne (0,552). En concordancia con lo observado en otros estudios ([Agerri and Agirre, 2022](#)), los modelos entrenados específicamente para el español no son capaces de mejorar los resultados de los modelos multilingües comparables.
- En el leaderboard para inglés, el mejor resultado es el de un modelo monolingüe (roberta-large, 0,587) seguido de cerca por el modelo multilingüe dominante en español (xlm-roberta-large, 0,565). Parece que en inglés, la ventaja de los modelos multilingües no se aprecia como en español, aunque en general son bastante parejos: discrimina más el tamaño del modelo que el hecho de que sea mono o multilingüe.
- Comparando con las baselines no lingüísticas, los modelos de lenguaje aportan mejoras más sustanciales en las tareas más difíciles. En español, por ejemplo, se producen mejoras relativas de más del 100 % respecto al baseline no lingüístico en las tres tareas con más dificultad intrínseca.

| MODELO                             |    | EXIST-2022-T1 | EXIST-2022-T2 | DIPROMATS-T1 | DIPROMATS-T2 | DIPROMATS-T3 | DIANN-2023 | EXIST-2023 T1 | EXIST-2023 T2 | EXIST-2023 T3 | SQUAD/SQAC-2024 | promedio |
|------------------------------------|----|---------------|---------------|--------------|--------------|--------------|------------|---------------|---------------|---------------|-----------------|----------|
| baselines                          | ES | 0,693         | 0,463         | 0,750        | 0,220        | 0,090        | 0,747      | 0,468         | 0,251         | 0,218         | 0,132           | 0,403    |
| xlm-roberta-base                   | ES | 0,740         | 0,500         | 0,790        | 0,360        | 0,099        | 0,840      | 0,616         | 0,402         | 0,320         | 0,369           | 0,504    |
| xlm-roberta-large                  | ES | 0,770         | 0,560         | 0,818        | 0,471        | 0,267        | 0,830      | 0,645         | 0,421         | 0,400         | 0,459           | 0,564    |
| bert-base-multilingual-cased       | ES | 0,720         | 0,470         | 0,780        | 0,350        | 0,103        | 0,840      | 0,600         | 0,369         | 0,333         | 0,323           | 0,489    |
| distilbert-base-multilingual-cased | ES | 0,720         | 0,470         | 0,750        | 0,340        | 0,092        | 0,780      | 0,570         | 0,361         | 0,287         | 0,221           | 0,459    |
| roberta-base-bne                   | ES | 0,740         | 0,560         | 0,814        | 0,420        | 0,121        | 0,750      | 0,628         | 0,396         | 0,369         | 0,406           | 0,520    |
| roberta-large-bne                  | ES | 0,750         | 0,570         | 0,817        | 0,440        | 0,238        | 0,820      | 0,641         | 0,403         | 0,380         | 0,464           | 0,552    |
| bertin-roberta-base-spanish        | ES | 0,730         | 0,490         | 0,759        | 0,360        | 0,078        | 0,730      | 0,621         | 0,390         | 0,333         | 0,417           | 0,491    |
| bert-base-spanish-wwm-cased        | ES | 0,710         | 0,540         | 0,791        | 0,440        | 0,135        | 0,810      | 0,625         | 0,392         | 0,374         | 0,412           | 0,523    |
| distilbert-base-spanish-uncased    | ES | 0,720         | 0,510         | 0,770        | 0,340        | 0,068        | 0,750      | 0,596         | 0,393         | 0,332         | 0,248           | 0,473    |
| baselines                          | EN | 0,674         | 0,437         | 0,707        | 0,210        | 0,080        | 0,665      | 0,437         | 0,220         | 0,206         | 0,125           | 0,376    |
| xlm-roberta-base                   | EN | 0,760         | 0,531         | 0,797        | 0,480        | 0,164        | 0,760      | 0,621         | 0,325         | 0,349         | 0,325           | 0,511    |
| xlm-roberta-large                  | EN | 0,810         | 0,124         | 0,807        | 0,550        | 0,391        | 0,780      | 0,635         | 0,363         | 0,387         | 0,416           | 0,568    |
| bert-base-multilingual-cased       | EN | 0,770         | 0,516         | 0,803        | 0,480        | 0,178        | 0,730      | 0,599         | 0,315         | 0,344         | 0,295           | 0,503    |
| distilbert-base-multilingual-cased | EN | 0,740         | 0,472         | 0,773        | 0,450        | 0,164        | 0,680      | 0,599         | 0,308         | 0,304         | 0,199           | 0,469    |
| roberta-base                       | EN | 0,780         | 0,572         | 0,807        | 0,520        | 0,196        | 0,750      | 0,629         | 0,329         | 0,377         | 0,375           | 0,534    |
| roberta-large                      | EN | 0,810         | 0,603         | 0,819        | 0,550        | 0,472        | 0,790      | 0,643         | 0,351         | 0,403         | 0,463           | 0,590    |
| distilbert-base-uncased            | EN | 0,770         | 0,509         | 0,782        | 0,471        | 0,164        | 0,660      | 0,616         | 0,339         | 0,368         | 0,267           | 0,495    |
| bert-base-cased                    | EN | 0,760         | 0,552         | 0,808        | 0,450        | 0,264        | 0,720      | 0,614         | 0,320         | 0,366         | 0,300           | 0,515    |

Tabla 12: Evaluación de modelos de lenguaje estándar sobre el Leaderboard ODESIA (Core Tasks). La métrica reportada es alguna variante de F1 en todas las tareas excepto las de EXIST 2023, en las que se utiliza ICM-soft normalizado.

## 5. Medición del gap: experimentación extendida con datasets públicos (Extended Tasks)

### 5.1. Medición de la brecha de efectividad español - inglés

Para afinar la brecha de efectividad inglés/español hemos utilizado todos los datasets bilingües disponibles en el Leaderboard ODESIA EXTENDED, que incorpora cuatro datasets bilingües adicionales contruidos en inglés y español de dominio público. Hemos aplicado con ellos la misma metodología que con los de ODESIA CORE. Los datasets adicionales son:

- **DIANN Task 2** La segunda tarea de DIANN consiste en localizar y determinar el alcance de las negaciones en los abstract biomédicos.
- **MLDoc** El Multilingual Document Classification Corpus (MLDoc) ([Schwenk and Li, 2018](#)) es un dataset de clasificación de noticias en varios idiomas, del que usamos el subconjunto de inglés y el subconjunto de español. La tarea tiene cuatro categorías: corporate/industrial, economics, government/social y markets.
- **MultiCONER 2022** ([Malmasi et al., 2022](#)) es un dataset multilingüe para reconocimiento de entidades nombradas complejas de seis categorías diferentes. Como en los demás casos, utilizamos los subconjuntos de español e inglés para nuestra experimentación. En este dataset, el conjunto de entrenamiento es un orden de magnitud más pequeño que el conjunto de evaluación (del orden de 5,000 vs 50,000).
- **STS-2017** ([Cer et al., 2017](#)) es un dataset multilingüe de similitud textual, en la que los sistemas deben predecir el grado de similitud entre un par de oraciones. Esta es la única tarea de regresión en nuestro diseño experimental, y también la única en la que se usa como métrica de evaluación la correlación Pearson entre las predicciones del sistema y las anotaciones manuales. De nuevo, utilizamos los subconjuntos de inglés y español.
- **SQUAD/SQAC**: en este caso se trata de dos datasets contruidos de forma independiente pero con la misma metodología. SQAC (Spanish Question Answering Corpus) ([Gutiérrez-Fandiño et al., 2021](#)) es un dataset de Question Answering extractivo, en el que, dada una pregunta y un párrafo asociado, el sistema debe localizar el span más pequeño que contiene la respuesta. La metodología para crearlo está basada en la de SQuAD v1.1 ([Rajpurkar et al., 2016](#)), por lo que se pueden considerar datasets equivalentes. De hecho, nuestro algoritmo baseline obtiene un rendimiento equivalente en ambos idiomas (0,53 en ambos casos), lo que hace el par SQAC/SQUAD un conjunto ideal para comparar modelos de lenguaje en ambos idiomas. Nótese que este dataset también se utiliza como conjunto de entrenamiento en el SQUAD/SQAC 2024, pero en este último el conjunto de test se ha desarrollado en ODESIA sobre un tipo de documentos ligeramente diferente.

Para todos estos datasets utilizamos los datos públicos de entrenamiento y test, a diferencia de los datasets del Leaderboard ODESIA, en los que los conjuntos de evaluación no están disponibles públicamente. El proceso de entrenamiento para las tareas extendidas es idéntico al de las tareas de ODESIA CORE. En conjunto, la medición del gap se realiza sobre quince tareas.

### 5.2. Métricas y baselines

#### 5.2.1. MLDoc

##### Métrica

Para la tarea de MLDoc, al ser una tarea de clasificación multi-categoría, hemos usado la métrica estándar F1 macro.

##### Baseline

Para este conjunto de datos usamos los algoritmos ya comentados de Regresión logística, Xgboosts y SVM. Al igual que en los anteriores baseline, vectorizamos evitando utilizar información lingüística.

### 5.2.2. MultiCONER 2022

#### Métrica

Al igual que DIANN, MultiCONER es una tarea NER (Name Entity Recongnition), así que hemos usado la misma métrica que para DIANN (F1 en la implementación del script Segeval).

#### Baseline

Para MultiCONER 2022, usamos el mismo algoritmo baseline que para DIANN, CRF.

### 5.2.3. STS 2017

#### Métrica

STS es una tarea de similitud, y para medir esa similitud entre dos frase, usamos la métrica de Pearson Correlation.

La correlación de Pearson se calcula con la siguiente fórmula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

donde:

- $r$  es el coeficiente de correlación de Pearson,
- $x_i$  e  $y_i$  son los valores individuales de las variables  $X$  e  $Y$ ,
- $\bar{x}$  y  $\bar{y}$  son los promedios de las variables  $X$  e  $Y$ ,
- $\sum$  denota la suma sobre todos los datos de muestra.

#### Baseline

En el caso de STS 2017, un problema de regresión, simplemente vectorizamos las dos frases de cada caso mediante la función `TfidfVectorizer` de `sklearn` (Pedregosa et al., 2011), y calculamos el coseno entre ambas representaciones como aproximación de su similitud.

### 5.2.4. SQAC/SQUAD

#### Métrica

Para evaluar la tarea de SQAC-SQUAD hemos usado una versión flexible de F1 (lenient frente a strict). Para calcular este F1 score, primero se hace un preprocesamiento de las predicciones y gold standard, luego, con cada par de predicción-gold standard, se tokenizan y se cuenta cuantas palabras coinciden. Con este dato de todos los pares, se calcula el F1.

**Baseline** El algoritmo baseline consiste en cotejar, mediante distancia coseno como en el caso anterior, cada frase del contexto con la pregunta, quedándose con toda la frase como respuesta candidata. Nótese que al tomar toda la frase como respuesta (para evitar cualquier tipo de proceso lingüístico), la puntuación del baseline es muy baja.

### 5.2.5. DIANN

#### Métrica

La métrica que hemos usado para evaluar la tarea es F1, para ello hemos usado Segeval,24 un framework en python que evalúa el etiquetado de secuencias

#### Baseline

Para obtener el baseline de DIANN Detección de negación utilizamos Conditional Random Fields (CRF), una clase de métodos de modelado estadístico que se aplican a menudo en el reconocimiento de patrones. Lo hemos usado como fórmula de predicción estructurada. La idea, al igual que en las demás experimentaciones, es que no haya información semánticas en los baselines.

## 5.3. Evaluación: resultados experimentales

En la Tabla 13 pueden verse los resultados de nuestra experimentación sobre las tareas del Leaderboard ODESIA EXTENDED. Disponemos de 15 mediciones del gap sobre cada una de las 15 tareas contempladas. En cada una de ellas, los baseline se han calculado como el promedio de varios algoritmos de

aprendizaje supervisado sin conocimiento lingüístico, como se ha explicado anteriormente. Las columnas *mejor ES* y *mejor EN* recogen la efectividad del modelo de lenguaje con mejores resultados en cada tarea en español e inglés, respectivamente.

Es interesante observar que, a pesar de la diversidad de datasets y tareas, en todos los casos menos uno (EXIST 2023 tarea 2, donde el rendimiento en ambos idiomas es similar) se ha medido una brecha favorable al inglés, lo que proporciona una fuerte evidencia estadística de la existencia de esa brecha. Puede apreciarse, también, que hay un valor muy diferente al resto en el caso de la tarea DIPROMATS Task 3. Esa tarea es la más difícil de las contempladas en nuestra experimentación, ya que es un problema de clasificación multiclase y multilabel con 13 clases distribuidas de forma muy desigual. Al ser la más difícil, también es la tarea en la que los modelos de lenguaje aportan una mejora más sustancial respecto a las aproximaciones baseline sin conocimiento lingüístico. Aunque es un resultado en el que merece la pena profundizar, a efectos del cálculo del gap debemos descartarlo por razones estadísticas, ya que se trata de un outlier ( $p < 0,01$  según el test de Grubbs<sup>25</sup>).

Una vez eliminado el outlier, el indicador de brecha de efectividad  $E_{1.a}$  se ha calculado según se describe en la sección 2.10 sobre las otras catorce tareas. **El resultado final (con su error estándar) es una estimación de la brecha porcentual del  $20 \pm 06$  a favor del inglés.** Aunque hay una variación apreciable entre tareas, el error estándar nos indica que, en cualquier caso, la brecha real promedio estará en una horquilla entre el 14 % y el 26 %. Estos datos son compatibles con los obtenidos el primer año.

| Tarea                     | baseline ES                   | Baseline EN | mejor ES | mejor EN | brecha % |
|---------------------------|-------------------------------|-------------|----------|----------|----------|
| <b>Core Tasks</b>         |                               |             |          |          |          |
| EXIST-2022 Task 1         | 0,693                         | 0,674       | 0,770    | 0,810    | 16,54    |
| EXIST-2022 Task 2         | 0,463                         | 0,437       | 0,570    | 0,580    | 9,803    |
| DIPROMATS-2023 Task 1     | 0,750                         | 0,707       | 0,818    | 0,819    | 11,60    |
| DIPROMATS-2023 Task 2     | 0,220                         | 0,210       | 0,471    | 0,550    | 47,80    |
| DIPROMATS-2023 Task 3     | 0,090                         | 0,080       | 0,267    | 0,472    | 293,33   |
| DIANN Task 1              | 0,747                         | 0,665       | 0,840    | 0,790    | 0,38     |
| EXIST-2023 Task 1         | 0,469                         | 0,437       | 0,645    | 0,643    | 9,52     |
| EXIST-2023 Task 2         | 0,251                         | 0,220       | 0,421    | 0,363    | -2,70    |
| EXIST-2023 Task 3         | 0,218                         | 0,206       | 0,400    | 0,403    | 12,10    |
| SQUAD/SQAC-2024           | 0,132                         | 0,125       | 0,464    | 0,463    | 18,60    |
| <b>Extended Tasks</b>     |                               |             |          |          |          |
| DIANN Task 2              | 0,878                         | 0,575       | 0,960    | 0,917    | 71,80    |
| MLDoc                     | 0,930                         | 0,883       | 0,970    | 0,980    | 39,86    |
| MultiCoNER 2022           | 0,523                         | 0,553       | 0,710    | 0,750    | 4,86     |
| STS-2017                  | 0,680                         | 0,707       | 0,800    | 0,860    | 14,76    |
| SQUAD/SQAC                | 0,533                         | 0,528       | 0,770    | 0,883    | 24,57    |
| <b>Brecha efectividad</b> | <b><math>20 \pm 06</math></b> |             |          |          |          |

Tabla 13: Cálculo de la brecha de efectividad. Los baseline se han calculado como el promedio de varias algoritmos de aprendizaje supervisado sin conocimiento lingüístico. Los LLM son la efectividad del modelo de lenguaje con mejores resultados en cada tarea. El indicador de brecha de efectividad  $E_{1.a}$  se ha calculado según se describe en la sección 2.10, eliminando del promedio el outlier *DIPROMATS Task 3* y reportando el error cuadrático medio sobre el resto de mediciones. Un número positivo indica una brecha porcentual a favor del inglés, y negativo a favor del español.

<sup>25</sup>Hemos utilizado <https://www.graphpad.com/quickcalcs/grubbs2/>

## 6. Medición del gap: experimentación extendida con modelos generativos

### 6.1. UNED ACCESO: Tests de conocimiento general de acceso a la universidad

Para esta tarea se ha usado el dataset UNED ACCESO 2024. Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset Exámenes UNED 2024", entregado con este informe.

Este dataset contiene 1003 preguntas de respuesta múltiple de varias asignaturas del Curso de acceso para Mayores de 25 años de la UNED, en español con su correspondiente traducción manual al inglés, de los siguientes grados: Administración y Dirección de Empresas, Biología, Bioquímica, Economía, Fundamentos de Informática, Lengua Castellana, Literatura, Matemáticas, Matemáticas Aplicadas a las Ciencias Sociales, Matemáticas Avanzadas y Psicología.

El objetivo es evaluar las competencias de los LLMs tanto propietarios (GPT-3.5, GPT-4 y Claude-3-Opus), como open-source (Llama 2, Mistral y Gemma). En particular queremos:

- Medir la brecha en el rendimiento de los modelos generativos en español respecto al inglés.
- Medir la efectividad según las competencias que requiere cada pregunta. Este año se comparará el rendimiento por disciplina. Como extensión para el próximo año, se clasificarán las preguntas según el tipo de competencias necesarias para contestarlas (memorización, relación o razonamiento) para evaluar a los modelos en estas dimensiones.
- Efecto de las variaciones en la consulta (prompt). Para evaluar cómo de sensibles al prompt son los modelos se realizan ligeras variaciones, como el formato con el que se enumeran las respuestas o el orden en el que se proporcionan. Esto se describe a continuación.
- Efecto de la contaminación. Las preguntas se han extraído de un repositorio privado para minimizar los problemas de contaminación (es decir, que los modelos hayan visto las respuestas correctas en su fase de entrenamiento). Sin embargo, es necesario comprobar el nivel de contaminación al que pueden estar expuestas, ya sea porque las propias preguntas se hayan sacado de fuentes públicas o porque alguien haya publicado las respuestas en sitios no oficiales. Para estudiar cómo varía el rendimiento entre datasets con y sin contaminación debemos cuantificar qué es contaminación en nuestro dataset, ya que se trata de conocimiento general que en muchos casos puede estar repetido en internet. De esta manera se podrá discernir aún con más precisión si el rendimiento de los modelos se debe tan solo a las capacidades de memorización o si existe cierto nivel de razonamiento.

#### 6.1.1. Métrica

Los resultados se presentan por separado con dos métricas distintas: la Precisión y el coeficiente Kappa de Cohen.

La Precisión se calcula como la proporción de aciertos sobre el número total de respuestas generadas, mientras que la Kappa se calcula a partir de la precisión según la siguiente fórmula:

$$\text{Kappa} = \frac{\text{observed accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}}$$

donde la precisión esperada se corresponde con la que se obtiene respondiendo al azar: 1/3 en el caso de las preguntas con 3 respuestas posibles y 1/4 en el caso de preguntas con 4 respuestas. Esta medida pondera el nivel de acierto de tal manera que, si se responde al azar, se obtenga un cero. Así se pueden comparar los resultados entre exámenes con distinto número de respuestas posibles y que, por tanto, dan resultados distintos para una contestación puramente aleatoria. Siendo los valores de la precisión esperada 1/3 y 1/4 en este caso, y dado que la precisión toma valores entre 0 y 1, el valor de la Cohen's Kappa oscila entre -1/2 y 1 (tomando valores negativos cuando la precisión observada es menor que la contestación aleatoria).

Tras obtener los valores de Precisión y Kappa para cada asignatura, finalmente se calcula el resultado de cada run (modelo + idioma + configuración) haciendo la media aritmética entre los valores obtenidos en las asignaturas. También se reporta la media ponderada según el número de preguntas por asignatura (de manera que asignaturas con un número muy pequeño de preguntas contribuyen menos al resultado final, y viceversa).

### 6.1.2. Baseline

Como las preguntas son las mismas en ambos idiomas y han sido traducidas manualmente, la dificultad intrínseca es la misma en inglés y en español, y por tanto no es necesario calcular baselines no lingüísticas: se pueden comparar directamente los resultados entre idiomas.

### 6.1.3. Descripción de experimentos y parametrización

A continuación se describen los modelos utilizados, los parámetros de configuración de los experimentos, y las distintas variaciones realizadas a lo largo de los mismos.

**Modelos.** Se ha realizado experimentación con 6 modelos generativos: tres de ellos propietarios (GPT-3.5, GPT-4, y Claude-3-Opus), y tres abiertos (Llama2, Gemma y Mistral). En concreto, estos son los modelos utilizados:

- **gpt-4-1106-preview.** GPT-4 (OpenAI et al., 2024) es un modelo multimodal de grandes dimensiones, disponible desde la API de OpenAI, optimizado tanto para chat como para la predicción de tokens tradicional. En concreto, el modelo utilizado es un GPT-4 Turbo mejorado para seguir instrucciones.
- **gpt-3.5-turbo-1106.** Modelo GPT-3.5 Turbo mejorado para seguir instrucciones. Al igual que GPT-4, está disponible desde la API de OpenAI.
- **claude-3-opus-20240229.** Claude-3<sup>26</sup> es la última familia de modelos propietarios lanzados por Anthropic, accesibles a través de API. En concreto, Claude-3-Opus es el modelo Claude-3 más potente.
- **google-gemma-7b-it.** Gemma es la familia de modelos abiertos basados en la tecnología de Gemini lanzada por Google<sup>27</sup>. Está disponible en HuggingFace, y en este caso también se ha elegido la variante mejorada para seguir instrucciones.<sup>28</sup>
- **meta-llama-Llama-2-7b-chat-hf.** Llama-2 (Touvron et al., 2023) es la familia de modelos abiertos lanzada por Meta. Está disponible en HuggingFace y se ha elegido la variante mejorada para instrucciones.<sup>29</sup>
- **mistralai-Mistral-7B-Instruct-v0.2.** Mistral (Jiang et al., 2023) es la familia de modelos abiertos *Sparse Mixture of Experts* de Mistral AI. Está disponible en HuggingFace y para los experimentos se ha utilizado la variante mejorada para seguir instrucciones.<sup>30</sup>

**Parámetros.** En cuanto a los parámetros, en todos los casos se ha configurado la temperatura a 0, lo cual proporciona respuestas más deterministas, como se requiere en este dataset. Las preguntas se suministran de una en una salvo en el caso de GPT-4 y GPT-3.5, en el que se proporcionan en lotes de 20 preguntas. El prompt es fijo, ligeramente distinto entre los modelos GPT y el resto de modelos, debido a la manera en que se suministran las preguntas (por los lotes en el caso de GPT) y a las propias diferencias que existen entre las APIs (GPT y Claude) y el entorno de HuggingFace. Como información

<sup>26</sup><https://www.anthropic.com/news/claude-3-family>

<sup>27</sup><https://ai.google.dev/gemma?hl=es-419>

<sup>28</sup><https://huggingface.co/google/gemma-7b-it>

<sup>29</sup><https://huggingface.co/google/gemma-7b-it>

<sup>30</sup><https://huggingface.co/mistralai/Mistral-8x7B-Instruct-v0.1>

extra, en el prompt se proporciona el nombre de la asignatura a la que pertenece la/s pregunta/s. El idioma del prompt es siempre el mismo que el idioma del dataset en el que se está evaluando (español o inglés).

A continuación se muestra el prompt en español e inglés para los modelos GPT-3.5 y GPT-4:

```
prompt_español = 'Eres un sistema de respuesta de preguntas.
Te voy a enseñar una serie de preguntas de la asignatura {}'.
Cada una empieza con un ID único.
Para cada pregunta quiero que respondas con el ID de la pregunta
seguido por un espacio y la letra de la respuesta correcta, como
en el siguiente ejemplo:
`User: E000012345A10F1-00 La capital de España es: A. Barcelona
B. Sevilla C. Madrid
E000012345A20F1-02 ¿Cuál es la opción correcta ortográficamente?:
A. Hay que ir ahí B. No hayaba la respuesta. C. Se calló al suelo.
Assistant: E000012345A10F1-00 C
E000012345A20F1-00 A`' '

prompt_inglés = 'You are a Question Answering system.
I will present to you a batch of questions that cover the field of {}'.
Each question starts with a unique ID.
For each question I want you to answer on one line with the ID followed
by a space and the letter of the correct answer, as in the following
example:
`User: E000012345A10F1-00 The capital of Spain is: A. Barcelona
B. Seville C. Madrid
E000012345A20F1-02 Which is the orthographically correct
choice: A. Hay que ir ahí B. No hayaba la respuesta.
C. Se calló al suelo.
Assistant: E000012345A10F1-00 C
E000012345A20F1-00 A`'
```

Este prompt se proporciona en el rol sistema, mientras que en el rol de usuario se proporciona el batch de 20 preguntas.

Para los modelos abiertos (Gemma, Llama-2 y Mistral) y Claude-3 se proporciona un prompt de sistema, uno de usuario y uno de asistente. El prompt del sistema es la petición inicial que se proporciona al modelo para solicitar una respuesta específica, mientras que el prompt del usuario es la entrada proporcionada por el usuario humano al interactuar con el modelo. El prompt de asistente se refiere al propio modelo de lenguaje, que actúa como un asistente virtual capaz de generar texto en respuesta a los prompts del sistema y del usuario.

```
system_prompts = {
    'es': 'Eres un sistema experto en responder preguntas de exámenes.',
    'en': 'You are an expert system for answering exam questions.',
}

user_prompts = {
    'es': "Responde a la siguiente pregunta de la asignatura {},
tan solo con la letra de la respuesta correcta.\nPregunta: {}",
    'en': "Answer the following question of the subject {}
only with the letter of the correct answer.\nQuestion: {}"
}

assistant_prompts = {
    'es': 'Letra de la respuesta correcta: ',
    'en': 'Letter of the correct answer: ',
}
```

Además de esto, para los modelos abiertos se da formato a la instrucción según las etiquetas con las que ha sido entrenado:

```
def _instruction_format(self, sys_message: str, query: str, assistant : str):
    if 'gemma' in self.model_id:
        prompt = f'''<start_of_turn>user
{sys_message}

{query}<end_of_turn>
<start_of_turn>model
{assistant}'''
```

```
return prompt
elif 'mistral' in self.model_id or 'mixtral' in self.model_id:
    return f'<s> [INST] {sys_message} [/INST]\n
    User: {query}\nAssistant: {assistant}'
elif 'llama' in self.model_id:
    return f'""<s>[INST] <<SYS>> {sys_message} <</SYS>>\n
    {query} [/INST] {assistant}""'
```

Finalmente la salida literal de los modelos se limpia y estructura antes del script de evaluación, ya que, en muchas ocasiones, la respuesta contiene justificaciones de la respuesta u otro tipo de material adicional.

**Configuraciones.** Se ha experimentado con distintos prompts para tener una idea más precisa de las capacidades intrínsecas de los modelos, ya que se sabe que variaciones del prompt pueden dar lugar a respuestas muy diferentes. Las distintas configuraciones, o variaciones de prompt que se han realizado en los experimentos son las siguientes:

1. Las preguntas se proporcionan por orden y sin mezclar asignaturas.
2. Se baraja el orden en el que se proporcionan las preguntas dentro de cada asignatura.
3. Se cambia el formato A. B. C. D. por A) B) C) D)
4. Se cambia el formato A. B. C. D. por E. F. G. H.
5. Se permuta el orden de las respuestas
6. Se cambia el orden de las respuestas de manera que la respuesta correcta siempre es la A.

En el caso de los modelos GPT-3.5 y GPT-4 se realizan las 6 variaciones, ya que debido a los costes de la API las preguntas se proporcionan en batches de 20. En cambio, en el resto de modelos (open source y Claude) las preguntas se proporcionan de una en una, por lo que se omite la configuración 2. En total se obtienen 64 resultados de (2 modelos GPT x 2 idiomas x 6 config) + (4 modelos x 2 idiomas x 5 config). Para obtener el resultado final de cada modelo en un idioma, se promedia entre las configuraciones 1, 3, 4 y 5. La configuración 2 se excluye porque no es aplicable a los modelos abiertos, y la configuración 6 se excluye porque proporciona resultados engañosos: al ser la respuesta correcta siempre la A, premia a los modelos que tienen un sesgo a priori que les hace escoger la letra A con más frecuencia.

#### 6.1.4. Resultados

En esta sección se discuten los resultados obtenidos en términos de la Cohen's Kappa. La Tabla 14 muestra los resultados para cada modelo promediados entre las configuraciones 1, 3, 4 y 5. Pueden hacerse las siguientes observaciones:

- Los modelos propietarios más potentes superan ampliamente a la muestra de modelos abiertos de nuestra experimentación. El mejor modelo en español (Claude-3-Opus) tiene una mejora porcentual del 99 % respecto al mejor modelo abierto (Mistral-7B): 0,78 frente a 0,39.
- Los dos mejores modelos (Claude-3-Opus y GPT-4) tienen ligeramente mejores resultados en español que en inglés. Teniendo en cuenta que las preguntas estaban originalmente en español, y que algunas soluciones han circulado online entre los estudiantes, es muy posible que esa inversión indique algún tipo de contaminación (y, quizás, algún artefacto de traducción), más que una ausencia real de brecha español-inglés.
- En los modelos abiertos, los resultados son siempre mejores en inglés, en distintos porcentajes.

A continuación se detallan algunos resultados según distintas dimensiones.

| Modelo        | Idioma | Cohen's Kappa |
|---------------|--------|---------------|
| Claude-3-Opus | es     | 0,7796        |
|               | en     | 0,7676        |
| GPT-4         | es     | 0,7786        |
|               | en     | 0,7629        |
| GPT-3.5       | es     | 0,5313        |
|               | en     | 0,5358        |
| Mistral-7B    | es     | 0,3914        |
|               | en     | 0,4492        |
| Gemma-7b      | es     | 0,3720        |
|               | en     | 0,4172        |
| Llama-2-7b    | es     | 0,2425        |
|               | en     | 0,2660        |

Tabla 14: Resultados por modelo e idioma. La Cohen's kappa ha sido promediada entre las configuraciones de prompt 1, 3, 4 y 5.

| Modelo        | Brecha % EN-ES |
|---------------|----------------|
| Claude-3-Opus | -1,54          |
| GPT-4         | -2,01          |
| GPT-3.5       | 0,86           |
| Mistral-7B    | 14,78          |
| Gemma-7b      | 12,13          |
| Llama-2-7b    | 9,67           |

Tabla 15: Brecha porcentual inglés-español por modelo.

**Cálculo de la brecha de efectividad** La Tabla 15 muestra la brecha porcentual de efectividad para cada modelo, calculada como  $100 \cdot (\text{EN-ES}) / \text{ES}$  a partir de las precisiones promediadas de la Tabla 14).

Como habíamos comentado, se observa una ligera diferencia a favor del español en el caso de Claude y GPT-4. En los modelos abiertos se observa una brecha promedio del 12,20 % (mayor en el modelo abierto de más rendimiento, y más baja en el peor). Estos resultados sugieren varias hipótesis: que en el caso de los modelos propietarios tenga un gran impacto la contaminación, y/o que las traducciones generen artefactos al estar las preguntas originales en español. En cualquier caso, es muy probable que haya contaminación y por tanto hay que asegurarse antes de promediar estos resultados en el cálculo global de la brecha de ODESIA.

**Resultados por disciplina.** La Tabla 16 muestra los resultados de cada modelo, idioma en cada asignatura promediados según las configuraciones 1, 3, 4 y 5, ordenados según el valor que toma la media de las asignaturas. Claude aprueba todas las asignaturas menos Matemáticas Avanzadas, lo que confirma que ni los modelos más avanzados son capaces de razonar como fenómeno emergente. Consigue resultados por encima del 0.8 en seis de las once asignaturas (ADE, Biología, Bioquímica, Economía, Fundamentos de Informática y Psicología). En Lengua Castellana y Literatura se mantiene en torno al 0.68 y el 0.80 respectivamente, y alrededor del 0.58 en Matemáticas y Matemáticas Aplicadas a las Ciencias Sociales. Los resultados de GPT-4 son muy similares. Con GPT-3.5, los resultados bajan en general al menos un par de décimas, salvo en las tres asignaturas de Matemáticas en las que los resultados caen hasta los 0.1-0.2 puntos (cerca de dar respuestas aleatorias). Los resultados siguen bajando con Mistral-7B, especialmente en las asignaturas de Lengua Castellana, Literatura y las de Matemáticas. Por último, los resultados más bajos se obtienen en general con Gemma y Llama-2, que en algunos casos se comportan aún peor que una contestación al azar (valores por debajo de cero).

| modelo        | idioma | ADE    | Biología | Bioquímica | Economía | F. Informática | Lengua Castellana | Literatura | Matemáticas | Matemáticas CCSS | Matemáticas Avanzadas | Psicología | Media  |
|---------------|--------|--------|----------|------------|----------|----------------|-------------------|------------|-------------|------------------|-----------------------|------------|--------|
| Claude-3-Opus | es     | 0.8577 | 0.9370   | 1.0000     | 0.8693   | 0.9259         | 0.6772            | 0.8022     | 0.6250      | 0.5971           | 0.4375                | 0.8468     | 0.7796 |
| GPT-4         | es     | 0.8232 | 0.9747   | 1.0000     | 0.9084   | 0.9576         | 0.6844            | 0.7143     | 0.5068      | 0.5013           | 0.6093                | 0.8844     | 0.7786 |
| Claude-3-Opus | en     | 0.8018 | 0.9496   | 1.0000     | 0.8759   | 0.9312         | 0.6773            | 0.7693     | 0.5736      | 0.5891           | 0.4375                | 0.8387     | 0.7676 |
| GPT-4         | en     | 0.8103 | 0.9590   | 1.0000     | 0.8888   | 0.9576         | 0.6241            | 0.7875     | 0.4863      | 0.5013           | 0.5469                | 0.8306     | 0.7629 |
| GPT-3.5       | en     | 0.6336 | 0.8487   | 0.8919     | 0.5816   | 0.8412         | 0.4255            | 0.4871     | 0.1575      | 0.1183           | 0.2500                | 0.6586     | 0.5358 |
| GPT-3.5       | es     | 0.6120 | 0.7416   | 0.8156     | 0.5424   | 0.8148         | 0.4964            | 0.5091     | 0.2243      | 0.2460           | 0.1562                | 0.6854     | 0.5312 |
| Mistral-7B    | en     | 0.5430 | 0.7069   | 0.7585     | 0.6078   | 0.7672         | 0.2589            | 0.3407     | 0.0549      | 0.2461           | 0.0313                | 0.6264     | 0.4492 |
| Gemma-7B      | en     | 0.4655 | 0.6974   | 0.8220     | 0.5294   | 0.7513         | 0.1454            | 0.2564     | 0.1576      | 0.1742           | -0.0156               | 0.6048     | 0.4172 |
| Mistral-7B    | es     | 0.4612 | 0.6785   | 0.6695     | 0.3986   | 0.7513         | 0.1737            | 0.3553     | 0.0394      | 0.2061           | 0.0000                | 0.5712     | 0.3914 |
| Gemma-7B      | es     | 0.3836 | 0.6533   | 0.7458     | 0.3072   | 0.6032         | 0.1028            | 0.3370     | 0.1319      | 0.1782           | 0.0938                | 0.5552     | 0.3720 |
| Llama-2-7B    | en     | 0.3965 | 0.5714   | 0.4216     | 0.2549   | 0.6085         | 0.0957            | 0.2857     | 0.0548      | 0.1423           | -0.0468               | 0.4368     | 0.2929 |
| Llama-2-7B    | es     | 0.2844 | 0.4012   | 0.3263     | 0.1699   | 0.5609         | 0.1064            | 0.3150     | 0.0548      | 0.1184           | -0.0624               | 0.3925     | 0.2425 |

Tabla 16: Resultados por modelo, idioma y asignatura promediados según las configuraciones 1, 3, 4 y 5.

La Figura 11 muestra la media aritmética de todos los modelos, idiomas y configuraciones para cada asignatura. Se aprecia el alto rendimiento en las asignaturas de Biología, Bioquímica, Fundamentos de Informática y Psicología, y los resultados bajos en las asignaturas de Matemáticas. Es probable que los resultados altos sean en preguntas con alto contenido memorístico, esperamos tener más detalles sobre este aspecto cuando clasifiquemos las preguntas en función de las competencias necesarias para contestarlas.

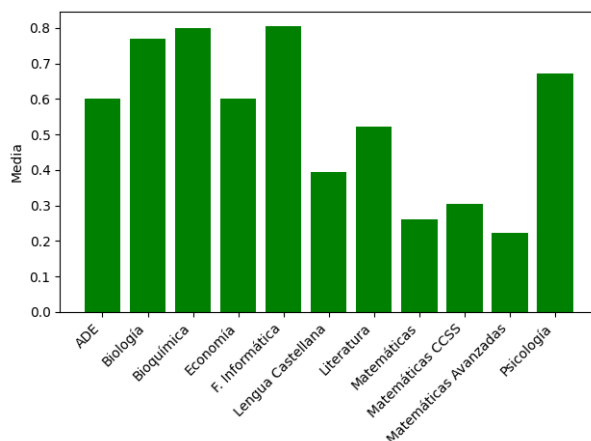


Figura 11: Media de cada asignatura para todos los modelos, idiomas y configuraciones

La Figura 12 muestra los resultados para cada modelo e idioma en las distintas disciplinas, en la configuración 1 (tomada como referencia en este caso por ser la configuración base). Lo que se observa de nuevo es cierta consistencia en cuanto que los modelos que obtienen mejores resultados lo hacen en todas las asignaturas en general, y viceversa. En el caso de modelos abiertos como Gemma y Llama-2 incluso se observan valores negativos, ya que obtienen resultados peores que contestando al azar.

Un caso llamativo es el de la asignatura de Bioquímica, en la que tanto Claude como GPT-4 consiguen acertar el 100 % de las preguntas en los dos idiomas y en todas las configuraciones.

**Variaciones.** La Figura 13 muestra los resultados para cada modelo, idioma y configuración, separados en bloques por colores. Lo que se observa es que los resultados son bastante consistentes a lo largo de las

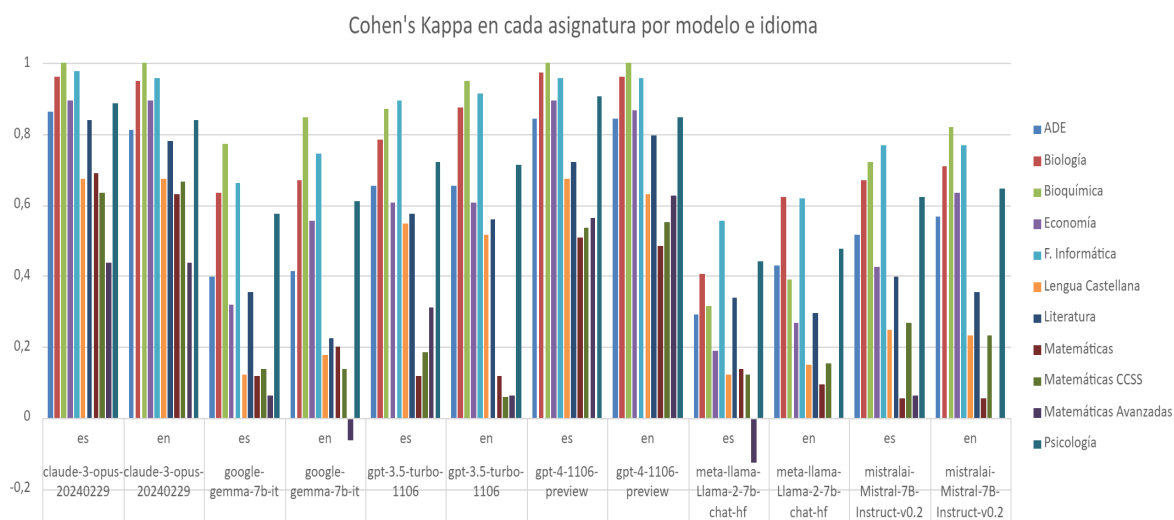


Figura 12: Media por modelo, idioma y configuración

configuraciones, y que las variaciones (cambios de orden y formato) no tienen un impacto drástico en los resultados.

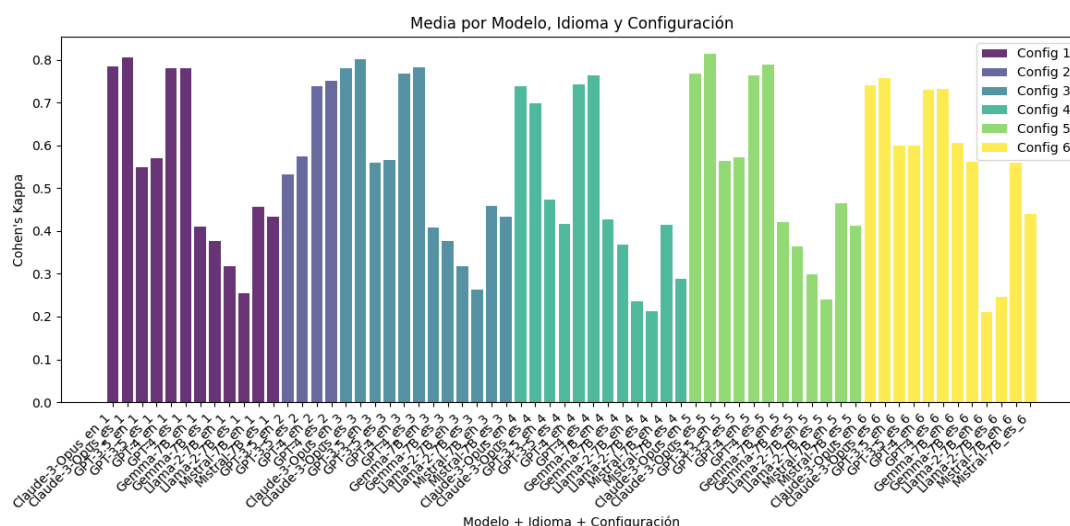


Figura 13: Media por modelo, idioma y configuración

Curiosamente, tanto Claude como GPT-4 alcanzan sus mejores resultados en la configuración 5, en la que se permuta el orden de las respuestas; esto indica que, al menos, no hay memorización del identificador de la respuesta correcta ni del orden en el que aparece (en caso de que haya contaminación). Por su parte, GPT-3.5, Llama-2 y Gemma alcanzan sus mejores resultados en la configuración 6, en la que la respuesta siempre es la A. Esto indica que hay un sesgo de los modelos a responder siempre la primera opción, y no un rendimiento mejor real en esa configuración.

Por otro lado, a modo de ejemplo la Figura 14 muestra el rendimiento por asignatura en cada configuración. Se observa que en las asignaturas en las que los resultados son altos (Bioquímica, Biología, Fundamentos de Informática, Economía o ADE) estos apenas varían a lo largo de las configuraciones, mientras que en las disciplinas en las que los resultados son más bajos, bien porque la tarea es más

compleja o bien porque hay menos preguntas de examen (como en las tres asignaturas de Matemáticas) los resultados presentan más oscilaciones al variar el tipo de consulta.

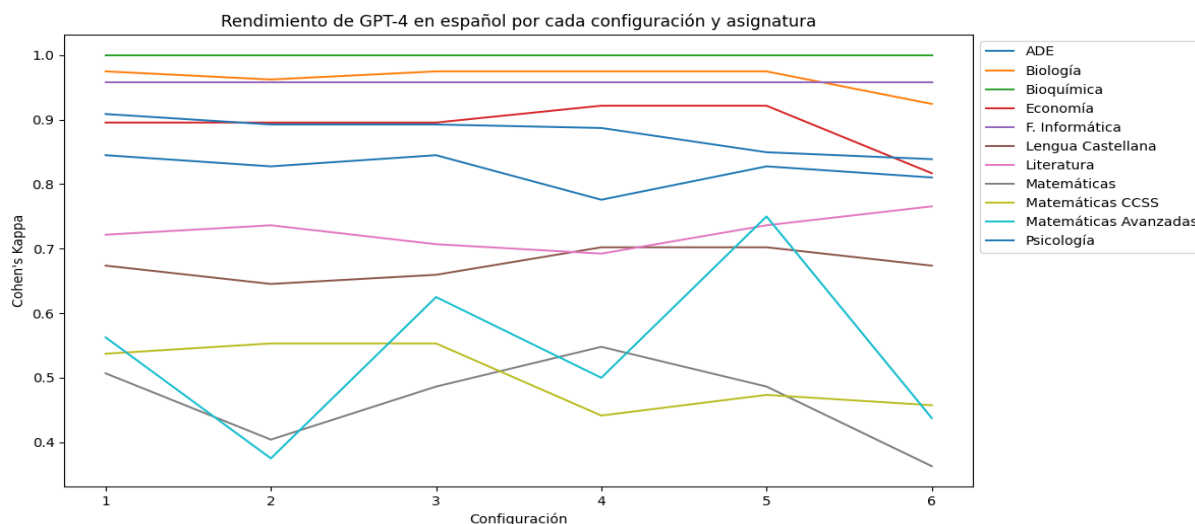


Figura 14: Precisión media por modelo, idioma y configuración

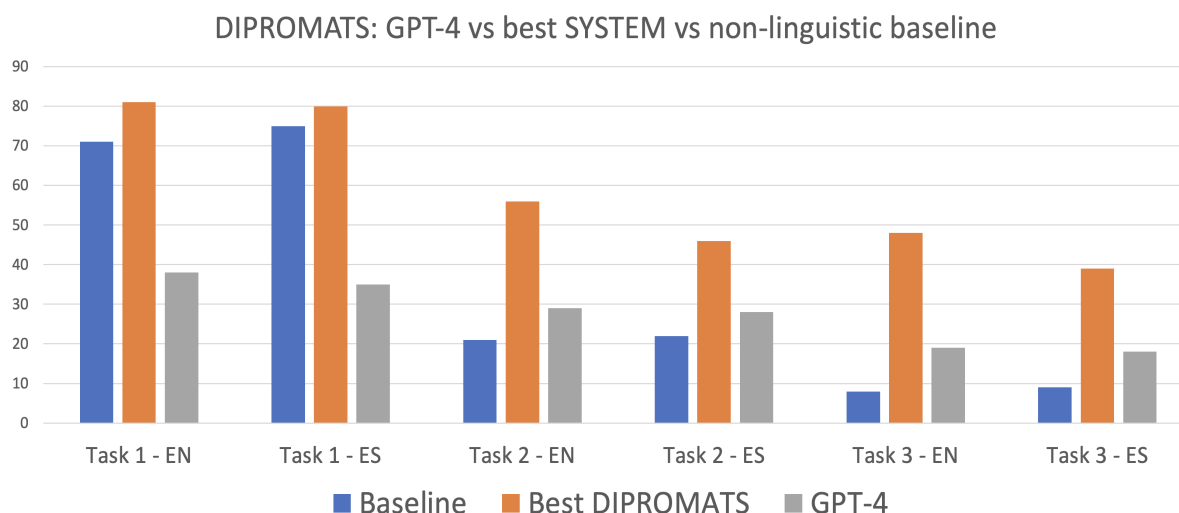


Figura 15: Resultados de GPT-4 en modo zero-shot sobre las tareas ODESIA CORE DIPROMATS

## 6.2. GPT-4 y DIPROMATS

En este segundo año hemos hecho también una primera experimentación del rendimiento de los modelos generativos más potentes (en concreto, GPT-4) en las tareas del leaderboard ODESIA en modo zero-shot o few-shot.

Es interesante aplicar los mejores modelos sobre las tareas ODESIA CORE, puesto que tenemos garantizada la no contaminación de los experimentos: el test set no ha sido distribuido públicamente y tampoco de manera restringida. Por otro lado, esperamos que el rendimiento no sea tan alto como el de los sistemas supervisados, aunque es difícil hacer predicciones con modelos tan potentes como GPT-4.

Hemos planteado la tarea de anotación del test set de DIPROMATS 1, 2 y 3 a GPT-4 en modo zero-shot, incluyendo en el prompt una versión ligeramente simplificada de la guía de anotación y pidiéndole que procesara los tweets por batches. Hay que destacar que las tareas 2 y 3 de DIPROMATS son especialmente difíciles en modo supervisado porque la información de entrenamiento está muy desequilibrada entre

clases, y hay muchas clases. Nos interesaba comprobar hasta qué punto resolverla en modo no supervisado podía ayudar a no depender de unos datos de entrenamiento sesgado.

En la Figura 15 se muestran los resultados de GPT-4 en comparación con las baselines no lingüísticas y los mejores sistemas de la competición DIPROMATS 2023. Los resultados son muy llamativos: en ausencia de datos de entrenamiento, GPT-4 tiende a identificar muchos falsos positivos en la tarea binaria de detección de propaganda; eso hace que sus resultados sean muy bajos (peor que aleatorios). Los sistemas supervisados (incluida la baseline no lingüística) han aprendido del training set que la clase positiva es menos frecuente que la negativa, y no cometen ese error. En la tarea dos (con cinco clases) y la tarea tres (con 13 clases), son problemas difíciles de abordar sin modelos de lenguaje, y en estos casos GPT-4 supera holgadamente a las baselines no supervisadas. Sin embargo, su rendimiento sigue siendo mucho peor que el de los mejores sistemas de la competición, lo que corrobora el principio de que, disponiendo de datos de entrenamiento, el uso de modelos discriminativos no sólo es mucho más barato y eficiente, sino también más eficaz que recurrir a los grandes modelos generativos.

En términos de rendimiento comparado entre inglés y español, los resultados son consistentes con los de la brecha en los modelos discriminativos: hay una brecha del 6,85 % en la primera tarea, del 10,82 % en la segunda, y del 37,50 % en la tercera, para una media de 18,40 %, dentro del rango del  $20 \pm 6$  % medido para los modelos discriminativos, a pesar de ser una muestra de tareas mucho más pequeña y más sesgada. Estos resultados son una primera indicación de que el paso de los modelos discriminativos a los grandes modelos generativos quizás no ha supuesto cambios drásticos en la brecha lingüística de rendimiento. En cualquier caso, será necesario realizar experimentación adicional para obtener datos consolidados.

### 6.3. CURIA: Generación de microresúmenes legales

Para medir el gap en el dominio legal con una tarea de generación de textos se ha creado el corpus CURIA 2024. El corpus está compuesto por sentencias de tribunales de justicia de la Unión Europea en inglés y español. Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset Exámenes UNED 2024".

En concreto, la tarea consiste en generar micro resúmenes (de unos 30 tokens en promedio) en lenguaje claro (y no especializado) a partir de sentencias que tienen una media aproximada de 9000 tokens en español y 8000 en inglés.

Si bien la creación del corpus se ha realizado dentro del Año 2, la evaluación se llevará a cabo dentro del tercer año. No obstante, sí se ha planificado cómo se evaluará la tarea y qué tipo de experimentos se realizarán, lo que se presenta a continuación.

#### 6.3.1. Métrica

La evaluación de una tarea de generación automática de resúmenes puede llevarse a cabo por medio de métricas intrínsecas o extrínsecas.

En el primer caso, la evaluación consistiría en analizar los resúmenes generados en función de características internas del texto como la *fluidez* (cada oración debe estar bien formada y libre de errores gramaticales), la *coherencia* (el resumen debe estar bien estructurado y no simplemente contener información), la *relevancia* (el resumen debe contener los aspectos más importantes del documento fuente y excluir el resto) o la *consistencia* (el resumen y el documento fuente deben ser objetivamente consistentes; es decir, no debe haber información nueva en el resumen que no esté en la fuente) (Kryscinski et al., 2019).

En segundo lugar, y de modo similar al caso de la traducción automática, la evaluación extrínseca consiste en la comparación de los resúmenes generados por los modelos generativos frente a referencias; es decir, comparar la salida del sistema con un *gold standard*. En esta experimentación usaremos evaluación extrínseca aplicando dos métricas con características diferentes:

- La métrica **ROUGE** (Lin, 2004), junto con sus variables (ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, ROUGE-SU), cuentan el número de n-gramas superpuestos entre el resumen generado y la referencia, y sus variantes consideran algunas alternativas como la subsecuencia común más larga

(LCS) entre la salida y la referencia, o el bigrama de salto entre la salida y la referencia; los bigramas de salto son pares de palabras que mantienen el orden de sus oraciones independientemente de las palabras que puedan intercalarse entre ellas. En implementaciones por defecto de ROUGE se utiliza WordNet para obtener sinónimos de sustantivos, verbos y adjetivos en inglés, y puede sustituirse por cualquier diccionario de sinónimos específico de un dominio o lengua.

- La segunda métrica que usaremos para evaluar el gap en generación de resúmenes en el dominio legal es **BertScore** (Zhang\* et al., 2020). Esta métrica se considera estado del arte y se fundamenta en una medida anterior, METEOR (Banerjee and Lavie, 2005), propuesta para superar las limitaciones derivadas de la exigencia de concordancia exacta de n-gramas que sufría ROUGE, así como la limitación a nivel semántico que mostraba. Con METEOR se permite la coincidencia entre frases estructuradas de forma diferente. Pero BertScore, además, utiliza la similitud coseno para comparar cada token o n-grama del resumen generado con la referencia. La idea subyacente es que la distancia coseno entre representaciones generadas por LLM pre-entrenados (e.g. BERT) aproximan a la similitud contextual entre las palabras. De este modo se trata de superar la limitación semántica que sufrían las métricas anteriores.

El valor BertScore supone entonces una media armónica entre la *Precision*, o promedio de similitud coseno entre cada token de la salida generada y su coincidencia más cercana en la referencia, y el *Recall*, similitud media del coseno entre cada palabra de la referencia y su coincidencia más próxima en el resumen generado. De este modo, BertScore tratar de estimar la similitud semántica basada en la distancia coseno dentro de su espacio de representación.

Se evaluará la generación de resúmenes en inglés y español, dentro del dominio legal, y a partir tanto de métricas clásicas como de aproximaciones que suponen el estado del arte en tareas de evaluación de traducción automática y generación de textos.

Para ello se pondrán los siguientes experimentos:

- Se aplicarán primero las diferentes variantes de las métricas ROUGE tanto a los resúmenes generados en español como en inglés. Usaremos las versiones de ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S y ROUGE-SU que no empleen Wordnet. Como primera aproximación, no incluiremos ninguna base de datos léxica a la hora de aplicar las métricas ROUGE.
- A continuación, aplicaremos la métrica BertScore tanto a los resúmenes generados en español como en inglés. Como esta métrica se basa en la aplicación del coseno sobre vectores de representación de palabras, éstos serán diferentes dependiendo del modelo pre-entrenado que se utilice.
  - Como baseline se considerará el modelo *BERT-base-multilingüe-cased* (Devlin et al., 2018b), tomando los embeddings de la primera y la última capa. De este modo evaluaremos el gap en función del mayor o menor grado de contextualidad en los vectores de representación de palabras.
  - Además, se considerará el modelo *MPNET: all-mpnet-base-v2* (Song et al., 2020) usando también los embeddings de la primera y la última capa. Este modelo ha sido corregido semánticamente y se trata de un modelo *sentence-Transformers* (Reimers and Gurevych, 2019), que representa frases como vectores de 768 dimensiones que pueden después utilizarse en tareas como el clustering o la búsqueda semántica, ya que al estar corregido semánticamente su espacio de representación, la distancia coseno aproxima mejor a la similitud semántica entre términos. La selección de este modelo viene dada por haber resultado como el mejor modelo multilingüe en la evaluación comparativa en Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023), que abarca 8 tareas relacionadas con estimación semántica de *embeddings* de textos.

Se emplearán modelos multilingües para evitar que la selección de modelos monolingües en inglés y español, con diferente calidad entre sí, pueda introducir sesgos en la evaluación final. Partimos de

una mismo modelo LLM multilingüe para generar los vectores de representación en ambas lenguas, de modo que la diferencia final en los valores de BertScore vendrá dada por el mejor o peor desempeño de los modelos generativos, en inglés y en español, ante una tarea de generación de resúmenes dentro del dominio legal, y no dependerá de la calidad del modelo pre-entrenado que se use.

### 6.3.2. Descripción de experimentos futuros

Se explorará la posibilidad de realizar experimentos futuros en las siguientes líneas:

- Evaluación por medio de métricas intrínsecas (*fluidez, coherencia, relevancia y consistencia*).
- Aplicación de la métrica MoverScore (Wei Zhao, 2019). A diferencia de BertScore, que solo comparaba por medio del coseno palabras una a una entre el resumen generado y la referencia, MoverScore permite comparar palabras dentro una secuencia con más de una palabra.; y lo hace resolviendo un problema de optimización que encuentra el camino mínimo para transformar un texto en otro. La idea es estimar la distancia que tendrían que recorrer las palabras para convertir una secuencia en otra.
- Aplicación de métricas de Simplificación de Texto (*Text Simplification*) (Alva-Manchego et al., 2019): BLEU, SARI, SAMSA, FKGL; tanto para inglés como para español.

## 6.4. PRON vs PROMPT: Generación de sinopsis

Como se ha explicado antes, el dataset consiste en 120 sinopsis para 60 títulos de películas imaginarias, 30 propuestos por GPT4 y 30 por un escritor de prestigio (Patricio Pron, premio Alfaguara de novela). Se solicitó a ambos, al escritor y a GPT-4, que escribieran sinopsis de aproximadamente 600 palabras para cada título, incluyendo tanto los propuestos por ellos mismos como por su contrincante. Además, los títulos se tradujeron manualmente al inglés, y se pidió a GPT-4 que escribiera sinopsis en inglés también para esos títulos.

### 6.4.1. Métrica

Los textos generados están (a fecha de finalización del año 2 del proyecto) siendo sometidos a una evaluación a ciegas por un panel de expertos, compuesto por críticos y académicos (la mitad bilingües), para garantizar una valoración objetiva de la calidad, creatividad y coherencia narrativa de las sinopsis. En general, las preguntas de evaluación se responden en una escala de 0 a 3, lo que permitirá establecer comparaciones intervállicas y de ratio (porcentuales).

Los resultados estarán disponibles a lo largo del tercer año del proyecto.

## 7. Conclusiones y trabajo futuro

En el segundo año del proyecto se ha avanzado notablemente en la evaluación de efectividad de modelos de lenguaje: se ha expandido el leaderboard original hasta 15 tareas que cubren una multitud de ámbitos de aplicación y tipos de problemas abstractos. Es, además, el primer leaderboard, hasta donde alcanza nuestro conocimiento, que incluye problemas en modo learning with disagreement, y se contribuye al estado del arte con una métrica que es la primera que puede aplicarse a problemas de clasificación multilabel y/o jerárquicos en learning with disagreement.

Además, fuera de los objetivos establecidos en el convenio se ha comenzado experimentación específica sobre modelos de lenguaje generativo: se han desarrollado dos datasets específicos para evaluar este tipo de modelos, un experimento adicional, y se han comenzado a probar los modelos generativos sobre las tareas del leaderboard.

Una limitación de nuestros resultados es que el leaderboard ODESIA CORE no tiene tanta diversidad temática como nos gustaría, y hay un sesgo hacia las tareas relacionadas con la desinformación y los contenidos tóxicos. Este desequilibrio se compensa con los datasets de ODESIA EXTENDED. Otra limitación es que nos hemos limitado a evaluar la efectividad de los modelos, y no otros aspectos relevantes como sus sesgos culturales o su eficiencia computacional.

## Agradecimientos

Este trabajo ha sido financiado por la Unión Europea - NextGenerationEU a través del “Plan de Recuperación, Transformación y Resiliencia”, por el Ministerio de Asuntos Económicos y Transformación Digital y por la UNED. Sin embargo, los puntos de vista y las opiniones expresadas son únicamente los del autor o autores y no reflejan necesariamente los de la Unión Europea o la Comisión Europea. Ni la Unión Europea ni la Comisión Europea pueden ser consideradas responsables de los mismos.

## Bibliografía

- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. [SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects](#). ArXiv:2309.07445 [cs].
- Rodrigo Agerri and Eneko Agirre. 2022. [Lessons learned from the evaluation of spanish language models](#).
- Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. [Mega: Multilingual evaluation of generative ai](#).
- Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. 2023. [An in-depth look at gemini’s language abilities](#).
- Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. [EASSE: Easier automatic sentence simplification evaluation](#). 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): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.
- Enrique Amigo and Agustín Delgado. 2022. [Evaluating Extreme Hierarchical Multi-label Classification](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland. Association for Computational Linguistics.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. [A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity](#).
- Valerio Basile, Livio Bioglio, Alessio Bosca, Cristina Bosco, and Viviana Patti. 2023. [UINAUIL: A unified benchmark for Italian natural language understanding](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 348–356, Toronto, Canada. Association for Computational Linguistics.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. [SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: An open platform for evaluating llms by human preference](#).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised cross-lingual representation learning at scale](#). *CoRR*, abs/1911.02116.

- Javier De la Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Maria Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *arXiv preprint arXiv:2207.06814*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Percy Liang et al. 2023. [Holistic evaluation of language models](#).
- Hermenegildo Fabregat, Juan Martínez-Romo, and Lourdes Araujo. 2018. [Overview of the DIANN task: Disability annotation task](#). In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 1–14. CEUR-WS.org.
- Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, Qianyu He, Rui Xu, Wenhao Huang, Jingping Liu, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, and Yanghua Xiao. 2024. [Xiezhi: An Ever-Updating Benchmark for Holistic Domain Knowledge Evaluation](#). ArXiv:2306.05783 [cs].
- Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring Massive Multitask Language Understanding](#). ArXiv:2009.03300 [cs].
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. [C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Léo Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. [Mixtral of experts](#).
- Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. [Neural text summarization: A critical evaluation](#). *CoRR*, abs/1908.08960.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale Reading Comprehension Dataset From Examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. [Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning](#).
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. [Holistic evaluation of language models](#).

- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#).
- Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022)*, pages 1412–1437.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. [Mteb: Massive text embedding benchmark](#).
- Dan Nielsen. 2023. [ScandEval: A benchmark for Scandinavian natural language processing](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kopic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [Gpt-4 technical report](#).
- Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442.

- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Laura Plaza, Jorge Carrillo-de-Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. [Overview of-EXIST 2023 – Learning with-Disagreement for-Sexism Identification and-Characterization](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 316–342, Cham. Springer Nature Switzerland.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. [NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- Alejandro Vaca Serrano, Guillem Garcia Subies, Helena Montoro Zamorano, Nuria Aldama Garcia, Doaa Samy, David Betancur Sanchez, Antonio Moreno Sandoval, Marta Guerrero Nieto, and Alvaro Barbero Jimenez. 2022. Rigoberta: A state-of-the-art language model for spanish. *arXiv preprint arXiv:2205.10233*.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. [Mpnet: Masked and permuted pre-training for language understanding](#).
- Aarohi et al Srivastava. 2022. [Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models](#). ArXiv:2206.04615 [cs, stat].
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. [Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them](#). ArXiv:2210.09261 [cs].
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. [Learning from Disagreement: A Survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019b. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018a. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018b. [GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding](#).
- Fei Liu Yang Gao Christian M. Meyer Steffen Eger Wei Zhao, Maxime Peyrard. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. [mt5: A massively multilingual pre-trained text-to-text transformer](#). *CoRR*, abs/2010.11934.
- Hui Zeng. 2023. Measuring Massive Multitask Chinese Understanding.
- Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. [M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models](#).
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. [AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models](#). ArXiv:2304.06364 [cs].

## **A. Apéndice: Aspectos técnicos relevantes en el desarrollo de proyectos software**

En este apéndice se muestran los aspectos técnicos y de documentación relevantes en el proceso de desarrollo de código. Aunque este aspecto queda, en cierto modo, fuera de los ámbitos del convenio, se ha intentado abordar, hasta donde llegan nuestras posibilidades como universidad, todos los puntos sugeridos.

### **A.1. Mantenibilidad**

Durante el proyecto se han seguido los principios básicos de buenas prácticas de programación, adaptadas en cada caso a su lenguaje de programación: Python y PHP. Entre otros, se han tenido en cuenta las convenciones de nombres de variables, clases, formatos de código, etc., así como estilo de comentarios, descripciones de los métodos, etc. Así mismo, se ha creado un repositorio de documentación en GitHub privado donde se ha ido documentando, en su mayor medida, todos los aspectos técnicos de la implementación de las aplicaciones ODESIA: su arquitectura, herramientas utilizadas en el desarrollo con versiones y dependencias, forma de despliegue, etc.

Por último, se ha habilitado un repositorio GitHub privado para el desarrollo de los proyectos de código, lo que permite el trabajo colaborativo de una forma segura y eficaz, a la vez que permite tener copias de seguridad en caso de posibles fallos.

Por último, y aunque todavía no en pleno desarrollo debido a problemas técnicos en nuestra universidad, se han montado dos servidores, uno para desarrollo y otro para producción. Esta diferenciación, que esperamos poder tener lista próximamente nos permitirá evitar errores de estabilidad en las aplicaciones finales, a la vez que agilizar los procesos de despliegue.

### **A.2. ENS: Esquema Nacional de Seguridad**

La UNED, como centro público de enseñanza, está trabajando desde hace tiempo en certificación de su esquema ENS. Tras una reunión con los responsables de ciberseguridad de nuestra universidad, nos indicaron que actualmente existe un borrador y se está a la espera de próxima aprobación. Por otro lado, los responsables nos indicaron que la infraestructura técnica de seguridad en la UNED supera los estándares y requisitos del esquema ENS. En este contexto, y dado que el desarrollo de este proyecto y el despliegue de las aplicaciones se realiza dentro de la infraestructura de la UNED, todas las aplicaciones ODESIA se encuentran sobre el paraguas de seguridad desarrollado por el servicio de ciberseguridad de la UNED.

Además, se han solicitado los certificados de seguridad, que penderán del certificado de seguridad raíz de la UNED, para poder incluir el protocolo https en las aplicaciones, y que esperamos tener próximamente. No obstante, todas las aplicaciones se ejecutan en una red virtualizada privada, a la que solo tienen acceso los contenedores que así se hayan configurados, con un único contenedor con acceso a Internet. Además, este último contenedor se encuentra protegido por el firewall UNED que da cobertura y seguridad a todos los servidores de la Universidad.

### **A.3. ENI: Esquema Nacional de Interoperabilidad**

La misión y objetivos del Esquema Nacional de Interoperabilidad queda un poco fuera del alcance de este proyecto, dadas las dimensiones del mismo y público objetivo. No obstante, todas las aplicaciones e infraestructuras se están desarrollando teniendo en cuenta que serán usadas en entornos de escritorio y con el navegador más popular: Chrome. Dicho esto, se está haciendo un esfuerzo, dentro de nuestras posibilidades, en adaptar todo lo posible las aplicaciones a otros entornos como móviles, tabletas, etc. Así mismo, se ha utilizado la especificación independiente Swagger para la comunicación entre elementos del proyecto ODESIA para conseguir un mejor control y mantenimiento de los mismos.

### **A.4. Reglamento General de Protección de Datos**

Al igual que en el ENS, el proyecto ODESIA se encuentra bajo el amparo del esquema de protección de datos de la UNED.

### **A.5. Informe de técnicas de Search Engine Optimization**

Las aplicaciones ODESIA son aplicaciones cuyo objetivo no es el posicionamiento en los buscadores, su uso por expertos cualificados, por lo que este punto no ha sido tenido en cuenta dentro del proyecto. No obstante, durante el desarrollo web de las aplicaciones se han usado los estándares de HTML5 con las mejores práctica de desarrollo web dentro de los lenguajes de programación y frameworks.

### **A.6. Diseño de la navegabilidad**

El desarrollo de las aplicaciones ODESIA se han centrado desde sus inicios en el desarrollo de aplicaciones amigables y fáciles de usar. Es por ello que la navegabilidad de las mismas está entre sus objetivos principales, proporcionando todas su funcionalidad mediante como mucho tres clicks.