E4.1 ANÁLISIS Y DEFINICIÓN DE DOMINIOS DE APLICACIÓN Y CASOS DE USO

KG4LLM - SERVICIOS PARA EL ENRIQUECIMIENTO DE MODELOS DE LENGUAJE CON GRAFOS DE CONOCIMIENTO (SER-21/23 OTT)

Resumen

Este entregable analiza y define los posibles dominios de aplicación y casos de uso a ejecutar en el proyecto bajo el prisma de la disponibilidad de recursos lingüísticos y de experimentación en el marco del proyecto, así como del posible impacto que este proyecto pueda generar en dichos dominios. El documento prioriza dominios y casos de uso en los que se encuentran recursos disponibles, en cuanto a corpora, grafos de conocimiento y datasets de evaluación. Por otro lado, este entregable también tiene en cuenta los requisitos de multilingüismo que deben aportar dichos recursos en cuanto a cobertura del español y lenguas oficiales.

José Manuel Gómez Pérez
Raúl Ortega

30 de Diciembre de 2023
Expert.ai Language Technology Research Lab

Calle Poeta Joan Maragall, 3-5, Escalera Izquierda, Planta 1ª, Derecha, 28020, Madrid
CIF: B-66425513, Inscrita en el Registro Mercantil de Madrid, en el Tomo 44.538, Folio 74, Hoja Número M-784613, Inscripción 1ª.

www.expert.ai

Historia de revisiones

Revisión	Fecha	Descripción	Autor
0.1	26/12/2023	Draft preliminar completo	Raul Ortega José Manuel Gómez
0.2	27/12/2023	Apéndice añadido	Raul Ortega
0.3	29/12/2023	Revisión, corrección y extensión	José Manuel Gómez
1.0	30/12/2023	Versión actualizada lista para enviar	José Manuel Gómez

Contenidos

1 Introducción y motivación

El objetivo principal de este entregable es proponer una serie de dominios de aplicación sobre los que llevar a cabo la experimentación necesaria con los métodos de inyección de conocimiento en LLM que serán desarrollados en los paquetes de trabajo técnicos del proyecto KG4LLM. Para ello, este entregable tiene en cuenta de manera fundamental criterios de disponibilidad de los recursos necesarios para alimentar dichos métodos y el posible impacto social o económico que su aplicación puede tener en los sectores de aplicación asociados a dichos dominios. De esta forma, este entregable pretende también informar el desarrollo posterior durante el proyecto de métodos de inyección de conocimiento en LLM, asegurando la selección de dominios que ofrezcan un escenario favorable para su desarrollo y evaluación.

Los métodos de inyección de conocimiento que se desarrollarán en este proyecto serán aplicables en distintas etapas del ciclo de vida de LLM. Entre ellos, debido a su potencial para la adaptación de LLM a dominios específicos, se estudiarán con especial interés aquellos métodos de inyección de conocimiento que se aplican o bien durante fases adicionales del preentrenamiento de los modelos o bien en tiempo de inferencia.

Métodos como K-Adapter [1] requieren una combinación de corpora documental y grafos de conocimiento para entrenar adaptadores lingüísticos y factuales que capturen conocimiento de dominio, combinándolo con las representaciones internas del modelo, que permanecen inalteradas. Otros métodos [2] basados en generación aumentada mediante recuperación (por sus siglas en inglés, RAG) utilizan conocimiento recuperado de una base de conocimiento documental o estructurada para enriquecer con información relevante el contexto facilitado al modelo al recibir una petición, influyendo así en el razonamiento llevado a cabo por el modelo. Finalmente, métodos como DPO (del inglés Direct Preference Optimization) [3] buscan optimizar el alineamiento del LLM con una serie de preferencias que cubren criterios como la ausencia de toxicidad, sesgos o alucinaciones y la presencia de información factual en el texto generado por el modelo. Para ello, DPO recurre a conjuntos de datos de preferencias construidos a partir de afirmaciones factuales (claims) procedentes de corpora documental, así como de hechos extraídos de grafos de conocimiento.

Alternativamente, métodos como Toolformer [4] permiten al LLM delegar parcialmente en herramientas externas accesibles mediante APIs con las que resolver tareas específicas, como preguntas y respuestas, cálculo matemático, procesamiento de fechas, búsqueda y traducción automática. Por otro lado, métodos como ToolLLM [5] utilizan miles de APIs disponibles a través de recursos como RapidAPI.com para generar un dataset, ToolBench, con el que entrenar un modelo basado en LLaMA que invoque esas APIs de manera orquestada.

Desde el punto de vista de la disponibilidad de recursos, un buen dominio para la experimentación con métodos de inyección de conocimiento en LLM será por tanto aquel que ofrezca corpora documental en abundancia, con contenido y lenguaje del dominio, así como recursos estructurados verticales para ese dominio, preferiblemente en forma de grafos de conocimiento. El long tail de dominios y funcionalidades en los servicios ofrecidos por recursos como ToolBench invita a considerar servicios quizá más genéricos pero ciertamente más reutilizables a la hora de incorporarlos en LLM mediante métodos como Toolformer. Por tanto, este entregable prioriza los criterios de disponibilidad de corpora documental y grafos de conocimiento sobre otros como la disponibilidad de APIs a servicios externos.

Uno de los objetivos de KG4LLM es proponer un marco de evaluación de factualidad generalizable que permita medir el impacto de las técnicas de inyección de conocimiento

propuestas, especialmente para la adaptación de LLM a dominios específicos, sobre la factualidad de las respuestas generadas. Sin embargo, también nos interesa medir el impacto de aplicar estos métodos de inyección de conocimiento no sólo en términos de su factualidad sino de su rendimiento a la hora de resolver tareas NLP complejas en esos dominios. Por tanto, se valorarán aquellos dominios que cuenten con conjuntos de datos que permitan la evaluación de tareas NLP verticales. Finalmente, debido al foco de KG4LLM en el español y lenguas cooficiales, será especialmente relevante que dichos recursos (corpora, grafos de conocimiento, datasets de evaluación) sean significativamente multilingües.

El resto de este entregable se estructura de la siguiente manera. La sección 2 está dedicada a cada uno de los dominios verticales que han sido identificados durante el trabajo prospectivo llevado a cabo para este entregable como dominios de interés de acuerdo con los criterios introducidos más arriba. Ejemplos de estos dominios incluyen el domino de salud o el legal. En concreto, para cada uno de esos dominios, esta sección da cuenta de corpora documental, grafos de conocimiento y recursos semánticos disponibles, así como de conjuntos de datos para evaluación en tareas NLP de posible utilidad para el proyecto. Mientras que la sección 2 se centra en dominios con entidad propia tanto desde el punto de vista lingüístico como semántico, la sección 3 identifica sectores que inherentemente combinan varios dominios. Un ejemplo de este tipo de sectores de aplicación puede ser el sector de seguros, en el que los dominios de salud y legal son particularmente relevantes. Por motivos de experimentación, resulta también interesante contar con datos y grafos de conocimientos generados sintéticamente a partir de grandes recursos multilingües previamente establecidos. La sección 4 propone una metodología de extracción de corpora y grafos de conocimiento multilingües para dominios verticales y sectores de aplicación transversales procedentes de Wikipedia y Wikidata. Finalmente, la sección **Error! Reference source not found.** ofrece una serie de recomendaciones sobre los dominios alrededor de los que estructurar la experimentación y el desarrollo de métodos para la inyección de conocimiento en LLM basada en el análisis llevado a cabo en este entregable.

2 Dominios primarios

Esta sección explora aquellos dominios que pueden ser de interés para INESData y en particular para el desarrollo de métodos de inyección de conocimiento en LLM en el marco del proyecto KG4LLM. Para ello, analiza la disponibilidad y relevancia de recursos verticales de utilidad para el proyecto: Corpora documental, grafos de conocimiento y recursos semánticos y conjuntos de datos para evaluación en tareas NLP de posible utilidad para el proyecto. Como dominio, entendemos aquellas áreas de conocimiento o disciplinas con entidad propia desde un punto de vista tanto lingüístico como semántico.

1.1 Salud

1.1.1 Corpora documental

PubMed[1] es un motor de búsqueda de libre acceso que permite consultar los contenidos de más de 36 millones de citas de la base de datos MEDLINE, provenientes tanto de revistas como de libros de la literatura biomédica. Basado en Pubmed, proponemos construir un dataset de

[1] https://pubmed.ncbi.nlm.nih.gov

abstracts de artículos revisados por pares que servirá tanto como corpus de dominio, como de dataset de verificación, con el que contrastar la factualidad de hechos médicos. Debido a su tamaño, y para facilitar su acceso como dataset de verificación en tiempo de ejecución, se filtrarán los artículos basados en su posible influencia, de manera que podamos procesar solo aquellos que hayan sido citados más de un número concreto de veces. Para ello haremos uso de **Semantic Scholar**[2], otro motor de búsqueda de libre acceso que contiene artículos de Pubmed y otras fuentes y que además otorga a los artículos una puntuación basada en su influencia, calculado mediante un algoritmo que trata de medir la influencia de cada referencia dentro de un artículo [6].

EC3[3] es un proyecto europeo que ha anotado y distribuido de manera libre un gran corpus de documentos clínicos en cinco lenguas europeas (español, euskera, inglés, francés e italiano). Las anotaciones incluyen información temporal y factual, además de entidades clínicas basadas en taxonomías médicas. Esto permite su uso tanto en análisis lingüístico como benchmarking, así como para el entrenamiento de sistemas de extracción de información. El corpus está organizado en tres diferentes capas, cada una de ellas anotadas con un diferente grado de supervisión y destinado a un uso concreto. La primera capa (~25K tokens), anotada de forma totalmente manual, aporta información temporal y de factualidad, y está destinada a su uso para análisis lingüístico y benchmarking. La segunda capa (50K-100K tokens) ofrece anotaciones sobre entidades clínicas usando un enfoque semi automático. Por último, la tercera capa (1M tokens) alberga texto sin anotar que puede ser explotado por enfoques semi-supervisados como input para ajustar LLM en tareas NLP específicas.

El corpus utilizado para entrenar el Biomedical and clinical language model for Spanish [7] es un conjunto de corpora públicos en español, incluido Wikipedia, Google Patents y PubMed, con 278K documentos clínicos y notas que han sido filtradas para obtener un único corpus de biomedicina limpio y un corpus clínico en crudo. Dentro de este filtrado, se ha aplicado separación de frases, detección de lenguaje, filtrado de frases malformadas y deduplicación. Este corpus puede ser utilizado para entrenar un LLM en el dominio biomédico en español.

1.1.2 Grafos de conocimiento y recursos semánticos

El **Unified Medical Language System (UMLS)**[4] es un conjunto de documentos y software que recoge múltiples vocabularios de varias lenguas (principalmente en inglés, pero también en árabe, euskera, chino, checo, holandés, francés, alemán, hebreo, húngaro, italiano, japonés, coreano, letón, noruego, polaco, portugués, ruso, español, sueco, turco y ucraniano) relacionados con los campos de la salud y la biomedicina. UMLS intenta promover la interoperabilidad de servicios y sistemas que trabajan con información biomédica. Puede ser utilizado para la extracción de entidades, para enlazar conocimiento de textos en diferentes idiomas o facilitar el desarrollo de sistemas de recuperación de datos.

[2] https://www.semanticscholar.org

[3] https://e3c.fbk.eu/about

[4] https://www.nlm.nih.gov/research/umls/index.html

1.1.3 Conjuntos de datos para evaluación

Para la evaluación de los LLMs proponemos una serie de tareas de resolución de preguntas multirespuesta en el dominio biomédico y de salud, que permitan conocer el grado de conocimiento que se ha logrado inyectar de forma efectiva en el modelo.

- **MedQA-USMLE** [8]**.** Basado en el Examen de Licencia Médica de los Estados Unidos, este dataset contiene preguntas tipo test en inglés (12K), además de chino simplificado (34K) y chino tradicional (14K).
- **PubmedQA** [9]. Contiene mil preguntas etiquetadas por profesionales, 61K sin etiquetar y 211K generadas de forma artificial. Las preguntas están enlazadas con párrafos de contexto, y tienen tres posibles respuestas: sí, no y quizás.
- **MMLU** [10]. Se trata de un marco de evaluación diseñado para medir el conocimiento adquirido por modelos en el marco de una tarea zero-shot o few-shot. Da cobertura a múltiples submaterias en ciencia e ingeniería, de las que proponemos utilizar aquellas relacionadas directamente con la salud y la biomedicina.
- **MedMCQA** [11]**.** Es una colección de más de 194K preguntas multirespuesta provenientes de los exámenes de acceso al instituto de ciencias médicas (AIIMS) y al posgrado (NEET PG) de India. Da cobertura a más de 2.4K temas relacionados con la salud.

El estado de la cuestión en el dominio biomédico también ofrece una serie de datasets para extracción de entidades médicas en español sobre los que evaluar. Además, el subset de evaluación del corpora utilizado para entrenar el Biodemedical and clinical language model for Spanish también puede ser usado para este mismo fin.

- **PharmaCoNER** [12]**.** Corpus en castellano con mil estudios clínicos anotados manualmente, enfocados principalmente en el reconocimiento de entidades que representen fármacos y otros químicos relacionados con la salud.
- **CANTEMIST (CANcer TExt Mining Shared Task – tumor named entity recognition)** [13]. Se trata de un dataset centrado en el reconocimiento de entidades relacionadas con el cáncer, anotadas por expertos en oncología.

1.2 Legal

1.2.1 Corpora documental

Multi-legal PILE[5] es una selección de corpus en las 24 lenguas oficiales de la Unión Europea procedentes de 17 jurisdicciones europeas. Combina Eurlex Resources[6], Pile of Law [14], Legal mC4[7] y Native Multi Legal Pile. Al incluir diversas fuentes de datos con diferentes licencias, MultiLegal PILE permite su uso para pre-entrenamiento de LLM, con licencias más permisivas para los subsets de Eurlex Resources y LegalmC4.

---

[5] https://www.researchgate.net/publication/371311164_MultiLegalPile_A_689GB_Multilingual_Legal_Corpus

[6] https://huggingface.co/datasets/joelito/eurlexresources

[7] https://huggingface.co/datasets/joelito/legal-mc4

El **Spanish Legal Domain Corpora**[8] es una colección de corpus procedentes del dominio legal en español. Se trata de un corpora de documentos de procesos penales, BOE, Europarlamento y otras fuentes oficiales. Su tamaño actual es de 2GB y se trata de documentos en texto plano sin anotar, por lo que puede ser utilizado para la especialización de un LLM en el dominio legal español.

1.2.2 Grafos de conocimiento y recursos semánticos

EuroVoc[9] es un tesauro multilingüe y multidisciplinario mantenido por la Oficina de Publicaciones de la Unión Europea. Contiene palabras clave en las veinticuatro lenguas de la UE, además de albanés, macedonio y serbio, organizadas en 21 campos temáticos y 127 subcampos, que sirven para describir el contenido de los documentos en EUR-Lex.

1.2.3 Conjuntos de datos para evaluación

Para la evaluación de los modelos en el dominio legal, la comunidad NLP ofrece una serie de marcos de evaluación específicos de dominio. Los dos primeros incluyen únicamente datasets en inglés, mientras que los dos siguientes son multilingües:

- **LEXGlue** [15]**.** Contiene tareas de clasificación y resolución de preguntas multirespuesta en inglés provenientes de la Corte Europea de los Derechos Humanos (ECtHR), la Corte Suprema de los Estados Unidos (SCOTUS) y subsets anotados de EUR-LEX, LEDGAR [16], UNFAIR-ToS [17] y CaseHOLD [18].
- **LegalBench**[10]**.** Se trata de un proyecto desarrollado por la Universidad de Stanford, dedicado al curado de tareas de evaluación que impulsen el razonamiento de LLM dentro del dominio legal en inglés. En la actualidad, el marco de evaluación contiene un total de 162 tareas en las que han participado 40 contribuyentes. Todas sus tareas consisten en pares de entrada-salida, abarcando tanto resolución de preguntas como clasificación de texto. Este marco de evaluación está en construcción, por lo que se pueden seguir añadiendo nuevas tareas a lo largo del proyecto.
- **FairLEX** [19]**.** Es una suite de cuatro datasets para la evaluación de LLM preentrenados dentro del ámbito legal. Estos datasets dan cobertura a cuatro jurisdicciones (Consejo Europeo, Corte Suprema de los Estados Unidos, Tribunal Supremo Federal de Suiza y Corte Suprema Popular de la República Popular de China), y cinco idiomas (inglés, alemán, francés, italiano y chino). En este marco de evaluación, los modelos son evaluados en base a cinco dimensiones relacionadas con la equidad (fairness) del model: género, edad, región, idioma y área legal.
- **LEXTREME** [20]. Consiste en la selección de 11 datasets que contienen texto del ámbito legal en los 24 idiomas de la Unión Europea. Los datasets provienen de las cortes brasileña, alemana, suiza, griega, rumana y europea, y consisten en clasificación de texto, predicción de decisiones legales, reconocimiento de entidades y evaluación del fairness en textos legales.

Además, para evaluación en este dominio también proponemos el uso de **CUAD** [21], un dataset de resúmenes de contratos legales procedentes del Atticus Project, y que incluye 13K

[8] https://zenodo.org/records/5495529

[9] https://eur-lex.europa.eu/browse/eurovoc.html

[10] https://hazyresearch.stanford.edu/legalbench

anotaciones. Otro posible marco de evaluación podría ser de nuevo **MMLU** [10], esta vez seleccionando las submaterias relacionadas con el dominio legal.

1.3 Finanzas

1.3.1 Corpora documental

El **Financial Reports EDGAR SEC**[11] contiene los reportes anuales de compañías públicas norteamericanas mediante el sistema SEC EDGAR desde 1993 a 2020. Estos reportes albergan una gran cantidad de texto en crudo en forma de frases (~72 millones) relacionadas con el dominio financiero. Además, cada una de esas piezas de texto tiene asociadas etiquetas de sentimiento, por lo que además de poder actuar como corpus de preentrenamiento y/o de verificación, este dataset también puede ser utilizado para evaluar LLM sobre esa tarea.

El **Corpus Reuters**[12], creado en el año 2000, contiene una colección de noticias tanto en inglés como en otros doce idiomas (holandés, francés, alemán, chino, japonés, ruso, portugués español, italiano, danés, noruego y sueco). Dentro de este corpus, proponemos el uso de los subsets **RCV2** (noticias en múltiples idiomas desde 1996 a 1997) y **TRC2** (noticias en inglés desde 2008 a 2009). En concreto, es de particular interés el subconjunto de noticias relacionadas con el mundo financiero. El acceso a este dataset solo está disponible mediante petición por formulario. Puesto que este dataset abarca un periodo de tiempo corto (dos años), sólo es viable como fuente de conocimiento para adaptar LLMs a este dominio y no como dataset de verificación.

1.3.2 Grafos de conocimiento y recursos semánticos

La **Financial Industry Business Ontology (FIBO)**[13] define una serie de vocabularios en inglés que son de interés en el dominio financiero. Ha sido desarrollado por el Enterprise Data Management Council (EDMC) y ha sido estandarizado por el Object Management Group (OMG). El primer nivel de ontologías estaría dividido en Business Entities (BE), Business Process (BP), Corporate Actions and Events (CAE), Derivatives (DER), Financial Business and Commerce (FBC), FIBO Foundations (FND), Indices and Indicators (IND), Loans and Mortgages (LOAN), Market Data (MD) y Securities & Equities (SEC). Los conceptos de FIBO han sido evaluados y consensuadas por compañías afiliadas al EDMC desde 2008, y es posible navegarlos mediante su plataforma online, por lo que sería necesario su procesamiento y refinado su uso como knowledge graph o para el entrenamiento de un entity linker orientado al dominio financiero y la inyección de conocimiento en LLM.

1.3.3 Conjuntos de datos para evaluación

Para la evaluación de los LLMs pre-entrenados en el dominio financiero, existen varias tareas de resolución de preguntas y de análisis de sentimiento:

- **Financial Phrase Bank (FPB)** [22]**.** Se trata de un dataset creado a partir de la anotación de 4845 frases procedentes de artículos del dominio financiero en inglés, y anotadas por 16 investigadores con conocimientos del dominio. Cada frase está anotada como

[11] https://huggingface.co/datasets/JanosAudran/financial-reports-sec

[12] https://trec.nist.gov/data/reuters/reuters.html

[13] https://spec.edmcouncil.org/fibo

positiva, negativa y neutral, según el tipo de sentimiento que genera al público. Su uso está extendido como marco de evaluación de LLM en el dominio financiero.

- **Conversational Finance Question Answering (ConvFinQA)** [23]**.** Dataset que permite evaluar el razonamiento complejo de modelos NLP a la hora de resolver problemas aritméticos. Este marco de evaluación ofrece un desafío importante para este tipo de modelos, ya que para su resolución es necesario seguir una serie de cadenas de razonamiento aritmético. El dataset cuenta con 3.8K textos de contexto, y 14K preguntas de respuesta abierta, creadas siguiendo un flujo conversacional que permite simular peticiones a LLM generativos.
- **Financial Opinion Mining and Question Answering (FiQA-2018)** [24]**.** Este marco de evaluación incluye tareas tanto de resolución de preguntas como de análisis de sentimiento. El dataset de la primera tarea (Aspect-based financial sentiment analysis) categoriza frases procedentes de artículos financieros en inglés, según un score de sentimiento. El dataset de la segunda (Opinion-based QA over financial data) contiene un conjunto de preguntas o consultas acerca de aspectos del dominio financiero, cuyas respuestas deben ser sintetizadas a partir de un corpus de documentos provenientes de distintas fuentes en inglés, como noticias, blogs o reportes financieros.

Finalmente, si nos centramos en submaterias relacionadas con el dominio financiero, **MMLU** [10] puede ser utilizado de nuevo como marco de evaluación en este dominio.

1.4 Ciencia e ingeniería espacial y Astronomía

1.4.1 Corpora documental

ArXiV[14] es un repositorio de artículos científicos en el campo de las matemáticas, física, ciencias de la computación y biología cuantitativa. Es posible filtrar estos artículos por materia, seleccionando aquellas publicaciones de la categoría “astro-ph" para formar un corpus de texto del dominio astrofísico, siguiendo el criterio utilizado para el entrenamiento de otro LLM del dominio como AstroLLaMA[15]. Al tratarse en su mayoría de preprints que no han pasado por un filtro de evaluación por pares, su uso como dataset de verificación es limitado, pero sí que puede ser utilizado como corpus de preentrenamiento para LLM.

NASA Technical Reports Server (NTRS)[16] es un repositorio que provee acceso a documentos y metadatos, incluido artículos, patentes e informes, además de imágenes y vídeos técnicos, de la Agencia Espacial Norteamericana (NASA). Estos textos en inglés, al ser de acceso público, pueden ser utilizados como corpus de preentrenamiento de LLM.

Dentro del ámbito europeo, también contamos con el EO Portal[17] **y CEOS**[18], dos portales de la Agencia Espacial Europea (ESA) con documentación en inglés sobre centenares de misiones e instrumental de satélites de observación terráquea. Esta documentación, pese a no ser lo

[14] https://arxiv.org

[15] https://arxiv.org/abs/2309.06126

[16] https://ntrs.nasa.gov

[17] https://www.eoportal.org/satellite-missions

[18] https://database.eohandbook.com/about.aspx

suficientemente numerosa como para ser utilizada para preentrenamiento de LLM, puede ser usada como parte de un dataset de verificación de alta calidad y centrada en dominio.

1.4.2 Grafos de conocimiento y recursos semánticos

El Unified Astronomy Thesaurus (UAT)[19] es un tesauro desarrollado y mantenido por la American Astronomical Society (AAS), que formaliza conceptos astronómicos en inglés y sus interrelaciones. Estas relaciones son meramente taxonómicas, por lo que su potencial como knowledge graph es limitado. La **Data Ontology**[20] desarrollada por ESA tampoco contiene relaciones no taxonómicas entre sus conceptos. Ambos están accesibles mediante sus respectivos portales.

1.4.3 Conjuntos de datos para evaluación

Para la evaluación de LLMs orientados al espacio y la astronomía, proponemos la tarea de **Detecting Entities in the Astrophysics Literature (DEAL)** [25], que contiene datasets con fragmentos de textos procedentes de artículos de astrofísica del NASA Astrophysical Data System, que han sido anotados manualmente con entidades de interés para el dominio, como pueden ser satélites o cuerpos celestes. El subset de evaluación contiene 1.3K de estos fragmentos. También podría ser relevante el uso de **NASA NTRS** como tarea de evaluación en el marco de una tarea de clasificación de documentos teniendo en cuenta su subdominio asociado, siguiendo la taxonomía del NASA Technology Tree con la que están anotados sus documentos.También proponemos el uso de **NASA NTRS** como tarea de evaluación de modelos, en el marco de una tarea de clasificación de documentos teniendo en cuenta su subdominio asociado, siguiendo la taxonomía del NASA Technology Tree con la que están anotados sus documentos.

1.5 Ciencias de la Tierra

1.5.1 Corpora documental

Scigraph[21] es un motor de búsqueda desarrollado por Springer Nature que permite consultar información y metadatos de sus publicaciones, incluyendo artículos, proyectos de investigación, revistas, conferencias y libros. Al igual que Pubmed, Scigraph permite construir un dataset de abstracts verificados que sirva de corpus de dominio y de dataset de verificación. Para ello sería necesaria la selección y filtrado de artículos del dominio de Earth Sciences, usando los códigos de clasificación de Fields of Research (FoR), de la Australian and New Zealand Standard Research Classification (ANZSRC)[22]. También es posible usar nuevamente **Semantic Scholar** para filtrar aquellos artículos que tengan una mayor relevancia, o incluso añadir más a la colección de verificación si fuera necesario.

[19] https://astrothesaurus.org

[20] https://data.esa.int/esado

[21] https://sn-scigraph.figshare.com

[22] https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/latest-release

1.5.2 Conjuntos de datos para evaluación

Para la evaluación de LLM centrados en ciencias de la tierra proponemos dos datasets:

- **Climate-Fever** [26]**.** Un dataset de verificación de noticias en inglés relacionadas con el cambio climático, generado siguiendo la misma metodología con la que se construyó Fever, otro dataset de claims procedentes de Wikipedia. Consta de 1.5K claims extraídos de internet, cada uno de ellos acompañado por 5 evidencias anotadas de forma manual por expertos. Estas evidencias pueden estar anotadas como "support" si la evidencia está alineada con lo que se menciona en el claim, "contradict" si la evidencia refuta el claim o "not enough information" si no existe suficiente evidencia para catalogarlo como ninguna de las otras dos etiquetas.
- **SciTail** [27]**.** Se trata de un dataset de entailment (o alineamiento lógico entre hipótesis y premisas) generado a partir de preguntas multirespuesta de exámenes de ciencias en inglés. Contiene 27K pares premisa-hipótesis, de los que en 10K la hipótesis está soportada por la premisa y en 17K está refutada.

3 Sectores de aplicación transversales

A diferencia de los dominios descritos anteriormente, en esta sección nos centraremos en sectores de actividad que no constituyen un dominio primario por sí mismos, sino que son el resultado de agregar conocimiento y recursos lingüísticos y semánticos procedentes de diversas disciplinas. Un caso paradigmático es el sector de seguros, que en esencia resulta de la combinación de los dominios legal y de medicina en el marco de un sector de negocio con cierta terminología específica. Mientras que un dominio como la medicina se centra en la comprensión y aplicación del conocimiento en un campo específico, un sector de aplicación como seguros se centra en las actividades económicas y comerciales relacionadas con un área de actividad como la gestión de riesgos.

1.6 Seguros

El uso de LLM en contextos relacionados con el sector de seguros resulta tan interesante como desafiante. Por un lado, los documentos del sector, como pólizas o reclamaciones, contienen terminología propia pero también procedente de dominios como salud, lo que hace necesario que estos LLM sean capaces de identificar y procesar entidades médicas, como lesiones, enfermedades o tratamientos. Por otro lado, al tratarse de documentos con vinculación legal, gran parte del texto en este sector sigue una estructura y terminología similar a la que podríamos encontrar en documentos del ámbito legal, por lo que también requiere adaptar el modelo a este tipo de dominio. La escasez de recursos específicos para este tipo de dominio transversal hace imprescindible la combinación de otros recursos procedentes de los dominios relacionados. Debido a la existencia de datos sensibles, el acceso a pólizas y otros contratos similares a gran escala está limitado, por lo que el entrenamiento y evaluación del LLM en el ámbito de seguros requiere hacer uso de recursos de los dominios relacionados.

1.7 Análisis de patentes

Las patentes son otro tipo de documento legal, orientado a la descripción técnica de lo que se desea registrar, con influencia de dominios específicos como la ingeniería, las ciencias en general y las humanidades. Para adaptar un LLM a este tipo de dominio transversal, además de incluir en su preentrenamiento y evaluación recursos relacionados con los subdominios asociados

(Semantic Scholar, Scigraph, MMLS, etc.), también contamos con una serie de corpora y datasets de evaluación de patentes tanto en inglés como en otros idiomas:

- **BigPatent** [28]**.** Se trata de un dataset con 1.3M de patentes en inglés pertenecientes al archivo de los Estados Unidos, junto a resúmenes escritos a mano por expertos. Tanto las patentes como los resúmenes pueden utilizarse para el preentrenamiento de un LLM, además de para su evaluación en tareas de resumen de documentos de gran tamaño o de categorizarlos sobre sus 9 categorías. También pueden ser utilizados como archivo de verificación, de manera que puedan ser consultados y comparados con otros documentos y patentes.
- **The Harvard USPTO Patent Dataset (HUPD)** [29]. Es un corpus a gran escala en inglés, que utiliza 4.5M de entradas del archivo de patentes de los Estados Unidos (USPTO) entre enero de 2004 y diciembre de 2014. Al no estar completamente actualizado, este recurso tiene posiblemente más sentido como corpus de preentrenamiento que como dataset de verificación.
- **ParaPat** [30]**.** Se trata de un dataset con 68M de frases y 800M de tokens en 74 pares de idiomas alineados procedente de Google Patents. La alineación de los idiomas fue realizada usando el algoritmo de Hunalign para los 22 pares de idiomas con más entradas, mientras que el resto fueron alineadas a nivel de abstract. El dataset contiene patentes en las siguientes lenguas: checo, alemán, griego, inglés, español, francés, húngaro, japonés, coreano, portugués, rumano, ruso, eslovaco, ucraniano y chino. Por su tamaño y su naturaleza multilingüe, este dataset puede ser utilizado tanto como corpus de preentrenamiento, como para base de datos de verificación.
- **EuroPat** [31]**.** Este proyecto trata de recoger, procesar y alinear patentes procedentes de múltiples oficinas (USPTO y EPO) para crear un corpus paralelo en múltiples formatos e idiomas. Este corpus no está disponible como recurso, y requiere de un procesamiento previo. EuroPat es válido para su uso tanto en preentrenamiento como para verificación.
- Como recurso lingüístico para este tipo de documento contamos con el **European Patent Register**[23], una base de datos de patentes europeas que cuenta con anotaciones acordes a la ontología del Linked Open EP Data[24]. No obstante, su acceso es limitado, y será necesaria una investigación más en detalle para considerar su uso en la inyección de conocimiento estructurado dentro del LLM.
- Finalmente, la Oficina Europea de Patente (EPO, por sus siglas inglés) también ofrece un gran corpus diseñado para cubrir las necesidades de **análisis de texto en el dominio de las publicaciones de patentes**.[25] El corpus contiene títulos, resúmenes, descripciones, afirmaciones e informes de búsqueda de publicaciones de patentes.

[23] https://www.epo.org/en/searching-for-patents/data/bulk-data-sets/register-data

[24] https://data.epo.org/linked-data/documentation/patent-ontology-overview.html

[25] https://www.epo.org/en/searching-for-patents/data/bulk-data-sets/text-analytics

4 Dominios y sectores sintéticos

1.8 Motivación

Debido a la escasez de dominios en los que contemos tanto con corpora como con grafos de conocimiento y conjuntos de datos para evaluación, el desarrollo de métodos de inyección de conocimiento en LLM puede verse restringido a sólo unos pocos dominios o sectores de actividad. Para generalizar y homogeneizar el estudio de estos métodos, recurrimos a técnicas que nos permiten segmentar por dominios recursos multilingües y multidominio ya preexistentes, entre los que destacamos Wikipedia y Wikidata.

1.9 Enfoque

El objetivo de este apartado es el desarrollo de un método que permita identificar y extraer subgrafos de conocimiento de Wikidata. Para ello utilizamos como semilla las categorías asociadas a los artículos de la Wikipedia, seleccionando aquellos que estén etiquetados con categorías o subcategorías de alguno de los dominios verticales mencionados con anterioridad.

Debido al método de etiquetado de estas categorías en Wikipedia[26], un artículo puede estar asociado a una subcategoría del dominio, pero no a la categoría padre. Por ello, para poder identificar todos los artículos de Wikipedia asociados a una categoría raíz, es necesario recorrer su árbol completo de subcategorías junto a sus artículos asociados. De manera heurística, fijamos la profundidad de este árbol en 4 subcategorías, ya que este valor ofrece un equilibrio razonable entre volumen/calidad de los datos extraídos y recursos computacionales necesarios para extraerlos. Las categorías raíz que hemos definido como nodo inicial para cada dominio son:

- *Category: Health* (salud)
- *Category: Law* (legal)
- *Category: Finance* (finance)
- *Category: Astronomy* (espacio)
- *Category: Earth_science* (ciencias de la tierra)
- *Category: Insurance* (seguros)
- *Category: Patent_law* (patentes)

1.10 Resultados: dominios emergentes

La Figura 1 muestra los resultados del análisis de los dominios verticales con respecto a los artículos de Wikipedia. Los datos muestran el sesgo hacia artículos de divulgación y ciencia de Wikipedia, siendo salud y ciencias de la tierra los dominios mejor representados en cuanto a número de artículos asociados. También se puede apreciar cómo tanto el dominio de patentes como el de seguros adolecen de un número reducido de artículos etiquetados, reforzando de nuevo el carácter transversal de estos dominios y la necesidad de la extracción de conocimiento desde otros dominios para poder adaptar LLM a ellos.

[26] https://en.wikipedia.org/wiki/Help:Category

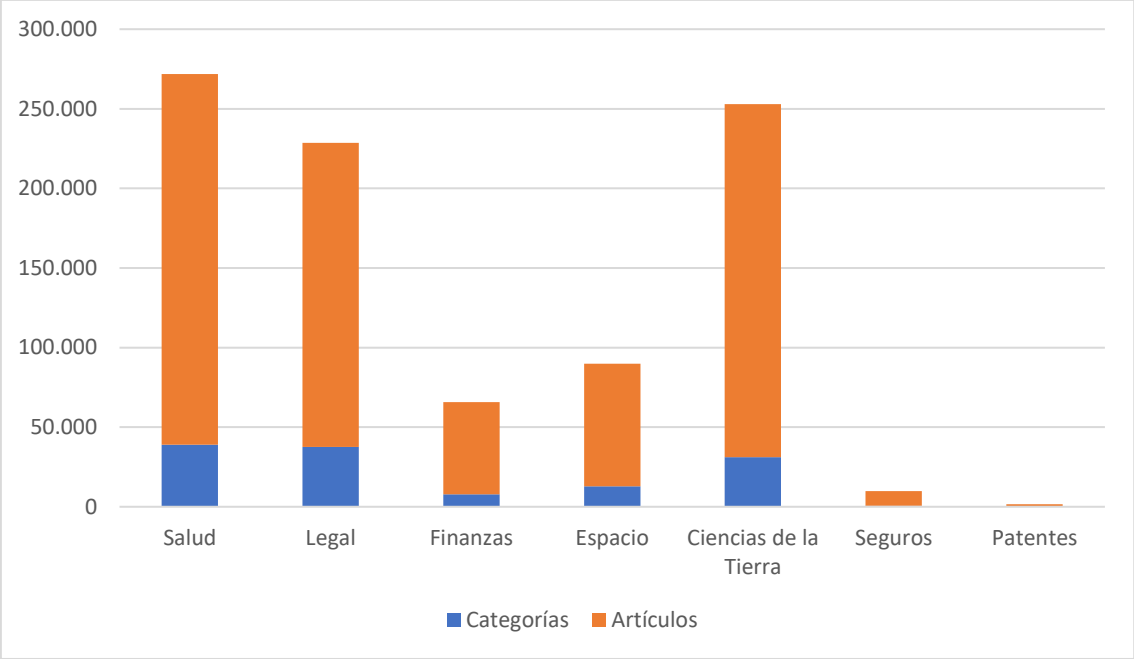

Figura 1. Distribución de artículos y categorías de Wikipedia según los diferentes dominios verticales. Los resultados mostrados corresponden a un análisis a diciembre de 2023, usando una profundidad de recorrido del árbol de subcategorías de 4.

La Figura 2 muestra la distribución de las categorías foráneas de cada dominio transversal, es decir, aquellas categorías que no forman parte del árbol recorrido para cada dominio, pero que sí son parte de las etiquetas de sus artículos asociados. Podemos ver que los artículos de Wikipedia relacionados con la categoría Patentes están muy orientados al dominio legal, con influencia también de la ingeniería, la ciencia y salud. Por otro lado, en los artículos bajo el dominio raíz de seguros parece que la salud es el dominio foráneo más importante, aunque también tienen un peso significativo el dominio de finanzas y el dominio legal.

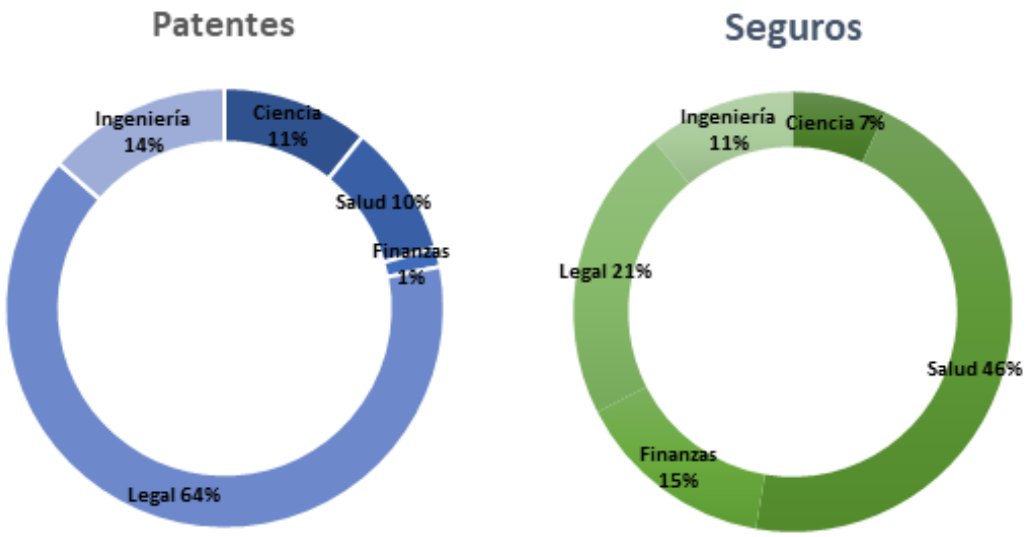

Figura 2. Distribución de otros dominios en artículos de la Wikipedia asociados a patentes y seguros

5 Conclusiones y recomendaciones

Próximamente, el entregable E1.1 (*Documento de estado de la cuestión, guías para la inyección de conocimiento en LLM y métricas*) hará un análisis detallado de los distintos métodos para la inyección de conocimiento en LLM que servirán de punto de partida para los métodos a desarrollar durante KG4LLM. Mientras tanto, este entregable se centra en el análisis de dominios potencialmente prometedores para los objetivos del proyecto, basado en gran medida en los recursos necesarios para desarrollar dichos métodos y su disponibilidad en los distintos dominios de posible impacto. La Tabla 1 muestra los recursos necesarios para algunos de los principales métodos de referencia en KG4LLM, que describimos a grandes rasgos a continuación:

- **K-Adapter** [1] requiere un corpus de dominio anotado con información procedente de tareas de extracción de conocimiento estructurado, como entity linking, object prediction o relation extraction, y uno o varios grafos de conocimiento alineados con ese corpus. El objetivo es entrenar adaptadores lingüísticos y factuales sobre esos recursos que capturen conocimiento de dominio, combinándolo con las representaciones internas del modelo, que permanecen inalteradas.
- **Retrieval Augmented Generation (RAG):** Métodos como el propuesto por Izacard et al. [2] utilizan conocimiento recuperado de una base de conocimiento documental (o, alternativamente, estructurada) para enriquecer con información relevante el contexto facilitado al modelo al recibir una petición, inyectando en ese momento información posiblemente valiosa para responder la petición.
- **Direct Preference Optimzation (DPO)** [3] busca optimizar el alineamiento del LLM con una serie de preferencias que por ejemplo favorezcan la generación de respuestas factuales sobre aquellas que no lo son. Para ello, DPO recurre a conjuntos de datos de preferencias construidos a partir de corpora documental sobre entidades de dominio procedentes de grafos de conocimiento u otros recursos como taxonomías o tesauros.

- **Otros métodos, basados en anchoring con KG:** Distintos métodos de comprensión de imágenes y texto como MAGMA [32] enriquecen las representaciones de modelos de lenguaje mediante la inyección de representaciones visuales y viceversa. Otros [33] utilizan un grafo de conocimiento para anclar las representaciones tanto visuales como textuales de un concepto, acercándolas entre sí en un espacio vectorial común, y favorecer el entrenamiento en tareas verticales, haciendo el modelo más resistente al olvido catastrófico. Inspirado en estos métodos, una variación suya podría centrarse en enriquecer las representaciones textuales generadas por LLM con representaciones de los conceptos correspondientes en un grafo de conocimiento, que permanecerían congeladas durante el entrenamiento. El uso de grafos de conocimiento multilingües (o grafos monolingües enlazados entre sí) permitiría transferir conocimiento a través de los distintos idiomas.

Tabla 1 Relación de métodos aplicables en distintas fases del ciclo de vida de LLM y recursos necesarios para implementar dichos métodos

Fase	**Métodos de referencia**	**Recursos necesarios**		
		Corpora de dominio alineado	**Grafo de conocimiento**	**Otros**
Preentrenamiento adicional	K-Adapter	sí, anotado con entidades y/o relaciones	sí, para entrenar las tareas de extrac. de conoc.	-
	Anchoring con KG	sí, anotado con entidades	sí, al que anclar las entidades	-
Inferencia	(Self)RAG	sí, con info. adicional en petición al LLM	alternativamente	-
Alineamiento con preferencias	DPO	sí, de verificación	sí, alt. taxonomías o tesauros	otros LLM

Como se ha mostrado en este documento, los distintos dominios y sectores de actividad ofrecen una cobertura desigual en términos de los recursos disponibles para experimentar y desarrollar métodos de inyección de conocimiento en LLM, como se puede observar en la tabla 2. El dominio mejor representado en ese sentido es el de Salud, con abundancia de corpora documental, útiles tanto para entrenar e inyectar conocimiento en LLM como para añadir información de contexto en tiempo de inferencia o verificar las afirmaciones generadas por LLM. El dominio de Salud también es rico en conjuntos de datos para evaluación en tareas NLP verticales y ofrece recursos en español y algunas lenguas cooficiales. Otros dominios bien representados son los dominios legal y financiero, así como el sector de análisis de patentes. Dominios científicos como Ciencias de la Tierra y Espacio tienen abundancia de corpora en inglés, pero escasez de recursos multilingües o datasets de evaluación.

Sin embargo, la falta de grafos de conocimiento específicos de dominio es generalizada, a excepción de Salud. Este dominio cuenta con UMLS, que por otro lado tiene una cobertura limitada en español y lenguas cooficiales. Otros dominios, como el dominio legal, cuentan con otro tipo de recursos semánticos, como el tesauro multilingüe EuroVoc, con una expresividad semántica muy limitada comparado con un grafo de conocimiento. EuroVoc también viene

acompañado de un gran corpus documental. El dominio financiero cuenta con FIBO, que sin embargo no parece idóneo para los objetivos del proyecto según nuestro análisis preliminar.

La limitación relacionada con la escasez de recursos semánticos, y en especial de grafos de conocimiento, en los distintos dominios reforzará el papel de los grandes grafos de conocimiento de propósito general como Wikidata para dominios específicos a lo largo del proyecto. El estudio de Wikipedia y Wikidata introducido en este entregable pretende informar la utilización de esos recursos para tal fin.

Por otro lado, también aumentará la relevancia, en los dominios que no cuenten con grafos de conocimiento específicos, de los métodos de inyección de conocimiento que no estén directamente afectados por esta limitación. Estos métodos incluyen aquellos basados en la recuperación de información procedente de corpora documental y su inyección en tiempo de inferencia, así como los basados en alineamiento con preferencias, por ejemplo, factuales o de sesgo. Para experimentación con métodos que sí están afectados directamente por esta limitación, como K-Adapter, se priorizarán dominios en los que sí existen esos recursos, como Salud, o se aprovechará la cobertura de dominio ofrecida por Wikidata.

Otra limitación es la escasez de corpora paralelo en español y lenguas cooficiales, que es menor en dominios como Salud, Legal (gracias al proyecto del Plan Nacional de Tecnologías de Lenguaje MarIA) y el sector de análisis de patentes. Finalmente, dominios como Seguros representan un reto en sí mismos debido a la ausencia de recursos específicos que, por otro lado, es interesante debido a lo significativo del sector desde un punto de vista de negocio y al interés latente en explorar nuevos métodos de inyección de conocimiento en LLM en escenarios de pocos o ausencia de recursos.

En resumen, bajo un criterio de disponibilidad de recursos y posible impacto en KG4LLM, los dominios de mayor interés entre los preseleccionados podrían incluir Salud, Legal, Patentes y Seguros. La selección final se hará en colaboración con la dirección de INESData una vez finalizado el entregable E1.1.

Tabla 2 Relación de dominios y sectores de interés. La disponibilidad de los recursos de cada tipo se indica mediante un código de color (verde: existen y son multilingües incluyendo español y/o lenguas cooficiales, naranja: existen sólo en inglés o con una cobertura limitada de otros idiomas y rojo: no existen). Esta tabla no incluye Wikipedia y Wikidata.

Dominio/sector	**Corpora**	**Grafo de conocimiento**	**Otros recursos estructurados**	**Conjuntos de datos para evaluación**
Salud/biomedicina				
Ciencias de la Tierra				
Espacio				
Finanzas				
Patentes				
Legal				
Seguros				

Referencias

[1] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, J. Ji, G. Cao, D. Jiang y M. Zhou, «K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters,» de *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

[2] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel y E. Gave, «Atlas: Few-shot Learning with Retrieval Augmented Language Models,» arXiv:2208.03299, 2022.

[3] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning y C. Finn, «Direct Preference Optimization: Your Language Model is Secretly a Reward Model,» de *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, Louisiana, United States, 2023.

[4] T. Shick y H. Schütze, «Generating datasets with pretrained language models,» de *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, 2021.

[5] Y. Quin, «ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs,» arXiv:2307.16789, 2023.

[6] M. A. Valenzuela-Escarcega, V. A. Ha y O. Etzioni, «Identifyng Meaningful Citations,» de *AAAI Workshop: Scholarly Big Data*, 2015.

[7] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre y M. Villegas, «Pre-trained Biomedical Language Models for Clinical NLP in Spanish,» de *BioNLP 2022 workshop*, Dublin, Ireland, 2022.

[8] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, F. Hanyi y P. Szolovits, «What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams,» *Association for the Advancement of Artificial Intelligence,* 2021.

[9] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen y X. Lu, «PubMedQA: A Dataset for Biomedical Research Question Answering,» de *EMNLP*, Hong Kong, China, 2019.

[10] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song y J. Steinhardt, «Measuring Massive Multitask Language Understanding,» de *ICLR*, Vienna, Austria, 2021.

[11] «MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering,» de *Conference on Health Inference and Learning (CHIL)*, 2022.

[12] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas y M. Krallinger, «PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track,» de *Workshop on BioNLP Open Shared Tasks*, Hong Kong, China, 2019.

[13] A. Miranda-Escalada, E. Farré-Maduell y M. Krallinger, «Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results,» de *Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*, Málaga, Spain, 2020.

[14] P. Henderson, M. S. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky y D. E. Ho, «Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset,» de *Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks*, New Orleans, Louisiana, United States, 2022.

[15] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz y N. Aletras, «LexGLUE: A Benchmark Dataset for Legal Language Understanding in English,» de *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, 2022.

[16] D. Tuggener, P. v. Däniken, T. Peetz y M. Cieliebak, «LEDGAR: A Large-Scale Multilabel Corpus for Text Classification of Legal Provisions in Contracts,» de *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, 2020.

[17] M. Lippi, P. Pałka, G. Contissa, F. Lagioia y H.-W. Micklitz, «CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service,» *Artificial Intelligence and Law,* vol. 127, pp. 117-139, 2019.

[18] L. Zheng, N. Guha, B. R. Anderson, P. Henderson y D. E. Ho, «When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53.000+ Legal Holdings,» de *Eighteenth International Conference for Artificial Intelligence and Law (ICAIL'21)*, Sao Paulo, Brazil, 2021.

[19] I. Chalkidis, T. Pasini, S. Zhang, L. Tomada, S. Schwemer y A. Søgaard, «FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing,» de *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, 2022.

[20] J. Niklaus, V. Matoshi, P. Rani, A. Galassi, M. Stürmer y I. Chalkidis, «LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain,» de *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023.

[21] D. Hendrycks, C. Burns, A. Chen y S. Ball, «CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review,» de *35th Conference on Neural Information Processing Systems (NeurIPS2021) Track on Datasets and Benchmarks*, 2021.

[22] P. Malo, A. Sinha, P. Korhonen, J. Wallenius y P. Takala, «Good debt or bad debt: Detecting semantic orientations in economic texts,» *Journal of the Association for Information Science and Technology,* vol. 65, nº 4, 2014.

[23] Z. Chen, S. Li, C. Smiley, Z. Ma, S. Shah y W. Y. Wang, «ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering,» de *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Abu Dhabi, United Arab Emirates, 2022.

[24] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk y A. Balahur, «WWW'18 Open Challenge: Financial Opinion Mining and Question Answering,» de *WWW '18: Companion Proceedings of the The Web Conference 2018*, Lyon, France, 2018.

[25] X. Dai y S. Karimi, «Detecting Entities in the Astrophysics Literature: A Comparison of Word-based and Span-based Entity Recognition Methods,» de *Proceedings of the 1st Workshop*

on Information Extraction from Scientific Publications (WISP), Copenhagen, Denmark, 2022.

[26] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita y M. Leippold, «Climate-Fever: A Dataset for Verification of Real-World Climate Claims,» de *Tackling Climate Change with Machine Learning Workshop at NeurIPS 2020*, 2020.

[27] T. Khot, A. Sabharwal y P. Clark, «SciTaiL: A Textual Entailment Dataset from Science Question Answering,» de *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, United States, 2018.

[28] E. Sharma, C. Li y L. Wang, «BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization,» de *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019.

[29] M. Suzgun, L. Melas-Kyriazi, S. K. Sarkar, S. D. Kominers y S. M. Shieber, «The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent Applications,» de *37th Conference on Neural Information Processing Systems (NeurIPS2023) Track on Datasets and Benchmarks*, New Orleans, Louisiana, United States, 2023.

[30] F. Soares, M. Stevenson, D. Bartolome y A. Zaretskaya, «ParaPat: The Multi-Million Sentences Parallel Corpus of Patents Abstracts,» de *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC2020)*, Marseille, France, 2020.

[31] «The EuroPat Corpus: A Parallel Corpus of European Patent Data,» de *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC2022)*, Marseille, France, 2022.

[32] C. Eichenberg, S. Black, S. Weinbach, L. Parcalabescu y A. Frank, «MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning,» de *In Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates, 2022.

[33] S. Monka, L. Halilaj, S. Schmid y A. Rettinger, «Learning Visual Models Using a Knowledge Graph as a Trainer,» de *International Semantic Web Conference*, NY, USA, 2021.

Apéndice A: Information cards

Corpora documental

Nombre	**Dominio o sector**	**Fuente (URL)**	**Idiomas (L/%)**	**#Tokens (M)**	**#Docs. (M)**	**Evidencia (s/n)**
Pubmed	Salud	https://ftp.ncbi.nlm.nih.g	Inglés	-	36	Sí

		ov/pubmed/baseline/				
EC3	Salud	https://github.com/hltfbk/E3C-Corpus	Inglés (20%), Francés (52%), Italiano (21%), Español (4%), Euskera (3%)	5	0.05	Solo la primera capa
Biomedical and clinical Spanish corpora	Salud	https://github.com/PlanTL-GOB-ES/lm-biomedical-clinical-es#corpora-	Español	1054	Más de 0.28	Solo los subsets pertenecientes a fuentes fiables (Pubmed)
Multi-legal PILE	Legal	https://huggingface.co/datasets/joelniklaus/Multi_Legal_Pile	Inglés (55%), portugués (17%), alemán (7%), español (6%), otros (15%)	8121	57	Sí
Spanish Legal Domain Corpora	Legal	https://github.com/PlanTL-GOB-ES/lm-legal-es#corpora-	Español	1373	-	Solo los subsets pertenecientes a fuentes fiables (BOE)
Financial Reports EDGAR SEC	Finanzas	https://huggingface.co/datasets/JanosAudran/financial-reports-sec	Inglés	1876	67	Sí
Corpora de Reuters	Finanzas	https://trec.nist.gov/data/reuters/reuters.html	Inglés, holandés, francés, aleman, chino, japonés ruso, portugués, español, italiano, danés, noruego y sueco	-	2.6	No
NASA NTRS	Espacio	https://www.sti.nasa.gov/	Inglés	-	0.074	Sí

		docs/ntrs-public-metadata.json.gz				
EoPortal	Espacio	https://www.eoportal.org/satellite-missions	Inglés	-	0.001	Sí
CEOS	Espacio	https://database.eohandbook.com/about.aspx	Inglés	-	0.001	Sí
BigPatent	Patentes	https://huggingface.co/datasets/big_patent	Inglés	-	1.3	Sí
HUPD	Patentes	https://huggingface.co/datasets/HUPD/hupd	Inglés	-	4.5	Sí
ParaPat	Patentes	https://huggingface.co/datasets/para_pat	Checo, alemán, griego, inglés, español, francés, húngaro, japonés, coreano, portugués, rumano, ruso, eslovaco, ucraniano y chino	800	68	Sí
EuroPat	Patentes	https://europat.net/	Inglés, alemán, español, francés, croata, noruego y polaco	2563	85	Sí
ArXiV	Múltiple	https://www.kaggle.com/datasets/Cornell-University/arxiv	Inglés	-	1.7	No
Scigraph	Múltiple	https://sn-scigraph.figs	Inglés	-	9.62	Sí

		hare.com/ndownloader/files/39016082				
Semantic Scholar	Múltiple	https://www.semanticscholar.org/product/api	Inglés	-	30	Sí siempre que el artículo pertenezca a una fuente fiable (Pubmed)

Grafos de conocimiento y otros recursos semánticos

Nombre	**Dominio o sector**	**Fuente (URL)**	**Tipo**	**Idiomas (L/%)**	**#Conceptos (M)**	**#Relaciones (M)**
UMLS	Salud	https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html	KG	Inglés y otros 21 idiomas con menor representación	0.9	12
Eurovoc	Legal	https://eur-lex.europa.eu/browse/eurovoc.html	Tesauro	Idiomas de la UE	0.007	-
FIBO	Finanzas	https://spec.edmcouncil.org/fibo/ontology	Ontología	Inglés	-	-
UAT	Espacio	https://astrothesaurus.org/	Tesauro	Inglés	-	-
ESA Data Ontology	Espacio	https://data.esa.int/esado	Tesauro	Inglés	0.015	-

European Patent Registry	Patentes	https://www.epo.org/en/searching-for-patents/data/bulk-data-sets/register-data	Base de datos anotada	Inglés	-	-

Tabla 4 Conjuntos de datos para evaluación

Nombre	**Dominio o sector**	**Fuente (URL)**	**Tarea**	**Idiomas (L/%)**	**#Ejemplos (k)**
MedQA-USMLE	Salud	https://github.com/jind11/MedQA	MCQA	Inglés (20%), chino simplificado (57%) y chino tradicional (23%).	60
PubmedQA	Salud	https://huggingface.co/datasets/pubmed_qa	MCQA	Inglés	211 (artificial), 1 (etiquetado) y 61.2 (sin etiquetar)
MedMCQA	Salud	https://huggingface.co/datasets/medmcqa	MCQA	Inglés	183 (train), 4.18 (validation) y 6.15 (test)
PharmaCoNER	Salud	https://github.com/TeMU-BSC/PharmaCoNER-Tagger	NER	Español	-
CANTEMIST	Salud	https://github.com/TeMU-BSC/cantemist-evaluation-library	NER	Español	-
LEXGlue	Legal	https://huggingface.co/datasets/lex_glue	Clasificación y MCQA	Inglés	235
LegalBench	Legal	https://huggingface.co/datas	162 tareas	Inglés	25

		ets/nguha/legalbench			
FairLEX	Legal	https://huggingface.co/datasets/coastalcph/fairlex	Clasificación	Inglés (Más del 20%), chino (50%), francés e italiano	209
LEXTREME	Legal	https://huggingface.co/datasets/joelniklaus/lextreme	Clasificación y NER	Idiomas de la UE	1330
QUAD	Legal	https://huggingface.co/datasets/cuad	Clasificación	Inglés	22.5 (train) y 4 (test)
FPB	Finanzas	https://huggingface.co/datasets/financial_phrasebank	Sentiment analysis	Inglés	2-5
ConvFinQA	Finanzas	https://github.com/czyssrs/ConvFinQA	Multihop QA y numerical reasoning	Inglés	4
FiQA-2018	Finanzas	https://huggingface.co/datasets/BeIR/fiqa	QA y Sentiment Analysis	Inglés	6
DEAL	Espacio	https://huggingface.co/datasets/adsabs/WIESP2022-NER	NER	Inglés	1.8 (train), 1.4 (validation) y 2.5 (test)
Climate-Fever	Ciencias de la Tierra	https://huggingface.co/datasets/climate_fever	Claim Analysis	Inglés	1.5
SciTail	Múltiple	https://huggingface.co/datasets/scitail	Entailment	Inglés	27
MMLU	Múltiple	https://github.com/hendrycks/test	57 tareas	Inglés	-