

E5.1 MARCO DE EVALUACIÓN Y RECURSOS ASOCIADOS

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

Resumen

Este entregable consiste en el marco de evaluación de la facticidad propuesto en el proyecto. Debido a las limitaciones en cuanto a la disponibilidad de conjuntos de datos de verificación en escenarios de adaptación a dominio multilingüe y en concreto en el marco de los dominios de interés del proyecto identificados en el entregable D4.1 y los idiomas oficiales de España, el presente marco de evaluación propone estrategias en las que la dependencia de ese tipo de recursos no sea un limitante para evaluar la facticidad de un LLM. Además de las ideas principales en las que se basa este marco de evaluación, el entregable incluye el software y recursos asociados, entre los que destacan conjuntos de datos destilados de recursos semiestructurados y grandes LLM para la evaluación de facticidad, y particularmente conjuntos de datos de verificación, en distintos dominios relevantes para el proyecto.

José Manuel Gómez Pérez
Cristian Berrío
Raul Ortega

30 de Junio de 2024
Expert.ai Language Technology Research Lab

Historia de revisions

Revision	Date	Description	Author (Organisation)
0.1	30/04/2024	Tabla de contenidos y estructura básica	Expert.ai
0.2	01/05/2024	Primera versión completa	Expert.ai
1.0	30/06/2024	Versión final	Expert.ai

Tabla de contenidos

1	Introducción	5
2	Enfoque.....	5
3	Marco de evaluación.....	6
4	Recursos de conocimiento a nivel de dominio	8
4.1	Terminología LGIT.....	8
4.2	UMLS.....	9
5	Datasets de evaluación.....	10
5.1	Esquema del dataset de evaluación	10
5.2	Descripción del contenido	11
5.3	Datasets de verificación	12
6	Escenarios de evaluación de factualidad	14
6.1	Estimador basado en la confianza del LLM	14
6.2	Estimador basado en referencias.....	14
7	Evaluación de factualidad de LLM en dominios piloto	15
8	Repositorio del marco de evaluación	15
9	Conclusiones y trabajo futuro.....	16
	Referencias.....	17
Apéndice A:	Prompts para obtener preguntas sobre entidades y resultados de ejemplo	26
Apéndice B:	Prompts y ejemplos para, dada una pregunta, generar una respuesta en forma de párrafo.....	26
Apéndice C:	Prompts y ejemplos para extraer hechos atómicos a partir de respuestas.....	30
Apéndice D:	Prompts para convertir hechos atómicos en preguntas y resultados de ejemplo	32

Índice de tablas

<i>Tabla 1 Estadísticas de los datasets.</i>	11
<i>Tabla 2 Entidades de test en el dominio de los seguros directamente mapeadas a artículos en Wikipedia.</i>	13
<i>Tabla 3 Entidades de test mapeadas a través de sinónimos y/o hiperónimos con artículos de Wikipedia</i>	13
<i>Tabla 4 Resultados de Llama-2 7b en el dominio de seguros y de la salud, con FactScore y con el estimador de la factualidad basada en la confianza del modelo.</i>	15
<i>Tabla 5 Prompts de GPT3.5 para convertir entidades en preguntas.</i>	26
<i>Tabla 6 Ejemplos de preguntas generadas a partir de entidades en el dominio de seguros y en el de salud (COVID-19).</i>	26
<i>Tabla 7 Prompt para dada una pregunta, responder en forma de párrafo.</i>	28
<i>Tabla 8 Ejemplos de respuestas generadas a partir de preguntas en el dominio de los seguros y en el de salud (COVID-19).</i>	30
<i>Tabla 9 Ejemplos de hechos atómicos extraídos a partir de respuestas en el dominio de seguros y en el de salud (COVID-19).</i>	32
<i>Tabla 10 Prompt de GPT 3.5 para convertir hechos atómicos en preguntas.</i>	35
<i>Tabla 11 Ejemplos de preguntas generadas a partir de hechos atómicos el dominio de seguros.</i>	35

Tabla de figuras

<i>Figura 1: Métodos para la evaluación de factualidad en LLM</i>	<i>6</i>
<i>Figura 2: Marco de evaluación: Pasos de evaluación de la factualidad para los métodos basados en la confianza del LLM y/o en referencias sobre un dataset de verificación</i>	<i>8</i>
<i>Figura 3: Distribución por grupos semánticos seleccionados en el dominio de salud.</i>	<i>9</i>
<i>Figura 4: Porcentaje de entidades contenidas en el subconjunto de UMLS para cada idioma. 10</i>	
<i>Figura 5 Niveles y subniveles de los datasets de evaluación, junto con sus respectivos tamaños para cada dataset de evaluación</i>	<i>11</i>

1 Introducción

Este entregable recoge los resultados de la tarea T5.1 del paquete de trabajo PT5 en el proyecto KG4LLM, sobre el desarrollo de métricas y marco de evaluación de factualidad previamente definido en la tarea T1.3 y que fue presentado en el entregable E1.1. El entregable E5.1 se basa en las guías definidas por E1.1 y tiene en cuenta los retos y limitaciones allí identificados, así como las recomendaciones basadas en dicho análisis.

El resto de este documento se estructura de la manera siguiente. La sección 2 introduce de manera general la filosofía detrás del desarrollo del marco de evaluación y las líneas generales que seguimos a la hora de presentarlo en este documento. La sección 3 presenta el marco de evaluación propiamente dicho y los distintos métodos de evaluación que ofrece. La sección 4 presenta los recursos utilizados para ilustrar la aplicación del marco de evaluación en los dominios seleccionados para este documento, que incluyen un glosario de términos para el dominio de seguros basado en la terminología LGIT y un grafo de conocimiento extenso, complejo y multilingüe como UMLS. Por otro lado, la sección 5 presenta el proceso de desarrollo de los datasets destilados a partir de estos recursos mediante el uso de otros LLM, para su uso durante la evaluación de factualidad, así como los datasets de verificación para cada dominio. A continuación, la sección 6 describe cómo usar el dataset dependiendo de si el estimador elegido para evaluar la factualidad del modelo está basado en referencias o en la confianza del LLM en el texto generado. La sección 7 ilustra el uso del marco de evaluación de la factualidad aplicándolo a un LLM en particular, LLaMa-2, en los dominios objetivo. La sección 8 presenta el repositorio desde el que se puede acceder al marco de evaluación para su uso por la comunidad de usuarios de INESData. Finalmente, en la sección 9 presentamos nuestras conclusiones y anticipamos parte del trabajo que vendrá a continuación.

2 Enfoque

Como se introdujo en E1.1, que recoge el estado de la cuestión, guías para la inyección de conocimiento en LLM y métricas, el proyecto KG4LLM apuesta por el desarrollo de un marco de trabajo para la evaluación de factualidad en LLM que sea lo más agnóstico posible de dominio e idioma. Esta decisión se basa en el análisis de una serie de dominios de interés, entre los que se encontraban salud, patentes, legal, ciencias de la tierra y seguros entre otros, para los idiomas oficiales de España llevado a cabo en el entregable E4.1. En dicho análisis se comprobó que prácticamente ningún dominio cuenta con los recursos semánticos o de corpora documental deseables para implementar en este escenario métodos convencionales de inyección de conocimiento en LLM.

El marco de evaluación de la factualidad aquí propuesto permite verificar el texto generado por el modelo contra datasets de verificación centrados en combinaciones específicas de dominio e idioma. Asimismo, también ofrece métodos intrínsecos basados en la confianza que el modelo tiene en el texto generado, de manera independiente de la existencia de un dataset de verificación, aliviando así la dependencia de este tipo de recursos para un idioma y dominio dados. Finalmente, este marco de evaluación se adapta a los recursos de conocimiento disponibles sobre los que se pretende evaluar la factualidad del modelo. Dichos recursos pueden variar desde un glosario de términos a un grafo de conocimiento completo.

Este marco de evaluación es por tanto válido para evaluar la factualidad de cualquier LLM para cualquier idioma y dominio. Con el fin de simplificar la experimentación, que ilustra el uso del marco de evaluación, en este entregable nos centramos en LLMs de la familia Llama y en

concreto Llama-2 7B, idioma inglés¹, y dos dominios antagónicos en cuanto a la disponibilidad de recursos. Por un lado, el dominio de seguros, para el que, como vimos en el entregable E4.1, no existen datasets de verificación, corpora o recursos, como grafos de conocimiento, públicos, y, por otro lado, el dominio de salud, para el que sí hay abundancia de recursos, entre los que se incluyen grafos de conocimiento multilingües como UMLS, corpora y datasets de verificación como PubMed. De esta manera, mostramos el comportamiento de nuestro método en dominios tanto de abundancia como de escasez de recursos.

3 Marco de evaluación

El marco propuesto se resume en la Figura 1, donde se describen los cuatro métodos de evaluación de la factualidad fundamentales. Cada uno de estos métodos permiten evaluar la factualidad del LLM con un enfoque diferente pero complementario.

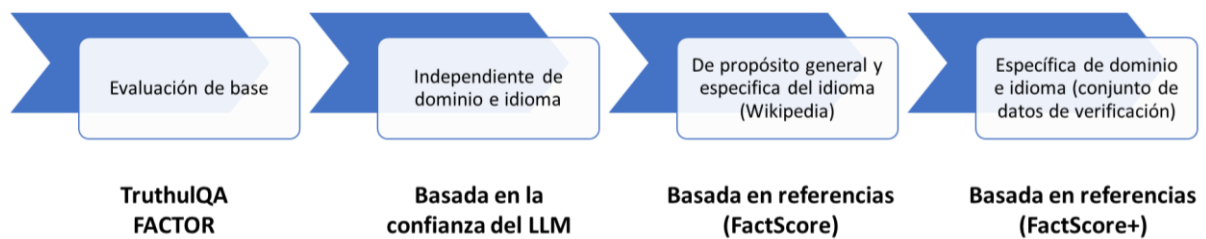


Figura 1: Métodos para la evaluación de factualidad en LLM

A continuación, desgranamos estos métodos.

1. **Evaluación de base.** Se propone utilizar conjuntos de datos de evaluación previamente disponibles en el estado del arte para evaluar la veracidad del LLM, como TruthfulQA (Lin & Hilton et al., 2022) y FACTOR (Muhlgay et al., 2023). Dichos conjuntos de datos son de propósito general y normalmente sólo están disponibles en inglés, pero la evaluación sobre ellos puede aportar una perspectiva inicial interesante sobre la factualidad del modelo en el escenario más favorable posible (monolingüe y de propósito general). Este tipo de evaluación también permite medir el impacto de los métodos de inyección de conocimiento y adaptación a dominio a nivel general. Idealmente, dicho impacto debería ser positivo o al menos no perjudicar la factualidad del modelo en los escenarios evaluados por estos datasets. De manera similar, la evaluación de base incluye marcos de evaluación generales para LLM más allá de aspectos relacionados con la factualidad, como MMLU (Hendrycks et. al, 2021), con los que se permite monitorizar el impacto de los métodos desarrollados en el proyecto sobre las capacidades básicas de los LLM.
2. **Evaluación basada en la confianza del LLM, independiente de dominio e idioma.** La siguiente fase es completamente independiente de dominio e idioma y se basa en el método originalmente propuesto por Tian et al. (2023b). En ella explotamos la correlación demostrada en (Kadavath et al., 2022; Tian et al., 2023a) entre la confianza

¹ Aunque el método es independiente del idioma y tan sólo requiere que el LLM a evaluar soporte dicho idioma y la existencia de recursos de conocimiento de partida en ese idioma, que como indicamos puede variar entre un glosario de términos hasta un grafo de conocimiento.

de un LLM en el texto generado y la probabilidad de que dicha respuesta sea factual. Este método primero extrae hechos atómicos del texto utilizando un LLM fuerte, como GPT-3.5/4. A continuación usa ese mismo modelo para convertir cada hecho atómico en una pregunta sin ambigüedad con la que evaluar el conocimiento del LLM sobre ese hecho en particular y muestrea una respuesta del LLM bajo evaluación un número de veces (20, en el trabajo original) con el fin de estimar la incertidumbre del modelo sobre la respuesta generada. La proporción de las apariciones de la respuesta más frecuente sobre el total de respuestas obtenidas es la puntuación de veracidad final asignada a cada hecho, que esencialmente representa la confianza del modelo en su factualidad. Este método es extremadamente flexible al no depender de la existencia de un dataset de verificación para el dominio e idioma en los que nos proponemos evaluar el LLM. Por el contrario, confía en el conocimiento del dominio y las capacidades que para ese idioma puede tener el LLM utilizado para generar hechos atómicos y las consiguientes preguntas sobre ellos.

3. **Evaluación basada en referencias sobre conjuntos de datos generales y específica de idioma.** El método de evaluación basado en la confianza del LLM descrito en el punto anterior se complementa con otro tipo de estimadores, basados en referencias, como FactScore. Estos métodos suelen ofrecer los mejores resultados comparados con el método basado en la confianza del modelo. Sin embargo, dependen de la existencia de un dataset de verificación, basado normalmente en distintas versiones de Wikipedia para cada idioma como base de conocimiento general y reutilizable para los distintos dominios e idiomas objetivo.
4. **Evaluación basada en referencias sobre conjuntos de datos de verificación específicos de dominio e idioma.** De manera similar al punto anterior, el propósito de este método de evaluación es aprovechar los conjuntos de datos de verificación disponibles para ciertos dominios, como se vio en E4.1. Para ello, se plantea una adaptación del método FactScore que sustituya conjuntos de datos basados en Wikipedia por esos conjuntos de datos de verificación. La idea es que la verificación será de mayor calidad al contar este tipo de datasets de verificación específicos de dominio con una mayor profundidad y especialización a nivel de dominio que un recurso de propósito general como Wikipedia.

En el marco de evaluación aquí propuesto, nos centramos en métodos basados en confianza del modelo y basados en referencia, cuando haya disponible un dataset de verificación para el dominio e idioma indicados. La *Figura 2* resume los pasos a tomar para evaluar la factualidad de un LLM, tanto basado en un tipo de estimador como en el otro.

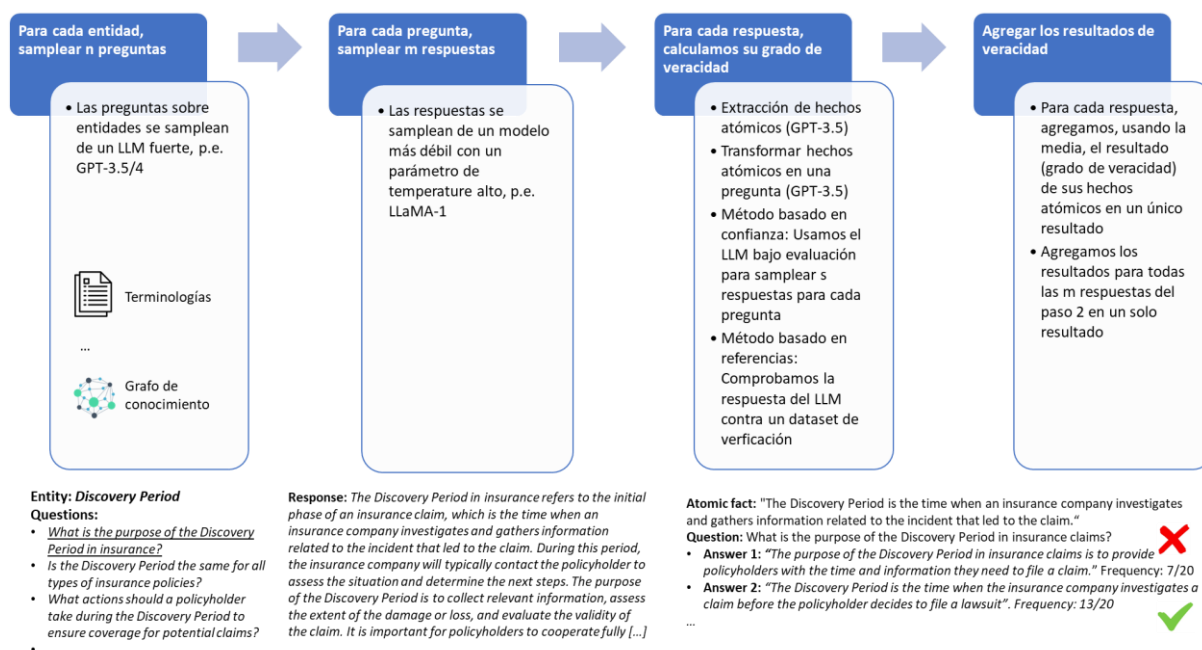


Figura 2: Marco de evaluación: Pasos de evaluación de la factualidad para los métodos basados en la confianza del LLM y/o en referencias sobre un dataset de verificación

4 Recursos de conocimiento a nivel de dominio

Para poder llevar a cabo el primer paso definido en nuestro marco de evaluación de factualidad para LLM descrito en la Figura 2, es necesario partir de un recurso de conocimiento para el dominio de interés en el idioma indicado en el que se describan las entidades propias del dominio y las relaciones entre ellas. En su expresión más sencilla, ese recurso puede ser una lista o glosario de términos. También podría tratarse de una taxonomía con información de hiperonimia entre entidades e idealmente de un grafo de conocimiento con relaciones explícitamente representadas entre las entidades de dominio.

Como veremos en la sección 5, esa información se utilizará para generar preguntas sobre esas estructuras de conocimiento de dominio en el idioma objetivo mediante el uso de un LLM fuerte. Posteriormente, esas preguntas serán respondidas por un LLM débil con un parámetro de temperatura alto, con el fin de generar una gran diversidad de respuestas, que contendrán distintas distribuciones de hechos correctos e incorrectos.

4.1 Terminología LGIT

El Fondo de Seguros para Gobiernos Locales (LGIT, por sus siglas en inglés) fue establecido en 1987 por la Asociación de Condados de Maryland (MACo) y la Liga Municipal de Maryland (MML) para abordar los desafíos que enfrentaban los gobiernos locales en Maryland al intentar asegurar seguros accesibles. LGIT es una asociación propiedad de sus miembros, autorizada por ley estatal y gestionada por los propios gobiernos locales, diseñada para ofrecer programas conjuntos de auto-seguro para ciudades, pueblos y condados en Maryland. En lugar de pagar primas a compañías de seguros, los gobiernos locales contribuyen a un fondo común utilizado para cubrir reclamaciones, pérdidas y gastos. Este modelo garantiza una cobertura de seguros estable y predecible, independientemente de las fluctuaciones en el mercado de seguros privados, que anteriormente dejaban a muchos gobiernos locales luchando con costos elevados y pérdida de cobertura durante ciclos económicos difíciles. LGIT

ofrece un Glosario de Términos de Seguros.² Esta terminología constituye una guía de referencia para proporcionar al usuario definiciones rápidas y fáciles de términos comunes de seguros. Actualmente el glosario de LGIT contiene 114 términos y sus definiciones en inglés.

4.2 UMLS

El Unified Medical Language System³ (UMLS) es un recurso de conocimiento en salud que recoge múltiples vocabularios en varios idiomas (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.

UMLS puede ser utilizado para la extracción de entidades, enlazar conocimiento de textos en diferentes idiomas o facilitar el desarrollo de sistemas de recuperación de datos. Para el propósito de este entregable, nos quedamos con un subconjunto de UMLS centrado en temas relacionados con COVID-19, que alineamos con un dataset de verificación construido con 4.5 millones de abstracts de papers científicos en áreas relevantes (ver sección 5.3.2).

En concreto, dicho subconjunto de UMLS cuenta con 291 entidades en 23 idiomas, cubriendo 10 dominios representados en más de un 3.4% de artículos del conjunto de datos de verificación y un long tail con el 60% restante, para un total de 73 subdominios del ámbito de salud, entre los que se encuentran *Enfermedad o Síndrome*, *Fáctor Inmunológico* o *Proceso Terapéutico* (ver *Figura 3*). El porcentaje de esas entidades cubiertas para cada idioma está representado en la *Figura 4*. De las lenguas oficiales españolas sólo el Castellano y el Euskera están representados, con 61.09% y 1.37% respectivamente.

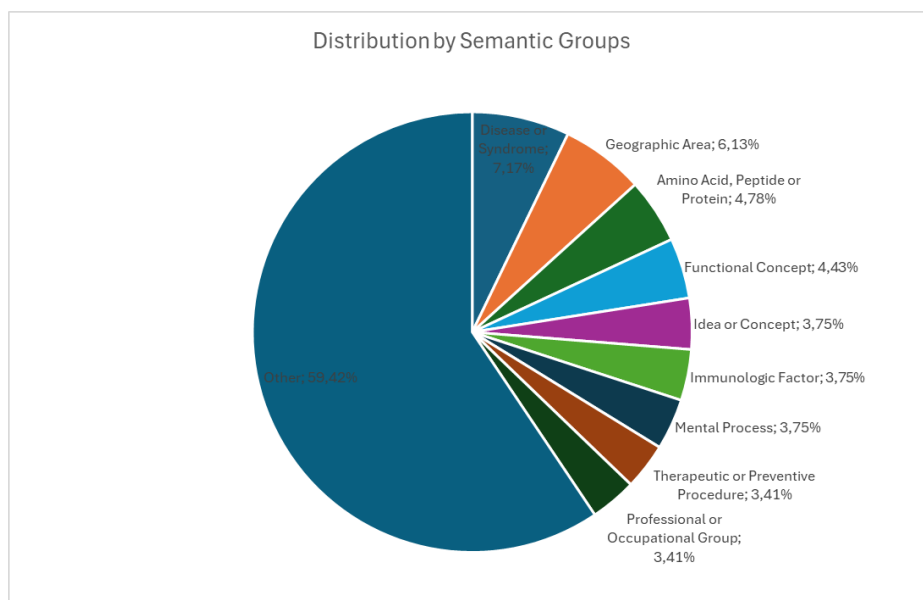


Figura 3: Distribución por grupos semánticos seleccionados en el dominio de salud.

² <https://lgit.org/611/Glossary-of-Insurance-Terminology>

³ <https://www.nlm.nih.gov/research/umls>

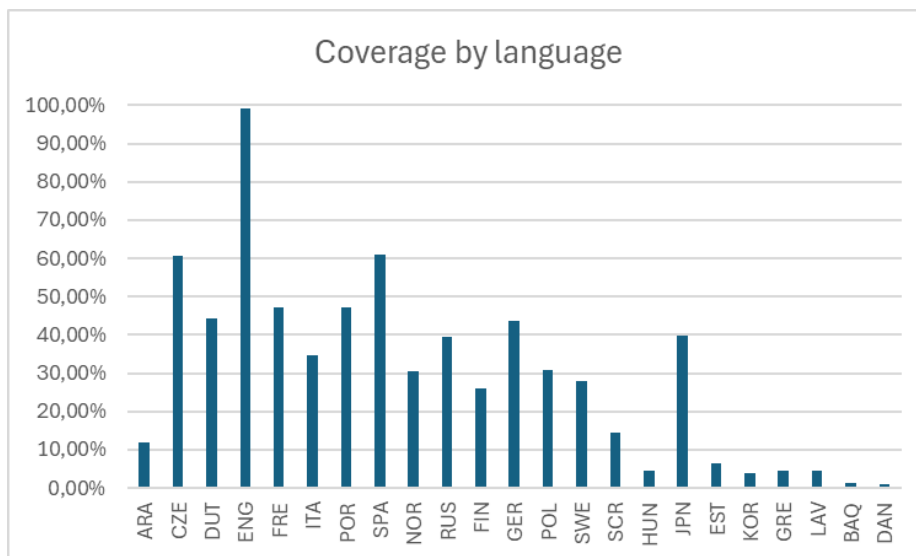


Figura 4: Porcentaje de entidades contenidas en el subconjunto de UMLS para cada idioma.

5 Datasets de evaluación

En esta sección se recoge información acerca de los dataset de evaluación generados para el dominio de seguros, Insurance QA-test, y para el dominio de salud, Medical QA-test. A lo largo de esta sección, se expone el esquema seguido por estos datasets y se aporta información sobre su contenido, así como de los datasets de verificación en los que se apoyan para la evaluación de la factualidad basada en referencias.

5.1 Esquema del dataset de evaluación

En ambos dominios, los datasets de evaluación están compuestos por varios niveles y subniveles. En un primer nivel están las entidades provenientes de los recursos de conocimiento a nivel de dominio explicados en la sección anterior. Estas entidades pueden ser extendidas con otros artefactos provenientes de los recursos utilizados, como, por ejemplo, relaciones entre tanto a nivel de propiedades de dominio como taxonómicas, dependiendo de la expresividad del recurso de conocimiento utilizado. En un segundo subnivel están las preguntas generadas a partir de dichas entidades. Después, en un tercer subnivel están las respuestas a estas preguntas. En un cuarto subnivel están los hechos atómicos extraídos de las respuestas, y, por último, en el quinto subnivel están las preguntas generadas a partir de los hechos atómicos extraídos. En la [Figura 5](#), se pueden ver los diferentes niveles y subniveles que componen el dataset de evaluación, además del número de recursos generados en cada paso del proceso.

	Insurance QA-test	Medical QA-test
Entidad	11	27
Pregunta (entidades)	66	162
Respuesta	396	972
Hecho atómico	2258	4588
Pregunta (hechos atómicos)	2258	4588

Figura 5 Niveles y subniveles de los datasets de evaluación, junto con sus respectivos tamaños para cada dataset de evaluación

5.2 Descripción del contenido

El primer paso para la evaluación la factualidad, tanto basada en la confianza del modelo como la basada en referencias, es generar una lista de preguntas centradas en el dominio objetivo, usando las entidades como semilla.

En la *Tabla 1*, se detalla el tamaño de cada dataset. Para el caso de Insurance QA, el dataset generado a partir de la terminología LGIT, se cuenta con un total de 113 entidades, de las cuales el 10% (11) son usadas para test, y de las 102 restantes, 91 entidades se usarán para entrenamiento y otras 11 para validación. Para Medical QA, por su parte, se han seleccionado 291 entidades, de las cuales 235 se utilizan para entrenar y 29 para validar. En ambos dominios se generan 6 preguntas por cada entidad, y 6 respuestas por cada pregunta.

Dataset	Entidades [train, test]	Preguntas por entidad	Respuestas por pregunta	Ejemplo de preguntas
Insurance QA	113 [102 (91 train, 11 validation), 11]	6	6	What is excess liability insurance and how does it differ from primary liability coverage?
Medical QA	291 [264 (235 train, 29 validation), 27]	6	6	How effective are COVID-19 vaccines in preventing infection and transmission of the virus?

Tabla 1 Estadísticas de los datasets.

En el Apéndice A: , se incluyen los prompts utilizados para generar las preguntas a partir de las entidades de cada dominio, así como algunos ejemplos de preguntas generadas por GPT 3.5 a partir de una entidad en particular.

Una vez generadas las preguntas en cada dominio, el siguiente paso es responder tales preguntas. En este caso, de manera inicial se ha utilizado un Llama-2 7B para generar las respuestas. En el caso de la evaluación de factualidad del modelo se ha empleado una temperatura de 0.6. En el Apéndice B: se recogen los prompts utilizados para responder a las preguntas, así como algunos ejemplos de respuestas obtenidas. Como se puede ver, se trata de un prompt *few-shot* en el que se le pasan al modelo una serie de preguntas y respuestas de ejemplo, junto a un separador, para mostrar al modelo el formato de una respuesta esperada, y cómo se debe marcar el final de la respuesta mediante el separador.

Para evaluar la factualidad de las respuestas generadas en el paso anterior, es necesario extraer los hechos atómicos a partir de las respuestas. Esto se hace con GPT 3.5, haciendo uso del prompt que mostramos en el Apéndice C: Dentro del prompt también se indica la entidad en cuestión sobre la cual se genera el párrafo. Esto resulta de utilidad, ya que permite al LLM sustituir posibles referencias a la entidad por el nombre explícito, logrando de este modo que los hechos atómicos tengan la suficiente información para poder ser verificados por ellos mismos. Algunos ejemplos de extracción de hechos atómicos a partir de respuestas se pueden observar en el Apéndice C:

Por último, en el Apéndice D, se encuentran los prompts para convertir un hecho atómico en una pregunta. También se incluye algunos ejemplos de preguntas generadas.

5.3 Datasets de verificación

Como ya se ha indicado previamente, la evaluación de la factualidad basada en referencias depende de un dataset de verificación que sirva para evaluar si un hecho atómico está soportado por alguna evidencia dentro del dataset de verificación. A continuación, se presenta cada uno de los datasets empleados para el dominio de seguros, y para el dominio de salud.

5.3.1 Dataset de verificación para seguros basado en Wikipedia

Como dataset de verificación para el dominio de seguros utilizamos el dataset de verificación desarrollado por Min et al. (2023b), el cual consiste en un volcado de Wikipedia en inglés con fecha del 01/04/2023. Para la extracción de este dataset también aplicamos los métodos de extracción de subgrafos de conocimiento resultantes de la tarea T2.1, que fueron presentados con anterioridad en el entregable E4.1.

Una de las limitaciones de utilizar Wikipedia para verificar la factualidad, es que no todas las entidades de test se encuentran directamente como artículos, por lo que es necesario realizar un mapeo entre las entidades en el dominio de seguros, y los artículos donde se pueda encontrar información para verificar los hechos sobre estas entidades. Este mapeo consiste en primer lugar en comprobar de forma manual si la entidad aparece en Wikipedia, en caso de ser así, el mapeo es directo. Sin embargo, si la entidad no aparece como tal, se realiza una expansión de términos, donde para cada entidad se obtiene a través de ChatGPT una lista de sinónimos e hiperónimos. Después se hace la búsqueda de los términos expandidos en Wikipedia y, si existe el artículo, se añade al mapeo.

Del total de las 11 entidades de test, 6 han sido directamente mapeadas y las 5 entidades restantes están parcialmente mapeadas a través de sinónimos y/o hiperónimos. En la Tabla 2 se recogen las entidades mapeadas directamente, mientras que en la Tabla 3 aparecen las entidades que han sido mapeadas en artículos de Wikipedia a través de sinónimos y/o hiperónimos.

Indemnification Agreement	Indemnity
Loss Ratio	Loss ratio
Negligence	Negligence
Insurance	Insurance
Self Insurance	Self-insurance
Umbrella Liability Insurance	Umbrella insurance

Tabla 2 Entidades de test en el dominio de los seguros directamente mapeadas a artículos en Wikipedia.

Excess Liability Insurance	synonyms: Umbrella Liability Insurance	Umbrella insurance
	hypernyms: Liability Insurance Commercial Liability Insurance Business Insurance Insurance Policies	Liability insurance Commercial general liability insurance Insurance Insurance policy
Special Events Insurance	hypernyms: Liability Insurance Commercial Insurance Property and Casualty Insurance Business Insurance Casualty Insurance	Liability insurance Commercial general liability insurance General insurance Insurance Casualty insurance
Endorsements	synonyms: Riders	Insurance policy
Advertising Injury	hypernyms: Liability Coverage Commercial Liability Professional Liability Legal Liability	Liability insurance Commercial general liability insurance Professional liability insurance Legal liability

Tabla 3 Entidades de test mapeadas a través de sinónimos y/o hiperónimos con artículos de Wikipedia

5.3.2 Dataset de verificación para salud basado en PubMed

Para el dominio de salud, hemos utilizado la base de datos de 4.5 millones de artículos de PubMed presentada en el Entregable E4.1. Dicha base de datos cuenta con numerosos abstracts relacionados con el estudio de COVID-19, por lo que nos va a permitir usarla como dataset de verificación para los hechos atómicos generados por los diferentes modelos del lenguaje que queramos evaluar. Por ello, el método utilizado para alinear estos abstracts de verificación con los hechos atómicos es diferente al explicado en el apartado anterior para el dominio de seguros. En este caso, en lugar de utilizar las entidades semilla de cada uno de los hechos atómicos para buscar una serie de artículos de Wikipedia que sirvan para soportarlos o refutarlos, usaremos directamente el hecho atómico para buscar artículos relacionados en nuestra base de datos.

De este modo, tras haber indexado los 4.5 millones de artículos en un índice de Elasticsearch, hemos lanzado peticiones para encontrar 5 artículos relacionados para cada hecho del dataset de evaluación, usando como query el propio hecho, para posteriormente añadirlos al contexto que nos permite discernir si un hecho atómico está soportado o no, siguiendo el método desarrollado por Tian et al. (2023b). Este cambio de criterio a la hora de elegir el dataset de verificación deberá ser tomado en cuenta a la hora de comparar los resultados de los modelos en ambos dominios, ya que tanto la forma de recuperación de estos textos de verificación, como la naturaleza de los mismos (abstracts de artículos científicos de dominio vs artículos de Wikipedia) difieren de forma notable y pueden repercutir en los resultados finales de evaluación.

6 Escenarios de evaluación de factualidad

Dado un dataset de evaluación, se contemplan dos escenarios de evaluación de la factualidad diferentes. Dependiendo del tipo de estimador que se utilice, se usará el dataset de evaluación al completo (el basado en la confianza del modelo), o tan sólo una parte del dataset (el basado en referencias).

6.1 Estimador basado en la confianza del LLM

En este escenario, se utiliza el dataset de evaluación al completo, partiendo de las preguntas generadas a partir de hechos atómicos. Este estimador se basa en la hipótesis de que un modelo más factual tendrá una confianza mayor al responder las preguntas de un dataset de evaluación. El LLM objeto de estudio será evaluado utilizando estas preguntas y comprobando cómo de consistente es a la hora de responder a cada una de ellas un número M de veces. Este dataset de evaluación de preguntas generadas a partir de hechos atómicos resulta por tanto reutilizable para cualquier LLM del que se quiera evaluar la factualidad, ya que en este marco únicamente se evalúa la consistencia o la confianza del modelo a la hora de responder, no si la respuesta es correcta según un dataset de verificación. Esto último se evalúa con el siguiente estimador, basado en referencias.

6.2 Estimador basado en referencias

En este caso, la verificación de la factualidad se realiza a nivel de hechos atómicos, comparándolos con una base de datos de referencia que permita verificarlos. Por ello, para cada LLM que se quiera evaluar, será necesario generar una nueva colección de hechos. Por este motivo del dataset de evaluación mencionado con anterioridad, únicamente recuperaríamos los recursos hasta el nivel de preguntas generadas a partir de entidades, regenerando el resto de los recursos usando el modelo objetivo. Una vez contemos con los

hechos atómicos formados por el LLM, se procederá a la búsqueda de textos con los que poder verificarlos y al cálculo de un score que permita evaluar la veracidad de las generaciones de dicho modelo para un dominio en concreto.

7 Evaluación de factualidad de LLM en dominios piloto

Los resultados de la evaluación de la factualidad, tanto basada en la confianza del modelo como la basada en referencias (FactScore), de un Llama-2 7B base para los diferentes dominios se recogen en la Tabla 4.

Dominio	LLM	FactScore	Factualidad basada en la confianza del modelo
Seguros	Llama-2 7b	0.8116	0.6816
Salud (COVID-19)	Llama-2 7b	0.765	0.4233

Tabla 4 Resultados de Llama-2 7b en el dominio de seguros y de la salud, con FactScore y con el estimador de la factualidad basada en la confianza del modelo.

Estos resultados, si bien quizás no aportan una información definitiva acerca del desempeño del modelo para los dominios seleccionados, sí que sirven como piedra angular para la evaluación de diferentes modelos en dichos dominios. Más adelante se añadirán diferentes modelos a esta tabla, lo que seguramente nos dará una visión más global acerca de la factualidad de los LLMs.

8 Repositorio del marco de evaluación

El código del marco de evaluación se encuentra disponible en <https://github.com/oeg-upm/inesdata-fact-eval>. En el repositorio se encuentran dos carpetas principales: “model_confidence” y “FActScore”.

La carpeta “model_confidence” contiene los scripts en Python necesarios para evaluar la factualidad de modelos siguiendo el procedimiento:

1. generate_questions_dataset.py: Genera el dataset de prompt a partir de las entidades.
2. sample_model.py/sample_model_vllm.py: Muestra el modelo con los prompts previamente generados para generar respuestas a las preguntas generadas en el paso anterior.
3. extract_claims.py: Extrae los hechos atómicos de las respuestas del modelo.
4. claims_to_questions.py: Convierte los hechos atómicos en preguntas.
5. answer_questions.py/answer_questions_vllm.py: Responde a las preguntas con el modelo que se quiera evaluar.
6. truthfulness_score.py: Calcula la puntuación de factualidad de cada respuesta del modelo, a partir de las respuestas del paso anterior.
7. truthfulness_score_mean.py: Calcula la puntuación agregada de todas las puntuaciones de todas las respuestas.

Para evaluar la factualidad de un nuevo modelo en el dominio de los seguros o de salud, dado el dataset de evaluación, los cuatro primeros pasos pueden omitirse, y se partiría del punto 5.

La carpeta “FactScore” contiene un *fork* del repositorio original⁴, aunque con algunas modificaciones, como el soporte de Llama 3 instruct como modelo de verificación de los hechos con las referencias. También se soporta la verificación usando varios artículos, lo cual es necesario cuando se realiza la expansión de términos para ciertas entidades en el dominio de los seguros. Para la evaluación con FactScore, nos apoyamos en algunos scripts de la carpeta “model_confidence”, en concreto el procedimiento es el siguiente:

1. Generar el dataset de preguntas a partir de las entidades. (Si no se ha generado ya anteriormente)
2. Muestrear el modelo con los prompts previamente generados.
3. Extraer los hechos atómicos de las respuestas del modelo.
4. `factscorer.py/factscorer_es.py`: Calcula el FactScore a partir de los hechos atómicos. La versión “`factorer_es.py`” utiliza Elasticsearch para la búsqueda de párrafos relacionados con los hechos atómicos, y es lo que se ha usado para guardar el dataset de referencia para el dominio de la salud.

Para evaluar la factualidad de un nuevo modelo en el dominio de los seguros o de salud, dado el dataset de evaluación, se puede omitir el primer paso, y se partiría del segundo paso.

9 Conclusiones y trabajo futuro

Este entregable presenta el marco de evaluación de la factualidad de INESData, con el que es posible evaluar, usando distintos estimadores la factualidad de cualquier LLM, en cualquier idioma y para cualquier dominio. Una de las principales características de este marco de evaluación es su flexibilidad, que incluye la posibilidad de utilizar estimadores basados en la confianza del modelo sin necesidad de un dataset de verificación para el dominio e idioma en cuestión. Además de presentar el marco de evaluación, este entregable ilustra su aplicación en dominios antagónicos en cuanto a la disponibilidad de recursos. Por un lado, seguros, en el que sólo contamos con un glosario de términos, del que nos nutrimos para proporcionar a nuestro método las entidades necesarias, y un dataset de verificación genérico basado en Wikipedia. Por otro, el dominio de salud, en el que contamos con un grafo de conocimiento multilingüe mucho más rico y un dataset de verificación ad-hoc. Una de las primeras observaciones que obtenemos es que los resultados de factualidad obtenidos por los distintos LLM deben ser comparados dentro de cada dominio, ya que vienen determinados por la calidad y tamaño de estos recursos. En el futuro, aplicaremos el marco de evaluación a más modelos en distintos idiomas y dominios. Finalmente, este marco de evaluación es de crucial importancia porque nos permite generar datasets de preferencias con los que alinear los LLM para un objetivo de factualidad, como se muestra en el entregable E2.1, así como evaluar el impacto de dichos métodos de inyección de conocimiento factual en LLM, comparándolos con los LLM de partida como baselines.

⁴ <https://github.com/shmsw25/FActScore>

Referencias

- Ahn, J., & Oh, A. (2021). **Mitigating language-dependent ethnic bias in BERT**. arXiv preprint arXiv:2109.05704.
- Almazrouei E., Alobeidli H., Alshamsi A., Cappelli A., Cojocaru R., Debbah M., Goffinet É., Hesslow D., Launay J., Malartic Q., Mazzotta D., Nouné B., Pannier B., & Penedo G. (2023). **The Falcon Series of Open Language Models**. arXiv preprint arXiv:2311.16867
- Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). **Self-rag: Learning to retrieve, generate, and critique through self-reflection**. arXiv preprint arXiv:2310.11511.
- Bender, EM. (2019). **A typology of ethical risks in language technology with an eye towards where transparent documentation can help**.
- Benjamins, V. R. (1993). **Problem Solving Methods for Diagnosis**. PhD thesis, University of Amsterdam, Amsterdam, The Netherlands.
- Benjamins, V. R., & Fensel, D. (1998). **Problem-solving methods**. International Journal of Human-Computer Studies, 49(4), 305-313.
- Berquand, A., Ladeira A. V. (2022). **From Mission Description to Knowledge Graph: Applying Transformer-based models to map knowledge from publicly available satellite datasets**.
- Bommasani, R., Liang, P., & Lee, T. (2023). **Holistic Evaluation of Language Models**. Annals of the New York Academy of Sciences.
- Bordia, S., & Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. arXiv preprint arXiv:1904.03035.
- Brown, TB. et al. (2020). **Language Models are Few-Shot Learners**. ArXiv, abs/2005.14165.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). **Semantics derived automatically from language corpora contain human-like biases**. Science, 356(6334), 183-186.
- Chen, J., Sriram, A., Choi, E., & Durrett, G. (2022). **Generating Literal and Implied Subquestions to Fact-check Complex Claims**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3495–3516). Association for Computational Linguistics.
- Chen S., Zhao Y., Zhang J., Chern I., Gao S., Liu P., & He J. (2023). **FELM: Benchmarking Factuality Evaluation of Large Language Models**. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Track on Datasets and Benchmarks*
- Chern, I. et al. (2023). **Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios**. CoRR, abs/2307.13528.
- Chowdhery, A.- et al. (2022). **Palm: Scaling language modeling with pathways**. URL: <https://arxiv.org/abs/2204.02311>
- Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., & Amodei, D. (2017). **Deep Reinforcement Learning from Human Preferences**. ArXiv, abs/1706.03741.
- Chuang, Y., Xie, Y., Luo, H., Kim, Y., Glass, J.R., & He, P. (2023). **DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models**. ArXiv, abs/2309.03883.
- Cobbe K., Kosaraju V., Bavarian M., Chen M., Jun H., Kaiser L., Plappert M., Tworek J., Hilton J., Nakano R., Hesse C., & Schulman J. (2021). **Training Verifiers to Solve Math Word Problems**. arXiv preprint arXiv:2110.14168
- Costa-Jussa, Marta R. et al. (2022). **No Language Left Behind: Scaling Human-Centered Machine Translation**. ArXiv, abs/2207.04672
- Dai, D. et al. (2022). **Knowledge Neurons in Pretrained Transformers**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.
- Davison, J., Feldman, J., and Rush, A. (2019). **Commonsense knowledge mining from pretrained models**. 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)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1109>

Delobelle, P., Tokpo, E. K., Calders, T., & Berendt, B. (2022). **Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models**. In NAACL 2022: the 2022 Conference of the North American chapter of the Association for Computational Linguistics: human language technologies (pp. 1693-1706).

Denaux, R., Gomez-Perez, JM. (2020). **Linked Credibility Reviews for Explainable Misinformation Detection**. In: J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), The Semantic Web – ISWC 2020, Springer International Publishing, Cham, 2020, pp. 147–163.

Denaux, R. and Gomez-Perez, JM. (2019). **Vecsigrafo: Corpus-based Word-Concept Embeddings. Bridging the Statistic-Symbolic Representational Gap in Natural Language Processing**. Semantic Web Journal 10, 5 (2019), 881–908. <https://doi.org/10.3233/SW-190361>

Devlin, J. et al. 2019. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol.1 (Long and Short Papers), pp. 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K. W., & Gupta, R. (2021, March). **Bold: Dataset and metrics for measuring biases in open-ended language generation**. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 862-872).

Dhingra, B., Cole, JR., Eisenschlos, JM., Gillick, D., Eisenstein, J., and Cohen, WW. (2022). **Time-aware language models as temporal knowledge bases**. Transactions of the Association for Computational Linguistics, 10:257–273.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., and Williams, A. (2020). Multidimensional gender bias classification. ArXiv.

Dolci, T., Azzalini, F., & Tanelli, M. (2023). **Improving Gender-Related Fairness in Sentence Encoders: A Semantics-Based Approach**. Data Science and Engineering, 1-19.

Driess, D. et al. (2023). **PaLM-E: An Embodied Multimodal Language Model**. International Conference on Machine Learning.

Fensel, D.A. (2000). **Problem-Solving Methods: Understanding, Description, Development, and Reuse**. Lecture Notes in Computer Science 1791, Springer 2000, ISBN 3-540-67816-6

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Derroncourt, F., ... & Ahmed, N. K. (2023). **Bias and fairness in large language models: A survey**. *arXiv preprint arXiv:2309.00770*.

García-Silva, A., Berrio, C., Gómez-Pérez, JM. (2023). **Textual Entailment for Effective Triple Validation in Object Prediction**. The Semantic Web – ISWC 2023, Springer International Publishing, Cham, 2020, to appear.

Gao L., Biderman S., Black S., Golding L., Hoppe T., Foster C., Phang J., He H., Thite A., Nabeshima N., Presser S., & Leahy C. (2020). **The Pile: An 800GB Dataset of Diverse Text for Language Modeling**. *arXiv preprint arXiv:2101.00027*.

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., & Neubig, G. (2022). **PAL: Program-aided Language Models**. ArXiv, abs/2211.10435.

Gao, L., et al. (2023). **RARR: Researching and Revising What Language Models Say, Using Language Models**. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Gao, S., et al. (2024). **Efficient Tool Use with Chain-of-Abstraction Reasoning**. arXiv:2401.17464

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). **Retrieval-augmented generation for large language models: A survey**. arXiv preprint arXiv:2312.10997.

Geva M., Khashabi D., Segal E., Khot T., Roth D., & Berant J. (2021). **Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies**. *arXiv preprint arXiv:2101.02235*

Gómez-Pérez, JM., Ortega, R. (2023). **E4.1 Análisis y Definición de Dominios de Aplicación y Casos de Uso**. KG4LLM Technical Report.

Gómez-Pérez, JM., García-Silva, A., Leone, R., Albani, M., Fontaine, M., Poncet, C., Summerer, L., Donati, A., Roma, I., Scaglioni, S. (2023). **Artificial Intelligence and Natural Language Processing and**

Understanding in Space: A Methodological Framework and Four ESA Case Studies. Engineering Applications of Artificial Intelligence (to appear).

Gomez-Perez, JM., Denaux, R., Garcia-Silva, A. (2020) **A Practical Guide to Hybrid Natural Language Processing - Combining Neural Models and Knowledge Graphs for NLP.** Springer, Cham. DOI: <https://doi.org/10.1007/978-3-030-44830-1>

Gómez-Pérez, JM., Ortega, R. (2020). **ISAAQ - Mastering Textbook Questions with Pre-trained Transformers and Bottom-Up and Top-Down Attention.** 5469-5479. 10.18653/v1/2020.emnlp-main.441. Empirical Methods to Natural Language Processing (EMNLP) 2020.

Gomez-Perez, Jose Manuel. (2010). **Acquisition and understanding of process knowledge using problem solving methods.** Studies on the Semantic Web, IOS Press, 978-1-60750-600-3 (print) | 978-1-61499-341-4 (online). DOI: <https://doi.org/10.3233/978-1-61499-341-4-i>

Guo, W., & Caliskan, A. (2021, July). **Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases.** In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 122-133).

He, B. et al. (2020). **BERT-MK: Integrating graph contextualized knowledge into pre-trained language models.** In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2281–2290, Online, Nov. 2020. Association for Computational Linguistics.

He, X., Tian, Y., Sun, Y., Chawla, N., Laurent, T., LeCun, Y., Bresson, X., & Hooi, B. (2024). **G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering.** ArXiv, abs/2402.07630.

Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., & Steinhardt J. (2021). **Measuring Massive Multitask Language Understanding.** *arXiv preprint arXiv:2009.03300*.

Hoffmann J. et al. (2022). **Training Compute-Optimal Large Language Models.** *arXiv preprint arXiv:2203.15556*

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q.D., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). **Parameter-Efficient Transfer Learning for NLP.** *International Conference on Machine Learning*.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Yu, J.A., Joulin, A., Riedel, S., & Grave, E. (2022). **Few-shot Learning with Retrieval Augmented Language Models.** ArXiv, abs/2208.03299.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). **Unsupervised dense information retrieval with contrastive learning.** *arXiv preprint arXiv:2112.09118*.

Ji, Z. et al. 2022. **Survey of hallucination in natural language generation.** *ACM Computing Surveys*.

Ji, H., Grishman, R.: **Knowledge Base Population: Successful Approaches and Challenges.** In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1148–1158. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), <https://aclanthology.org/P11-1115>

Kadavath, S., Conerly, T., Askell, A., Henighan, T.J., Drain, D., Perez, E., Schiefer, N., Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T.B., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., & Kaplan, J. (2022). **Language Models (Mostly) Know What They Know.** ArXiv, abs/2207.05221.

Kamoi, R., Goyal, T., Rodriguez, J.D., & Durrett, G. (2023). **WiCE: Real-World Entailment for Claims in Wikipedia.** ArXiv, abs/2303.01432.

Kaneko, M., & Bollegala, D. (2022, June). **Unmasking the mask—evaluating social biases in masked language models.** In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 11, pp. 11954-11962).

Kandpal, N., Deng, H., Roberts, A., Wallace, E., Raffel, C. (2022). **Large Language Models Struggle to Learn Long-Tail Knowledge.**

Kirkpatrick, J. et al. 2017. **Overcoming catastrophic forgetting in neural networks.** Proceedings of the national academy of sciences, 114(13):3521–3526.

- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). **Large language models are zero-shot reasoners**. *Advances in neural information processing systems*, 35, 22199-22213.
- Komeili, M., Shuster, K., and Weston, J. (2022). **Internet-augmented dialogue generation**. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.
- Krystian, W., McCann, B., Xiong, C., and Socher, R. (2020). **Evaluating the Factual Consistency of Abstractive Text Summarization**. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online. Association for Computational Linguistics.
- Kuhn, L., Gal, Y. and Farquhar, S. (2023) **Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation**.
- Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). **Measuring bias in contextualized word representations**. arXiv preprint arXiv:1906.07337.
- Laurençon, H. et al. (2022). **The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset**. In *Advances in Neural Information Processing Systems* (pp. 31809–31826). Curran Associates, Inc.
- Lauscher, A., Majewska, O., Ribeiro, L., Gurevych, I., Rozanov, N., Glavaš, G. (2020). **Common Sense or World Knowledge? Investigating Adapter-Based Knowledge Injection into Pretrained Transformers**. 43-49. 10.18653/v1/2020.deeplio-1.5.
- Lawrence, Peter. (2024). **Text-to-Graph via LLM: pre-training, prompting, or tuning?**. https://medium.com/@peter.lawrence_47665/text-to-graph-via-llm-pre-training-prompting-or-tuning-3233d1165360
- Lawrence, Peter. (2023). **Large Language Model = Knowledge Graph Store? Yes, by Fine-Tuning LLM With KG**. <https://betterprogramming.pub/large-language-model-knowledge-graph-store-yes-by-fine-tuning-llm-with-kg-f88b556959e6>
- Lehmann, J. et al. (2015). **DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia**. *Semantic Web*, 6, 167-195.
- Levine, Y. et al. (2020). **SenseBERT: Driving some sense into BERT**. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 2020.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). **Retrieval-augmented generation for knowledge-intensive nlp tasks**. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
- Li, K., Patel, O., Viégas, F., Pfister, H., & Wattenberg, M. (2023). **Inference-Time Intervention: Eliciting Truthful Answers from a Language Model**. ArXiv, abs/2306.03341.
- Lin et al. (2022). **Few-shot Learning with Multilingual Generative Language Models**. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Lin, S., Hilton, J., and Evans, O. 2022. **TruthfulQA: Measuring How Models Mimic Human Falsehoods**. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Liu, Y. et al. (2019). **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. ArXiv, abs/1907.11692.
- Liu, Y., Fabbri, A., Liu, P., Zhao, Y., Nan, L., Han, R., Han, S., Joty, S., Wu, C.S., Xiong, C., & Radev, D. (2023). **Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4140–4170). Association for Computational Linguistics.
- Logan IV, R. L., Liu, N. F., Peters, M. E., Gardner, M., & Singh, S. (2019). **Barack's wife hillary: Using knowledge-graphs for fact-aware language modeling**. arXiv preprint arXiv:1906.07241.
- Malaviya C, Lee S., Chen S., Sieber E., Yatskar M., & Roth D. (2023). **ExpertQA: Expert-Curated Questions and Attributed Answers**. arXiv preprint arXiv:2309.07852.
- Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. 2023. **When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories**. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Manakul, P., Liusie, A. and Gales, M. 2023. **SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models**. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9004–9017, Singapore. Association for Computational Linguistics.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). **On measuring social biases in sentence encoders**. arXiv preprint arXiv:1903.10561.

Maynez, J., Narayan, S., Bohnet, B. and McDonald, R. (2020). **On Faithfulness and Factuality in Abstractive Summarization**. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919, Online. Association for Computational Linguistics.

Mcdermott, J., 1988. **Preliminary Steps Toward a Taxonomy of Problem-Solving Methods**. Springer US, Boston, MA. pp. 225–256. DOI: https://doi.org/10.1007/978-1-4684-7122-9_8

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). **Locating and Editing Factual Associations in GPT**. In *Advances in Neural Information Processing Systems* (pp. 17359–17372). Curran Associates, Inc.

Mesquita, F., Cannavicchio, M., Schmidek, J., Mirza, P., and Barbosa, D. (2019). **KnowledgeNet: A Benchmark Dataset for Knowledge Base Population**. 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), pages 749–758, Hong Kong, China. Association for Computational Linguistics.

Miller, GA. (1994). **WordNet: A Lexical Database for English**. In Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.

Min S., Shi W., Lewis M., Chen X., Yih W., Hajishirzi H., & Zettlemoyer L. (2023). **Nonparametric Masked Language Modeling**. arXiv preprint arXiv:2212.01349

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.T., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. (2023). **Factscore: Fine-grained atomic evaluation of factual precision in long form text generation**.

Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., Abend, O., Leyton-Brown, K., Shashua, A., & Shoham, Y. (2023). **Generating Benchmarks for Factuality Evaluation of Language Models**. ArXiv, abs/2307.06908.

Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: **Measuring stereotypical bias in pretrained language models**. arXiv preprint arXiv:2004.09456.

Nadeem, M., Bethke, A., and Reddy, S. (2021). **StereoSet: Measuring stereotypical bias in pretrained language models**. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371, Online. Association for Computational Linguistics.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. (2020). **CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models**. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

Névéol, A., Dupont, Y., Bezançon, J., & Fort, K. (2022, May). **French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English**. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8521-8531).

A. Newell, J. C. Shaw, and H. A. Simon. **Report on a general problem solving program**. In IFIP congress, volume 256, page 64. Pittsburgh, PA, 1959.

A. Newell, H. A. Simon, et al. **Human problem solving**. Prentice-Hall, 1972.

Ni J., Qu C., Lu J., Dai Z., Ábrego GH., Ma J., Zhao VY., Luan Y., Hall KB., Chang M., & Yang Y. (2021). **Large Dual Encoders Are Generalizable Retrievers**. arXiv preprint arXiv:2112.07899

Nozza, D., Bianchi, F., & Hovy, D. (2021). **HONEST: Measuring hurtful sentence completion in language models**. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

O'Hern, M.S. & Rindfleisch, A. (2010). **Customer Co-Creation: A Typology and Research Agenda**. In: Review of Marketing Research, vol. 6, p. 84-106.

Opitz, J. (2019). **Argumentative relation classification as plausibility ranking**. In Preliminary Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers, pages 193–202, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright CL, Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J., & Lowe R. (2022). **Training language models to follow instructions with human feedback**. *arXiv preprint arXiv:2203.02155*.

Parisi, A. & Zhao, Y. & Fiedel, N. (2022). **TALM: Tool Augmented Language Models**. 10.48550/arXiv.2205.12255.

Patel, A., Bhattamishra, S., and Goyal, N. (2021). **Are NLP models really able to solve simple math word problems?** In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2080–2094, Online. Association for Computational Linguistics.

Penedo, G. et al. (2023). **The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only**. ArXiv, abs/2306.01116.

Peters, ME., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., and Smith, N. (2019). **Knowledge Enhanced Contextual Word Representations**. 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), pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Petroni, F. et al. (2019). **Language Models as Knowledge Bases**. 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), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Puri, R., & Catanzaro, B. (2019). **Zero-shot Text Classification With Generative Language Models**. ArXiv, abs/1912.10165.

Qian, R., Ross, C., Fernandes, J., Smith, E., Kiela, D., & Williams, A. (2022). **Perturbation augmentation for fairer nlp**. *arXiv preprint arXiv:2205.12586*.

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., ... & Chen, H. (2022). **Reasoning with language model prompting: A survey**. *arXiv preprint arXiv:2212.09597*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). **Language Models are Unsupervised Multitask Learners**.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., & Finn, C. (2023). **Direct Preference Optimization: Your Language Model is Secretly a Reward Model**. ArXiv, abs/2305.18290.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). **Squad: 100,000+ questions for machine comprehension of text**. *arXiv preprint arXiv:1606.05250*.

Rehm, G., & Way, A. (2023). **Strategic Research, Innovation and Implementation Agenda for Digital Language Equality in Europe by 2030**. European Language Equality. Springer Cham. https://doi.org/10.1007/978-3-031-28819-7_45

Rehm G. et al. (2023). **European Language Grid A Language Technology Platform for Multilingual Europe**. Springer Cham. <https://doi.org/10.1007/978-3-031-17258-8>

Sanh, Victor & Webson, Albert & Raffel, Colin & Bach, Stephen & Sutawika, Lintang & Alyafeai, Zaid & Chaffin, Antoine & Stiegler, Arnaud & Scao, Teven & Raja, Arun & Dey, Manan & Bari, M & Xu, Canwen & Thakker, Urmish & Sharma, Shanya & Szczechla, Eliza & Kim, Taewoon & Chhablani, Gunjan & Nayak, Nihal & Rush, Alexander. (2021). **Multitask Prompted Training Enables Zero-Shot Task Generalization**.

Sap, M. et al. (2019). **ATOMIC: an atlas of machine commonsense for if-then reasoning**. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in

Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19). AAAI Press, Article 372, 3027–3035. <https://doi.org/10.1609/aaai.v33i01.33013027>

Santhanam K. et al. (2022). **ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction**. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Scao, TL. et al. (2022). **BLOOM: A 176B-Parameter Open-Access Multilingual Language Model**. ArXiv, abs/2211.05100.

Schick, T., et al. (2023). **Toolformer: Language Models Can Teach Themselves to Use Tools**. ArXiv, abs/2302.04761.

Schick, T., Udapa, S., and Schutze, H. (2021). **Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP**. Transactions of the Association for Computational Linguistics. https://doi.org/10.1162/tacl_a_00434

Schick, T., Schutze, H. (2021). **Generating datasets with pretrained language models**. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.emnlp-main.555>

Schreiber, G. et al. (1994). **CommonKADS: A comprehensive methodology for KBS development**. IEEE Expert. 9. 28-37. 10.1109/64.363263.

Schreiber, G. (2000). **Knowledge engineering and management: the CommonKADS methodology**. MIT press.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). **Proximal policy optimization algorithms**. arXiv preprint arXiv:1707.06347.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). **Retrieval Augmentation Reduces Hallucination in Conversation**. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Speer, R., Chin, J., and Havasi, C. (2017). **ConceptNet 5.5: an open multilingual graph of general knowledge**. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17). AAAI Press, 4444–4451.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2020). **ERNIE 2.0: A continual pre-training framework for language understanding**. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8968–8975. AAAI Press, 2020.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). **FEVER: a Large-scale Dataset for Fact Extraction and VERification**. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Thoppilan, R. et al. (2022). **LaMDA: Language Models for Dialog Applications**. ArXiv, abs/2201.08239.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C.D. (2023). **Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback**. ArXiv, abs/2305.14975.

Tian, K., Mitchell, E., Yao, H., Manning, C.D., & Finn, C. (2023). **Fine-tuning Language Models for Factuality**. ArXiv, abs/2311.08401.

Touileb, S., Øvrelid, L., & Velldal, E. (2022, July). **Occupational biases in Norwegian and multilingual language models**. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP) (pp. 200-211).

Touvron, H. et al. (2023). **Llama 2: Open Foundation and Fine-Tuned Chat Models**. ArXiv, abs/2307.09288.

Vashishtha, A., Ahuja, K., & Sitaram, S. (2023). **On evaluating and mitigating gender biases in multilingual settings**. arXiv preprint arXiv:2307.01503.

Vaswani, A. et al. (2017). **Attention is All you Need**. NIPS.

- Wang, R. et al. (2021). **K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters**. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1405–1418, Online. Association for Computational Linguistics.
- Wang, A., Cho, K., and Lewis, M. (2020). **Asking and Answering Questions to Evaluate the Factual Consistency of Summaries**. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5008–5020, Online. Association for Computational Linguistics.
- Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R.K., & Lim, E. (2023). **Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models**. Annual Meeting of the Association for Computational Linguistics.
- Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A., Arunkumar, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Pal, K., Patel, M., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P., Verma, P., Puri, R., Karia, R., Doshi, S., Sampat, S., Mishra, S., Reddy A, S., Patro, S., Dixit, T., & Shen, X. (2022). **Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks**. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 5085–5109). Association for Computational Linguistics.
- Wang Y., Kordi Y., Mishra S., Liu A., Smith NA., Khashabi D., & Hajishirzi H. (2023). **Self-Instruct: Aligning Language Models with Self-Generated Instructions**. *arXiv preprint arXiv:2212.10560*.
- Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K. W., & Lim, E. P. (2023). **Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models**. *arXiv preprint arXiv:2305.04091*.
- Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., ... & Petrov, S. (2020). **Measuring and reducing gendered correlations in pre-trained models**. *arXiv preprint arXiv:2010.06032*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). **Chain-of-thought prompting elicits reasoning in large language models**. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). **Emergent abilities of large language models**. *arXiv preprint arXiv:2206.07682*.
- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. **The FAIR Guiding Principles for scientific data management and stewardship**. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
- Xiong, W., Du, J, Wang, W., and Stoyanov, V. (2020). **Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model**. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.
- Yao, Y., Huang, S., Zhang, N., Dong, L., Wei, F., & Chen, H. (2022). **Kformer: Knowledge Injection in Transformer Feed-Forward Layers**. *ArXiv*, abs/2201.05742.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). **Tree of thoughts: Deliberate problem solving with large language models**, may 2023. *arXiv preprint arXiv:2305.10601*, 14.
- Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. **Generative Knowledge Graph Construction: A Review**. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1–17, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yu, D. et al. (2022). **KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering**. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4961–4974.
- Zamani, H., Diaz, F., Dehghani, M., Metzler, D., and Bendersky, M. (2022). **Retrieval-Enhanced Machine Learning**. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 2875–2886. <https://doi.org/10.1145/3477495.3531722>
- Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M, and Liu, Q. (2019). **ERNIE: Enhanced language representation with informative entities**. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1441–1451, Florence, Italy, July 2019. Association for Computational Linguistics.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., ... & Sun, L. (2023). **A comprehensive survey on pretrained foundation models: A history from bert to chatgpt**. *arXiv preprint arXiv:2302.09419*.

Apéndice A: Prompts para obtener preguntas sobre entidades y resultados de ejemplo

Entidad de seguros a Pregunta	Write 6 different questions about "{ENTITY}" in the insurance domain that have to be answered with a short paragraph containing factual knowledge. Write only the questions (without the answer) following the format "-- question_1\n-- question_2\n-- question_n"
Entidad de UMLS a Pregunta	Write 6 different questions about "{ENTITY}" in the covid-19 domain that have to be answered with a short paragraph containing factual knowledge. Write only the questions (without the answer) following the format "-- question_1\n-- question_2\n-- question_n"

Tabla 5 Prompts de GPT3.5 para convertir entidades en preguntas.

SEGUROS	
Advertising Injury	What is considered "advertising injury" in the insurance domain?
Blanket Limit	What are the advantages of having a blanket limit in insurance policies?
Certificate of Insurance	What information is typically included on a Certificate of Insurance?
SALUD (COVID-19)	
ACE2 Gene	How does the ACE2 gene play a role in viral entry into cells in COVID-19?
Diabetes	What are the potential complications for individuals with diabetes who also contract COVID-19?
Obesity	What is the relationship between obesity and COVID-19 severity?

Tabla 6 Ejemplos de preguntas generadas a partir de entidades en el dominio de seguros y en el de salud (COVID-19).

Apéndice B: Prompts y ejemplos para, dada una pregunta, generar una respuesta en forma de párrafo

Respuesta a pregunta sobre entidad de seguros	INPUT: Answer the following question in a single paragraph: "What is liability insurance?" OUTPUT: Liability insurance is a type of insurance coverage that provides financial protection to individuals or businesses in the event they are held legally responsible for causing harm or damage to others. It typically covers the costs associated with legal defense, as well as any judgments or settlements that may arise from lawsuits. Liability insurance is crucial for protecting assets and mitigating financial risks, as it helps policyholders manage the financial
---	--

	consequences of third-party claims, such as bodily injury or property damage, arising from their actions or negligence. This type of insurance is common in various fields, including auto insurance, business insurance, and homeowner's insurance, offering a safeguard against the potential financial burdens resulting from legal liabilities. ---- INPUT: Answer the following question in a single paragraph: "What types of liabilities are typically covered by insurance policies?" OUTPUT: Insurance policies generally cover a range of liabilities, including but not limited to property damage, bodily injury, and personal liability. Property and casualty insurance often protect against financial losses resulting from damage to or loss of physical assets, such as homes, cars, or businesses. Liability insurance, on the other hand, typically addresses legal responsibilities arising from bodily injury or property damage caused by the policyholder or their possessions. This can include coverage for medical expenses, legal fees, and settlement costs. Additionally, professional liability insurance may protect individuals or businesses against claims related to errors, negligence, or malpractice in their professional services. The specific types of liabilities covered depend on the type of insurance policy and its terms and conditions. ---- INPUT: Answer the following question in a single paragraph: "How is liability coverage different from other types of insurance?" OUTPUT: Liability coverage distinguishes itself from other types of insurance by primarily focusing on providing financial protection for the policyholder against claims or lawsuits filed by third parties due to injuries or property damage for which the insured is deemed responsible. Unlike property or health insurance, which primarily safeguard the policyholder's personal belongings or well-being, liability coverage steps in to cover legal expenses, court judgments, and settlements arising from incidents where the insured is found at fault. This type of insurance is crucial in shielding individuals or businesses from the potentially significant financial repercussions of legal actions, emphasizing protection against liabilities rather than the direct safeguarding of personal assets or well-being. ---- INPUT: Answer the following question in a single paragraph: "{question}" OUTPUT:
Respuesta a pregunta sobre entidad de UMLS	INPUT: Answer the following question in a single paragraph: "What are mRNA vaccines and how do they work?" OUTPUT: Messenger RNA (mRNA) vaccines are a new type of vaccine to protect against infectious diseases. To trigger an immune response, many vaccines put a weakened or inactivated germ into our bodies. Not mRNA vaccines. Instead, they teach our cells how to make a protein that triggers an immune response inside our bodies. That immune response, which produces antibodies, is what protects

	us from getting infected if the real virus enters our bodies. ---- INPUT: Answer the following question in a single paragraph: "Are COVID-19 vaccines safe for pregnant women and breastfeeding mothers?" OUTPUT: The European Medicines Agency (EMA) and the World Health Organization (WHO) advise pregnant women to be vaccinated, although they may need a different dose from what is used in the general population. The vaccine is not recommended during the first trimester of pregnancy, but can be given at any other time in pregnancy or when breastfeeding. The agencies recommend pregnant women should be informed of the benefits and risks of vaccination before they make their decision. Pregnancy loss has been reported in women who received the Pfizer/BioNTech or Moderna vaccines, and no such reports have been made for those vaccinated with the AstraZeneca/Oxford vaccine. It is unknown how many of these reported pregnancy losses were associated with vaccination. ---- INPUT: Answer the following question in a single paragraph: "How long does immunity from COVID-19 vaccines last?" OUTPUT: The duration of protection conferred by vaccines against COVID-19 is not yet known. Scientists are currently monitoring protection over the medium- and longer-term. This includes the possibility of booster doses to maintain protection, which is being studied. As countries progressively vaccinate their populations, immunity to COVID-19 in general is expected to increase. In addition, the proportion of the population that has some degree of immunity to COVID-19 is expected to increase with time as more people are vaccinated, people develop natural immunity over time and/or recover from infection ---- INPUT: Answer the following question in a single paragraph: "{question}" OUTPUT:
--	---

Tabla 7 Prompt para dada una pregunta, responder en forma de párrafo.

SEGUROS	
What is considered "advertising injury" in the insurance domain?	In the insurance domain, advertising injury refers to any advertising-related damage caused by an insured person, such as libel, slander, or infringement of copyright or trademark. In such instances, liability insurance can provide coverage for financial losses incurred by the insured person, such as legal expenses, settlements, and court judgments. This type of coverage is commonly found in professional liability insurance policies, as businesses are particularly vulnerable to advertising-related risks. Advertising injury can also extend to other forms of promotion, such as sponsorship and product endorsements, which can be costly if something goes wrong. By providing protection against these risks, liability insurance can help businesses manage

	the financial consequences of advertising-related legal claims and avoid bankruptcy or closure.
What are the advantages of having a blanket limit in insurance policies?	A blanket limit in an insurance policy refers to a single limit set for the maximum amount of coverage available to an individual or entity. This limit is often stated in a dollar amount or a percentage of the total value of the assets or liabilities covered. Advantages of having a blanket limit in insurance policies include: * Easy to calculate: The insurance company determines the total value of the covered assets or liabilities and sets a blanket limit equal to that amount, making it easy for the insured to calculate their maximum coverage. * No need for regular updates: Once the blanket limit is set, there is no need to regularly update it as the value of the covered assets or liabilities changes. * Clarity of coverage: The blanket limit provides a clear understanding of the maximum coverage that will be provided in the event of a claim, which can help prevent misunderstandings or disputes. * Cost-effective: Because the blanket limit is based on the total value of the assets or liabilities covered, it can be more cost-effective than setting individual limits for each type of coverage. * Flexibility: The blanket limit can be adjusted or replaced as needed, providing the insured with the flexibility to adapt to changes in their risk profile or insurance needs.
What information is typically included on a Certificate of Insurance?	A Certificate of Insurance (COI) is a document issued by an insurance company to verify the existence and extent of liability insurance coverage for a policyholder. This document is essential for businesses because it provides a snapshot of the policy's key information, such as the policyholder's name, the insurer, and the type of coverage. A COI also outlines the specific coverages and limits of liability, ensuring that the policyholder is aware of their level of protection. Additionally, a COI may be required by contracts or partnerships, and it is an essential tool for ensuring compliance with legal and regulatory requirements. Overall, a COI is a crucial document for any business seeking to mitigate risks and manage potential financial liabilities.
SALUD (COVID-19)	
How does the ACE2 gene play a role in viral entry into cells in COVID-19?	When the COVID-19 virus enters a human cell, it attaches itself to a molecule called ACE2, which acts as an entry point for the virus. This is how the virus can penetrate the body. ACE2 is most commonly expressed in lung tissue, as well as kidney and heart tissue. The COVID-19 virus appears to be more likely to bind with ACE2 in the lungs, making it possible for the virus to replicate there. ACE2 is also expressed on the surfaces of other tissues in the body, including the mouth and nose.
What are the potential complications for	Several studies have shown that people with diabetes have an increased risk of experiencing severe illness from COVID-19, including death, compared with people without diabetes. There is evidence that diabetes can make the symptoms of COVID-19 more severe and that it also increases the likelihood of

individuals with diabetes who also contract COVID-19?	developing long-term complications such as post-intensive care syndrome, diabetic ketoacidosis and stroke. People with diabetes should adhere to strict COVID-19 infection control measures and stay vigilant in monitoring their blood glucose levels.
What is the relationship between obesity and COVID-19 severity?	A growing body of evidence shows that obesity is a risk factor for developing more severe COVID-19 disease, but that the risk varies between age groups. While obesity has been shown to increase the risk of developing more severe disease in all age groups, the link between obesity and risk of death is particularly strong in the elderly. It is important to remember that other health conditions such as diabetes, heart disease or chronic respiratory illness may also increase the risk of severe disease. In addition, people with obesity who contract COVID-19 are also more likely to require hospitalisation, intensive care and mechanical ventilation.

Tabla 8 Ejemplos de respuestas generadas a partir de preguntas en el dominio de los seguros y en el de salud (COVID-19).

Apéndice C: Prompts y ejemplos para extraer hechos atómicos a partir de respuestas.

Prompt utilizado para extraer hechos atómicos a partir de respuestas:

Extract a list with all the atomic facts about "{ENTITY}" extracted from the following paragraph. At all times when a pronoun is used instead of "{ENTITY}", replace the pronoun with "{ENTITY}".

"{INPUT}"

Write only the facts using the format "-- fact\n-- fact\n-- fact".

SEGUROS	
Párrafo	Hechos atómicos
In the insurance domain, advertising injury refers to any advertising-related damage caused by an insured person, such as libel, slander, or infringement of copyright or trademark. In such instances, liability insurance can provide coverage for financial losses incurred by the insured person, such as legal expenses, settlements, and court judgments. This	Advertising Injury refers to any advertising-related damage caused by an insured person, such as libel, slander, or infringement of copyright or trademark.
	Liability insurance can provide coverage for financial losses incurred by the insured person, such as legal expenses, settlements, and court judgments related to Advertising Injury.

type of coverage is commonly found in professional liability insurance policies, as businesses are particularly vulnerable to advertising-related risks. Advertising injury can also extend to other forms of promotion, such as sponsorship and product endorsements, which can be costly if something goes wrong. By providing protection against these risks, liability insurance can help businesses manage the financial consequences of advertising-related legal claims and avoid bankruptcy or closure.	Professional liability insurance policies commonly include coverage for Advertising Injury, as businesses are particularly vulnerable to advertising-related risks.
	Advertising Injury can extend to other forms of promotion, such as sponsorship and product endorsements.
	Liability insurance can help businesses manage the financial consequences of advertising-related legal claims and avoid bankruptcy or closure.
A blanket limit in an insurance policy refers to a single limit set for the maximum amount of coverage available to an individual or entity. This limit is often stated in a dollar amount or a percentage of the total value of the assets or liabilities covered. Advantages of having a blanket limit in insurance policies include: * Easy to calculate: The insurance company determines the total value of the covered assets or liabilities and sets a blanket limit equal to that amount, making it easy for the insured to calculate their maximum coverage. * No need for regular updates: Once the blanket limit is set, there is no need to regularly update it as the value of the covered assets or liabilities changes. * Clarity of coverage: The blanket limit provides a clear understanding of the maximum coverage that will be provided in the event of a claim, which can help prevent misunderstandings or disputes. * Cost-effective: Because the blanket limit is based on the total value of the assets or liabilities covered, it can be more cost-effective than setting individual limits for each type of coverage.	A blanket limit in an insurance policy refers to a single limit set for the maximum amount of coverage available to an individual or entity.
	The limit is often stated in a dollar amount or a percentage of the total value of the assets or liabilities covered.
	Having a blanket limit in insurance policies makes it easy for the insured to calculate their maximum coverage.
	Once the blanket limit is set, there is no need to regularly update it as the value of the covered assets or liabilities changes.
	The blanket limit provides a clear understanding of the maximum coverage that will be provided in the event of a claim.
	Because the blanket limit is based on the total value of the assets or liabilities covered, it can be more cost-effective than setting individual limits for each type of coverage."

* Flexibility: The blanket limit can be adjusted or replaced as needed, providing the insured with the flexibility to adapt to changes in their risk profile or insurance needs.	The blanket limit can be adjusted or replaced as needed, providing the insured with flexibility.
SALUD (COVID-19)	
Párrafo	Hechos atómicos
When the COVID-19 virus enters a human cell, it attaches itself to a molecule called ACE2, which acts as an entry point for the virus. This is how the virus can penetrate the body. ACE2 is most commonly expressed in lung tissue, as well as kidney and heart tissue. The COVID-19 virus appears to be more likely to bind with ACE2 in the lungs, making it possible for the virus to replicate there. ACE2 is also expressed on the surfaces of other tissues in the body, including the mouth and nose.	ACE2 gene is a molecule that acts as an entry point for the COVID-19 virus.
	ACE2 gene is most commonly expressed in lung tissue, kidney tissue, and heart tissue.
	The COVID-19 virus appears to be more likely to bind with ACE2 gene in the lungs.
	ACE2 gene is also expressed on the surfaces of other tissues in the body, including the mouth and nose.
Several studies have shown that people with diabetes have an increased risk of experiencing severe illness from COVID-19, including death, compared with people without diabetes. There is evidence that diabetes can make the symptoms of COVID-19 more severe and that it also increases the likelihood of developing long-term complications such as post-intensive care syndrome, diabetic ketoacidosis and stroke. People with diabetes should adhere to strict COVID-19 infection control measures and stay vigilant in monitoring their blood glucose levels.	People with diabetes have an increased risk of experiencing severe illness from COVID-19, including death.
	Diabetes can make the symptoms of COVID-19 more severe.
	Diabetes increases the likelihood of developing long-term complications such as post-intensive care syndrome, diabetic ketoacidosis, and stroke.
	People with diabetes should adhere to strict COVID-19 infection control measures.
	People with diabetes should stay vigilant in monitoring their blood glucose levels.

Tabla 9 Ejemplos de hechos atómicos extraídos a partir de respuestas en el dominio de seguros y en el de salud (COVID-19).

Apéndice D: Prompts para convertir hechos atómicos en preguntas y resultados de ejemplo

Hecho atómico sobre	I will provide a statement containing one atomic fact about the insurance concept "Equipment Breakdown Insurance". Please rephrase the following statement into a specific question testing knowledge of the key fact in the
---------------------	--

seguros a pregunta	statement. For example: Statement: Motor failures are included in the coverage, such as engine, transmission, or alternator issues. Question: Equipment breakdown insurance usually covers motor failures such as what? Statement: Equipment Breakdown Insurance (EBI) covers equipment failure or mechanical breakdown due to internal failure or mechanical malfunction. Question: Equipment Breakdown Insurance (EBI) covers equipment failure or mechanical breakdown due to what? Statement: Specific endorsements or riders may be required to be added to the policy for equipment breakdown insurance coverage. Question: What may be required to be added to the policy for equipment breakdown insurance coverage? I will provide a statement containing one atomic fact about the insurance concept "Intentional Acts". Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example: Statement: This exclusion is common in both life and health insurance policies, as well as in other forms of insurance. Question: Are intentional acts a common exclusion in life and health insurance policies? Statement: Intentional acts, such as fraud or forgery, are typically not covered by insurance policies. Question: Are intentional acts, such as fraud or forgey typically covered by insurance policies? I will provide a statement containing one atomic fact about the insurance concept "Theft, Disappearance and Destruction". Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example: Statement: Theft coverage may include coverage for the cost of replacing stolen property or the value of any damage caused to the item Question: Theft coverage may include coverage for the cost of what? Statement: Theft, Disappearance and Destruction can be excluded in insurance policies for intentional damage caused by the insured. Question: Theft, Disappearance and Destruction can be excluded in insurance policies for intentional damage caused by whom? I will provide a statement containing one atomic fact about the insurance concept {ENTITY}. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example: Statement: {STATEMENT}
-----------------------------------	---

	Question:
Hecho atómico sobre salud a pregunta	I will provide a statement containing one atomic fact about the medical concept "Vaccines" in the context of COVID-19. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example: Statement: COVID-19 vaccines can help people prevent mild and moderate illness from COVID-19 and greatly reduce the risk of severe illness that may lead to hospitalization and death. Question: How effective are COVID-19 vaccines in preventing the infection of the virus? Statement: Vaccination against COVID-19 by Pfizer, Moderna, and AstraZeneca requires a 21-day gap between doses. Question: What is the recommended spacing between doses of the COVID-19 vaccines? I will provide a statement containing one atomic fact about the medical concept "ACE2 gene" in the context of COVID-19. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example: Statement: ACE2 gene may allow entry of SARS-CoV-2 into the cells of the airway epithelium. Question: How does the ACE2 gene play a role in viral entry into cells in COVID-19? Statement: Having a high level of ACE2 in the airways is one of the risk factors for severe COVID-19 disease. Question: Which is the relation between ACE2 and the level of severity of a COVID-19 infection? I will provide a statement containing one atomic fact about the medical concept "Mental Depression" in the context of COVID-19. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example: Statement: The global prevalence of depression and anxiety symptoms increased during the pandemic. Question: How does the prevalence of mental depression changed among the general population during the COVID-19 pandemic? Statement: The disruptions and associated economic impacts are likely to increase the risk of developing depression in some people. Question: What are some risk factors that can contribute to the development of mental depression in individuals during the COVID19 crisis. I will provide a statement containing one atomic fact about the medical concept {ENTITY}. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example:

	Statement: {STATEMENT}
	Question:

Tabla 10 Prompt de GPT 3.5 para convertir hechos atómicos en preguntas.

SEGUROS	
Hecho atómico	Pregunta
Advertising Injury refers to any advertising-related damage caused by an insured person, such as libel, slander, or infringement of copyright or trademark.	What does Advertising Injury refer to in insurance terms?
Liability insurance can provide coverage for financial losses incurred by the insured person, such as legal expenses, settlements, and court judgments related to Advertising Injury.	What types of financial losses can liability insurance cover related to Advertising Injury?
Professional liability insurance policies commonly include coverage for Advertising Injury, as businesses are particularly vulnerable to advertising-related risks.	Is Advertising Injury commonly included in professional liability insurance policies?
SALUD (COVID-19)	
Hecho atómico	Pregunta
ACE2 gene is also expressed on the surfaces of other tissues in the body, including the mouth and nose.	Where else in the body is the ACE2 gene expressed, aside from the airway epithelium?
People with diabetes have an increased risk of experiencing severe illness from COVID-19, including death.	How does having diabetes impact the severity of illness from COVID-19?
People with diabetes should adhere to strict COVID-19 infection control measures.	What precautionary measures should individuals with diabetes follow to reduce the risk of contracting COVID-19?

Tabla 11 Ejemplos de preguntas generadas a partir de hechos atómicos el dominio de seguros.