E5.2 DISEÑO DEL DEMOSTRADOR

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

Resumen

Este entregable de diseño recoge el análisis de requisitos y casos de uso de un demostrador que permita visualizar los resultados de la experimentación con los métodos de inyección de conocimiento en LLM desarrollados en el proyecto, tanto de manera general como particularmente en el ámbito de su aplicación a los casos de uso y dominios verticales. El demostrador permitirá comparar los resultados de cada LLM sobre el marco de evaluación con los resultados de otros modelos, estableciendo así una tabla de clasificación (del inglés, leaderboard). Junto al diseño de la interfaz de usuario del demostrador también presentaremos el flujo de trabajo para realizar esta evaluación, que consistirá en la publicación de un ejecutable que permita la evaluación de forma automática de cada LLM, así como el envío de sus resultados a la plataforma que los publicará.

José Manuel Gómez Pérez
Raul Ortega

23 de septiembre de 2024
Expert.ai Language Technology Research Lab

Calle Henri Dunant, 17, Planta 0, 28036, Madrid
CIF: B-66425513, Inscrita en el Registro Mercantil de Madrid, en el Tomo 44.538, Folio 74, Hoja Número M-784613, Inscripción 1ª.

www.expert.ai

Historia de revisions

Revision	Date	Description	Author (Organisation)
0.1	23/09/2024	Tabla de contenidos y estructura básica	Raúl Ortega (expert.ai)
0.2	26/09/2024	Versión completa sin revisar	Raúl Ortega (expert.ai)
1.0	01/10/2024	Versión completada verificada	Raúl Ortega (expert.ai) José Manuel Gómez-Pérez (expert.ai)

Tabla de contenidos

Tabla de figuras

1 Introducción

Este entregable presenta el diseño de un demostrador que permita comparar el desempeño en términos de factualidad de diferentes modelos. A lo largo del documento se desarrollará el contenido de las dos componentes principales del demostrador: el interfaz de usuario y el marco de evaluación. La primera consistirá en una aplicación web con la que los usuarios puedan interactuar para comparar modelos o ver en detalle los resultados en cada uno de los dominios verticales e idiomas para los que ha sido evaluado. La segunda componente consiste en una serie de servicios que permitirán a los usuarios tanto evaluar de forma autónoma sus modelos como enviar los resultados obtenidos para que formen parte del demostrador. Por último, daremos unas cuantas notas acerca del trabajo futuro.

2 Interfaz de usuario

El objetivo principal de la interfaz de usuario del demostrador es exponer de forma clara e intuitiva todos los posibles aspectos de la evaluación de la factualidad de modelos, incluyéndolos en una misma vista en forma de clasificación, de manera que se puedan comparar entre sí y comprobar en detalle cada uno de sus resultados. A lo largo de este apartado, iremos desgranando el funcionamiento de la interfaz de usuario del demostrador, cómo interactuar con ella y qué podemos esperar cuando se lleve a cabo la evaluación real de factualidad en modelos de lenguaje.

La interfaz ha sido diseñada como una aplicación web, y una vez esté habilitada, será accesible a través de la URL https://labdemos.expertcustomers.ai/kg4llm/leaderboard. Contará con una tabla de clasificación con la que interactuar, que nos servirá para desplegar una serie de gráficos y datos relacionados con cada uno de los modelos evaluados. También será posible filtrar los resultados mostrados en la clasificación en base a las diferentes dimensiones en las que han sido evaluados: dominio e idioma.

Figura 1. Primera versión de la interfaz de usuario del demostrador.

En la Figura 1 podemos ver una primera versión de la interfaz de usuario del demostrador. En ella se distinguen tres paneles que proveen diferentes tipos de datos al usuario: tabla de clasificación o leaderboard en la sección izquierda de la interfaz, gráficos comparativos en la sección superior derecha y tabla detallada de resultados en la sección inferior derecha. Cada uno de estos apartados nos proporciona una perspectiva diferente de la evaluación, y su interacción y forma de uso será detallada a continuación.

2.1 Tabla de clasificación

Se trata del elemento principal de la interfaz. Tal y como muestra la Figura 2, su función es la de mostrar el listado de modelos evaluados siguiendo los criterios seleccionados por el usuario, situando aquellos con mayor puntuación en la parte superior y los de menor en la inferior. Esta puntuación ha sido definida como IMF (índice medio de factualidad), y su obtención y funcionamiento se explicarán en más detalle en la sección 3 de este mismo entregable.

Leaderboard + Filtros

POS	MODELO	IMF
1	gnu_terry	56.87
2	super_llm_epoch_42	55.99
3	large_r2d2	53.52
4	avengers_assemble	52.52
5	doom_slayer_llm	51.31
6	rogue_ai	51.21
7	loser_llm	50.54
8	not_skynet_at_all	48.01
9	noclip_llm	45.60
10	depressing_marvin	39.31

Figura 2. Tabla de clasificación o leaderboard.

El usuario puede personalizar y filtrar los resultados mostrados por esta tabla de clasificación siguiendo las dos dimensiones en las que han sido evaluados: dominio e idioma. De este modo,

Commented [JG1]: Quizá en lugar de "PUNTUACIÓN" habría que poner algo más específico, como "factualidad media" o un acrónimo tipo IMF (Indice Medio de Factualidad)

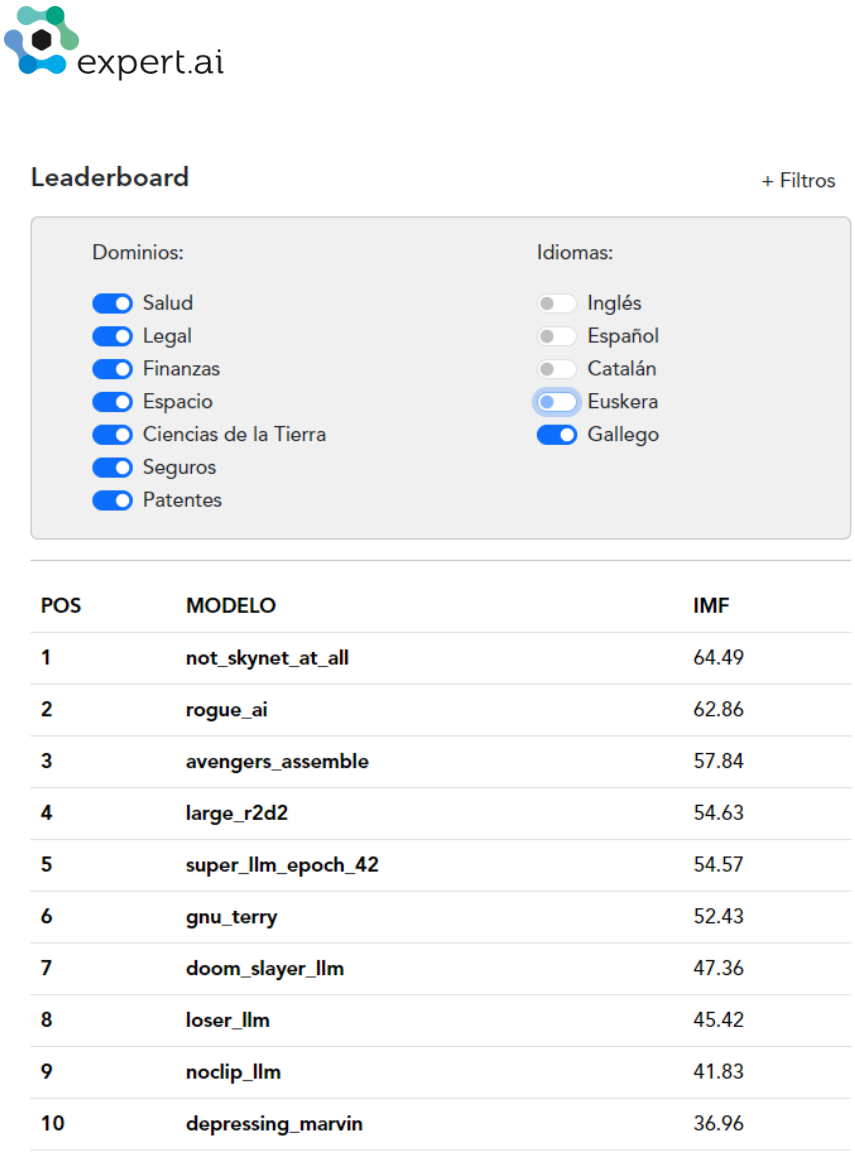
Commented [RO2R1]: Listo, ya he modificado las capturas, y he incluido el término Índice Medio de Factualidad a las referencias en el texto. También he añadido el párrafo del "qué estamos evaluando" en lugar del lorem ipsum

al pulsar en el botón “+ Filtros”, se desplegará una serie de interruptores que permitirán al usuario elegir bajo qué criterios le gustaría clasificar los modelos de la tabla de clasificación, tal y como muestra en la Figura 3.

POS	MODELO	IMF
1	gnu_terry	56.87
2	super_llm_epoch_42	55.99
3	large_r2d2	53.52
4	avengers_assemble	52.52
5	doom_slayer_llm	51.31
6	rogue_ai	51.21
7	loser_llm	50.54
8	not_skynet_at_all	48.01
9	noclip_llm	45.60
10	depressing_marvin	39.31

Figura 3. Tabla de clasificación con los filtros desplegados.

Por defecto, la clasificación se realiza extrayendo la métrica de factualidad de media entre todos los idiomas y dominios evaluados. Al desactivar cualquiera de los interruptores, se anulará la medición en ese dominio o idioma para todos los modelos, y se reordenará de forma automática la tabla de clasificación siguiendo este nuevo criterio, asignando a cada uno de ellos el valor medio de factualidad filtrado por los dominios e idiomas aún activos. Siguiendo este procedimiento, se podría, por ejemplo, obtener la clasificación de los modelos en un idioma para cualquier dominio tan solo desactivando el resto de las lenguas en el desplegable de filtros, tal y como muestra la Figura 4. Este filtrado se puede llevar a cabo con cualquier combinación de idiomas o dominios que el usuario requiera, siempre que al menos se seleccione un idioma y un dominio.



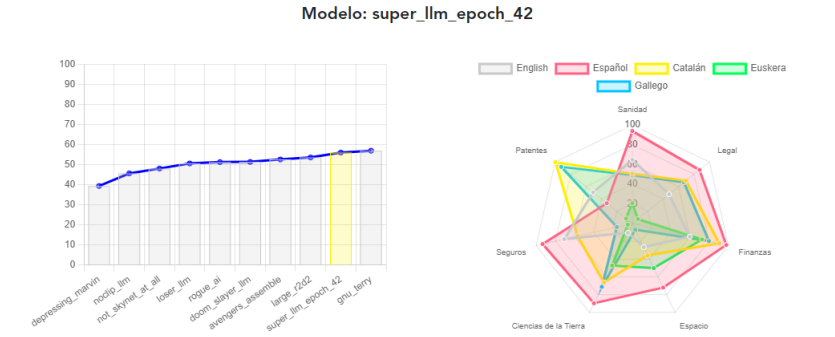
POS	MODELO	IMF
1	not_skynet_at_all	64.49
2	rogue_ai	62.86
3	avengers_assemble	57.84
4	large_r2d2	54.63
5	super_llm_epoch_42	54.57
6	gnu_terry	52.43
7	doom_slayer_llm	47.36
8	loser_llm	45.42
9	noclip_llm	41.83
10	depressing_marvin	36.96

Figura 4. Tabla de clasificación en gallego para cualquier dominio.

A su vez, el usuario podrá interactuar con cada uno de los modelos de la tabla, desplegando información asociada al mismo en los gráficos informativos y en la tabla detallada de resultados.

2.2 Gráficos informativos y tabla detallada de resultados

Al interactuar con cualquiera de los modelos de la tabla de clasificación, se desplegará a su derecha una serie de paneles con información relacionada para dicho modelo, tal y como muestra la Figura 5.



	Inglés	Español	Catalán	Euskera	Gallego	Avg
Salud	64.76	93.81	50.47	20.63	49.09	55.75
Legal	47.61	87.20	69.61	7.26	67.04	55.74
Finanzas	59.77	97.72	90.37	72.49	79.61	79.99
Espacio	26.61	72.06	35.86	50.11	6.90	38.31
Ciencias de la Tierra	10.35	89.71	66.33	47.29	71.20	56.98
Seguros	70.76	93.31	56.95	4.67	16.06	48.35
Patentes	50.83	33.08	99.80	8.13	92.12	56.79
Avg	47.24	80.99	67.06	30.08	54.57	**55.99**

Figura 5. Gráficos informativos y tabla detallada de resultados para un modelo ficticio.

Estos paneles incluyen la gráfica comparativa de evaluación, el gráfico radar bidimensional y la tabla detallada de resultados.

2.2.1 Gráfica comparativa de evaluación

Esta gráfica muestra la posición del modelo con respecto al resto de los evaluados. Se trata de la representación visual de lo mostrado en la tabla de clasificación. Por tanto, cualquier cambio en los filtros aplicados a la tabla también se verán reflejados en esta gráfica, tanto en el orden como en el valor de los datos.

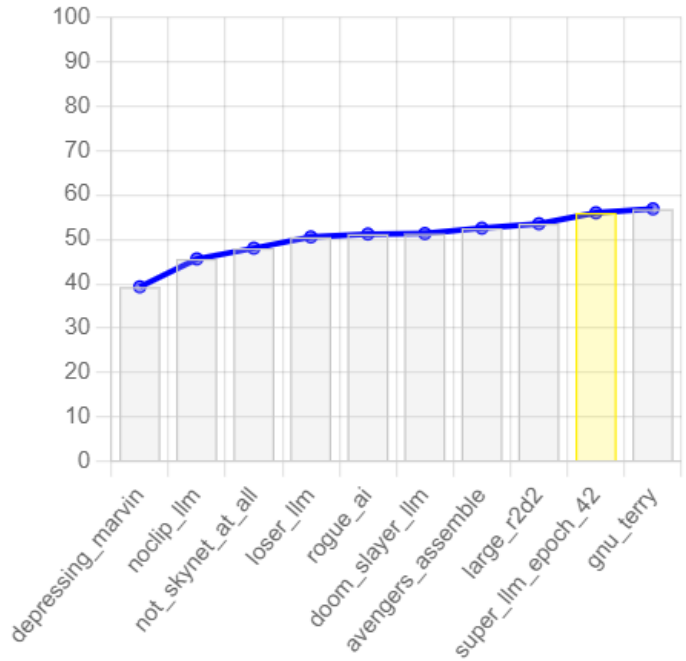


Figura 6. Gráfica comparativa de evaluación centrada en el modelo ficticio "super_llm_epoch_42"

Como muestra la Figura 6, se trata de un gráfico de barras junto a unos puntos y líneas que refuerzan visualmente la clasificación de los modelos. El eje horizontal representa los modelos evaluados, mientras que el eje vertical representa el resultado de la evaluación para los criterios aplicados en el desplegable de filtros. La barra de color amarillo destacará la posición del modelo que se pretende comparar con el resto.

2.2.2 Gráfico radar bidimensional

Este gráfico trata de presentar de manera sencilla toda la complejidad de los resultados de evaluación para el modelo que se haya seleccionado en la tabla de clasificación. Esto implica mostrar el desempeño del modelo tanto en la dimensión del idioma como en el del dominio en el que haya sido evaluado. Tal y como muestra la Figura 7, el gráfico radar tiene siete vértices correspondientes a cada uno de estos dominios en los que se estudia la factualidad: Sanidad, Legal, Finanzas, Espacio, Ciencias de la Tierra, Seguros y Patentes. Al mismo tiempo, superpone cinco capas con cinco colores, representando cada uno de ellos el idioma evaluado. De este modo, la gráfica volcará los valores de evaluación para cada dominio e idioma, acercando a los vértices aquellos resultados que se aproximen al máximo valor de evaluación (100).

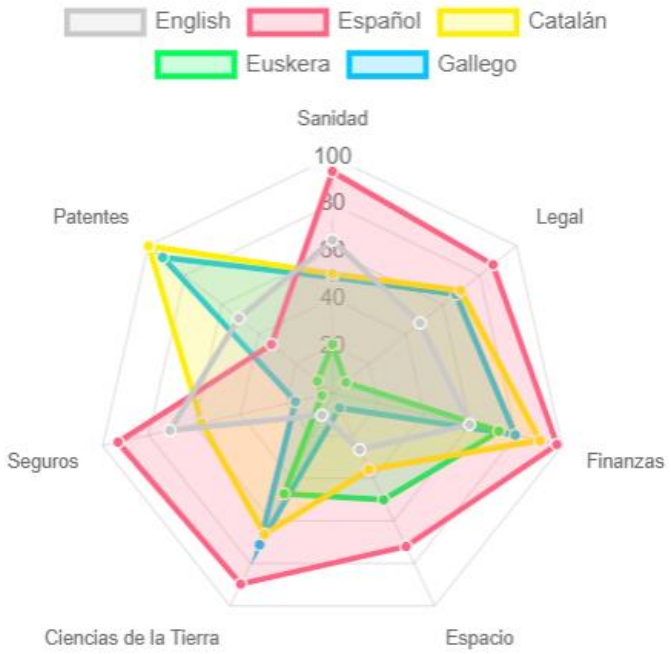


Figura 7. Gráfico radar bidimensional para un modelo ficticio.

Además, y para facilitar la visualización de los datos, el gráfico permite eliminar de la vista cualquiera de los idiomas, tan solo pulsando su texto en la leyenda (Figura 8).

2.2.3 Tabla detallada de resultados

Esta tabla muestra los resultados del modelo seleccionado en la tabla de clasificación para la evaluación de la factualidad en cada dominio e idioma. Su contenido es el que podemos ver representado visualmente en el gráfico radar bidimensional. Además, como se puede ver en la Figura 9, la tabla muestra los valores medios para cada idioma y dominio, y el valor medio total de la evaluación.

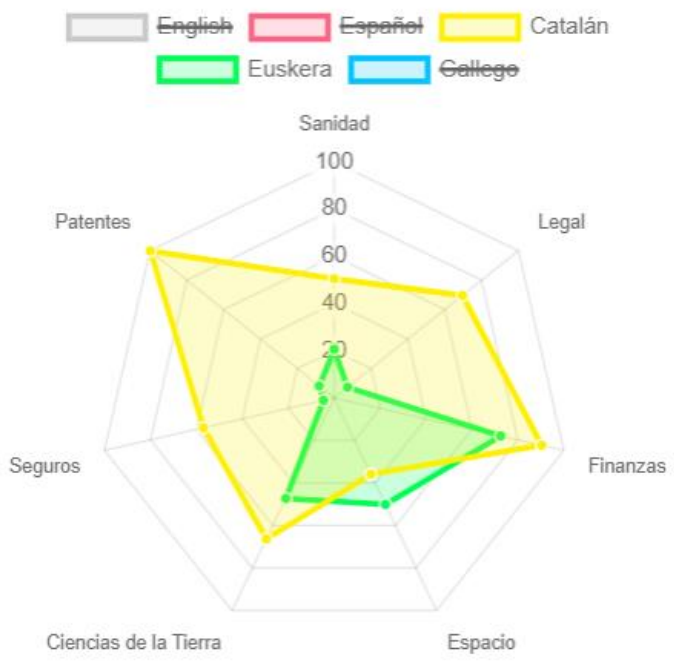


Figura 8. Gráfico radar bidimensional de un modelo ficticio para los idiomas catalán y euskera.

Además, en esta tabla quedarán sobresaltados aquellos resultados que formen parte de los criterios aplicados en el desplegable de filtros, tal y como muestra la Figura 10.

	Inglés	Español	Catalán	Euskera	Gallego	Avg
Salud	64.76	93.81	50.47	20.63	49.09	55.75
Legal	47.61	87.20	69.61	7.26	67.04	55.74
Finanzas	59.77	97.72	90.37	72.49	79.61	79.99
Espacio	26.61	72.06	35.86	50.11	6.90	38.31
Ciencias de la Tierra	10.35	89.71	66.33	47.29	71.20	56.98
Seguros	70.76	93.31	56.95	4.67	16.06	48.35
Patentes	50.83	33.08	99.80	8.13	92.12	56.79
Avg	47.24	80.99	67.06	30.08	54.57	**55.99**

Figura 9. Tabla de resultados detallados de evaluación para los diferentes dominios e idiomas.

	Inglés	Español	Catalán	Euskera	Gallego	Avg
Salud	92.31	90.56	47.58	70.40	37.16	67.60
Legal	85.02	16.05	47.15	56.78	4.31	41.86
Finanzas	29.69	93.21	11.84	75.88	66.26	55.38
Espacio	92.74	48.55	51.34	26.91	51.96	54.30
Ciencias de la Tierra	75.82	32.90	80.79	32.34	58.92	56.15
Seguros	54.08	58.15	89.68	33.55	57.75	58.64
Patentes	48.60	54.49	85.84	41.24	90.63	64.16
Avg	68.32	56.27	59.17	48.16	52.43	**56.87**

Figura 10. Tabla de resultados detallados de evaluación para los diferentes dominios e idiomas de un modelo ficticio. Se sobresalta aquellos resultados para todos los dominios y el idioma gallego.

3 Marco de evaluación

Como hemos mencionado en el anterior apartado, la evaluación que formará parte del demostrador estará basada en la puntuación de factualidad o IMF (índice medio de factualidad) para cada uno de los idiomas y dominios del estudio. Esto significa que cualquier modelo que se pretenda incluir en la tabla de clasificación deberá pasar por un proceso de evaluación, tal y como explicamos en el entregable E5.1 (Gómez-Pérez et al., 2024) de este proyecto. En esencia, los modelos evaluados obtendrán una puntuación de factualidad asociada para cada par dominio-idioma, que irá de 0 a 100, y esos valores serán los que asociaremos al modelo en cuestión en la tabla detallada del demostrador.

A lo largo del proyecto, estas evaluaciones serán llevadas a cabo tanto de manera interna como de manera externa. Esto significa que una serie de modelos base van a ser evaluados por parte del equipo de trabajo del proyecto siguiendo este criterio, al mismo tiempo que se habilitará un sistema de evaluación y publicación automática de resultados, para aquellos que deseen comprobar la factualidad de sus modelos, ejecutando la evaluación en sus propios sistemas. Este último proceso es el que vamos a desarrollar a lo largo de este apartado, explicando tanto sus requisitos previos como los pasos necesarios para su instalación y ejecución. Por último, también daremos más detalles acerca de cómo se realizará el proceso de publicación los resultados en la tabla de clasificación del demostrador.

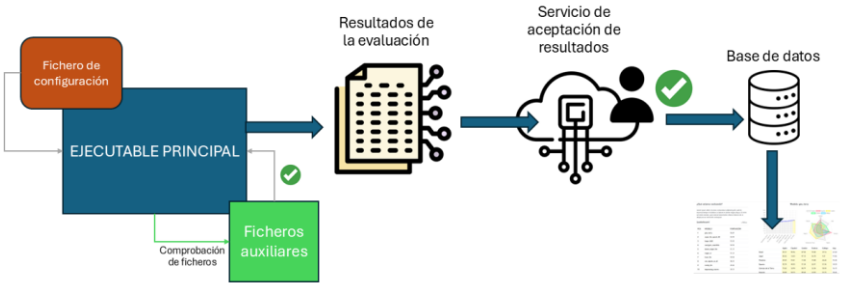


Figura 11. Flujo de trabajo del marco de evaluación del demostrador.

El marco de evaluación que vamos a desplegar para poder evaluar de forma autónoma la factualidad de los modelos consistirá esencialmente en tres partes bien diferenciadas, tal y como muestra la Figura 11:

- **Fichero de configuración.** Se trata de un fichero de texto en formato .ini que incluirá todas las posibles variables que permiten configurar la ejecución de la evaluación. Esto incluye tanto la información relativa al modelo que se pretende evaluar (tipo de arquitectura, dirección del archivo de pesos, hiperparámetros, etc.), como el tipo de evaluación a realizar (completa, en base a ciertos dominios y/o idiomas, etc.) o si se desea publicar como parte del demostrador o no. En el caso de que se desee publicar los resultados, también se tendrá que añadir una serie de metadatos que permitan identificar al autor de la evaluación. La correcta configuración de archivo será requisito indispensable para el correcto funcionamiento de la evaluación, y será el único que será necesario modificar para lanzar el ejecutable principal.

- **Programas auxiliares.** Se trata en una serie de ficheros necesarios para el lanzamiento del ejecutable principal. Su modificación o eliminación total o parcial invalidará cualquier tipo de evaluación que se desee publicar en el demostrador.
- **Ejecutable principal.** Tal y como se puede observar en la Figura 11, el ejecutable principal es el encargado de cargar el fichero de configuración, comprobar el estado de los programas auxiliares, lanzar la evaluación y enviar los resultados al servidor que los almacena para su posterior publicación en la tabla de clasificación del demostrador. Todo este proceso generará unos archivos con los resultados de la evaluación que pueden ser enviados para su publicación en el demostrador. Su funcionamiento requerirá tanto de una correcta configuración del fichero de configuración, como del cumplimiento de los requisitos que describiremos en el próximo apartado.

3.1 Instalación y requisitos previos

Para facilitar el uso de este marco de evaluación, se valorará el uso de herramientas de automatización de instalación mediante contenedores, como Docker. En posteriores entregables se entrará más en detalle acerca de los requisitos software necesarios para la instalación del marco. Como requisitos hardware, se recomienda tener acceso a tarjetas gráficas de al menos 40GB de memoria VRAM, que permitan utilizar en inferencia LLMs de tamaño medio, además de tener actualizados sus controladores.

3.2 Envío de resultados y agregado al demostrador

Como indica la Figura 11, si la configuración del marco ha sido realizada de forma satisfactoria, los resultados de la evaluación serán enviados de forma automática a un servicio web junto a una traza que acredite su autenticidad. Estos resultados se almacenarán de forma temporal en uno de los servidores del proyecto, a la espera de la aprobación manual por parte de los miembros del equipo. Una vez aprobados, los resultados se almacenarán en una base de datos a la que se accederá desde el interfaz de usuario de forma dinámica, y se podrán consultar y comparar con los de los demás modelos en el demostrador.

4 Conclusiones y trabajo futuro

En este entregable hemos presentado el diseño del interfaz de usuario y el marco de evaluación de un demostrador que compare modelos del lenguaje en términos de factualidad. Hemos mostrado el aspecto de su aplicación web, y hemos explicado sus diferentes componentes y cómo interactuar con ellos. También hemos dado detalles acerca del conjunto de ficheros y programas necesarios para realizar esta evaluación de forma autónoma. En posteriores entregables daremos más detalles acerca de este último punto, en referencia a los pasos necesarios para su instalación y puesta en marcha.

Referencias

Gómez-Pérez, JM., Berrio C., Ortega, R. (2024). **E5.1 Marco de Evaluación y Recursos Asociados**. KG4LLM Technical Report.