

E3.1V1 MÉTODOS DE USO DE HERRAMIENTAS EXTERNAS POR LLM

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

Resumen

Este entregable consiste en la primera versión del software resultante de la implementación de los métodos desarrollados para el uso de herramientas externas por parte de LLM (T3.1 y T3.2) y su extensión de forma que también tengan en cuenta las propiedades no funcionales de las herramientas (T3.3). Esta iteración se centra en los resultados preliminares del desarrollo de un método para habilitar llamadas externas mediante el uso de estrategias de prompting. Este enfoque elimina la necesidad de generar conjuntos de datos masivos y de utilizar grandes recursos para el entrenamiento. La estrategia ha sido evaluada en el dominio médico y específicamente en la generación de diagnósticos mediante un agente basado en un LLM. El modelo ha sido guiado mediante prompts para resolver tareas utilizando métodos de resolución de problemas que prevén el uso de llamadas externas a través de funciones de transferencia. La evaluación desarrollada demuestra que la estrategia de prompting implementada puede inducir la capacidad de autoreflexión en un modelo, permitiéndole comprender, solicitar y obtener de manera autónoma la información necesaria a través del uso de herramientas externas. Este proceso es esencial para llegar a la resolución del diagnóstico médico.

José Manuel Gómez Pérez
Flavio Merenda

31 de Julio de 2024
Expert.ai Language Technology Research Lab

Expert.ai. Calle Henri Dunant, 17, Escalera Izquierda, 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

Revision History

Revision	Date	Description	Author (Organisation)
0.1	10/07/2024	Tabla de contenidos y estructura básica	Expert.ai
0.2	26/07/2024	Primera versión completa	Expert.ai
1.0	05/08/2024	Revisión del document	Expert.ai

Tabla de contenidos

1	Introducción	4
2	Enfoque	4
3	Habilitación de herramientas externas mediante prompting.....	5
4	Evaluación	8
5	Repositorio	14
6	Conclusiones y trabajo futuro.....	14
	References	16

1 Introducción

Este entregable forma parte del paquete de trabajo PT3, cuyo objetivo es desarrollar los métodos necesarios para habilitar a los LLMs con el uso de llamadas a funciones externas, como APIs de motores de búsqueda, calculadoras o calendarios. Este documento presenta la primera versión del software resultante de la implementación de los métodos desarrollados para el uso de herramientas externas por parte de los LLMs en las tareas T3.1 y T3.2, así como su extensión para considerar también las propiedades no funcionales de dichas herramientas en la tarea T3.3. Este proceso de desarrollo permitirá mitigar desafíos identificados en E1.1, como alucinaciones y sesgos, dotando a estos modelos de herramientas que puedan compensar sus limitaciones en capacidades de razonamiento y en el uso de sus memorias paramétricas.

El resto de este entregable se estructura de la siguiente manera. En la sección 2 se analiza la selección de una solución específica en el marco de este proyecto para permitir que los LLMs utilicen herramientas externas, considerando tanto las estrategias existentes para habilitar su uso como los recursos necesarios para su implementación. La sección 3 detalla el método elegido, incluyendo sus especificaciones, implementación y aplicación práctica. Los resultados de este proceso se presentan en las secciones 4 y 5, donde se informa sobre el rendimiento de esta metodología en un caso de uso específico del proyecto y se comparte el repositorio que contiene el código que implementa esta solución. Finalmente, la sección 6 expone las conclusiones y propuestas para el futuro desarrollo de metodologías para el uso de herramientas externas.

2 Enfoque

Como se describe en E1.1, los LLMs exhiben habilidades significativas gracias a su preentrenamiento sobre conjuntos de datos masivos, lo que les proporciona capacidades emergentes de razonamiento. Por su arquitectura, estos modelos pueden reproducir sesgos y errores, especialmente en tareas que requieren capacidades de razonamiento simbólico más complejas y usar información específica o no disponible para el propio modelo. Diversos métodos presentados en E1.1, como PAL (Gao et al., 2022), TALM (Parisi et al., 2022) y Toolformer (Schick et al., 2023), han explorado la posibilidad de enriquecer estos modelos mediante el uso de ayudas externas a través de llamadas a herramientas que puedan soportar habilidades de razonamiento complejo, como operaciones matemáticas o consultas a internet para obtener información actualizada o a bases de datos para información más factual. Este enfoque ha demostrado mejorar significativamente los resultados de estos modelos, reduciendo los sesgos asociados con su entrenamiento y el conocido problema de las alucinaciones, donde un modelo, aunque no sepa la respuesta, produce una solución que tal vez no sea factual.

Estas estrategias implican la incorporación de pasos adicionales al clásico flujo de trabajo end-to-end de los LLMs, que generalmente prevén una interacción simple de entrada y salida. La inclusión de estos pasos adicionales permite descomponer una tarea en subtareas más sencillas que pueden resolverse con ayuda de herramientas externas, creando lo que se denominan agentes y sus flujos de trabajo. Cuando un agente recibe una entrada, debe decidir cuándo y cómo utilizar información y herramientas adicionales. En el estado del arte, se han experimentado diferentes estrategias para habilitar el uso de herramientas externas, destacándose principalmente las estrategias de fine-tuning, como Toolformer (Schick et al., 2023), y aquellas que inducen este comportamiento mediante few-shot self-reflection prompting, como ReAct (Yao et al., 2023). En la primera estrategia, se genera un conjunto de datos de manera semiautomática para instruir al modelo sobre cuándo puede ser útil realizar una llamada externa, qué herramienta usar para dicha llamada y cómo estructurarla. Una vez generado este conjunto de datos, se utiliza para entrenar el modelo sobre casos de uso de dominio general. En la segunda estrategia, durante el mismo proceso de inferencia, se muestra al modelo, con un par de ejemplos insertados en el mismo prompt, cómo descomponer una tarea compleja en subtareas, generando pasos intermedios y acciones relacionadas para resolver la tarea con la ayuda de llamadas externas. Inicialmente, este método no requiere fine-tuning, pero los autores

de los experimentos demuestran cómo esta metodología puede beneficiarse adicionalmente de pequeños entrenamientos específicos.

En el marco del proyecto se ha optado por experimentar con estrategias que no requieren entrenamientos adicionales, intentando extraer capacidades de razonamiento directamente del modelo de lenguaje mediante estrategias de prompting e induciendo al modelo a solicitar la información adicional necesaria para continuar con el proceso. En la próxima sección, se describe este proceso de habilitación de llamadas externas, desarrollado en un contexto de resolución de tareas extremadamente complejas en el campo del diagnóstico médico, que habitualmente necesitan integrar información generada por herramientas externas como pruebas de imagen, ultrasonidos y análisis entre otras, para poder continuar con el proceso de razonamiento y producir un diagnóstico y un plan de tratamiento médico.

3 Habilidad de herramientas externas mediante prompting

Como se mencionó en la sección anterior, en el marco del proyecto se ha decidido experimentar con la habilitación de llamadas externas mediante prompting principalmente para superar dos limitaciones. Por un lado, la necesidad de crear conjuntos de datos de entrenamiento suficientemente grandes y representativos que consideren todas las posibles llamadas externas necesarias para todos los casos de uso y dominios enumerados en el entregable E4.1. Por otro lado, los recursos computacionales requerido para los entrenamientos necesarios para adaptar estos modelos a las tipologías de llamadas, que en trabajos previos como Toolformer (Schick et al., 2023) han requerido hasta 25K ejemplos por API externa y el uso simultáneo de 8 GPUs NVIDIA A100 40GB. Habilitar el uso de herramientas externas mediante prompting, por el contrario, ofrece más flexibilidad al no requerir la generación de datos de entrenamiento específicos, lo que puede ser limitante si no se considera de antemano toda la diversidad de llamadas posibles, y permite evitar procesos de entrenamiento masivos.

El uso de llamadas externas rompe el flujo clásico de modelos probabilísticos end-to-end, abriendo la posibilidad de crear flujos de trabajo más complejos que permiten la descomposición de tareas complejas en subtareas. Esto permite al modelo de lenguaje comportarse como un agente. Ejemplos de esta tipología de arquitectura son *Tree of Thoughts* (Yao et al., 2023), *Reflexion* (Shinn y Labash, 2023) y *Chain of Hindsight* (Liu et al., 2023). Dentro del proyecto KG4LLM, se han identificado los métodos de resolución de problemas (PSM, por sus siglas en inglés, *Problem Solving Methods*) como una posible tipología de flujo de trabajo para mejorar la capacidad de razonamiento de un modelo de lenguaje mediante la descomposición de tareas en subtareas más pequeñas y sencillas. Dentro de la literatura de PSM, algunos métodos de resolución de problemas prevén la necesidad de que un agente se comunique con una fuente de información externa para recuperar la información necesaria que le permita resolver un problema en concreto. Estas interacciones forman parte de las funciones de transferencia identificadas por Schreiber et al. (1994) en la metodología *CommonKADS*, y son fundamentales en tareas como el diagnóstico, la clasificación y el monitoreo, entre otras.

El dominio médico es uno de los dominios de interés para el proyecto KG4LLM identificados en el entregable E4.1. El soporte ofrecido por sistemas automáticos de razonamiento para diagnóstico médico recientemente ha atraído una atención creciente con el fin de reducir la carga de trabajo de los médicos y democratizar el acceso a la atención médica (Fansi Tchango, A. et al., 2022). Por otro lado, el diagnóstico médico es una aplicación natural de PSM y generalmente requiere información externa adicional para llevar a cabo la tarea. Por este motivo, se ha elegido el diagnóstico médico como primer caso de uso para experimentar con estrategias de llamada a herramientas externas basadas en PSM.

En este trabajo, se ha integrado un PSM de diagnóstico dentro del proceso de razonamiento de un modelo de lenguaje mediante prompts que inducen al modelo a replicar los pasos necesarios para resolver la tarea. Para que un modelo pueda entender qué información necesita para solucionar una tarea, debe generar de alguna manera una representación de la misma y habilitar una capacidad que podemos definir como reflexión. Esto se puede lograr induciendo al modelo

a descomponer el problema y utilizar las soluciones de cada subtarea como datos de entrada para la siguiente tarea en el flujo de trabajo previamente definido. De esta manera, el flujo permite generar representaciones intermedias que, a pesar de la arquitectura end-to-end del modelo, pueden replicar una metodología de razonamiento similar a la que emplea un médico. El flujo de trabajo seguido es el que está formalizado en la Figura 5.

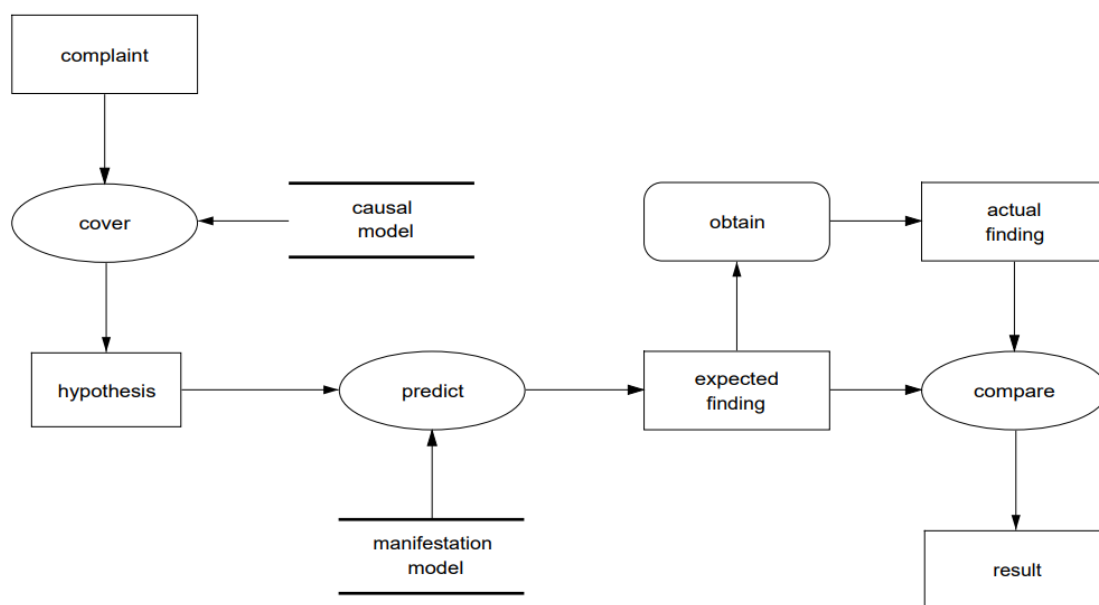


Figura 1. Estructura de inferencia para una aplicación de diagnóstico simple (Schreiber, G. et al., 1994).

Para seguir este proceso de razonamiento y permitir que el modelo solicite de manera autónoma la información externa necesaria (“actual finding”), el modelo debe identificar la información requerida (“expected finding”) y solicitar y obtener los datos externos necesarios (“obtain”). Un proceso de diagnóstico implica la generación de posibles hipótesis (“cover”) (Fansi Tchango, A. et al., 2022) y su verificación (“compare”), de ser necesario, mediante el uso de herramientas externas como análisis, ultrasonidos, radiografías, entre otras. A lo largo de este proceso, el primer input del modelo son los datos del paciente, incluyendo su historia clínica y sus síntomas. Una vez recibida esta información, se solicita al modelo que genere posibles hipótesis de la enfermedad y las pruebas médicas necesarias para verificar estas hipótesis. Los prompts utilizados para inducir estos pasos de razonamiento se muestran en la Figura 2 y la Figura 3.

```
Your task is to conduct a diagnostic process.
Given the JSON-structured data provided:

- Cover each complaint or symptom and include a potential solution
  (thus a fault) in the differential of possible hypotheses.
  In this context, a fault refers to any plausible medical
  condition that could explain the given symptoms.

Ensure the output is structured as a JSON object, meticulously
populating all the fields according to the following structure:

class Hypothesis(BaseModel):
    symptom: str
    fault: str

class Differential(BaseModel):
    differential: List[Hypothesis]

Provide only the JSON object. Exclude comments and extra text.
Be as concise as possible.
```

Figura 2. Prompt de sistema utilizado para generar hipótesis médicas.

```
Your task is to conduct a diagnostic process.
Given the JSON-structured data provided:

- Select each possible hypothesis from the differential
  and specify the observables needed to verify it.
  In this context, an observable refers to any procedure used to
  investigate the given hypothesis, including medical tests or surgeries.

Ensure the output is structured as a JSON object, meticulously
populating all the fields according to the following structure:

class Observable(BaseModel):
    symptom: str
    fault: str
    observable: List[str]

class Observables(BaseModel):
    observables: List[Observable]

Provide only the JSON object. Exclude comments and extra text.
Be as concise as possible.
```

Figura 3. Prompt de sistema utilizado para especificar observables.

Una vez obtenidos los resultados de estos pasos de razonamiento, que incluyen las posibles hipótesis de enfermedades y las pruebas necesarias para validarlas, se utiliza este output como nuevo input para el modelo de lenguaje. En esta etapa, se solicita al modelo que indique las pruebas específicas necesarias para continuar con el proceso de razonamiento, con el fin de validar o descartar las hipótesis generadas que están presentes en el conjunto de posibles enfermedades. El prompt utilizado para inducir al modelo a solicitar la prueba necesaria corresponde a la función de transferencia 'obtain' (Schreiber et al., 1994). Esta función está estructurada como un JSON donde el modelo, si no puede extraer la información requerida del contexto, reporta el valor 'OBTAIN'. Este valor funciona como un comando para obtener información a través de herramientas externas y el prompt utilizado para inducir este

comportamiento se muestra en la Figura 4. El resultado de esta etapa del proceso, que incorpora diversas hipótesis de enfermedades y sus correspondientes evidencias obtenidas mediante herramientas externas, servirá como input para que el modelo genere un posible diagnóstico.

```
Your task is to conduct a diagnostic process.
Given the JSON-structured data provided:

- From multiple lists of observables needed to validate various hypotheses,
  identify the most critical subset of findings. For these key observables,
  indicate the need to acquire their values by setting the result field to 'OBTAIN'.

Ensure the output is structured as a JSON object, meticulously
populating all the fields according to the following structure:

class Finding(BaseModel):
    reason: str
    observable: str
    result: str

class Findings(BaseModel):
    findings: List[Finding]

Provide only the JSON object. Exclude comments and extra text.
Be as concise as possible.
```

Figura 4. Prompt de sistema utilizado para habilitar llamadas a herramientas externas.

Como se demuestra en este flujo de razonamiento, exclusivamente mediante el uso de prompting, se ha podido utilizar la memoria paramétrica del modelo para generar hipótesis de enfermedades, teniendo en cuenta los datos iniciales relacionados con un paciente. Posteriormente, se ha producido un proceso de razonamiento, o de inferencia, simplemente descomponiendo la tarea principal de diagnóstico en diferentes pasos. Esta descomposición ha permitido inducir al modelo a solicitar externamente, a través de herramientas adicionales, la información necesaria para desarrollar sus hipótesis y seguir con el proceso de razonamiento.

4 Evaluación

Con el objetivo de evaluar la capacidad de los modelos de lenguaje generativo para habilitar herramientas externas mediante prompting en tareas de diagnóstico médico, se ha utilizado el dataset público multilingüe E3C¹. Este dataset contiene narrativas en las cuales los médicos documentan la historia de un caso clínico. Los datos están disponibles en los siguientes idiomas: italiano, inglés, francés, español y euskera. Para esta evaluación preliminar, se ha considerado el idioma inglés, que incluye 84 narrativas médicas. Se ha decidido utilizar el conjunto de datos completo, a pesar de su división original en entrenamiento (36) y test (48), dado que los experimentos no requieren ninguna forma de entrenamiento y las anotaciones originales del conjunto de datos están destinadas a una tarea de extracción de información.

¹ <https://live.european-language-grid.eu/catalogue/corpus/7618>

Como se mencionó en la sección anterior, se ha diseñado un flujo de trabajo que prevé la entrada y salida de datos estructurados en formato JSON para el LLM. En consecuencia, para adaptar y estructurar el dataset de manera que pueda ser utilizado como entrada, se ha empleado un LLM para traducir cada narrativa en un JSON con campos estandarizados y definidos. De hecho, cada narrativa médica consiste en un párrafo con una estructura variable que generalmente incluye una presentación del paciente, la razón de la admisión (que suele incluir los síntomas recientes), los procedimientos llevados a cabo para llegar a una formulación diagnóstica y el desarrollo de un plan de tratamiento. Se ha decidido estandarizar toda esta información variable y no estructurada para desarrollar un caso clínico bien organizado que permita formalizar un proceso de razonamiento e incluir una entrada, un proceso de inferencia y un resultado. El objetivo es que un modelo, dado un input, pueda replicar los pasos de inferencia y generar la salida correcta. Un ejemplo de una narrativa y su correspondiente JSON de entrada, procesado y estructurado por el LLM, se muestra en la Figura 5.

```
{
  ...
  "text": "We received a boy 3-year-old boy
with autistic disorder on hospital
of pediatric ward A at university
hospital Mohammed VI of Marrakesh.
He has no family history of illness
or autistic spectrum disorder.
The history revealed that at 2 years of age.
The child was diagnosed with a
''severe communication disorder'',
with social interaction difficulties
and sensory processing delay.
A composite follow-up of all previous
assessments and investigations was undertaken.
Blood work was normal
(thyroid-stimulating hormone (TSH),
hemoglobin, mean corpuscular volume
(MCV), and ferritin). Genetic testing
was unremarkable (normal karyotype,
negative for fragile X) and the magnetic
resonance imaging (MRI) in search of a
demyelinating attack of the white matter,
electroencephalography (EEG),
optometry assessment were also normal.
At the conclusion of this composite
assessment, the boy was given a primary
diagnosis of autistic spectrum disorder.
...
}
```

```
{
  "input": {
    "patient_data": [
      {
        "field": "Age",
        "record": "3 years old"
      },
      ...
    ],
    "anamnesis": {
      "history": [
        {
          "complaint": "Severe communication disorder",
          "reported": "Diagnosed at 2 years of age"
        },
        ...
      ],
      "symptoms": [
        {
          "complaint": "Gastrointestinal symptoms",
          "reported": "Presented at initial presentation"
        },
        ...
      ]
    },
    "physical_examination": [
      {
        "observable": "Height and weight",
        "observation": "Normal for age"
      },
      ...
    ]
  },
  ...
}
```

Figura 5. Narrativa médica y su estructuración como JSON de entrada.

Los prompts utilizados para que el LLM pudiera estructurar esta información que comprende entrada, inferencia y salida se detallan en la Figura 6. Este proceso de creación del dataset se detallará más a fondo en el entregable E2.1v2.



Figura 6. Prompts utilizados para crear el conjunto de datos a partir de narrativas médicas no estructuradas: a) entrada, b) inferencia, c) salida.

Para los experimentos de problem solving se ha utilizado un LLM entrenado para seguir instrucciones y resolver las tareas especificadas en los prompts. El modelo utilizado es Llama 3.1 de 8B de parámetros². El JSON en la Figura 5 (derecha) representa el input o "user prompt" que se pasa al LLM junto al "system prompt" con las instrucciones presentadas en la Figura 2, constituyendo el primer paso del flujo de trabajo del agente. Con la información del paciente, la anamnesis y la examinación física, se requiere que el modelo genere un conjunto de hipótesis

² <https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct>

de posibles enfermedades que serán utilizadas como entrada del prompt en Figura 3 para añadir los observables necesarios para verificarlas (Figura 7).

```
{
  "differential": [
    {
      "symptom": "Severe communication disorder",
      "fault": "Language Processing Disorder"
    },
    {
      "symptom": "Social interaction difficulties",
      "fault": "Social Anxiety Disorder"
    },
    {
      "symptom": "Sensory processing delay",
      "fault": "Sensory Processing Disorder"
    },
    {
      "symptom": "Unexplained fatigue",
      "fault": "Chronic Fatigue Syndrome"
    },
    {
      "symptom": "Gastrointestinal symptoms (bloating, constipation, diarrhea)",
      "fault": "Irritable Bowel Syndrome"
    },
    {
      "symptom": "Psychiatric symptoms (depressed mood, disproportionate anger, emotional lability)",
      "fault": "Bipolar Disorder"
    }
  ]
}

{
  "observables": [
    {
      "symptom": "Unexplained fatigue",
      "fault": "Chronic Fatigue Syndrome",
      "observable": [
        "Medical History and Physical Examination",
        "Sleep Study",
        "Laboratory Tests (e.g. complete blood count, electrolytes)"
      ]
    },
    {
      "symptom": "Gastrointestinal symptoms (bloating, constipation, diarrhea)",
      "fault": "Irritable Bowel Syndrome",
      "observable": [
        "Gastrointestinal Motility Study",
        "Endoscopy",
        "Imaging Studies (e.g. abdominal ultrasound, CT scan)"
      ]
    },
    {
      "symptom": "Psychiatric symptoms (depressed mood, disproportionate anger, emotional lability)",
      "fault": "Bipolar Disorder",
      "observable": [
        "Psychiatric Evaluation",
        "Mood Disorder Questionnaire",
        "Neuropsychological Evaluation"
      ]
    }
  ]
}
```

Figura 7. Conjunto de hipótesis y observaciones generadas por el modelo, que servirán como base para el proceso de razonamiento necesario para producir un diagnóstico médico.

Una vez generado el JSON con hipótesis y observables, se pasa como parte de un nuevo "user prompt" para una nueva iteración junto al "system prompt" presentado en la Figura 4, generando como output un nuevo JSON que, con la ayuda del token "OBTAIN", generado por el LLM cuando considera que necesita información adicional, será procesado posteriormente como función de transferencia para obtener información de herramientas externas (Figura 8).

En un caso médico como el presentado, puede ser necesario utilizar una amplia variedad de herramientas externas con distintos niveles de automatización. Por ejemplo, en la situación descrita en la Figura 8, el modelo idealmente debería tener la capacidad de integrarse con sistemas autónomos para realizar procedimientos como endoscopias, tomografías computarizadas y pruebas de laboratorio. En el caso de uso específico de la evaluación descrita en este documento, estos valores ya están presentes en el dataset y corresponden a la etapa de inferencia del conjunto de datos previamente anotado, que incluye todos los procedimientos médicos y los valores obtenidos a través suyo. Por tanto, más que la capacidad de recuperar información relevante para resolver un paso concreto del flujo de trabajo, nos centramos en la habilidad del LLM de identificar cuándo carece de información en su memoria paramétrica y pedir esa información a un servicio externo en su lugar.

La función de transferencia activada por el comando "OBTAIN" pasa nuevamente por un modelo de lenguaje que compara las pruebas médicas requeridas por el modelo de lenguaje que está

resolviendo el diagnóstico con la inferencia anotada del conjunto de datos actuando como un “mapper” (Figura 9). Si el modelo de lenguaje requiere una prueba que está presente en el conjunto de inferencia disponible, proporcionará los resultados del test específico, actuando como herramienta externa. Al modelo se le puede proporcionar la información requerida nuevamente en forma de JSON, si está disponible, la cual formará parte del nuevo input hasta llegar finalmente a la generación de un output con un diagnóstico (Figura 10).

```
{
  "findings": [
    {
      "reason": "Social interaction difficulties",
      "observable": "Social Skills Assessment",
      "result": "OBTAIN"
    },
    {
      "reason": "Sensory processing delay",
      "observable": "Sensory Processing Questionnaire",
      "result": "OBTAIN"
    },
    {
      "reason": "Unexplained fatigue",
      "observable": "Laboratory Tests (e.g. complete blood count, electrolytes)",
      "result": "OBTAIN"
    },
    {
      "reason": "Gastrointestinal symptoms (bloating, constipation, diarrhea)",
      "observable": "Endoscopy",
      "result": "OBTAIN"
    },
    {
      "reason": "Gastrointestinal symptoms (bloating, constipation, diarrhea)",
      "observable": "Imaging Studies (e.g. abdominal ultrasound, CT scan)",
      "result": "OBTAIN"
    }
  ]
}
```

Figura 8. los procedimientos necesarios que el modelo solicita externamente

```
Your task is to create connections between separate lists.  
Given the JSON-structured data provided:  
  
- Create a JSON object that maps each word from  
the source_words list to the corresponding word in  
the target_words list if they share identical or  
very similar meanings. Ensure the JSON format clearly  
represents the relationship between the two lists.  
  
Ensure the output is structured as a JSON object, meticulously  
populating all the fields according to the following structure:  
  
class MappedWords(BaseModel):  
    source_word: str  
    target_word: str  
  
class Map(BaseModel):  
    map: List[MappedWords]  
  
Provide only the JSON object. Exclude comments and extra text.  
Be as concise as possible.
```

Figura 9. Función de transferencia utilizada para mapear los observables con los procedimientos anotados.

El diagnóstico médico es una tarea compleja que, como se ha avanzado aquí, requiere el uso de herramientas externas que, de momento, funcionan con el soporte de una persona especializada. Sin embargo, lo relevante de este proceso es la capacidad del modelo para autoreflexionar y entender qué información es necesaria para seguir adelante con el razonamiento. Una vez que un modelo sea capaz de generar una solicitud estructurada, como un JSON, se puede pasar a cualquier tipo de función de transferencia que pueda comunicarse externamente con APIs de herramientas externas, como se ha visto en trabajos precedentes, incluyendo herramientas como calculadoras, motores de búsqueda, entre otras.

En los experimentos sobre las narrativas médicas en inglés, la función de transferencia delineada en la Figura 4 ha permitido seguir con el proceso de razonamiento en el 100% de los ejemplos, lo que ha permitido al agente llegar a un diagnóstico correcto en el 82% de los casos.

```
{
  "evidence": [
    {
      "observable": "Imaging Studies (e.g. MRI, CT scan)",
      "result": "Normal"
    },
    {
      "observable": "Laboratory Tests (e.g. complete blood count, electrolytes)",
      "result": "Normal"
    },
    {
      "observable": "Endoscopy",
      "result": "Total villous atrophy corresponding to a stage 4 of Marsh classification"
    }
  ]
}
```

```
{
  "diagnoses": [
    {
      "fault": "Celiac Disease",
      "evidence": [
        "Total villous atrophy corresponding to a stage 4 of Marsh classification"
      ]
    }
  ]
}
```

Figura 10. Información de evidencia proporcionada al modelo con diagnóstico final.

5 Repositorio

El código utilizado para resolver tareas complejas mediante metodologías de PSM y técnicas de prompting está disponible en el repositorio GitHub de INESDATA³. En este repositorio se encuentran los experimentos de diagnóstico médico mencionados en la sección 4, utilizando el dataset E3C⁴. El código en `psm\_diagnostic\_agent\_workflow.py` permite ejecutar el flujo de trabajo descrito en la sección anterior sobre este mismo dataset, integrando la función de transferencia `obtain\_findings`. Esta función habilita al modelo para auto reflexionar sobre las necesidades de información adicional y generar llamadas en formato JSON, facilitando el uso de herramientas externas para obtener la información necesaria y completar la tarea de diagnóstico.

6 Conclusiones y trabajo futuro

En este entregable hemos presentado una metodología para habilitar el uso de herramientas externas por un LLM. Teniendo en cuenta las necesidades multilingües del proyecto y la dificultad de generar datasets específicos para diferentes casos de uso en diversos dominios y posibles acciones que requieren información adicional, se ha optado por la experimentación de una modalidad que no implica procesos de entrenamiento. Esto permite superar el reto de generar datasets y organizar entrenamientos masivos, como se ha experimentado en trabajos previos. La posibilidad de romper el clásico flujo de trabajo end-to-end y abrir la opción de descomponer tareas y organizar pasos de razonamiento puede habilitar a un LLM para interactuar externamente dentro de este flujo de trabajo más complejo. En el marco del proyecto, se ha avanzado en la posibilidad de utilizar métodos de resolución de problemas (PSM) y, dentro de este esquema, activar las funciones de transferencia para habilitar el uso de herramientas externas. Inicialmente, se ha experimentado con el dominio médico y específicamente con la

³ <https://github.com/oeg-upm/inesdata-problem-solving>

⁴ <https://live.european-language-grid.eu/catalogue/corpus/7618>

tarea de diagnóstico, dada la necesidad frecuente de utilizar información adicional para continuar con el proceso de razonamiento y completar la tarea. Esta estrategia ha sido evaluada sobre un dataset público, adecuadamente modelado para la tarea, permitiendo interactuar con el agente de manera estructurada a través de esquemas JSON. La evaluación ha demostrado la capacidad de esta metodología para descomponer y reflexionar correctamente sobre los problemas, y habilitar la capacidad de solicitar y obtener autónomamente la información necesaria para solucionar un proceso de razonamiento y generar un diagnóstico médico.

References

- Fansi Tchango, A., Goel, R., Martel, J., Wen, Z., Marceau Caron, G., & Ghosn, J. (2022). **Towards trustworthy automatic diagnosis systems by emulating doctors' reasoning with deep reinforcement learning**. *Advances in Neural Information Processing Systems*, 35, 24502-24515.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). **React: Synergizing reasoning and acting in language models**. *arXiv preprint arXiv:2210.03629*
- 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., Deroncourt, 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., Giurghi, 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.
- Krscinski, 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.deelio-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., Vi'egas, 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., Udupa, 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*.