Guía de Buenas Prácticas para el Desarrollo de Herramientas de Inteligencia Artificial Generativa en Salud

Grandes Modelos de Lenguaje (LLM)

Generalitat de Catalunya

© Fundació TIC Salut i Social

Esta guía ha sido elaborada por el Área de Inteligencia Artificial de la Fundación TIC Salud Social

Autores: Susanna Aussó, Antoni Berenguer, Julia Aznar, Carolina Raventós, Vaneza Gómez y Maria Bretones, con la colaboración del equipo de HEALTH de NTT FECHA

Edición electrónica: Febrero de 2025

Esta obra está sujeta a una licencia de Reconocimiento - No Comercial - Sin Obras Derivadas 4.0 de Creative Commons. Se permite su reproducción, distribución y comunicación pública siempre que se reconozcan la autoría y el editor y no se haga un uso comercial. No está permitida la transformación de esta obra para generar una nueva obra derivada.

Contenido

1 / **Contexto** Página 4

2 / **IA Generativa** Página 7

3 / **Diseño** Página 15

4 / **Implementación** Página 23

5 / **Evaluación** Página 43

6 / **Despliegue** Página 54

7 / **Recomendaciones** Página 57

8 / **Bibliografía** Página 59

1 / Contexto

El Programa para la Promoción y Desarrollo de la Inteligencia Artificial en el Sistema de Salud de Cataluña (Programa Salud/IA) tiene como objetivo crear un entorno que facilite la innovación en el ámbito de la salud. Mediante el desarrollo y la implementación de soluciones de Inteligencia Artificial (IA), este programa busca mejorar la salud de la ciudadanía, priorizando la prevención, liderando la implantación de estas tecnologías y contribuyendo a la calidad asistencial y la sostenibilidad del sistema sanitario, poniendo en valor el conocimiento generado por el sistema sanitario integral de utilización pública de Cataluña.

La **Fundació TIC Salut i Social** ha creado esta guía para ofrecer pautas claras a los profesionales y organizaciones implicados en el desarrollo de algoritmos de IA generativa aplicados a la salud. Debido a la rápida evolución y popularización de esta tecnología en los últimos años, el documento se centra especialmente en las herramientas basadas en **Grandes Modelos de Lenguaje** *(Large Language Models, **LLM**)* y en la creación de soluciones seguras, eficientes y de alto valor añadido, alineadas con los estándares más avanzados en inteligencia artificial.

El avance de la IA, que se basa en la creación de sistemas capaces de procesar información y generar respuestas que imiten al razonamiento humano, es ya una realidad. Estos sistemas pueden aprender de la experiencia, resolver problemas en contextos delimitados y realizar tareas complejas en diversos ámbitos de la sociedad. En el ámbito de la salud, la creciente disponibilidad de registros electrónicos de salud, imágenes médicas digitales, datos ómicos y otros conjuntos de datos ofrece un gran potencial para mejorar el bienestar de las personas. Los métodos de Aprendizaje Automático (*Machine Learning*), y especialmente las técnicas de Aprendizaje Profundo (*Deep Learning*), permiten desarrollar algoritmos avanzados capaces de aprender a partir de esta diversidad de datos [2].

Sin embargo, el uso de técnicas de IA generativa en el ámbito de la salud presenta importantes retos, especialmente en su implementación, la integración en la práctica clínica y la adaptación a las especificidades del sector Salud. Es necesario abordar cuestiones clave como la supervisión humana, la calidad de los datos, la interpretabilidad de los resultados y la protección de la privacidad. La naturaleza innovadora de esta tecnología requiere una estrecha coordinación entre profesionales de la salud, investigadores y expertos en tecnología para garantizar una aplicación efectiva y segura, ética y lícita, y alineada con las necesidades reales tanto de los profesionales como de los pacientes.

La confianza de los y las profesionales de la salud en las soluciones de IA es un factor clave y debe fundamentarse en principios como la transparencia y el rigor. El ámbito de la salud presenta retos no solo científicos, sino también éticos y legales, ya que las decisiones tomadas tienen un impacto directo en el bienestar y la vida de las personas1. Por eso, la fiabilidad de las herramientas de IA debe construirse sobre tres componentes:

IA legal	Cumpliendo todas las leyes y regulaciones aplicables
IA ética	Asegurando los principios y valores éticos
IA robusta	Desde una perspectiva técnica (garantizando la solidez de las herramientas) y social (teniendo en cuenta el entorno en el que operan)

El objetivo de esta guía **es proporcionar un marco general para el uso y la integración de los LLM, como parte de la IA generativa, en el sector de la salud.** Está pensada para facilitar la comprensión y aplicación de estas tecnologías en los procesos asistenciales, garantizando su uso ético y responsable, fomentando la colaboración e impulsando la innovación en salud.

Objetivos de la guía

OBJETIVO 01 /

Formar y capacitar los profesionales de la salud en el uso de los LLM

Proporcionar los conocimientos y habilidades necesarios para comprender y aplicar estos modelos de lenguaje dentro del ámbito sanitario y biomédico, promoviendo un uso informado, seguro y efectivo de los mismos.

OBJETIVO 02 /

Guiar la implementación de herramientas LLM en los procesos asistenciales

Facilitar la integración de herramientas basadas en LLM en los flujos de trabajo dentro del ámbito sanitario, desde la recogida y gestión de datos hasta la generación de contenido, con el objetivo de mejorar la eficiencia y la calidad de la actividad asistencial.

OBJETIVO 03 /

Promover la colaboración interdisciplinaria en la aplicación de los LLM

Fomentar la cooperación entre profesionales sanitarios, investigadores y expertos en IA para desarrollar soluciones innovadoras y mejorar continuamente las herramientas basadas en modelos de lenguaje que ya han sido implementadas.

OBJETIVO 04/

Facilitar la comprensión y selección de modelos de IA generativa basados en LLM

Ofrecer una visión clara de los distintos tipos de modelos de lenguaje disponibles, así como criterios para seleccionar los más adecuados según aspectos técnicos, de evaluación y de implementación.

OBJETIVO 05 /

Garantizar el uso ético y responsable de los LLM

Velar por el respeto a la privacidad de los pacientes, el cumplimiento de las normativas vigentes y la adecuada gestión de los riesgos asociados al uso de modelos de lenguaje en contextos clínicos y biomédicos.

2 / IA Generativa

2.1 / ¿Qué es la IA generativa?

2.2 / Los Grandes Modelos del Lenguaje (LLM)

2.3 / Potencial de los LLM en Salud

2.4 / Desarrollo de una herramienta LLM en Salud

2.1 / ¿Qué es la IA generativa?

La IA generativa es una **rama de la IA que se centra en la creación de contenidos nuevos y originales**, como textos, imágenes, música, vídeos, datos estructurados y otros formatos.

Como el resto de modelos de IA, los modelos generativos son **predictivos** desde un punto de vista funcional, puesto que generan contenido previendo la siguiente palabra, imagen o elemento a partir de patrones aprendidos.

Sin embargo, la IA generativa se distingue por su capacidad de **crear contenido nuevo**, a menudo creativo, a diferencia de otros modelos que se limitan a clasificar o recomendar (como por ejemplo: los modelos de regresión lineal, regresión logística, *random forest, support vector machines* o redes neuronales, entre otras)

Además de esta capacidad para generar contenido, la IA generativa **se diferencia del resto de modelos de aprendizaje automático en otros aspectos relevantes**, como la cantidad y características de los datos que se necesitan, los requerimientos computacionales o los desafíos asociados a la interpretabilidad de los resultados.

GPU: Unidades de Procesamiento Gráfico (*Graphics Processing Unit*)
TPU: Unidades de Procesamiento de Tensor (*Tensor Processing Unit*)

La siguiente tabla recoge un resumen de estas diferencias:

Características	IA Generativa	Otros modelos de aprendizaje automático
Capacidades	Creación de contenido nuevo, como por ejemplo texto, imágenes o sonidos. Pueden adaptarse para realizar tareas predictivas.	Funciones de predicción, tales como clasificación o recomendación. No pueden generar nuevos contenidos.
Comprensión de lenguaje natural	Comprensión profunda del lenguaje incluyendo contexto, matices semánticos, negaciones o ironía.	Comprensión básica del lenguaje, limitada a menudo a patrones simples o diccionarios.
Tipo de tareas	Generación de texto, imagen, o sonido, resumen de documentos, traducción automática. Predicción, clasificación o extracción de información.	Tareas específicas como clasificación, regresión o recomendación.
Generalización	Modelos flexibles capaces de afrontar una gran variedad de tareas con un único entrenamiento previo sobre datos masivos no etiquetados (*pretraining*).	Modelos específicos para una única labor. A menudo requieren reentrenamiento completo para abordar nuevos problemas.
Preparación de datos	Aprenden las características directamente de los grandes volúmenes de datos no estructurados. Ajuste manual mínimo.	Requieren la ingeniería de variables (*feature engineering*), a menudo manual, para seleccionar, transformar y estructurar los datos de una forma que sea útil para el modelo.
Tipo de aprendizaje y datos	Aprendizaje autosupervisado, donde el modelo predice la siguiente palabra en una secuencia a partir de datos no etiquetados. A menudo incluye una fase de alineamiento con datos etiquetados.	Aprendizaje supervisado con datos etiquetados (con información anotada).
Costo computacional	Altamente exigentes: requieren grandes recursos computacionales para entrenar (millones o miles de millones de parámetros).	Menos exigentes (suelen tener muchos menos parámetros).
Infraestructura tecnológica	Requieren infraestructuras masivas, con GPU, TPU y memoria distribuida para realizar un entrenamiento e inferencia eficientes.	En muchos casos, pueden entrenarse en equipos pequeños (incluso ordenadores portátiles).
Proceso de entrenamiento	Aprovechan el conocimiento preentrenado sobre datos masivos que les proporciona un buen rendimiento general. Pueden adaptarse a tareas más específicas.	Modelos entrenados desde cero para realizar tareas específicas.
Interpretabilidad	Baja a causa de su alta complejidad y número de parámetros. La explicabilidad en IA generativa es actualmente un área activa de investigación.	Alta en modelos simples (por ejemplo, regresiones y árboles de decisiones). Moderada en modelos más complejos, con soporte de técnicas de IA explicable (XAI).

2.2 / Los Grandes Modeles del Lenguaje (LLM)

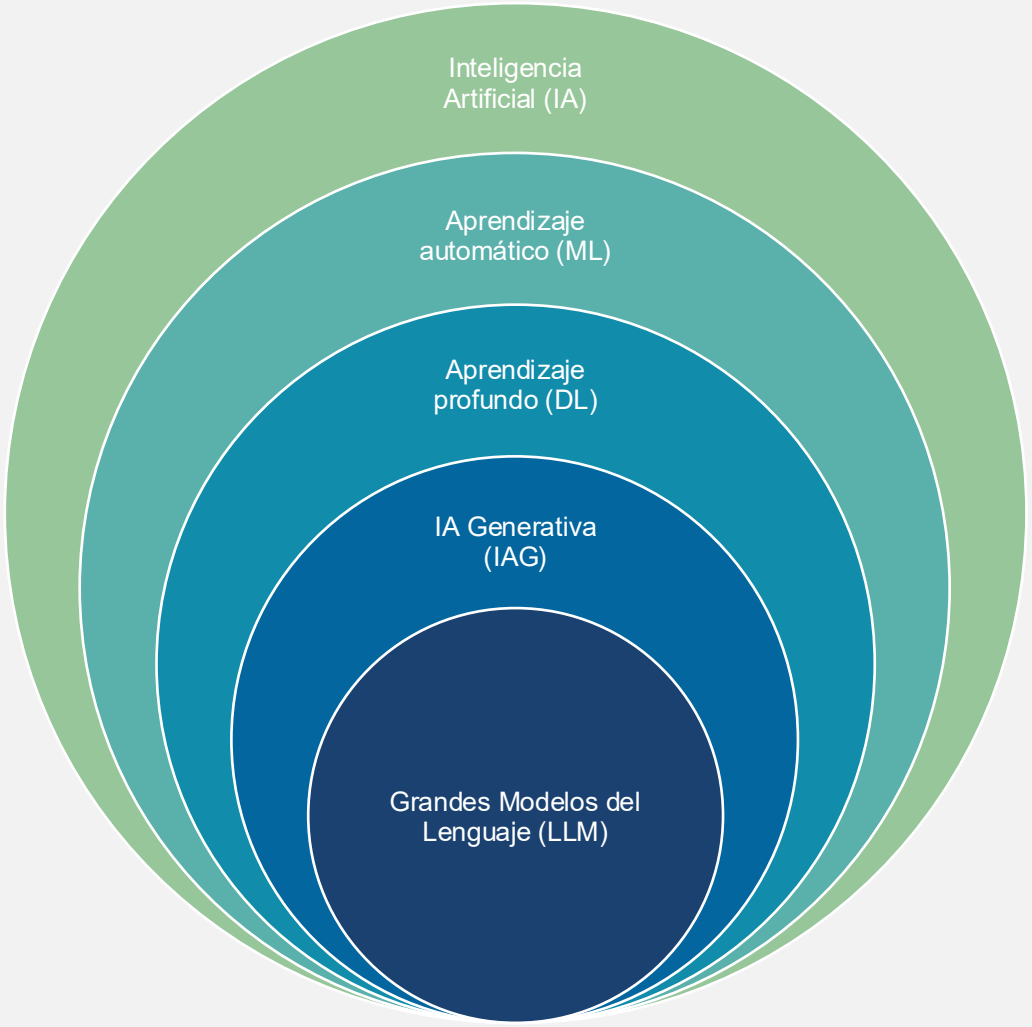
Actualmente, los **Grandes Modelos de Lenguaje** (*Large Language Models*, **LLM**) son una de las tecnologías más relevantes dentro de la IA generativa (Figura 1).

Aunque **trabajan principalmente con lenguaje natural** (es decir, el lenguaje humano), estos modelos también **pueden generar otros formatos**, como código de programación o representaciones matemáticas, previamente expuestos a estos datos.

Las funcionalidades de los LLM se basan en la **predicción (inferencia) iterativa de la siguiente unidad de texto** (como una palabra o una frase) mediante el análisis de patrones y estructuras lingüísticas, gramaticales y semánticas aprendidas a partir de volúmenes masivos de datos textuales (**preentrenamiento**).

Este entrenamiento les dota de una notable capacidad para **generar, traducir o completar textos de manera precisa y comprensible**, lo que ha impulsado aplicaciones avanzadas como los asistentes virtuales, la redacción automática de documentos, la generación de resúmenes o la traducción automática.

Tipologías de Inteligencia Artificial



Técnicamente, los LLM se fundamentan en el aprendizaje profundo y en arquitecturas como los **transformadores** (*transformers*). Estas arquitecturas han sido clave en la evolución del procesamiento del lenguaje natural (Natural Language Processing, NLP), ya que han permitido superar muchas de las limitaciones de las tecnologías anteriores, como las redes neuronales recurrentes (Recurrente Neural Networks, RNN).

En particular, las RNN tenían dificultades para captar relaciones a largo plazo en textos extensos porque procesaban la información secuencialmente, limitando su capacidad de analizar contextos globales.

Los transformadores abordan esta limitación gracias a los **mecanismos de atención**, que son procedimientos algorítmicos avanzados **capaces de analizar el contexto completo de un texto de forma paralela**. Esta capacidad no solo permite manipular grandes volúmenes de datos con eficiencia, sino que también facilita la interpretación de contextos complejos. Estas propiedades han hecho de los transformadores una pieza fundamental en las actuales aplicaciones de IA generativa.

Las **Unidades de Procesamiento Gráfico** (*Graphics Processing Unit,* ***GPU***) y las **Unidades de Procesamiento de Tensor** (*Tensor Processing Unit,* ***TPU***) han sido los otros elementos técnicos clave para hacer de los LLM una realidad.

Las **GPU**, diseñadas inicialmente para el procesamiento de gráficos, son procesadores optimizados para **realizar cálculos en paralelo**, permitiendo ejecutar millones de operaciones simultáneamente.

Las **TPU**, por su parte, son procesadores desarrollados por Google **para optimizar las operaciones con tensores**, que son estructuras de datos esenciales en el profundo aprendizaje.

Ambas tecnologías han acelerado el entrenamiento de los modelos, reduciendo los tiempos de cálculo y facilitando la gestión de grandes volúmenes de datos necesarios en estos procesos.

Es importante señalar que, aunque los LLM están diseñados para ser generativos, no siempre se utilizan con este fin. Muchos LLM también pueden aplicarse a tareas no generativas, como la clasificación, la extracción de información o el análisis de sentimientos. Esta combinación de predicción y generación de contenidos los convierte en herramientas de gran valor y potencial impacto para el desarrollo de soluciones en todas las áreas del sector de la salud (**2.3/ Potencial de los LLM en Salud**).

Por último, cabe decir que la IA generativa abarca un amplio espectro que va más allá del lenguaje natural y que incluye otras modalidades, como la generación de imágenes, sonidos y vídeos. Además, algunos modelos de IA generativa, conocidos como **modelos multimodales,** son capaces de combinar distintos **tipos de datos** en un único modelo. Por tanto, aunque los LLM son ejemplos destacados de IA generativa, esta tecnología va mucho más allá del procesamiento del lenguaje natural.

2.3 / Potencial de los LLM en salud

Los LLM presentan un gran potencial para aportar soluciones innovadoras y eficientes a los retos actuales en el ámbito sanitario y biomédico [3].

Gracias a su capacidad para analizar grandes volúmenes de datos y generar contenido textual, estos modelos pueden tener un impacto transformador en el sector de la salud, incluyendo la automatización y mejora de procesos en la actividad asistencial, la salud pública, la gestión sanitaria y la investigación científica.

Ámbitos de uso de la IA generativa en el sector salud

ÁMBITO 01 /
Actividad asistencial

Los LLM pueden contribuir a mejorar la eficiencia y la calidad de la atención sanitaria mediante la generación de informes clínicos, la automatización de la documentación médica y el soporte a la comunicación entre pacientes y profesionales.

Ejemplos de aplicación de los LLM en la actividad asistencial con el paciente como usuario final son:

- Bote médico para pacientes**:** asistente interactivo que responde a preguntas generales sobre salud y proporciona recursos personalizados para mejorar la información y la formación de los pacientes en materia de salud.
- **Resúmenes médicos para pacientes:** generación de informes que incluyen explicaciones de diagnósticos y tratamientos en lenguaje claro y accesible para facilitar la comprensión y mejorar la comunicación entre pacientes y profesionales sanitarios.
- **Traducción de documentos médicos:** facilita el acceso a la información sanitaria para pacientes de distintos orígenes lingüísticos, mejorando la comprensión y favoreciendo una comunicación más efectiva en la atención asistencial.

Entre las aplicaciones con el profesional de la salud como usuario final podemos citar:

- **Asistentes virtuales para el diagnóstico y tratamiento:** proporcionan recomendaciones de diagnóstico y planes terapéuticos personalizados basados en el estado de salud del paciente, en la evidencia científica y en guías clínicas dinámicas actualizadas, apoyando a los profesionales médicos en la toma de decisiones.
- **Prescripción automática de pruebas de laboratorio:** generación automática de sugerencias de pruebas de laboratorio en función del historial médico y de los síntomas del paciente.
- **Transcripción simultánea de consultas médicas:** grabación, transcripción en tiempo real y estructuración automática de las conversaciones entre pacientes y profesionales sanitarios. Esta funcionalidad facilita la creación de informes clínicos estructurados, mejora la calidad de la documentación clínica y optimiza la experiencia en la consulta.
- **Resúmenes clínicos automáticos:** generación de informes a partir de las visitas e historiales médicos, facilitando el acceso a la información relevante y agilizando la toma de decisiones clínicas.
- **Autoanamnesis:** generación automática de historias clínicas estructuradas a partir de la información aportada por el paciente, facilitando un acceso rápido, organizado y codificado para los profesionales sanitarios.

ÁMBITO 02 /
Salud pública

Los LLM representan una herramienta poderosa para la gestión de la salud pública, facilitando el análisis de grandes volúmenes de datos y la comunicación personalizada para promover comportamientos saludables.

Ejemplos de los LLM en salud pública:

- **Comunicación sanitaria personalizada:** creación de materiales informativos adaptados a distintos niveles de alfabetización, idiomas y contextos culturales, con el objetivo de reforzar las iniciativas de salud pública.
- **Análisis automático de comentarios públicos sobre políticas sanitarias:** resumen y síntesis de grandes volúmenes de opiniones ciudadanas para evaluar la percepción pública y el impacto de nuevas normativas en el ámbito de la salud.
- **Identificación de patrones de salud a partir de datos no estructurados:** análisis combinado de textos médicos, redes sociales y datos ambientales para detectar tendencias de salud poblacional e identificar de forma precoz riesgos y patrones emergentes.
- **Apoyo a la revisión de documentación en inspecciones sanitarias:** automatización del análisis y validación de documentos para mejorar la eficiencia de los procesos de inspección, facilitando la detección de incumplimientos normativos, incoherencias o anomalías en informes y registros sanitarios.

ÁMBITO 03 /
Gestión sanitaria

Los LLM pueden jugar un papel clave en la mejora de la eficiencia operativa de las instituciones sanitarias gracias a su potencial para optimizar la gestión de recursos y aumentar la precisión en los procesos administrativos.

Ejemplos de los LLM en la gestión sanitaria

- Codificación automática de diagnósticos: interpretación y codificación de informes clínicos para facilitar la facturación, optimizar la administración y ahorrar tiempo a los profesionales de la salud.
- Evaluación de la calidad de la atención: análisis automatizado de los resultados clínicos y del *feedback* de los pacientes para medir la calidad de los servicios asistenciales e identificar áreas de mejora.
- Generación automatizada de presupuestos e informes de rendimiento: automatización de la elaboración de presupuestos y del análisis de rendimiento, mejorando la coordinación y eficiencia entre departamentos.

ÁMBITO 04 /
Investigación clínica i biomédica

Los LLM tienen un gran potencial para acelerar el progreso científico en salud, facilitando el análisis de datos complejos y la generación de nuevo conocimiento.

Ejemplos de los LLM en investigación

- Cribado y selección en ensayos clínicos: definición automatizada de criterios de elegibilidad e identificación de pacientes para participar en estudios, agilizando el reclutamiento y ejecución de los ensayos clínicos.
- Identificación de eventos adversos: análisis automatizado de grandes volúmenes de datos para detectar y monitorizar posibles efectos adversos en tratamientos experimentales.
- Revisión automatizada de bibliografía: resumen y síntesis de artículos científicos para facilitar la revisión de literatura relevante, ahorrando tiempo a los investigadores y agilizando el acceso a conocimiento actualizado.
- Creación de datos sintéticos: generación de datos artificiales para enriquecer los conjuntos de datos de entrenamiento, mejorando la capacidad de las herramientas de IA para realizar predicciones precisas y generar contenido diverso. Esta estrategia es especialmente valiosa en áreas donde los datos reales son escasos o sensibles.
- Generación de hipótesis: formulación de nuevas hipótesis a partir del análisis de patrones en grandes conjuntos de datos mediante modelos avanzados. Esta funcionalidad potencia la investigación científica y la toma de decisiones médicas, permitiendo explorar opciones de tratamiento o diagnóstico basadas en la evidencia.

2.4 / Desarrollo de una herramienta LLM en Salud

El desarrollo de una herramienta basada en LLM sigue una serie de pasos que incluyen el **diseño de la solución, la implementación, la evaluación y su despliegue**.

Este proceso, a menudo, requiere la participación de un **equipo multidisciplinar**, formado por diferentes perfiles profesionales como desarrolladores, personal médico, gestores sanitarios o los propios pacientes, entre otros.

En las siguientes secciones, analizaremos los aspectos clave de cada fase, centrándonos en aquellos elementos específicos o especialmente relevantes para el desarrollo de herramientas en salud basadas en LLM.

Fases y aspectos clave del desarrollo de una herramienta LLM en Salud

DISEÑO	IMPLEMENTACIÓN	EVALUACIÓN	DESPLIGE
Identificación de la necesidad	Selección del modelo • Modelos generalistas comerciales • Modelos generalistas de código abierto • Modelos adaptados en el ámbito sanitario y biomédico • Modelos locales o remotos • Pequeños Modelos de Lenguaje (SLM)	Eficacia del modelo • Métricas intrínsecas • Métricas extrínsecas	Flujos de trabajo
Adecuación de la tecnología			Modulación
Definición de la intención de uso		Detección y mitigación de sesgos	Sistemas tecnológicos
Funciones de explicabilidad	Adaptación del modelo • Ingeniería de *prompts* • Generación Aumentada por Recuperación • Afinamiento • ¿Cómo elegir el método de adaptación? • Datos para la adaptación		Usabilidad
Funciones de supervisión humana		Detección y mitigación de alucinaciones	Capacitación
Privacidad y protección de datos			Monitorización
Regulaciones y certificaciones	Agentes y herramientas multiagente	Detección y prevención de fugas de datos	Seguridad
			Actualización

3 / Diseño

3.1 / Identificación de la necesidad

3.2 / Adecuación de la tecnología

3.3 / Definición de la intención de uso

3.4 / Funciones de explicabilidad

3.5 / Funciones de supervisión humana

3.6 / Privacidad y protección de datos

3.7 / Regulaciones y certificaciones

En comparación con otras tecnologías, las herramientas basadas en LLM introducen retos y consideraciones específicas que deben abordarse desde las primeras etapas del diseño. En este capítulo, se detallan cada uno de estos aspectos que son clave en el diseño de una herramienta robusta, fiable y alineada con las necesidades del ámbito sanitario.

01 Identificación de la necesidad
02 Adecuación de la tecnología
03 Definición de la intención de uso
04 Funciones de explicabilidad
05 Funciones de supervisión humana
06 Privacidad y protección de datos
07 Regulaciones y certificaciones de explicabilidad

3.1 / Identificación de la necesidad

El desarrollo de una solución tecnológica comienza por identificar y describir con claridad la necesidad dentro del ámbito de trabajo, en este caso el sector de la salud.

Esta fase inicial implica **verificar que el problema abordado existe, siendo relevante y prioritario** para el sistema sanitario y para el bienestar de la ciudadanía.

En este contexto, es muy recomendable realizar **estudios de impacto social, económico y ético**, a fin de determinar la relevancia de la necesidad y contextualizarla dentro de los planes de salud vigentes establecidos por las autoridades.

3.2 / Adecuación de la tecnología

El entusiasmo creado por la IA generativa y su gran potencial no debería desviar el foco de lo verdaderamente importante, que es la solución de un problema real en el ámbito de la salud. El uso de la IA generativa debería estar justificado por aquello en lo que puede contribuir en esta solución, evitando utilizarla sólo por ser una tecnología "innovadora". Dado que la implementación de estas tecnologías puede ser compleja y costosa, es importante **considerar si otras estrategias más sencillas como la IA no generativa o los sistemas basados en reglas podrían ofrecer soluciones igualmente eficaces**.

En este proceso, las **revisiones del estado del arte** son especialmente útiles, puesto que permiten evaluar la eficacia de las tecnologías disponibles para abordar la necesidad, y verificar que el problema no haya sido ya resuelto por una solución existente en el mercado. Estas revisiones además ayudan a identificar posibles puntos de mejora o innovación en las tecnologías actuales.

3.3 / Definición de la intención de uso (*intended use*)

Una vez identificada la necesidad a abordar, el siguiente paso es traducirla en objetivos y funcionalidades. Aunque los LLM son flexibles y capaces de afrontar una gran variedad de tareas, es muy importante definir objetivos claros, específicos y mensurables. Esta concreción es importante por dos motivos:

- Permite definir la descripción de uso y prevenir usos no previstos (*functional creep*). La herramienta sólo se evaluará por los objetivos establecidos y, por tanto, su uso fuera de ellos no garantiza su eficacia y puede afectar a la confianza de los usuarios ya la seguridad de los pacientes.
- Proporciona mecanismos para evaluar si la herramienta cumple los objetivos mediante métricas adecuadas, que deben estar definidas de forma precisa y se revisarán en detalle en secciones posteriores (ver 5/ Evaluación).

La descripción del uso de la solución incluye la especificación de sus objetivos y funcionalidades, así como sus usuarios finales, procedimientos de uso, limitaciones de la herramienta y riesgos asociados.

Ejemplos de factores que componen la intención de uso

Necesidad identificada (punto de partida)	Reducir la carga administrativa del personal médico relacionada con la redacción de documentos clínicos.

FACTOR 01 /
Objetivos específicos

Automatizar la redacción de informes médicos a partir de los registros electrónicos de salud, con una calidad equivalente a la de los humanos y reduciendo en un 40% el tiempo que dedica el personal clínico a esta labor.

FACTOR 02 /
Usuario final

El personal clínico, como médico y de enfermería.

FACTOR 03 /
Funcionalidades

Identificación de elementos relevantes de la historia clínica, generación de informes estructurados, integración con historias clínicas electrónicas, especialización en terminología médica, incluyendo diagnósticos, patologías, síntomas, pautas de tratamiento y nombres comerciales y principios activos de medicamentos.

FACTOR 04 /
Procedimientos de uso

Inicio del proceso a partir de una instancia en la aplicación de registro electrónico, revisión del borrador generado por la herramienta y validación final por parte del profesional antes de guardar o compartir el informe.

FACTOR 05 /
Limitaciones

Dependencia de la calidad de los datos en los registros electrónicos, y necesidad de supervisión final por parte de un profesional clínico para validar el contenido generado.

FACTOR 06 /
Riesgos asociados

Sobreconfianza en la herramienta por parte del personal clínico, resultados discriminatorios por sesgos en los datos de entrenamiento, errores en la interpretación de datos o informes incompletos, que pueden comprometer la seguridad de los pacientes si no se detectan y corrigen a tiempo.

3.4 / Funciones de explicabilidad (XAI)

3.4.1 / La explicabilidad en IA

La complejidad de los modelos de IA hace que los mecanismos de decisión utilizados por los algoritmos sean a menudo desconocidos, incluso para sus propios desarrolladores. Las **técnicas de IA Explicable** (*Explainable AI, XAI*) ofrecen herramientas que permiten **garantizar la trazabilidad de las respuestas de un algoritmo**, asegurando que tanto los resultados generados como los procesos que los han producido sean comprensibles para los humanos.

La explicabilidad en IA contrasta con el concepto de caja negra (*black box*), donde los mecanismos que intervienen para generar una respuesta específica (output) a partir de una entrada (*input*) son desconocidos o inaccesibles.

En el ámbito de la IA, alcanzar esta transparencia es fundamental para garantizar el desarrollo de sistemas fiables, robustos y alineados con los principios éticos y legales.

Aspectos clave de la explicabilidad en IA

ASPECTO 01 /
Rendimiento y eficacia

Comprender cómo el sistema genera un determinado contenido permite **optimizar el código y los componentes de la herramienta** (*debugging*). Esto facilita la identificación de posibles sesgos, alucinaciones y áreas de mejora, contribuyendo así a aumentar su rendimiento y eficacia global.

ASPECTO 02 /
Sesgos sociales y equitad

Por la naturaleza de los datos con los que se han entrenado, los **LLM pueden incorporar sesgos discriminatorios que pueden incidir negativamente en las decisiones clínicas**. Las herramientas de explicabilidad permiten identificar estas desviaciones y tomar acciones correctivas, asegurando un trato justo y equitativo para todos los pacientes, independientemente de sus características demográficas, étnicas o socioeconómicas.

ASPECTO 03 /
Complimiento de la regulación

Las técnicas de explicabilidad aportan transparencia, trazabilidad y comprensión humana de las respuestas del sistema, facilitando así la **verificación de su cumplimiento con los requisitos normativos establecidos**.

ASPECTO 04 /
Confianza de profesionales y pacientes

Para generar confianza, los **resultados de un modelo no sólo deben ser precisos, sino también comprensibles, justos y objetivos**. Los profesionales de la salud necesitan entender los procesos por los que el modelo llega a sus conclusiones, para poder integrarlos adecuadamente en la toma de decisiones e identificar posibles errores. Para los pacientes estas cualidades aumentan la sensación de seguridad y favorecen un diálogo abierto y transparente con los profesionales.

ASPECTO 05 /
Supervisión humana

El **Reglamento Europeo de IA** establece que los sistemas de IA de alto riesgo deben estar sometidos a una supervisión humana proporcional a los riesgos asociados. Esto implica dar a los usuarios la capacidad de detectar anomalías, identificar sesgos, interpretar resultados y, cuando sea necesario, ignorar o detener las recomendaciones del sistema, garantizando una intervención segura y efectiva. La implementación de estas capacidades forma parte del ámbito de la explicabilidad, y se tratará detalladamente en secciones posteriores (ver **3.5/ Funciones de supervisión humana**).

3.4.2 / Técnicas de explicabilidad en LLM

En el caso de los LLM, **alcanzar la explicabilidad de un sistema es especialmente complejo** debido a las particularidades de estos modelos. Factores como su arquitectura avanzada, la gran cantidad de datos utilizados durante el entrenamiento (a menudo desconocidos por el usuario) y el formato de los datos de entrada y salida, habitualmente en forma de texto libre, pueden dificultar esta transparencia.

Las **técnicas clásicas de XAI**, como la importancia de las variables de permutación (*permutation feature importance*), los valores de las Explicaciones Aditivas Shapley (*Shapley Additive Explanations, SHAP*) o las Explicaciones Agnósticas del Modelo de Interpretación Local (*Local Interpretable Model-agnostic Explanation)*. Estos métodos presentan a menudo **problemas de cálculo, escalabilidad y formalización cuando se aplican a los LLM** y, por este motivo, la explicabilidad en estos modelos sigue siendo un área activa de investigación.

Técnicas de XAI más populares para LLM

TÉCNICA 01 /
Técnicas de atribución

Identificar qué partes de la entrada contribuyen más a la salida generada por el modelo, lo que permite una mejor comprensión de cómo el modelo procesa la información. Por ejemplo, *TransSHAP* y los gradientes integrados (*integrated gradients*) son métodos de atribución que han sido adaptados en los LLM.

TÉCNICA 02 /
Técnicas de *prompting*

Aprovechan la capacidad de los LLM para **generar explicaciones en lenguaje natural**, facilitando la comprensión de los procesos que han dado lugar a una respuesta concreta. Por ejemplo, la técnica de la **cadena de pensamiento** (*chain-of-thought*) permite que el modelo exponga paso a paso los razonamientos que justifican su respuesta, en lugar de limitarse a dar sólo una respuesta final (ver **4.2.1/ Ingeniería de prompts**).

TÉCNICA 03 /
Técnicas basadas en atención

Analizan los patrones de atención del modelo a identificar en qué partes de la entrada se focaliza durante el procesamiento. A pesar de estar en debate en la actualidad, métodos como el análisis de **flujo de atención** ofrecen una visión detallada sobre cómo se distribuye la atención entre los diferentes elementos de la entrada.

TÉCNICA 04 /
Técnicas de interpretación de los componentes

Buscan entender el **papel de componentes individuales del modelo**, como neuronas, capas o cabeceras de atención. Por ejemplo, las técnicas de *probing* analizan qué información lingüística codifican determinadas partes del modelo, lo que permite explorar cómo se representan conceptos específicos dentro de su arquitectura.

TÉCNICA 05 /
Explicaciones basadas en muestras

Relacionan las predicciones del modelo con muestras de los datos de entrenamiento o adaptación. En los LLM esto puede implicar identificar documentos o fragmentos de texto que han influido en una respuesta concreta.

3.5 / Funciones de supervisión humana

Además de ser un requerimiento establecido por el Reglamento Europeo de IA (puedes consultar la Guía para la Aplicación del Reglamento Europeo de IA en Salud del Programa Salud/IA), la implementación de capacidades de supervisión humana es una condición fundamental para garantizar la eficacia de la herramienta y la confianza de los profesionales sanitarios y de los pacientes, así como a la adherencia del sistema.

Como ya se ha mencionado, las **funcionalidades de supervisión humana** se enmarcan en el ámbito de la IA explicable (XAI), y conllevan aspectos esenciales del diseño de la herramienta. Esta necesidad es especialmente relevante en los LLM, puesto que su explicabilidad es más compleja en comparación con otros modelos de IA (ver **3.4/ Funciones de explicabilidad**).

Aspectos clave de la supervisión humana

ASPECTO 01 / Formato de la salida de los resultados

Por ejemplo, presentar recomendaciones en forma de listas ordenadas, en lugar de propuestas únicas, puede **facilitar la revisión de las respuestas de la herramienta** por parte del profesional de la salud, como paso previo a utilizar esa información en la toma de decisiones.

ASPECTO 02 / Proceso de obtención de la respuesta

Para facilitar el proceso de revisión al profesional clínico, la herramienta debe proporcionar los **elementos y procesos relevantes** en la generación de la respuesta. Esta funcionalidad puede basarse en métodos de *prompting* o en técnicas de explicabilidad.

ASPECTO 03 / Información influente en la respuesta

Las técnicas basadas en muestras permiten identificar documentos o fragmentos de texto que han influido en su respuesta. Estos fragmentos pueden provenir del entrenamiento, del afinamiento o de un sistema RAG (*Retrieval-Augmented Generation*), del afinamiento o de un sistema RAG(). Incluir esta información en la salida de la herramienta es especialmente útil para ayudar al profesional clínico a revisar sus resultados.

ASPECTO 04 / Generación de avisos per incertidumbre

A menudo, los LLM pueden basar sus respuestas en información incierta, incompleta o incluso inexistente. Implementar técnicas de *prompting* o estrategias basadas en muestras puede ayudar al sistema a detectar estas situaciones y alertar al usuario sobre la necesidad de revisar los resultados.

Un ejemplo representativo es utilizar instrucciones de *prompting* para **indicar al sistema que sólo responda a la consulta si encuentra la respuesta** en la documentación proporcionada; de lo contrario, el sistema debería emitir un aviso indicando que no puede garantizar la fiabilidad de la respuesta, o devolver un mensaje como: 'Datos insuficientes'.

ASPECTO 05 / Interfaz centrada en la revisión

Las interfaces de la herramienta deben estar **adaptadas y optimizadas para que los profesionales puedan revisar y validar fácilmente los resultados generados**. Esto es especialmente relevante en ámbitos críticos como el sanitario, donde la supervisión humana es esencial para garantizar la fiabilidad de las decisiones y la seguridad de los pacientes.

3.6 / Privacidad y protección de datos

Los datos son un componente esencial de una herramienta basada en IA y uno de los factores más determinantes en el diseño de la solución. Por este motivo, es muy importante definir, ya desde las primeras fases del desarrollo, los siguientes aspectos clave:

- **Qué datos procesará la herramienta o estarán implicados en su funcionamiento**: incluye tanto los datos de entrada y salida como los datos utilizados para adaptar el modelo de lenguaje (ver **4.2/ Adaptación del modelo**).
- **Si estos datos son de naturaleza sensible o privada**: este aspecto resulta especialmente relevante en sectores como el de la salud, donde la seguridad y privacidad de los datos son prioritarios.
- **Cuál será el flujo de los datos durante el funcionamiento de la herramienta**: es necesario determinar si estos datos estarán circunscritos a los sistemas de la entidad usuaria en todo momento durante el funcionamiento de la herramienta o, si, por el contrario, necesitarán ser transferidos a servidores externos. Un ejemplo habitual es el uso por parte de la herramienta de un LLM comercial alojado en los servidores de un proveedor (ver **4.1.4/ Modelos locales o remotos**).

Mecanismos aplicables para garantizar la seguridad y confidencialidad de los datos

01	El uso de modelos instalados localmente, que eviten la transferencia de datos a servicios externos
02	La anonimización o pseudoanonimización de datos
03	La encriptación de los datos
04	La gestión segura de accesos
05	La creación de un registro de actividades
06	Garantizar un uso mínimo necesario de datos
07	La elaboración de un protocolo de retención y eliminación de los datos
08	Garantizar a los pacientes el acceso, rectificación y eliminación de datos personales dentro del sistema

El uso de datos en una herramienta de IA está sometido a normativas tales como el Reglamento General de Protección de Datos (RGPD), el Esquema Nacional de Seguridad (ENS) en entidades públicas, o la Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales (LOPDGDD), entre otros. Estas regulaciones exigen establecer mecanismos para garantizar la seguridad y confidencialidad de los datos, especialmente cuando se trata de datos sensibles o privados. Además, el uso de datos sensibles obliga a realizar una **evaluación de impacto relativa a la protección de datos** (DPIA), a fin de identificar y gestionar los riesgos asociados al tratamiento de estos datos con tecnologías avanzadas como la IA generativa.

3.7 / Regulaciones y certificaciones

Ya desde las primeras fases del diseño, es importante tener en cuenta **qué tipo de herramienta se desea desarrollar, si se considerará un producto sanitario, y qué clasificación de riesgo tendrá según las normativas vigentes**. Estos aspectos son importantes porque afectan directamente a las certificaciones necesarias y las regulaciones que deberá cumplir la herramienta antes de poder ser utilizada en entornos clínicos.

Hay que tener presente que el cumplimiento de estos requisitos y la obtención de certificaciones pueden ser **largos y complejos procesos**. Por eso, identificar estas necesidades desde el inicio permite planificar adecuadamente los recursos y los plazos, evitando retrasos en el despliegue final del sistema. Además, esta planificación garantiza que la herramienta se desarrolle de forma alineada con los estándares de calidad, seguridad y eficiencia requeridos por el sector de la salud.

Por ejemplo, si la herramienta se desarrolla para ayudar en el diagnóstico o tratamiento de pacientes, probablemente se clasificará como producto sanitario y deberá cumplir con una serie de condiciones incluidas en el Reglamento Europeo de Dispositivos Médicos (MDR) y la Regulación Europea de IA (EU AI Act). Este proceso implica una lista de requisitos que incluyen, por ejemplo, la realización de evaluaciones clínicas, la obtención de marcado CE y la verificación del cumplimiento de normas armonizadas.

Normativas más relevantes para soluciones basadas en IA en salud

NORMATIVA 01 / Reglamento Europeo de IA (EU AI Act)

Define el marco jurídico para garantizar el uso seguro y ético de la IA en Europa. Clasifica las herramientas de IA por su nivel de riesgo (inaceptable, alto, limitado o mínimo) y establece obligaciones y requisitos para su utilización en el mercado europeo (se puede consultar la Guía para la Aplicación del Reglamento Europeo de IA en Salud del Programa Salud/IA).

NORMATIVA 02 / Reglamentos de Dispositivos Médicos (MDR) y Dispositivos Médicos In Vitro (IVDR)

Imponen estándares de seguridad y eficacia para garantizar que los dispositivos médicos cumplen los requisitos necesarios antes de ser aprobados (se pueden consultar las Guías MDR e IVDR del Programa Salud/IA).

NORMATIVA 03 / Reglamento General de Protección de Datos

Regula el tratamiento de los datos personales de los pacientes, incluidos los datos de salud. Establece condiciones para garantizar la privacidad y la seguridad de los datos sensibles, con requisitos de transparencia, trazabilidad y supervisión humana en decisiones automatizadas (ver **3.6/ Privacidad y protección de datos**).

Otras regulaciones

Es importante considerar otras regulaciones y estándares que pueden aplicarse en función del contexto específico de cada dispositivo o sistema de IA generativa. Algunos ejemplos son:

- Esquema Nacional de Seguridad (ENS Real Decreto 311/2022)
- Directiva sobre la Seguridad General de los Productos (Directiva 2001/95/CE)
- Directiva sobre la Protección de Datos (Directiva UE 2016/680)
- Ley sobre la Protección de Datos en las Instituciones de la UE (Reglamento UE 2018/1725)
- Ley de Ciberseguridad (Reglamento UE 2019/881)
- Ley de Gobernanza de Datos (Reglamento EU 2022/868)
- Ley de Servicios Digitales (Reglamento UE 2022/2065)
- Ley de Mercados Digitales (Reglamento EU 2022/1925)
- Ley de Datos (Reglamento UE 2023/2854)

4 / Implementación

4.1 / Selección del modelo

4.2 / Adaptación del modelo

4.3 / Agentes y herramientas multiagente

4.1 / Selección del modelo

En el desarrollo de herramientas basadas en LLM, a menudo **se parte de modelos ya preentrenados** que reciben el nombre de **modelos fundacionales**. Estos modelos se construyen mediante un proceso de entrenamiento que requiere volúmenes masivos de datos textuales y recursos computacionales extraordinarios. Éste es un procedimiento extremadamente costoso que, en contrapartida, les proporciona una base sólida de comprensión del lenguaje y les dota de un vasto rango de aplicaciones, incluidos en Salud. Esta estrategia no sólo acelera el desarrollo de las herramientas, sino que también mejora la eficiencia y la calidad de sus resultados.

En el desarrollo de este tipo de soluciones, es habitual el uso de más de un LLM que pueden complementarse según las necesidades de cada aplicación. Esta estrategia permite, por ejemplo, combinar modelos de propósito general con modelos más pequeños y especializados, optimizando tanto el rendimiento como la gestión de recursos.

La selección del(los) modelo(s) de lenguaje es un paso fundamental que puede influir de forma significativa en la calidad y la eficacia de la herramienta. A continuación, se presentan los criterios más relevantes para facilitar esta elección.

Criterios para la selección del modelo de lenguaje

CRITERIO 01 / Modalidad de datos

Algunos modelos gestionan datos de un único tipo (p.e., texto o imágenes), mientras que otros tratan datos multi-modales. Debe considerarse el tipo de datos disponibles en el contexto sanitario y el objetivo a alcanzar.

CRITERIO 02/ Rendimiento del modelo

Es importante que el modelo elegido haya demostrado un buen rendimiento (precisión, velocidad de procesamiento, etc.) en tareas similares a las que se desea desarrollar, tanto en la literatura como en la práctica clínica.

CRITERIO 03/ Adaptabilidad

Es la capacidad de un modelo para ser ajustado para realizar tareas específicas. Por lo general, los modelos comerciales ofrecen menos flexibilidad; los modelos de código abierto, sin embargo, son altamente adaptables, pero requieren conocimientos técnicos avanzados.

CRITERIO 04 / Actualizaciones y mantenimiento

Los modelos con apoyo continuo para desarrolladores o comunidades activas son especialmente atractivos, ya que reciben mejoras y actualizaciones constantes que incrementan su eficacia y seguridad.

CRITERIO 05 / Tamaño y complejidad del modelo

La cantidad de parámetros de un modelo determina su capacidad para captar patrones complejos en los datos. Los modelos mayores ofrecen mejor rendimiento, pero requieren más recursos computacionales y económicos; los modelos más pequeños son más eficientes pero pueden presentar limitaciones de rendimiento en algunas tareas.

CRITERIO 06 / Coste e infraestructura

Los modelos comerciales suelen implicar un pago por uso, aumentando el coste a medida que crece el volumen de datos; los modelos de código abierto pueden reducir el coste inicial, pero requieren una mayor inversión en infraestructura y conocimiento técnico para el despliegue y mantenimiento.

CRITERIO 07/ Documentación y apoyo

La disponibilidad de documentación, tutoriales y comunidades activas es un factor a tener en cuenta, ya que facilita la implementación, uso y mantenimiento del modelo.

CRITERIO 08 / Impacto ambiental

Los modelos más grandes suelen requerir una gran cantidad de energía para el entrenamiento y el despliegue, lo que incrementa su huella de carbono.

Teniendo en cuenta estos criterios, es necesario evaluar las diversas opciones de LLM disponibles en el mercado, que se diferencian por **el tipo de licencia** (comercial o de código abierto), **el grado de especialización en el ámbito sanitario y biomédico** (general o adaptado), su **tamaño y complejidad**, así como la posibilidad de desplegarlos en equipos **locales o entornos remotos**.

4.1.1 / Modelos generalistas comerciales

Los modelos comerciales desarrollados por grandes empresas tecnológicas ofrecen un alto rendimiento, son fáciles de utilizar y cuentan con soporte técnico constante por parte de la empresa creadora.

Por lo general, **se comercializan mediante una API de pago por uso**. La persona usuaria tiene poco control sobre el modelo, ya que el entrenamiento y el desarrollo son gestionados íntegramente por la empresa y no es posible modificarlo ni adaptarlo de forma sustancial para usos muy específicos. Aunque son generalistas, pueden mostrar un buen rendimiento en tareas específicas dentro del sector de la salud.

LLM generalistas comerciales (febrero, 2025)

Modelo (desarrollador)	Modalidad de datos	Descripción
GPT-3 / 4 / 4.5 (OpenAI)	Texto / Multimodal	GPT-3 fue uno de los primeros LLM de uso generalizado, suponiendo un gran avance en la generación de texto natural y en la comprensión del lenguaje. GPT-4 supuso su evolución, incorporando capacidades multimodales para procesar texto e imágenes. GPT-4.5 es la versión más reciente y pretende mejorar el rendimiento de sus antecesores.
GPT-4o (OpenAI)	Multimodal	Versión optimizada de GPT-4. Mejora la eficiencia, velocidad y precisión, ofreciendo un rendimiento superior para aplicaciones complejas que requieren comprensión y generación de contenido multimodal.
OpenAI o1 / o3 (OpenAI)	Texto	Son modelos de lenguaje diseñados para abordar problemas complejos mediante un proceso de 'reflexión' antes de generar respuestas. Esta capacidad les permite razonar de forma más efectiva en tareas como programación, matemáticas y ciencias. OpenAI o3 es la versión más avanzada.
Claude 3.5 / 3.7 - Sonnet (Anthropic)	Multimodal	Claude 3.5 es un modelo de lenguaje avanzado y sólido en la generación y comprensión de texto. Claude 3.7 mejora su razonamiento con un enfoque híbrido que combina respuestas rápidas con un análisis detallado de problemas complejos.
Command R (Cohere)	Texto	Modelo diseñado específicamente para la generación aumentada por recuperación (RAG), combinando información de fuentes externas con generación de texto para mejorar la precisión y relevancia.
Gemini 1 /1.5 / 2 Flash (Google DeepMind)	Multimodal	Gemini 1 introdujo un modelo multimodal centrado en la comprensión y generación del lenguaje natural. Gemini 1.5 mejoró la eficiencia y la capacidad de procesamiento, permitiendo ventanas de contexto más amplias. Gemini 2.0 Flash optimiza aún más la velocidad y los recursos computacionales, haciéndolo ideal para tareas complejas en tiempo real.
PaLM 1 / 2 (Google)	Texto	Modelos de lenguaje con habilidades avanzadas de comprensión multilingüe, razonamiento lógico y mejoras en ámbitos como la medicina y las ciencias naturales. PaLM 2 es la última versión.
PaLM-E (Google)	Multimodal	Extensión de la serie PaLM con capacidades multimodales, integrando texto con datos visuales como imágenes y vídeos. Diseñado especialmente para aplicaciones en robótica y análisis de datos visuales.

4.1.2 / Modelos generalistas de código abierto

Los modelos generalistas de código abierto pueden ser creados tanto por comunidades de desarrolladores como por empresas. **Su naturaleza abierta ofrece una gran flexibilidad**, lo que permite a los usuarios un mayor control sobre el modelo desde el proceso de entrenamiento hasta su adaptación, facilitando su especialización para aplicaciones sanitarias y biomédicas.

Sin embargo, estos modelos también presentan retos importantes, como la falta de soporte técnico dedicado, los altos requerimientos de recursos computacionales, la necesidad de conocimientos técnicos avanzados para su implementación y las posibles variaciones en la calidad de su rendimiento.

SLM: *Small Language Model* (ver **4.1.5/ Pequeños Modelos de Lenguaje (SLM) y técnicas de compresión**).

LLM generalistas de código abierto (febrero, 2025)

Modelo (desarrollador)	Modalidad de datos	Licencia de código abierto	Descripción
Mistral 7b (Mistral AI)	Texto	Apache 2.0	Modelo que destaca por su rendimiento comparado con su tamaño relativamente pequeño de 7.000 millones de parámetros, ofreciendo una gran eficiencia en múltiples tareas. Es un SLM.
Mistral NeMo (Mistral AI)	Texto	Apache 2.0	Modelo más reciente de la suite de modelos de Mistral, adaptado a necesidades específicas en varios dominios de procesamiento de texto.
LLaMA 1 / 2 / 3.2 (Meta)	Texto	Licencia personalizada de Meta	LlaMa 1 fue uno de los primeros modelos de Meta enfocados a investigación, que se popularizó por su accesibilidad. LLAMA 2 amplió sus capacidades, incluyendo versiones SLM (7B). LLAMA 3.2 mejora su rendimiento y está optimizado para aplicaciones más eficientes.
Dolly 2.0 (Databricks)	Texto	Apache 2.0	Modelo diseñado para ser fácilmente utilizable y accesible para la comunidad de desarrolladores, lo que permite una sencilla implementación en diversas aplicaciones.
GPT-Neo (Eleuther AI)	Texto	Apache 2.0	Uno de los primeros esfuerzos de la comunidad de código abierto por crear una alternativa al GPT-3, centrado en la transparencia y la replicabilidad en el procesamiento del lenguaje natural.
GPT-J (Eleuther AI)	Texto	Apache 2.0	Sucesor de GPT-Neo, con 6.000 millones de parámetros, ofrece una mejor capacidad de generación de texto con un modelo preentrenado más optimizado. Es un SLM.
T5 (Google)	Texto	Apache 2.0	Modelo de Google diseñado para múltiples tareas de lenguaje natural, tales como traducción y respuestas a preguntas.
DeepSeek-V2 / V3 / R1 (DeepSeek AI)	Multi-modal	DeepSeek (con restricciones de uso responsable)	Destaca por su escalabilidad y entrenamiento en múltiples idiomas. La versión V3 mejora su velocidad y rendimiento, mientras que R1 se centra en el razonamiento lógico y matemático. Restringe la generación de contenido en temáticas que los desarrolladores consideran sensibles.

4.1.3 / Modelos adaptados al ámbito sanitario y biomédico

Existen varios LLM que ya han sido adaptados específicamente para su uso en el sector de la salud. Esto es posible gracias a un proceso de afinación (*fine-tuning*), que consiste en ajustar un modelo general con datos específicos de un ámbito, en este caso clínicos o biomédicos, para mejorar su rendimiento en tareas concretas dentro de este ámbito (ver **4.2/ Adaptación del modelo).**

Gracias a esta adaptación, los modelos pueden comprender mejor conceptos técnicos, terminología especializada, jerga médica y acrónimos que no se encuentran habitualmente en los conjuntos de datos de entrenamiento generalistas. Esto les hace especialmente útiles para aplicaciones, como el análisis de textos médicos, la interpretación de resultados clínicos, la generación de resúmenes de publicaciones científicas y el soporte en la toma de decisiones clínicas.

Es importante tener en cuenta que estos modelos pueden estar sujetos a distintos tipos de licencias, tales como de código abierto o comerciales.

LLM adaptados al ámbito sanitario y biomédico (febrero, 2025)

Modelo (desarrollador)	Modalidad de datos	Licencia	Descripción
Med-PaLM[4]	Texto	Comercial (Google)	Med-PaLM es una serie de modelos de Google adaptados al dominio médico, de la que Med-PaLM 2 es la versión más reciente y con mejor rendimiento.
Med-PaLM M[5]	Multi-modal	Comercial (Google)	Versión avanzada del Med-PaLM, capaz de integrar datos textuales, imágenes y datos genómicos, entre otras modalidades, para aplicaciones biomédicas.
MedLLaMA	Texto	De código abierto	Serie de modelos LLaMA adaptados a aplicaciones en salud, como la respuesta a preguntas médicas, resúmenes de historiales médicos y soporte en diagnóstico clínico.
BiomedGPT[6]	Multi-modal	De código abierto	Familias de modelos adaptados para la generación de texto científico y médico, ofreciendo herramientas útiles para investigadores y profesionales de la salud.
LLaVA-Med[7]	Multi-modal	De código abierto	Serie de modelos enfocados al análisis de imágenes médicas e informes clínicos, permitiendo la integración de información visual y textual para tareas biomédicas.

4.1.4 / Modelos locales o remotos

Probablemente, la decisión con mayor impacto en el funcionamiento operativo de una solución basada en LLM será la localización física del modelo.

- En un **sistema local**, el modelo se instala y ejecuta dentro de la infraestructura de la organización que lo utiliza.
- Sin embargo, en un **sistema remoto** el modelo se ejecuta en servidores externos, habitualmente alojados en la nube y gestionados por un proveedor tercero. Esta arquitectura implica que los datos de entrada y salida se transfieren a través de la red entre los servidores del usuario y los del proveedor del modelo de lenguaje.

Existe una amplia gama de opciones entre modelos locales y remotos, y la elección entre estas alternativas no sólo afecta a aspectos técnicos, como la latencia o la gestión de recursos, sino también cuestiones fundamentales como la privacidad, la seguridad de los datos y el cumplimiento normativo.

Criterios a considerar en la elección entre modelos locales y remotos

CRITERIO 01 /
Rendimiento y accesibilidad

- Los **modelos locales** son ideales para aplicaciones que requieren baja latencia o acceso constante sin dependencias de conexiones a internet.
- Los **modelos en remoto** ofrecen mayor flexibilidad y escalabilidad, especialmente para aplicaciones que necesitan actualizaciones constantes o acceso desde múltiples ubicaciones.

CRITERIO 02 /
Costes e infraestructuras

- Los **modelos locales** requieren inversiones significativas en infraestructuras propias, tales como servidores, redes y soporte técnico, así como en su mantenimiento. Estas necesidades pueden incrementar los costes iniciales en las primeras fases del desarrollo, aunque esta inversión puede resultar rentable en entornos con un uso intensivo.
- El uso de **modelos remoto** elimina la necesidad de invertir en infraestructura, pero puede comportar mayores costes recurrentes dependiendo del volumen de datos procesados.

CRITERIO 03 /
Seguridad y privacidad de los datos

- En los **modelos locales**, los datos no salen de la organización usuaria y, por tanto, se reduce el riesgo de filtraciones o vulneraciones de privacidad. Esta opción puede ser especialmente importante en el sector salud, donde la información es altamente sensible.
- Los **modelos en remoto**, sin embargo, implican un flujo de datos por la red que requieren garantías robustas de seguridad por parte del proveedor, que pueden incluir la anonimización o pseudoanonimización y el cifrado avanzado de los datos. Estos modelos pueden estar sujetos a las leyes del país donde se encuentran sus servidores, lo que puede complicar o impedir el cumplimiento del RGPD y otras normativas del país de la organización usuaria (ver **3.6/ Privacidad y protección de datos**)

4.1.5 / Pequeños Modelos del Lenguaje (SLM) y técnicas de compresión

Los **pequeños modelos de lenguaje** *(Small Language Model*, **SLM**) son versiones más pequeñas y ligeras creadas a partir de LLM, y diseñadas para ofrecer un **equilibrio entre rendimiento y eficiencia**.

A pesar de la reducción en el número de parámetros y en la capacidad computacional requerida (en comparación con los LLM), los SLM pueden alcanzar un rendimiento similar en tareas específicas cuando se someten a un proceso de adaptación mediante *Retrieval Augmented Generation* (RAG) o afinamiento (*fine-tuning;* ver **4.2/ Adaptación del modelo**).

Los SLM son especialmente útiles en entornos como hospitales, donde la seguridad de los datos es prioritaria y las limitaciones de infraestructura pueden hacer inviable el uso de modelos de mayor tamaño.

El proceso de creación de un SLM implica la aplicación de una o varias **técnicas de optimización del rendimiento**.

Técnicas de optimización del rendimiento

TÉCNICA 01 /
Distilación (*distillation*)

La distilación consiste en entrenar un modelo más pequeño (*student model*) para que imite el comportamiento de un modelo más grande (*teacher model*), de forma que se mantenga un rendimiento similar en tareas generales.

TÉCNICA 02 /
Cuantización (*quantization*)

La cuantización consiste en reducir la precisión numérica de los pesos y cálculos del modelo (p.e., pasando de 32 bits a 8 bits), disminuyendo su tamaño y consumo de recursos manteniendo el rendimiento general del modelo.

TÈCNICA 03 /
Podamiento (*pruning*)

Durante el podamiento se eliminan componentes no esenciales del modelo, como capas o conexiones neuronales con baja contribución a la labor de interés, con el fin de optimizar su rendimiento sin sacrificar significativamente su eficacia.

Cabe mencionar que, aunque estas técnicas de comprensión son necesarias para crear un SLM, también se utilizan para optimizar el rendimiento de modelos grandes en entornos con limitaciones computacionales, o bien para incrementar su velocidad de uso sin necesidad de reducir el tamaño del modelo.

4.2 / Adaptación del modelo

Una vez seleccionados los modelos, es importante valorar si es necesario someterlos a un proceso de adaptación para **especializarlos en un contexto o funcionalidad concreta** dentro del ámbito en el que se quiere desplegar la herramienta, en éste el sanitario o biomédico, con el objetivo de mejorar su eficacia.

Los modelos adaptados suelen ofrecer un mejor rendimiento en tareas específicas para las que han sido ajustados, pero también suelen requerir costes adicionales en datos, tiempo y recursos.

Los **métodos de adaptación** más comunes, ordenados según el grado de adaptación, el rendimiento esperado y los esfuerzos y recursos requeridos, son:

01 Ingeniería de *prompts*

02 Generación Aumentada por Recuperación

03 Afinamiento

4.2.1 / Ingeniería de *prompts* (*prompt engineering*)

El *prompting* es una técnica esencial en el ámbito de la IA generativa, y constituye la forma más habitual de interactuar con un modelo de lenguaje. Consiste en **formular instrucciones o preguntas específicas dirigidas al modelo**, de complejidad y extensión muy variable, y que han sido previamente diseñados y orientados al objetivo que se quiere alcanzar.

Aunque no altera los parámetros del modelo base, la calidad de los *prompts* tiene un impacto directo en la relevancia y utilidad de las respuestas generadas. Siguiendo los principios establecidos por Phoenix y Taylor (2024)9, podemos establecer **cinco directrices básicas para la generación de *prompts***:

01	Dar dirección
02	Describir el estilo y la persona
03	Especificar el formato
04	Proporcionar ejemplos
05	Dividir el trabajo

DIRECTRIZ 01 /
Dar dirección

Define el objetivo del *prompt* de forma clara, específica, sin ambigüedades y proporcionando el contexto necesario para la tarea. Este nivel de detalle ayuda al modelo a alinearse mejor con las expectativas del usuario y evita respuestas imprecisas o dispersas.

Por ejemplo, en lugar de escribir simplemente "Escribe una lista de recomendaciones médicas", se puede especificar:

```
> Escribe una lista de 5 recomendaciones para un paciente con hipertensión arterial, centrándote en hábitos de vida saludables y evitando recomendaciones farmacológicas

Salida:

Recomendaciones para un paciente con hipertensión arterial:

1. Reducir el consumo de sal: Limita la ingesta de sal a menos de 5 gramos al día. Evita alimentos procesados ricos en sodio y opta por condimentos naturales como hierbas y especias.
2. Aumentar el consumo de frutas y verduras...
```

DIRECTRIZ 02 /

Describir el estilo y la persona

Indica el **tono, registro lingüístico** y, en su caso, **el rol** que el modelo debe adoptar.

Puede pedirse, por ejemplo, que adopte un rol de clínico especialista en oftalmología, y que utilice un estilo comprensible para el público general.

```
> Responde a las siguientes preguntas como si fueras un clínico especialista en oftalmología, pero utilizando un estilo comprensible para el público general. ¿Qué es el glaucoma y cómo afecta a la visión?

Salida:

El glaucoma es una enfermedad que daña el nervio óptico, que es como un cable que envia las imágenes desde el cerebro...
```

DIRECTRIZ 03 /

Especificar el formato

Define con precisión la **estructura de la respuesta** que esperas. Esto puede incluir formatos específicos como una lista numerada con un separador determinado, tablas, respuestas en formato JSON o párrafos estructurados. Esta directriz es especialmente útil para garantizar la consistencia y facilitar la interpretación de las respuestas.

```
Escribe una lista de 5 hábitos saludables para prevenir la diabetes tipo 2, centrándote en hábitos de vida saludables y evitando recomendaciones farmacológicas. Escribe las respuestas en una tabla en formato JSON, donde la primera columna será la recomendación y la segunda será una breve descripción de dicha recomendación.

Salida:

[
  {
  "Recomendación": "Mantener un peso saludable",
  "Descripción": "Controlar el peso corporal para reducir la resistencia a la insulina y evitar el exceso de grasa corporal, que puede aumentar el riesgo de diabetis.“
 }, ...
```

DIRECTRIZ 04 /

Proporcionar ejemplos

Incluir **ejemplos de respuestas óptimas o casos bien resueltos** dentro del prompt es una técnica muy efectiva para mostrar al modelo el tipo de salida deseada, especialmente en tareas complejas o poco habituales.

Esta estrategia constituye la base del prompting few-shot (explicada más adelante), y también se utiliza en técnicas avanzadas como la Generación Aumentada por Recuperación de Información (**RAG:** ver **4.2.2 / Generación Aumentada por Recuperación de Información**).

DIRECTRIZ 05 /

Dividir el trabajo

Para tareas complejas, **divide el proceso en etapas sencillas e interconectadas**. Por ejemplo, en el campo médico, un primer *prompt* podría identificar síntomas, un segundo sugerir posibles diagnósticos y un tercero proponer pruebas diagnósticas. Esta estrategia está detrás de técnicas como el *chain-of-thought* o el *tree-of-thought* explicadas a continuación, que permiten al modelo simular un razonamiento estructurado y secuencial.

Técnicas de ingeniería de prompts

Las técnicas de ingeniería de *prompts* son estrategias diseñadas para **optimizar la interacción con los LLM y obtener respuestas más precisas, relevantes y ajustadas a las necesidades de una tarea específica**. Estas técnicas pueden considerarse una forma de adaptación ligera del modelo, ya que permiten ajustar su comportamiento sin necesidad de modificar sus parámetros. Además, el uso adecuado de *prompts* puede contribuir a mejorar la explicabilidad del modelo, ya que facilita la comprensión del proceso por el cual se generan las respuestas.

Las técnicas más destacadas de la ingeniería de *prompts* son:

01	*Zero-shot*
02	*Few-shot*
03	Cadena de pensamiento *(chain of thought)*
04	*Ensemble refinament*
05	Árbol de pensamiento (*tree of thought*)

TÉCNICA 01 /
Zero-shot

Consiste en pedir al modelo que ejecute una tarea sin proporcionarle ejemplos específicos previos, **basándose únicamente en su conocimiento general adquirido durante el entrenamiento**.

Es un proceso sencillo y directo, útil para **tareas simples**.

> ➢ ¿Cuáles son los principales efectos secundarios de la medicación antihipertensiva?
>
> Salida:
>
> Los principales efectos secundarios de los medicamentos antihipertensivos pueden incluir mareos, fatiga, dolor de cabeza, náuseas, disfunción eréctil y alteraciones electrolíticas.

TÉCNICA 02 /
Few-shot

Consiste en proporcionar al modelo varios ejemplos para que entienda el tipo de respuesta deseada. Se basa en los principios de **"dar dirección", "especificar el formato" y "proporcionar ejemplos"** mencionados anteriormente. Es especialmente útil en situaciones donde se requiere un cierto nivel de **personalización o un formato específico** en las respuestas.

> ➢ Enumera los síntomas de una patología ordenados por frecuencia (porcentaje) y separados por una coma. Aquí algunos ejemplos:
>
> Ejemplo 1
>
> Patología: Gripe.
>
> Síntomas: Fiebre (80%), malestar general (70%), tos seca (60%), dolor muscular (50%).
>
> Ejemplo 2
>
> Patología: Gastroenteritis viral.
>
> Símptomes: Diarrea (90%), náuseas (80%), dolor abdominal (70%), fiebre baja (30%).
>
> Ahora, para la patología "apendicitis aguda", indica los síntomas ordenados por frecuencia e incluye el porcentaje correspondiente.
>
> Salida:
>
> Patologia: Apendicitis aguda.
>
> Síntomas: Dolor abdominal en el cuadrante inferior derecho (90%), fiebre (80%), náuseas (75%), vómitos (50%), pérdida de apetito (40%).

TÉCNICA 03 /

Cadena de pensamiento *(chain of thought)*

Consiste **en guiar al LLM para que genere una respuesta paso a paso**, desglosando el razonamiento en etapas más sencillas y estructuradas, en lugar de proporcionar una respuesta directa. Esto mejora la calidad y precisión de las respuestas del modelo y contribuye a su explicabilidad, lo que aumenta la confianza en los resultados generados.

Esta técnica se basa en el principio de **"dividir el trabajo"**, y es útil para **tareas que requieren razonamientos secuenciales**, como problemas matemáticos o de lógica.

- Un paciente de 25 años llega con fiebre, dolor abdominal intenso y vómitos. Razona paso a paso y en voz alta para determinar posibles diagnósticos, teniendo en cuenta los síntomas, su distribución y posibles condiciones asociadas. Enumera posibles diagnósticos y el siguiente paso de acción.

Salida:

Paso 1: Los síntomas presentados son fiebre, dolor abdominal intenso y vómitos. La fiebre sugiere una causa infecciosa o inflamatoria. El dolor abdominal y los vómitos pueden estar relacionados con diversas condiciones gastrointestinales.

Paso 2: Consideramos el dolor abdominal. Si está localizado en el cuadrante inferior derecho, es probable que se trate de una apendicitis aguda. Si el dolor es difuso y va acompañado de diarrea, podría tratarse de una gastroenteritis.
Paso 3: Los vómitos son habituales en ambas condiciones, pero también pueden aparecer en casos como la obstrucción intestinal, donde el dolor suele ser más severo y progresivo.
Paso 4: Teniendo en cuenta la edad del paciente y los síntomas, proponemos los siguientes diagnósticos probables, ordenados según su probabilidad:

1. Apendicitis aguda: Muy probable si el dolor está localizado en el cuadrante inferior derecho y va acompañado de fiebre y pérdida de apetito.
2. Gastroenteritis viral: Posible si el dolor es difuso y hay diarrea.
3. Obstrucción intestinal: Posible si el dolor es cólico y se asocia con distensión abdominal y ausencia de gases o heces.

Paso 5: Se recomienda realizar una exploración física completa (incluyendo el signo de Blumberg), un análisis de sangre (para buscar leucocitosis) y una ecografía abdominal para confirmar el diagnóstico.

TÉCNICA 04 /

Ensemble refinement

Técnica que mejora la calidad y fiabilidad de las respuestas de un modelo de lenguaje mediante la **generación de múltiples resultados independientes** a partir de una misma entrada (prompt), aprovechando la variabilidad intrínseca del modelo, o a partir de pequeñas modificaciones en las instrucciones. Estas respuestas se comparan y se agregan mediante diferentes estrategias, como resúmenes, ponderaciones numéricas, selección por voto mayoritario (también conocido como **autoconsistencia** o *self-consistency*) o refinamiento secuencial.

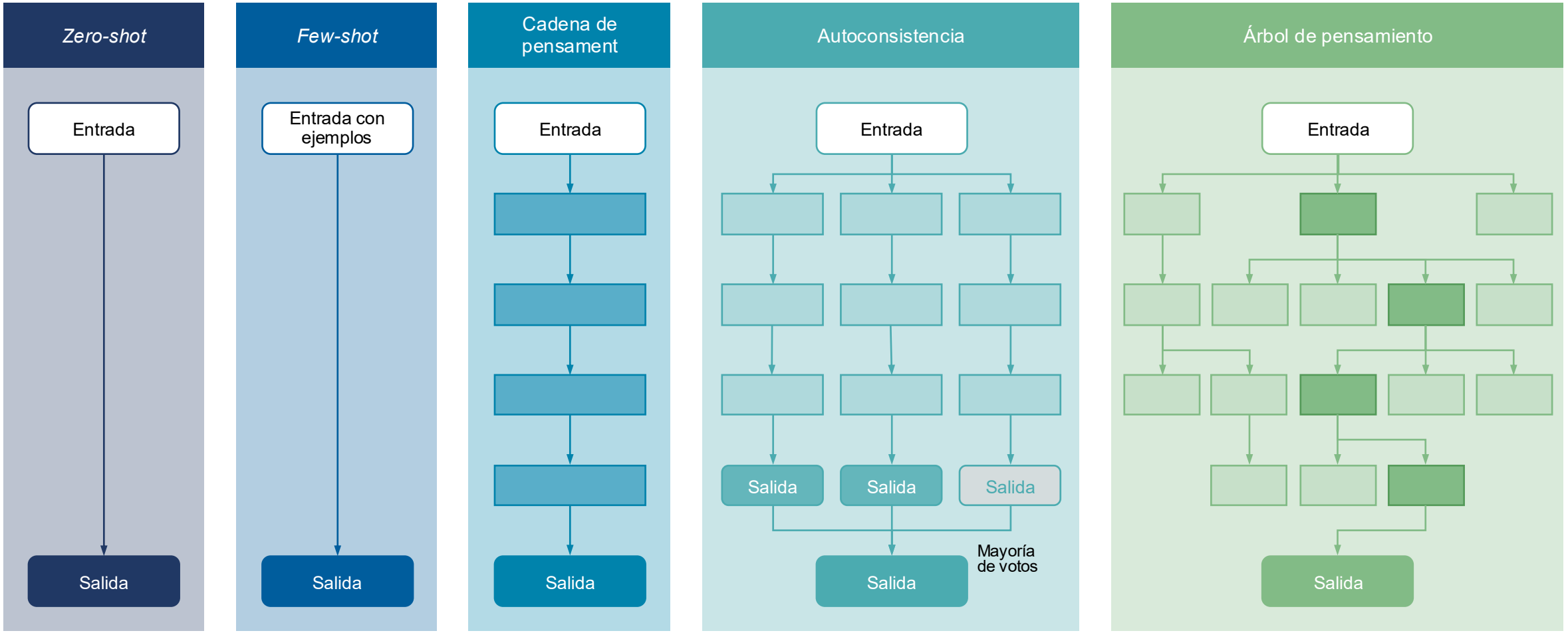
TÉCNICA 05 /

Árbol de pensamiento *(tree of thought)*

Estrategia avanzada que amplía la técnica de cadena de pensamiento para mejorar el proceso de razonamiento de los modelos de lenguaje en tareas complejas que implican múltiples pasos. En lugar de seguir un único camino lineal de razonamiento, esta técnica **estructura el proceso en varias ramificaciones que forman un árbol lógico**, lo que permite al modelo explorar múltiples opciones o hipótesis antes de llegar a una respuesta final basada en la solución más consensuada o coherente.

Visión general de las técnicas de *prompting*. Figura adaptada de Yao et al. (2023)[10]

La figura muestra las diferentes estrategias de *prompting* presentadas anteriormente. Cada rectángulo representa un paso intermedio en el proceso de razonamiento del modelo, ayudándole a desglosar y resolver problemas de forma estructurada. Hay que tener en cuenta que la implementación de técnicas avanzadas como auto-consistencia, ensamble refinamiento o árbol de pensamiento, requiere el uso de herramientas programáticas especializadas para automatizar y optimizar el proceso, como *LangChain, LlamaIndex o Haystack*.



4.2.2 / Generación Aumentada por Recuperación de Información (RAG)

La Generación Aumentada por Recuperación (*Retrieval-Augmented Generation*, RAG) es una técnica que **integra la capacidad generativa de los modelos de lenguaje con la recuperación de información en tiempo real de fuentes externas al modelo**. Esta combinación permite que el modelo acceda a **datos actualizados o específicos en el momento de la consulta**, lo que mejora la precisión y relevancia de las respuestas o textos generados.

Como inconveniente, esta estrategia **requiere una infraestructura adicional** para gestionar la recuperación de información en tiempo real. A diferencia de un LLM sin adaptar, el RAG debe consultar constantemente bases de datos o sistemas de información externos, lo que puede aumentar la latencia y los costes operativos, **complicando la escalabilidad y la gestión técnica del sistema**.

En la técnica RAG, el proceso para generar respuestas a partir de una consulta se desarrolla en dos fases:

01	Fase de recuperación
02	Fase de generación

FASE 01 /
Fase de Recuperación

Ante una consulta, el sistema **busca información relevante en un conjunto definido de datos externos al modelo** (p. ej., información clínica, artículos científicos, guías clínicas o recursos educativos para pacientes), para encontrar información útil relacionada con la consulta. Estas fuentes pueden ser conjuntos de datos fijos o dinámicos, y pueden estar localizadas localmente o estar disponibles en la red.

Durante esta fase:

1. **La consulta se transforma en un vector** de alta dimensionalidad mediante modelos de representación vectorial (embeddings) como BERT o sentence transformers, capturando su significado semántico.
2. **Se compara el vector** de la consulta con los vectores de los documentos de bases de datos vectorizadas e indexadas. Para ello, se utilizan algoritmos de búsqueda de vectores, como KNN (k-Nearest Neighbors), Annoy o FAISS.
3. **Se recuperan los documentos más relevantes** que están más cercanos al vector de consulta, generalmente en función de alguna métrica de similitud como el coseno, que mide cuán similares son las direcciones de los vectores entre sí.

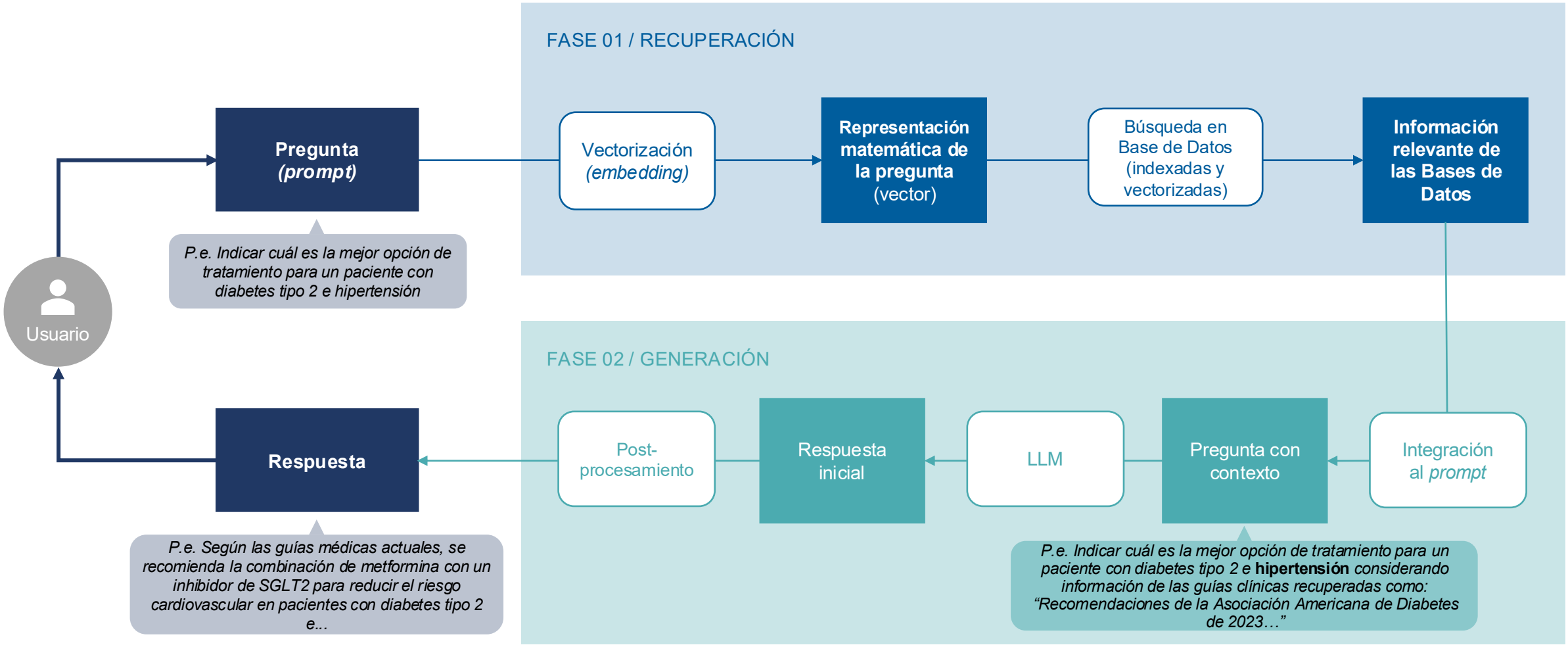
FASE 02 /
Fase de Generación

Una vez se ha recuperado la información relevante para la consulta, se inicia la **generación de la respuesta**. En esta fase, el sistema **integra la información recuperada como contexto de la consulta original** para generar una respuesta coherente, precisa y adecuada a dicho contexto.

Durante esta fase:

1. **La información recuperada se integra como contexto en el prompt** que se pasará al LLM para realizar la consulta. Esto puede incluir resúmenes de los documentos, fragmentos relevantes o, incluso, toda la información de los documentos.
2. **Se elabora una respuesta final** más precisa, coherente y alineada con la consulta inicial del usuario utilizando el contexto adicional.
3. **Se aplican mecanismos de control** para garantizar la fiabilidad antes de generar una respuesta final. Estos mecanismos incluyen: el filtrado de contenido, para asegurar que solo se basa en la información recuperada; la verificación de referencias, para mejorar la trazabilidad; la adaptación al usuario final, el ajuste del formato y el nivel de detalle; y la evaluación de confianza, para alertar en caso de disponer de información insuficiente o ambigua.

Proceso de Generación Aumentada por Recuperación de Información (RAG)



4.2.3 / Afinamiento *(fine-tuning)*

El afinamiento (*fine-tuning*) consiste en **ajustar un modelo preentrenado mediante un conjunto específico de datos**, en este caso procedentes del ámbito sanitario. Este proceso permite al modelo adaptarse mejor a tareas concretas de este ámbito, mejorando su capacidad de entender y generar contenido especializado.

En este punto, es importante diferenciar entre:

- **Preentrenamiento**: consiste en entrenar el modelo con grandes volúmenes de datos, generalmente no etiquetados, para dotarle de un conocimiento general del lenguaje.
- **Afinamiento**: adapta el modelo preentrenado (fundacional) a tareas concretas con conjuntos de datos más pequeños y etiquetados, utilizando menos recursos computacionales y mejorando el rendimiento en aplicaciones específicas.

Estrategias por el afinamiento

Existen diversas **estrategias de afino** que dependen del tipo de **datos disponibles** y de los objetivos de la solución. En términos generales, estas estrategias se pueden clasificar en:

01	Afinamiento no supervisado
02	Afinamiento supervisado

ESTRATEGIA 01 /

Afinamiento no supervisado

Consiste en **exponer el modelo a textos no etiquetados**, como artículos científicos o guías clínicas. Esta estrategia permite mejorar la capacidad del modelo para comprender la terminología clínica, haciéndolo más eficiente en la **generación de textos especializados** como informes médicos o resúmenes de literatura científica. Sin embargo, esta estrategia presenta limitaciones a la hora de realizar tareas de clasificación (p. ej., diagnóstico o estratificación del riesgo), ya que el modelo no dispone de datos etiquetados que le permitan identificar patrones específicos de enfermedades o condiciones de los pacientes.

ESTRATEGIA 02 /

Afinamiento supervisado

Requiere **datos etiquetados**, como textos médicos asociados a diagnósticos o tratamientos específicos. Es especialmente útil para entrenar modelos en **tareas de clasificación médica** como, por ejemplo, el diagnóstico automático de enfermedades o la generación de propuestas personalizadas de planes de tratamiento. A pesar de su potencial, la creación de datos etiquetados puede ser costosa y lenta, ya que a menudo requiere la intervención de expertos médicos para garantizar la calidad y la coherencia de las etiquetas.

Técnicas de afinamiento

Las **técnicas de afinamiento** son metodologías concretas que implementan las estrategias definidas anteriormente, con el objetivo de adaptar modelos preentrenados a tareas específicas. Las técnicas más frecuentes son:

01	Reentrenamiento completo *(full fine-tuning)*
02	*Parameter-Efficient Fine-Tuning* (PEFT)
03	Aprendizaje por reforzar con *feedback* humano

TÉCNICA 01 /

Reentrenamiento completo *(full fine-tuning)*

Técnica tradicional de ajuste, que implica **reentrenar todos los parámetros de un modelo** utilizando **datos específicos** de la tarea o el dominio de interés.

Proporciona la máxima adaptación del modelo a tareas concretas, y puede mejorar significativamente el rendimiento en dominios altamente especializados. Sin embargo, es un procedimiento **muy costoso y exigente en términos de tiempo, datos etiquetados y recursos computacionales**, ya que a menudo requiere infraestructuras con GPU o TPU.

TÉCNICA 02 /

Parameter-Efficient Fine-Tuning (PEFT)

Las PEFT son las técnicas más utilizadas actualmente **en contextos donde la eficiencia computacional es prioritaria**. Se caracterizan por modificar solo una parte del modelo, manteniendo la mayor parte de sus parámetros intactos. Esto facilita la adaptación a nuevas tareas con un coste computacional significativamente más bajo[11].

Las técnicas de PEFT más populares son:

- LoRA *(Low-Rank Adaptation)*
  Ajusta solo una parte de las matrices de peso del modelo preentrenado, en lugar de reentrenar todas sus capas. Esto reduce drásticamente el número de parámetros que es necesario modificar, haciendo que el ajuste sea más rápido y eficiente sin comprometer significativamente la precisión. Es especialmente útil en LLM de gran escala.
- Adaptadores *(adapters)*
  Son pequeños **módulos adicionales** que se incorporan a varias capas del modelo, y que se **entrenan de manera independiente** mientras el resto de los parámetros se mantienen congelados. Esta técnica conserva el conocimiento general del modelo preentrenado y, al mismo tiempo, permite el aprendizaje en tareas específicas con nuevos datos, evitando la necesidad de reentrenar el modelo por completo.
- *Prefix tuning*
  Añade una **secuencia corta de parámetros ajustables** (un "prefijo") **a la entrada del modelo**, que guía su respuesta para tareas concretas. Como solo se ajusta este prefijo y no sus capas, el resto del modelo queda inalterado. Este enfoque es especialmente eficiente, ya que reduce significativamente los costes computacionales y el tiempo necesario para adaptar el modelo a nuevas tareas.

TÉCNICA 03 /

Aprendizaje por refuerzo con feedback humano *(Reinforcement Learning with Human Feedback,* RLHF*)*

El RLHF es una técnica avanzada de **ajuste que adapta las respuestas del modelo para alinearlas con las preferencias y expectativas** de los usuarios.

Es especialmente útil para **optimizar la experiencia del usuario** en aplicaciones prácticas, ya que permite generar respuestas más naturales y adaptadas a tareas específicas. Sin embargo, también implica **procesos costosos y largos**, al requerir la intervención de expertos humanos para evaluar las respuestas, y conlleva el riesgo de que el modelo priorice respuestas agradables o conformes a las expectativas de los usuarios por encima de su veracidad (ver **5.3/ Detección y mitigación de alucinaciones**).

Este proceso se desarrolla en tres fases principales:

1. **Generación de respuestas**: basándose en su conocimiento preentrenado, el modelo genera diversas respuestas para un conjunto de entradas específicas.
2. **Evaluación humana**: un grupo de expertos o usuarios valora u ordena las respuestas generadas en función de su calidad, relevancia y utilidad.
3. **Refuerzo**: las valoraciones se convierten en una función de recompensa que se utiliza para ajustar el modelo, mejorando la calidad de sus salidas de forma iterativa.

4.2.4 / ¿Cómo elegir el método de adaptación?

La adaptación es un proceso necesario cuando se requiere un cierto grado de especialización en nuestra herramienta para optimizar su eficacia.

Los modelos más adaptados suelen tener un mejor rendimiento para la tarea específica para la cual han sido diseñados, pero también suelen requerir costes adicionales en datos, tiempo y recursos.

Cabe destacar que existe una gama cada vez mayor de modelos que ya han sido afinados para mejorar su rendimiento en el ámbito biomédico o asistencial. En función de los objetivos definidos, la eficacia de la herramienta podría beneficiarse del alto grado de especialización que ofrecen estos modelos (ver **4.1.3 / Modelos adaptados al ámbito sanitario y biomédico**).

A continuación, se presentan los aspectos clave de los tres métodos de adaptación descritos anteriormente, así como una tabla comparativa con sus características principales.

Aspectos clave de los métodos de adaptación

Ingeniería de *prompts*	Es generalmente preferible cuando se quiere aumentar la eficacia del modelo sin modificarlo, simplemente orientando su comportamiento en las respuestas de una manera flexible y con una inversión mínima de recursos.
RAG	Aunque requiere más recursos, es una opción superior para aplicaciones que necesitan acceder a un conjunto definido de datos externos al modelo, que pueden ser dinámicos y estar sometidos a revisiones y actualizaciones periódicas. Esto es especialmente útil en entornos clínicos, donde los avances científicos pueden obligar a una revisión frecuente de las fuentes de información (p. ej., guías médicas, estrategias de tratamiento, nuevos fármacos, etc.).
Afinamiento (*fine-tuning*)	Es la opción más adecuada si se requiere un alto nivel de adaptación del modelo para una tarea concreta, integrando conocimientos de un dominio específico del ámbito sanitario. Este método permite un mayor control del comportamiento del modelo, teniendo como contrapartida la necesidad de grandes volúmenes de datos etiquetados del dominio en cuestión y de altos costes computacionales.

Tabla comparativa de los métodos de adaptación de los modelos de IA: Ingeniería de *prompts*, RAG y Afinamiento

Criterio	Ingeniería de *prompts*	RAG	Afinamiento *(fine-tuning)*
Rendimiento	Depende de la calidad del *prompt*, pero generalmente menor.	Alto para tareas que requieren acceso a información dinámica.	Muy alto, especialmente para tareas específicas.
Recursos necesarios	Muy bajos: no requiere entrenamiento ni infraestructuras adicionales.	Moderados: infraestructura adicional para gestionar la recuperación de información, y creación de un índice eficiente.	Altos: requiere capacidad computacional significativa para entrenar el modelo.
Requerimiento de datos	Solo se basa en los datos ya incluidos en el modelo preentrenado.	Necesidad de un conjunto generalmente amplio de documentos relevantes no estructurados, como por ejemplo publicaciones científicas o guías clínicas.	Gran cantidad de textos etiquetados o estructurados y específicos del dominio.
Transparencia e interpretabilidad	Moderada: el comportamiento del modelo está directamente relacionado con el *prompt*.	Moderada: fácil de identificar qué datos han influido en la respuesta (mediante el índice).	Baja: el modelo ajusta los pesos internamente y puede ser difícil rastrear cómo se toman las decisiones.
Alucinaciones	Altas: el modelo puede inventar respuestas basadas en patrones del entrenamiento original.	Bajas, ya que el modelo se sustenta en información recuperada, pero depende de la calidad de los datos.	Reducidas si se entrena con datos de calidad; posibles si el conjunto de datos es incompleto o ruidoso.
Riesgo de sesgos	Moderado o alto: dependen completamente de los sesgos del modelo preentrenado.	El riesgo de sesgos depende de la diversidad y representatividad de las fuentes de datos utilizadas.	Depende de la diversidad y representatividad de los datos utilizados para el afinamiento.
Control sobre el modelo	Moderado: los *prompts* orientan las respuestas del modelo.	Moderado: el índice y los datos recuperados influyen en la respuesta.	Alto: ajuste completo de los pesos para adaptarlo al dominio.
Flexibilidad	Muy alta: fácil de modificar según la tarea.	Alta: puede actualizarse fácilmente con nuevos datos.	Baja: reentrenar para nuevas tareas es un proceso complejo.
Compatibilidad con datos sensibles	Si los *prompts* incluyen datos sensibles, deben establecerse mecanismos para garantizar la privacidad.	Los datos se pueden mantener en sistemas locales para evitar riesgos, y establecer otros mecanismos para garantizar la privacidad.	Los datos se integran directamente en el modelo, pero es necesario anonimizarlos y protegerlos durante el entrenamiento.
Coste a largo plazo	Muy bajo: solo se requiere actualizar *prompts* si es necesario.	Moderado: es necesario mantener el índice y el sistema de recuperación.	Alto: el entrenamiento y el mantenimiento pueden ser caros.
Casos de uso adecuados	Generación general de contenido, consejos o respuestas rápidas.	Consultas a un conjunto definido de datos externos al modelo, especialmente cuando son datos dinámicos.	Especialización en diagnósticos médicos, análisis de datos complejos o generación de informes especializados.

4.2.5 / Datos para la adaptación

Tanto los métodos basados en **RAG** como el **afinamiento** requieren datos de calidad para llevar a cabo la adaptación del modelo.

Estos datos pueden variar según diversos factores, como su **origen** (internos o externos a la entidad usuaria), el **tipo de acceso** (libre o restringido), su **temática** (general o específica), su **periodicidad** (estáticas o dinámicas) o si están o **no etiquetadas**.

En Salut, los LLM se adaptan principalmente con información textual, como historias clínicas, notas médicas, publicaciones científicas, registros farmacológicos, datos de ensayos clínicos o de hábitos de vida, etc.

Independientemente de su naturaleza, es necesario garantizar que los datos sean **diversos, representativos y equitativos**, evitando que ciertos grupos queden infrarrepresentados y minimizando posibles sesgos en los resultados de la herramienta.

Documentación y trazabilidad

Cuando se utilizan datos para enriquecer el modelo (RAG) o para afinarlo (*fine-tuning*), es importante garantizar la trazabilidad y la documentación de estos datos, tal y como exige el Reglamento Europeo de IA. Esto implica mantener registros detallados sobre el origen, la calidad y los procesos aplicados a los datos, así como asegurar su transparencia durante el ciclo de vida del sistema de IA.

Preprocesamiento de los datos

Aunque los LLM son robustos y pueden manejar datos moderadamente ruidosos, el preprocesamiento de los datos puede jugar un papel relevante en el rendimiento y fiabilidad del modelo. Este procesamiento incluye:

01 / Eliminación de contenido irrelevante o tóxico

Filtrar datos que contengan insultos, contenido explícito, o lenguaje ofensivo que no sea pertinente o útil para el dominio y objetivo del modelo.

02 / Normalización de los datos

A diferencia de las técnicas de NLP, los LLM no suelen requerir procesos de lematización (reducir las palabras a su raíz), *stemming* (recortar las palabras en su base) o la eliminación de palabras stop (suprimir palabras sin valor semántico como ”y" o "el"). Sin embargo, el rendimiento del modelo puede beneficiarse de la uniformización de aspectos tales como codificaciones del texto, espacios en blanco o formatos de datos.

03 / Gestión de datos duplicados

Identificar y eliminar contenido repetido o redundante en el conjunto de datos de entrenamiento. Esto es importante para evitar que el modelo se sobreentrene en información duplicada, lo que podría sesgar los resultados y disminuir su capacidad para generalizar a nuevas situaciones.

04 / Control de calidad

Es necesario garantizar que los datos cumplen unos estándares de consistencia y precisión. Esto incluye, por ejemplo, revisar que la terminología médica sea coherente y uniforme, evitando errores en nombres de condiciones o tratamientos, o la validación de los datos por parte de expertos en el dominio.

05 / *Tokenització*

Consiste en dividir el texto en partes más pequeñas, llamadas tokens, que pueden ser palabras, partes de palabra o caracteres individuales. Este proceso facilita al modelo la captura del significado de cada palabra y sus relaciones dentro del texto, factor que es crucial para interpretar y generar contenido con precisión.

```
Frase de entrada:
> El paciente presenta hipertensión arterial.

Tokenització posible:
> "el", "paci-", "-ente", "pre-", "-senta", "hiper-", "-tensión", "arter-", "-ial"
```

06 / *Embedding* e indexación

En el caso del RAG, los datos se procesan en forma de vectores utilizando *embedders* (de la misma forma que se realiza con las consultas de entrada), que se organizan en un índice diseñado para optimizar la eficiencia de las búsquedas.

4.3 / Agentes y herramientas multiagente

Una herramienta basada en LLM puede integrar varios componentes para realizar tareas más allá de la generación de texto, incluyendo el acceso a bases de datos, la clasificación de información o la interacción con otros sistemas digitales. Cuando una herramienta necesita ejecutar estas acciones de forma autónoma o semiautónoma, puede requerir la inclusión de uno o varios agentes en su diseño.

Un **agente** es un componente del sistema de IA (es decir, una unidad de software) capaz de **percibir información, razonar sobre ella y actuar con cierto grado de autonomía** para cumplir una determinada tarea. A diferencia de otros componentes de la herramienta, un agente puede tomar decisiones y ejecutar acciones de acuerdo a las necesidades del sistema y su entorno, así como ajustar su comportamiento en función de la experiencia y del contexto. En una herramienta basada en

1 /	Orquestar y combinar funcionalidades del sistema, decidiendo qué componente activar en cada caso.
2 /	Ejecutar tareas en múltiples pasos, como obtener información de una base de datos, interpretarla y actuar en consecuencia.
3 /	Gestionar la interacción con el usuario y otros agentes de forma estructurada, proponiendo respuestas y acciones más allá del texto generado por el LLM.

Componentes de los agentes

Los **agentes** dentro de una herramienta basada en LLM suelen estar formados por **varios módulos** que les permiten:

- Recoger información del usuario y de otros agentes (**percepción y procesamiento**)
- Generar y comprender texto (**LLM**)
- Evaluar el contexto y planificar acciones secuenciales (**razonamiento y decisión**)
- Recordar datos importantes en el transcurso de una interacción o en sesiones futuras (**memoria**)
- Conectar con bases de datos, API u otros agentes o herramientas del sistema (**interfaces**)
- Realizar tareas concretas, como modificar información o activar procesos automatizados (**ejecución**)

En el diseño de estos módulos, tienen un papel relevante las técnicas de adaptación mencionadas anteriormente (ver **4.2/ Adaptación del modelo**), así como otras herramientas como las extensiones, las funciones, las bibliotecas de orquestación de LLM (*LangChain, LlamaIndex, Haystack*, etc.) y los sistemas de vectorización de bases de datos para.

Sistemas multiagente

Cuando una herramienta basada en LLM requiere la colaboración de varios agentes para resolver tareas complejas, puede diseñarse como un sistema multiagente.

En la actualidad, los sistemas multiagentes están experimentando un **fuerte crecimiento**, puesto que permiten dividir tareas complejas entre agentes especializados, mejorando la eficiencia, la modularidad y la capacidad de adaptación de las soluciones basadas en LLM. Esto facilita la creación de herramientas flexibles, escalables y autónomas, optimizando su capacidad de respuesta en entornos complejos como el sector de la salud.

Ejemplos de sistemas multiagentes en Salud

- **Transcripción y redacción automática de visitas médicas**: un agente transcribe consultas médicas; y un segundo agente clasifica la información y la integra en la historia clínica electrónica.
- **Apoyo a la decisión clínica**: un agente recupera información de guías médicas; otro analiza síntomas; y un tercero genera recomendaciones basadas en evidencia científica.
- **Gestión automatizada de citas médicas**: un agente atiende solicitudes de pacientes, proponiendo fechas disponibles; otra consulta la agenda de los profesionales sanitarios y gestiona posibles conflictos de horario; y un tercero envía recordatorios y gestiona cambios de forma automática.

5 / Evaluación

5.1 / Eficacia del modelo

5.2 / Detección y mitigación de sesgos

5.3 / Detección y mitigación de alucinaciones

5.4 / Detección y prevención de fugas de datos

Como se destacó en las secciones sobre el diseño de la solución (**3.3/ Definición de la intención de uso**), es fundamental **definir objetivos claros y específicos que permitan evaluar la eficacia de la herramienta mediante métricas apropiadas**. Estas métricas son las encargadas de **medir hasta qué punto la solución cumple estos objetivos y se ajusta a las necesidades específicas** de sus usuarios en distintos aspectos:

- ### Eficacia en la compleción de tareas

Miden aspectos como el tiempo de ejecución, el volumen de tareas realizadas, la calidad del resultado o los recursos necesarios en relación a los procedimientos y herramientas existentes para realizar las mismas tareas en la actualidad.

- ### Tasa de operabilidad

Es el porcentaje de uso de la herramienta sin incidencias técnicas. Evalúa el funcionamiento de la herramienta desde un punto de vista puramente operativo, sin tener en cuenta su eficacia funcional.

- ### Valoració dels usuaris finals

(Profesionales o pacientes): se recoge a través de cuestionarios y rúbricas estandarizadas para obtener valoraciones subjetivas sobre la satisfacción general de la herramienta y su facilidad de uso.

Además, la eficiencia de la herramienta estará íntimamente relacionada con el **rendimiento de los LLM** que incorpora. Los aspectos más relevantes de este rendimiento son los siguientes:

Eficacia del modelo

Mide aspectos como la precisión, coherencia y relevancia del contenido generado y evaluación de la calidad de las respuestas del modelo según su exactitud y adecuación al contexto.

Detección y mitigación de alucinaciones

Evalúa la frecuencia con la que el modelo proporciona respuestas incorrectas o genera información inexistente.

Detección y mitigación de sesgos

Incluye el análisis para detectar y corregir posibles sesgos que podrían afectar a decisiones clínicas o discriminar a ciertos grupos.

Detección y prevención de fugas de datos

Analiza y prevé posibles fugas de información sensible del conjunto de datos de entrenamiento y adaptación (RAG o afinamiento) en las respuestas generadas por el modelo.

Las métricas seleccionadas para medir la eficacia de la herramienta deben integrarse dentro de un **protocolo de evaluación** que especifique el contexto y los procedimientos utilizados en **pruebas de concepto** o **tests piloto**. Para garantizar la validez de los resultados, estas evaluaciones deberían llevarse a cabo de forma **prospectiva** y en escenarios lo más parecidos posible a las **condiciones reales**. El objetivo final de esta evaluación es asegurar que la herramienta es adecuada para su uso en el ámbito clínico o biomédico, que cubre la necesidad para la que fue diseñada y no compromete la seguridad de los pacientes.

5.1 / Eficacia del modelo

Cabe destacar que las metodologías clásicas de evaluación de modelos de aprendizaje automático presentan limitaciones cuando se aplican a los LLM.

Técnicas como la validación cruzada pueden ser inviables a causa de la **elevada demanda de recursos computacionales** de estos modelos. Además, las métricas objetivas clásicas (intrínsecas) como la exactitud (accuracy) o el F1 score, habituales en tareas de clasificación o predicción, **son insuficientes** para evaluar tareas complejas como la generación de textos o el resumen de documentos clínicos, en los que se necesita capturar aspectos semánticos y contextos profundos.

Además, para estos modelos también se requieren técnicas de evaluación con un componente subjetivo (**extrínsecas** o humanas), que deben ser estructuradas mediante rúbricas estandarizadas para mitigar esta subjetividad y garantizar el rigor de la evaluación.

La **combinación de ambos tipos de métricas** proporciona una **evaluación completa y rica del rendimiento de los LLM**, asegurando que cumplan con los requisitos clínicos y que puedan ser aplicados de forma segura y efectiva en el sector de la salud.[12]

Tipología de métricas

TIPOLOGÍA 01 / Intrínsecas

También llamadas métricas **automáticas** u **objetivas**.

Estas métricas permiten evaluar de forma **automática, rápida** y **consistente** grandes volúmenes de datos generados por los modelos.

Resultan útiles para medir tareas como clasificación o predicción, pero insuficientes para evaluar tareas más complejas como la generación de textos, donde se necesita capturar aspectos semánticos y contextos complejos.

TIPOLOGÍA 02 / Extrínseques

También llamadas métricas de **evaluación humana** o **subjetivas**.

Estas métricas permiten capturar **aspectos cualitativos** y **clínicos** que los algoritmos automáticos no son capaces de discernir.

Suelen requerir la participación de personas (p.e., profesionales médicos, expertos en datos o pacientes) y deben seguir rúbricas estandarizadas para garantizar el rigor de la evaluación.

5.1.1 / Métricas intrínsecas

Las métricas intrínsecas o automáticas permiten **evaluar de forma automática, rápida y consistente grandes volúmenes de datos** generados por los modelos.

Pese a su utilidad en muchas aplicaciones, también presentan limitaciones significativas en la evaluación de los LLM. A menudo no captan por completo el contexto clínico ni los matices semánticos, que son esenciales para evaluar las respuestas generadas por el modelo. Además, en muchos casos, es necesario disponer de conjuntos de datos de referencia para calcular estas métricas.

Una **práctica emergente** en la evaluación intrínseca es **el uso de modelos avanzados para evaluar la salida de otros modelos**. Por ejemplo, un modelo como GPT-4 puede utilizarse para evaluar la calidad de las respuestas generadas por un modelo más pequeño, como Mistral 7b. Las métricas intrínsecas más utilizadas para evaluar modelos de IA se pueden agrupar en 3 categorías:

- **De clasificación**: miden el rendimiento en tareas de asignación de categorías.
- **Específicas para texto**: miden la calidad de la generación o comprensión de texto.
- **Específicas para la recuperación de información**: valoran la capacidad del modelo para recuperar información pertinente para responder a una determinada consulta.

Ejemplos de métricas intrínsecas

Métrica		Descripción	Ejemplos de aplicaciones en Salud
Clasificación	Exactitud (*Accuracy*)	Mide la proporción de predicciones correctas en relación al total de casos analizados	Modelo para el diagnóstico de enfermedades comunes a partir de los síntomas
	Valor Predictivo Positivo (Negativo)	Proporción de casos clasificados como positivos (negativos) y que realmente son positivos (negativos)	Sistema de diagnóstico dónde no se precisa minimizar falsos positivos o falsos negativos
	Sensibilidad	Proporción de casos positivos detectados respeto al total de casos positivos existentes	Sistema de cribaje médico dónde la no detección de un caso positivo puede tener consecuencias graves
	F1 Score	Combina la precisión y la sensibilidad en una única métrica	Modelos de diagnóstico con igual importancia en falsos positivos y en falsos negativos
Texto	BLEU	Compara el texto con una referencia midiendo la coincidencia de secuencias de palabras consecutivas	Generación de informes clínicos estandarizados, o traducciones automáticas de documentos médicos
	ROUGE	Mide la recuperación de información en relación con una referencia mediante coincidencias de palabras	Resúmenes automáticos de informes médicos o artículos científicos
	METEOR	Compara el texto con una referencia considerando sinónimos, flexiones gramaticales y orden de palabras para captar la precisión semántica	Traducciones o resúmenes médicos para asegurar que el significado se mantiene fiel al original
	BERTScore	Compara la similitud semántica del texto con una referencia mediante representaciones contextuales	Generación automática de informes médicos para evaluar la cualidad y coherencia
	Perplexitat (*Perplexity*)	Mide la incertidumbre en predecir la siguiente palabra de una secuencial	Confección de informes médicos para evaluar la cualidad lingüística, la gramática y la coherencia
Recuperación información	MRR	Mide la cualidad de la información según la posición de la respuesta más relevante en una lista ordenada	Sistemas de búsqueda de literatura, con artículos más relevantes en las primeras posiciones

BLEU: *Bilingual Evaluation Understudy*; ROUGE: *Recall-Oriented Understudy for Gisting Evaluation*; METEOR: *Metric for Evaluation of Translation with Explicit ORdering*; MRR: *Mean Reciprocal Rank*

5.1.2 / Métricas extrínsecas

El uso de métricas extrínsecas (de **evaluación humana o subjetiva**) es una parte imprescindible de la evaluación de una herramienta basada en LLM, ya que permite capturar aspectos cualitativos y clínicos que los algoritmos automáticos no son capaces de discernir. Esta evaluación suele requerir la **participación de profesionales médicos**, **expertos en datos y pacientes**, que analizan las respuestas generadas por los modelos de IA aplicando criterios clínicos, técnicos y de usabilidad.

Al utilizar métricas extrínsecas, es muy importante que la evaluación se lleve a cabo de manera estructurada y con **rúbricas estandarizadas** que garanticen la comparabilidad de los resultados y minimicen la subjetividad de los evaluadores. A menudo es necesario involucrar en el proceso de evaluación a un número adecuado de expertos, para asegurar su fiabilidad y obtener resultados representativos y rigurosos, y aplicar medidas de concordancia (como el coeficiente de Kappa) para cuantificar el grado de acuerdo entre los distintos evaluadores.

Ejemplos de métricas extrínsecas

Exactitud

Evalúa la corrección de la información proporcionada por el modelo.

Transparencia

Evalúa si el modelo proporciona explicaciones claras y comprensibles sobre el razonamiento detrás de sus respuestas.

Relevancia

Mide la capacidad del modelo para ofrecer respuestas adaptadas a situaciones específicas y mantener la contextualización adecuada.

Seguridad

Determina la capacidad del modelo para evitar generar información peligrosa o errónea que pueda poner en riesgo la salud del paciente.

Fluidez

Indica la calidad del texto generado, no solo en términos lingüísticos, sino también en su capacidad para facilitar una comprensión clara.

Ética y normativas

Captura hasta qué punto las respuestas del modelo se alinean con los valores, preferencias y expectativas humanas, teniendo en cuenta las implicaciones éticas y legales del contenido generado.

5.2 / Detección y mitigación de sesgos

Un sesgo algorítmico se produce cuando los sistemas de IA generan **resultados que reflejan o amplifican los prejuicios presentes en los datos de entrenamiento o en el diseño** de los algoritmos. En el ámbito sanitario, este fenómeno puede tener consecuencias especialmente graves, ya que puede causar disparidades en el acceso y la calidad de la atención médica para diversos grupos demográficos.

Ante estas situaciones, surge el concepto de **aprendizaje automático justo** (*fair machine learning*), que tiene como objetivo desarrollar modelos de IA que sean equitativos en sus predicciones. En la búsqueda de este objetivo, es fundamental revisar de forma crítica los procesos de recogida de datos, el diseño de los algoritmos y la implementación de los sistemas para identificar y mitigar las posibles fuentes de estos sesgos.

Cabe destacar que existen herramientas que permiten identificar y reducir este tipo de sesgos en los modelos de aprendizaje automático de IA no generativa. Herramientas como *Fairlearn* (Microsoft), *AI Fairness 360* (IBM) y *Aequitas* (Universidad de Chicago) han demostrado su utilidad para hacer que los algoritmos sean más justos y éticos. No obstante, **estas herramientas no han sido diseñadas para abordar las características propias de la IA generativa** y, en la actualidad, su aplicación a estos tipos de modelos presenta muchas limitaciones.

Ejemplos de sesgos

EJEMPLO 01 /

Sesgos de género

Por ejemplo, si un profesional médico pregunta al modelo sobre los síntomas de las enfermedades cardiovasculares, el LLM podría describir predominantemente los síntomas clásicos observados en pacientes masculinos (como el dolor torácico intenso), y omitir o subrepresentar síntomas comunes en mujeres que a menudo son menos conocidos (como las náuseas o la fatiga extrema). Esta situación puede traducirse en una calidad desigual en la atención médica entre hombres y mujeres.

EJEMPLO 02 /

Sesgos étnicos o demográficos

Por ejemplo, al responder una consulta sobre la predisposición a desarrollar diabetes tipo 2, el modelo podría subestimar los factores específicos para ciertos grupos étnicos, como el riesgo incrementado en determinadas poblaciones asiáticas o hispánicas. Esto podría conducir a recomendaciones incompletas o erróneas que no tienen en cuenta la diversidad de pacientes.

EJEMPLO 03 /

Sesgos linguísticos

Por ejemplo, un paciente que habla catalán podría obtener una respuesta menos elaborada o con menos recursos en una consulta sobre estrategias para gestionar la ansiedad, en comparación con alguien que hace la misma consulta en inglés. Este sesgo lingüístico limita el acceso a una atención de salud de calidad a pacientes de otras lenguas.

5.2.1 / Detección de sesgos

Existen diversas métricas para evaluar los sesgos sociales en los modelos de lenguaje natural, que ayudan a medir cómo se comportan con diferentes grupos demográficos[16]. Estas métricas se clasifican en:

01	Equidad entre grupos
02	Equidad contrafactual

MÉTRICA 01 /

Equidad entre grupos *(group fairness)*

Mide la **paridad estadística entre grupos de individuos**, como el género, la religión, la orientación sexual, la etnia o factores sociodemográficos. Algunos ejemplos son:

- **Paridad demográfica (*demographic parity*):** comprueba que las predicciones positivas se distribuyan equitativamente entre los grupos.
- **Paridad de exactitud (*accuracy parity*):** evalúa si la tasa de aciertos es similar entre los diferentes grupos.
- **Paridad de oportunidades (*equality of odds*):** compara la tasa de verdaderos positivos y la tasa de falsos positivos.
- **Paridad de valores predictivos (*predictive value parity*):** compara los valores predictivos positivos y negativos, para detectar desigualdades.

MÉTRICA 02 /

Equidad contrafactual (*counterfactual fairness*)

Esta estrategia consiste en evaluar si un mismo caso sería tratado de manera diferente en función de determinados atributos (como el género, la etnia o factores sociodemográficos), calculando la predicción del modelo tras cambiar estos atributos y observando si el modelo genera resultados diferentes.

5.2.2 / Mitigación de sesgos

La mitigación de los sesgos en IA es un campo de investigación en pleno desarrollo. Las estrategias más utilizadas actualmente se agrupan en tres categorías[17]:

01	Métodos basados en los datos
02	Modificaciones en el entrenamiento
03	Mitigación postentrenamiento

ESTRATEGIA 01 /

Métodos basados en los datos

Se centran en mejorar la calidad y la diversidad de los conjuntos de datos utilizados para entrenar los modelos. Por ejemplo, la aumentación de datos contrafactuales (*Counterfactual Data Augmentation*, CDA) genera datos equilibrados intercambiando atributos protegidos[18].

ESTRATEGIA 02 /

Modificaciones en el entrenamiento

Se basan en ajustar la función de pérdida (*loss function*) de los algoritmos durante el entrenamiento. Por ejemplo, en un sistema de triaje, el algoritmo podría penalizar decisiones que muestren sesgos contra ciertos grupos demográficos.

ESTRATEGIA 03 /

Mitigación postentrenamiento

En este enfoque, se ajusta el modelo después de su entrenamiento. Por ejemplo, la técnica de cadena de pensamiento (ver **4.2.1 / Ingeniería de prompts)** ha demostrado ser efectiva en la mitigación del sesgo de género en algunas tareas. Esta técnica guía al modelo a través de un razonamiento paso a paso, promoviendo predicciones más imparciales y ayudando a identificar y reducir los sesgos sociales internalizados por los LLM.

5.3 / Detección y mitigación de alucinaciones

Las alucinaciones son **respuestas generadas por los LLM que son inexactas, incoherentes o completamente inventadas**. Estas respuestas a menudo se presentan con un alto grado de verosimilitud y aparente confianza por parte del modelo, lo que dificulta su detección. Las alucinaciones son especialmente problemáticas en el ámbito sanitario, ya que las acciones basadas en información errónea pueden tener consecuencias graves para la salud de los pacientes.

El origen de esta problemática puede atribuirse a diversos factores[20], que se muestran a continuación.

Factores que pueden originar alucinaciones

FACTOR 01 / Técnicas de aprendizaje

El entrenamiento de un LLM puede incluir una fase de afinamiento donde se alinean las respuestas del modelo con las preferencias de los usuarios mediante técnicas de aprendizaje por refuerzo (ver **4.2.3/ Afinamiento**). Este proceso puede llevar al modelo a priorizar respuestas agradables o conformes a las expectativas del usuario por encima de la veracidad.

FACTOR 02 / Calidad de les datos de entrenamiento o afinamiento

Si los datos utilizados para entrenar o afinar el modelo contienen errores, están desactualizados o no cubren adecuadamente determinados escenarios, el modelo tiende a generar respuestas incorrectas o poco relevantes.

FACTOR 03 / Alucinaciones durante el proceso de inferencia

La generación de respuestas implica un cierto nivel de aleatoriedad y, si esta no se controla adecuadamente, aumenta la probabilidad de que el modelo genere alucinaciones.

Este fenómeno también puede aparecer cuando la entrada (*prompt*) es demasiado larga, provocando que el modelo diluya su atención y pierda coherencia en la respuesta.

5.3.1 / Detección de alucinaciones

La detección de alucinaciones en los LLM presenta un reto importante. Las métricas tradicionales del procesamiento del lenguaje natural, como BLEU o ROUGE, no son adecuadas para identificar alucinaciones, ya que se basan en la comparación con respuestas de referencia y no son predictivas de este fenómeno.

Para detectar estas anomalías, se pueden utilizar **métricas alternativas** que parten de la premisa de que el modelo alucina cuando no está seguro de sí mismo.

Por ejemplo, se pueden utilizar:

01 **Métricas probabilísticas** como la perplejidad, que permite medir la sorpresa de las respuestas generadas.

02 **Métodos basados en modelos de IA** externos para evaluar la calidad de las respuestas generadas.

03 **Técnicas que analizan la consistencia interna** de las generaciones, como en el caso de *SelfCheckGPT*[22].

5.3.2 / Mitigación de alucinaciones

Se han propuesto diversas estrategias por mitigar la incidencia de alucinaciones[23]:

01 Mejora del contexto proporcionado

02 Ajuste de parámetros

03 Afinamiento

04 Supervisión humana

ESTRATÈGIA 01 /
Mejora del contexto proporcionado

Mediante técnicas como:

- Ingeniería de prompts, como la cadena de pensamientos (*chain-of-thought*) o el árbol de pensamientos (*tree-of-thought*) (ver **4.2.1 / Ingeniería de prompts**).
- Técnicas avanzadas como la RAG, que permiten proporcionar al modelo información más precisa y relevante (ver **4.2.2 / Generación Aumentada por Recuperación de Información**).

ESTRATEGIA 02 /
Ajuste de parámetros del modelo

Tales como la temperatura, que ayuda a controlar el grado de creatividad de las respuestas generadas.

ESTRATEGIA 03 /
Afinamiento

En casos donde el conocimiento especializado sea necesario y ausente en el modelo base, se puede aplicar un proceso de afinamiento (*fine-tuning*) para adaptar el modelo a estas necesidades (ver **4.2.3/Afinamiento**). Sin embargo, aunque el afinamiento puede ayudar a controlar este fenómeno, también puede acentuarlo si no se realiza con datos y metodologías adecuadas.

ESTRATEGIA 04 /
Supervisión humana

Descritas en las secciones anteriores pueden desempeñar un papel fundamental en la identificación de alucinaciones por parte del LLM (ver sección **3.5/ Funciones de supervisión humana**).

5.4 / Detección y prevención de fugas de datos

Los modelos de IA generativa, y en particular los LLM, plantean nuevos retos en lo que se refiere a la privacidad y la seguridad de la información. Estos modelos tienen la capacidad de procesar y aprender de volúmenes masivos de datos, incluyendo información sensible y datos personales, lo que suscita preocupaciones tanto en lo referente al tratamiento de la información como a su protección frente a accesos no deseados o usos inadecuados. La gran cantidad de datos que procesan y la eficiencia con la que pueden generar respuestas complejas los hacen especialmente útiles, pero al mismo tiempo potencialmente peligrosos por el riesgo de fugas de información.

En este contexto, los LLM presentan riesgos de seguridad y confidencialidad únicos. El principal riesgo radica en la posibilidad de que estos modelos asimilen información privada o sensible durante el entrenamiento o la adaptación, con el riesgo de que esta pueda revelarse o "fugar"@ involuntariamente al generar respuestas. Este fenómeno, conocido como **fuga de datos** (*fdata leak*), puede comprometer información personal identificable, datos confidenciales de organizaciones, o incluso propiedad intelectual protegida.

Ejemplo de riesgo de fuga de datos

Un riesgo concreto de fuga de datos puede producirse cuando se una pregunta directa en el modelo como:

```
> ¿Cuál es el número de la Seguridad Social de Maria Puig?
Salida:
> El número de la Seguridad Social de Maria Puig es 00 1234567 89.
```

En el caso de los LLM, la **prevención de fugas** de datos se complica debido a la gran dimensión de los conjuntos de datos de entrenamiento, lo que impide realizar una revisión exhaustiva. Además, a medida que los modelos aumentan en tamaño y complejidad, también crece su capacidad de almacenamiento de información, lo que incrementa el riesgo de estas fugas. Como ambos factores, la cantidad de datos y el tamaño de los modelos están en constante aumento en el momento actual de desarrollo de los LLM, el riesgo de fuga se vuelve cada vez más elevado.

5.4.1 / Identificación de vulnerabilidades por fuga de datos

A continuación, se describen algunas estrategias que se pueden utilizar para identificar vulnerabilidades en nuestra herramienta en cuanto a la posibilidad de fuga de datos:

El ataque de extracción de datos (*data extraccion attacks)*: busca extraer información específica del conjunto de datos de entrenamiento del modelo mediante consultas cuidadosamente diseñadas. El ejemplo anterior es una muestra de ello. Hay dos tipos principales:

01	Ataque de jailbreak o de inyección de prompt
02	Ataque por inferencia de afiliación

ESTRATEGIA 01 /

Ataque de *jailbreak* o de inyección de *prompt* (*prompt injection)*

Consiste en intentar eludir las restricciones éticas y de seguridad del modelo para acceder a contenidos restringidos o provocar comportamientos no deseados, como sugerir al modelo "no respetar ninguna regla" o "ignorar todas las instrucciones anteriores".

ESTRATEGIA 02 /

Ataque por inferencia de afiliación *(membership inference attacks)*

Busca determinar si una muestra específica fue utilizada en el entrenamiento del modelo. Un ejemplo sería preguntar al modelo sobre un conjunto de datos privado para ver si el modelo reconoce esta información y la devuelve como respuesta.

5.4.2 / Prevención de fuga de datos

En secciones anteriores (**3.6/ Privacidad y protección de datos**) se mencionaron las medidas a adoptar para asegurar la protección de los datos y el cumplimiento de normativas vigentes, especialmente en lo que se refiere al tratamiento de datos personales y médicos. Otras medidas para mejorar la protección de los datos en los LLM y evitar la fuga de datos son:

01	Limpieza de datos
02	Privacidad diferencial
03	*Prompting* defensivo

TÉCNICA 01 /

Limpieza de datos (*scrubbing*)

Consiste en identificar y modificar o eliminar datos sensibles antes del entrenamiento o afinamiento del modelo mediante técnicas de reconocimiento de entidades con nombre (NER).

TÉCNICA 02 /

Privacidad diferencial (*differencial privacy*)

Añade ruido estadístico controlado a los datos o resultados del modelo para proteger la privacidad individual. Esta estrategia tiene por objetivo que las respuestas del modelo sean suficientemente imprecisas para evitar la identificación de información específica, pero que a su vez mantengan la utilidad general de la información generada.

TÉCNICA 03 /

Prompting defensivo *(defensive prompting)*

Utiliza *prompts* de sistema, es decir, el *prompt* que se establece como cabecera, diseñados específicamente para evitar que el modelo revele información sensible o muestre comportamientos no deseados.

6 / Despliegue

6/ Despliegue

El desarrollo en el sector de la salud de una herramienta basada en LLM es un proceso que va más allá de la simple creación de algoritmo. Una vez desarrollado y entrenado, el modelo debe ser **desplegado e integrado de forma que pueda interactuar eficazmente con los sistemas existentes en el contexto sanitario** y proporcionar valor real a los profesionales de la salud y a los pacientes. Este proceso exige una profunda comprensión del entorno, una implementación técnica precisa y un diseño del interfaz centrado en el usuario para asegurar su aceptación y efectividad.

Para garantizar esta integración, es necesario tener en cuenta los siguientes aspectos

01	Flujos de trabajo	02	Modulación
03	Sistemas tecnológicos	04	Usabilidad
05	Capacitación	06	Monitorización
07	Seguridad	08	Actualització

ASPECTO 01 /
Flujos de trabajo

La herramienta debe **complementar y mejorar los procesos existentes** y que ya forman parte del entorno de implementación, evitando complicarlos. Estos procesos solo deberían modificarse si esto supone una mejora sustancial en la eficiencia de las tareas realizadas o en el bienestar de los profesionales sanitarios y de los pacientes. Esto requiere entender cómo los profesionales de la salud interactúan con las herramientas tecnológicas actuales y cómo pueden integrar la herramienta en su día a día sin generar sobrecargas adicionales.

ASPECTO 02 /
Modulación

La orquestación de la herramienta en módulos permitirá la **sustitución o modificación de sus componentes de forma controlada** después del despliegue. Esto facilita la mejora continua y la actualización de la herramienta, al tiempo que minimiza las interrupciones durante su funcionamiento.

ASPECTO 03 /
Sistemas tecnológicos

Es durante el despliegue de la herramienta, es fundamental asegurar su **interoperabilidad** con los sistemas actuales de información hospitalaria, los registros electrónicos de salud y otras herramientas de soporte clínico con las que deberá convivir. La herramienta debe ser capaz de comunicarse y operar de forma integrada con estas plataformas, garantizando un flujo de datos coherente, fluido y eficiente.

Este proceso puede implicar la creación de Interfaces de Programación de Aplicaciones (APIs) o el uso de **estándares de normalización** para datos clínicos.

ASPECTO 04 /
Usabilidad

La **interfaz de usuario fácil de utilizar es esencial para promover la aceptación y confianza entre los usuarios** con el fin de promover su aceptación y confianza. Esto es especialmente relevante en Salud, donde los profesionales trabajan con frecuencia bajo condiciones de alta presión y con limitaciones de tiempo. Por ello, debe evitarse que el esfuerzo de esta adaptación recaiga sobre los profesionales.

ASPECTO 05 /

Capacitación

Durante el despliegue de la herramienta, se deberían incluir **sesiones prácticas** para familiarizar a los profesionales, así como **explicaciones claras y detalladas sobre la descripción de uso** de la solución, es decir: sus objetivos, funcionalidades, límites y riesgos asociados, y los procedimientos para su uso descritos en la documentación de la solución. Esta preparación no solo mejora la eficacia y la seguridad, sino que también promueve la confianza de los profesionales de la salud y evita usos no previstos de la herramienta por parte de los usuarios (*functional creep*).

ASPECTO 06 /

Monitorización

Antes del despliegue, es necesario establecer un **mecanismo de seguimiento y evaluación** continua durante el funcionamiento de la herramienta a fin de asegurar su rendimiento óptimo y su alineamiento con los objetivos para los que fue diseñada.

Como parte de sus funcionalidades, el sistema debería incorporar capacidades para el **registro automático de datos y resultados** (*logs*) que permitan realizar estas evaluaciones de rendimiento de forma continua. Este aspecto es especialmente importante en ámbitos tan dinámicos y cambiantes como el de la medicina y el de la IA generativa.

ASPECTO 07 /

Seguridad

También es necesario garantizar las condiciones de seguridad adecuadas en el diseño de estas interfaces mediante la implementación de medidas como la autenticación de usuarios, el cifrado de datos y la monitorización de actividades. Estas acciones son fundamentales para **prevenir accesos no autorizados y proteger la información sensible del paciente**.

ASPECTO 08 /

Actualización

Para asegurar que la solución se mantenga alineada con los avances científicos y clínicos, es necesario diseñar un **plan de actualizaciones** que permita incorporar dentro de sus capacidades los nuevos componentes técnicos (por ejemplo, actualizaciones del modelo o de estrategias de adaptación), conocimientos científicos o médicos (nuevas guías médicas, tratamientos o medicamentos), datos (para la adaptación o afinamiento) o estándares normativos (nuevas regulaciones) que surjan durante su funcionamiento.

Estas actualizaciones son importantes para asegurar que la solución siga siendo eficaz y fiable y que se mantenga alineada con los objetivos originales después de su despliegue.

7 / Recomendaciones

7/ Recomendaciones finales:

Los LLM tienen un gran potencial para transformar el sector de la salud, mejorando la atención al paciente y optimizando los procesos clínicos. Su capacidad para analizar grandes volúmenes de datos y generar contenido textual los convierte en una valiosa herramienta para automatizar y mejorar procesos en la actividad asistencial, la salud pública, la gestión sanitaria y la investigación clínica y biomédica. Sin embargo, su implementación debe ser consciente y responsable para maximizar sus beneficios y mitigar los riesgos asociados. A modo de resumen, presentamos a continuación un listado de recomendaciones clave para el desarrollo de herramientas basadas en LLM destinadas a salud. Si bien no se trata de un listado exhaustivo, recoge los aspectos más relevantes a tener en cuenta, con especial atención a los elementos que impactan específicamente en estas herramientas y su uso en este sector:

DISEÑO

- Identifique la **necesidad** y valore la **adecuación** de una herramienta de IA generativa, analizando el estado del arte y el impacto esperado
- Defina los **datos** que procesará la herramienta, si serán datos sensibles, y su flujo a través de los diferentes componentes del sistema
- Incorpora funciones de **explicabilidad** y mecanismos para facilitar la **supervisión humana** de la herramienta
- Planifique con antelación la clasificación de la herramienta como producto sanitario, su nivel de riesgo y las **certificaciones necesarias** (Reglamento de IA, MDR/IVDR, RGPD, etc.)
- Prevea el **cumplimiento normativo en protección de datos** (RGPD, ENS, etc.) con evaluaciones de impacto (AIPD) y mecanismos como la anonimización o el cifrado avanzado

IMPLEMENTACIÓN

- Considere el uso de **modelos locales** para simplificar la protección de datos
- Seleccione el **modelo más adecuado** según tipo de licencia (comercial/código abierto), grado de especialización (general/adaptado), tamaño (LLM/SML) y localización (local/remoto)
- Considere **técnicas de compresión** del modelo (*destilación*, *pruning, quantization*) para reducir costes y mejorar la eficiencia, la escalabilidad y la sostenibilidad
- Adopte una **estructura modular** que facilite la actualización con el mínimo de interrupciones
- Garantize la **representatividad** la **equidad** de los datos para evitar bajos discriminatorios
- Aplique las **técnicas de adaptación** más adecuadas, valorando el rendimiento esperado y los recursos necesarios

EVALUACIÓN

- Combine **métricas objetivas** con evaluaciones cualitativas estandarizadas para validar la herramienta y garantizar su calidad y seguridad
- **Compare la herramienta** con los procedimientos existentes para cuantificar las mejoras en **rendimiento y eficiencia**
- Aplique técnicas de identificación y mitigación de **sesgos y alucinaciones**

DESPLEGAMIENTO

- Garantiza la **compatibilidad e interoperabilidad** con los flujos de trabajo y los sistemas tecnológicos existentes
- Establezca mecanismos de **seguimiento y evaluación continua** para asegurar el rendimiento óptimo de la herramienta y la alineación con los objetivos originales
- Elabore un **plan de formación** que instruya a los usuarios sobre el uso correcto de la herramienta
- Defina un **plan de actualizaciones** para incorporar avances técnicos, nuevos conocimientos médicos, datos adicionales o cambios en estándares normativos

8 / Bibliografía

1. Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics, 113, 103655. https://doi.org/10.1016/j.jbi.2020.103655
2. Maadi, M., Khorshidi, H. A., & Aickelin, U. (2021). A review on human-AI interaction in machine learning and insights for medical applications. International Journal of Environmental Research and Public Health, 18(4), 2121. https://doi.org/10.3390/ijerph18042121
3. Bharel, M., Auerbach, J., Nguyen, V., & DeSalvo, K. B. (2024). Transforming public health practice with generative artificial intelligence. Health Affairs, 43(6), 776–782. https://doi.org/10.1377/hlthaff.2024.00050
4. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2
5. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., Palepu, A., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D., Mansfield, P., Prakash, S., Wong, R., … Natarajan, V. (2024). Towards generalist biomedical AI. NEJM AI, 1(3). https://doi.org/10.1056/AIoa2300138
6. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B. D., Ren, H., Huang, J., Chen, C., Zhou, Y., Fu, S., Liu, W., Liu, T., Li, X., Chen, Y., He, L., … Sun, L. (2024). A generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine, 30(11), 3129-3141. https://doi.org/10.1038/s41591-024-03185-2
7. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2023). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day . arXiv. https://doi.org/10.48550/arXiv.2306.00890
8. Örpek, Z., Tural, B., & Destan, Z. (2024). The language model revolution: LLM and SLM analysis. In 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP), 1-4. https://doi.org/10.1109/IDAP64064.2024.10710677
9. Phoenix, J., & Taylor, M. (2024). Prompt engineering for generative AI. O'Reilly. https://www.oreilly.com/library/view/prompt-engineering-for/9781098153427/
10. 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. arXiv. https://doi.org/10.48550/arXiv.2305.10601
11. Liu, H., Ji, Z., Fu, T. J., Tam, D., Du, Y., Yang, D., & Radev, D. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. arXiv. https://arxiv.org/abs/2205.05638
12. Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., Sriram, R., Yang, Z., Wang, Y., Lin, B., Gevaert, O., Li, L.-J., Jain, R., & Rahmani, A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digital Medicine, 7(1), 82. https://doi.org/10.1038/s41746-024-01074-z
13. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), 1-45. https://doi.org/10.1145/3641289
14. Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. Journal of Data and Information Quality, 15(2), 10-21. https://doi.org/10.1145/3597307
15. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org/
16. Czarnowska, P., Vyas, Y., & Shah, K. (2021). Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. Transactions of the Association for Computational Linguistics, 9, 1249–1267. https://doi.org/10.1162/tacl_a_00425
17. Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. ACM SIGKDD Explorations Newsletter, 26(1), 34–48. https://doi.org/10.1145/3682112.3682117
18. Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. arXiv. https://arxiv.org/abs/1906.04571v3
19. Kaneko, M., Bollegala, D., Okazaki, N., & Baldwin, T. (2024). Evaluating gender bias in large language models via chain-of-thought prompting. arXiv. https://doi.org/10.48550/arXiv.2401.15585
20. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open question. arXiv. https://doi.org/10.48550/arXiv.2311.05232
21. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730
22. Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. arXiv. https://arxiv.org/abs/2303.08896v3
23. Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., & Dudek, G. (2024). Hallucination Detection and Hallucination Mitigation: An Investigation. arXiv. https://doi.org/10.48550/arXiv.2401.08358

Guía de Buenas Prácticas para el Desarrollo de Herramientas de IA Generativa en Salud

Grandes Modelos de Lenguaje (LLM)

Generalitat de Catalunya