

Proyecto Integrado de Inteligencia Artificial

# **Sinergia de Agentes: Diversidad en Decisiones Médicas Telefónica**

Carla Atienza

Nils Duran

Alba Figueras

Lola Monroy

FIB UPC

Grado en Inteligencia Artificial

27 de mayo de 2025

# Índice

<b>1. Executive Summary</b>	<b>3</b>
2. Motivación y Descripción del Reto	4
3. Objetivos y alcance del proyecto	5
3.1 Objetivos	5
3.2. Diseño del sistema	6
3.2.1 Investigación previa	6
3.2.2 Design Thinking	7
Lean Canvas	9
Stakeholders	10
Benchmarking	10
User Journey - UX of AI	12
<b>4. Arquitectura del Sistema</b>	<b>14</b>
4.1. Módulo de agentes especializados	14
4.2. Módulo de interacción y coordinación de agentes	14
4.3. Módulo de benchmarking y evaluación	15
4.3.1. Evaluación objetiva con preguntas cerradas	15
4.3.2. Evaluación cualitativa con casos clínicos abiertos:	16
4.4. Módulo de interfaz de usuario	17
<b>5. Evaluación del Sistema</b>	<b>19</b>
<b>6. Principales Dificultades Encontradas y Solución</b>	<b>23</b>
<b>7. Organización del Equipo y Tareas</b>	<b>24</b>
<b>8. GANTT y PERT</b>	<b>25</b>
<b>9. Presupuesto y Viabilidad Económica</b>	<b>26</b>
9.1 Estimación de Costes por Tarea	26
9.2 Reservas	29
9.3 Presupuesto Total del Proyecto	29
9.4 Análisis de Viabilidad Financiera	29
9.5 Seguimiento y Control del Presupuesto	30
<b>10. Gestión de Riesgos</b>	<b>32</b>
10.1 Asignación de Roles y Responsabilidades	32
10.2 Definición de Umbrales de Riesgo	33
10.3 Identificación de Riesgos	34
10.3.1 RIESGOS TÉCNICOS	34
10.3.2 RIESGOS DE ORGANIZACIÓN	34
10.3.3 RIESGOS DE GESTIÓN	35
10.3.4 RIESGOS ÉTICOS Y LEGALES	35
10.3.5 RIESGOS FINANCIEROS	35
10.4 Valoración de Riesgos	36
RISK MATRIX	36
10.5 Estrategias de Mitigación y Activadores de Riesgo	37
10.6 Monitorización del Registro de Riesgos y Actualización	37
ACTION PLAN TABLE	38
<b>11. Impacto ambiental, económico y social</b>	<b>40</b>
11.1 Matriz de Sostenibilidad	40
11.1.1 Ámbito Económico	40

11.1.2. Medioambiental	41
11.1.3. Ámbito Social	42
11.2 Análisis de Protección de los Datos	43
11.3 Consideraciones Éticas	44
<b>12. Conclusiones</b>	<b>46</b>
<b>13. Referencias</b>	<b>47</b>

## **1. Executive Summary**

Este proyecto tiene como objetivo principal investigar si la colaboración entre múltiples agentes de inteligencia artificial, desarrollados con perfiles diversos, puede generar unas mejores decisiones en comparación con agentes homogéneos. Al tratarse de un trabajo de investigación, nuestro foco no está en crear un producto para el mercado, sino en programar un sistema que nos permita evaluar la hipótesis de los beneficios de la diversidad en los modelos de lenguaje.

Con la intención de dar forma a esta propuesta más bien abstracta, decidimos acotar el alcance en un dominio más específico: el ámbito médico. Este sector cuenta con una gran cantidad de datos objetivos etiquetados según su ámbito e informes abiertos de sus diferentes especialidades, lo cual lo hace idóneo para la simulación y evaluación de agentes expertos artificiales. Así, diseñamos un sistema compuesto por cinco agentes ajustados para simular diferentes especialidades médicas: medicina general, ciencias básicas (anatomía, fisiología), patología-farmacología, cirugía y pediatría-ginecología. La idea era analizar si su colaboración resultaba en diagnósticos médicos más precisos.

Para el desarrollo de los agentes expertos, optamos por realizar un fine-tuning con LoRA sobre el mismo modelo base con unos datos específicos para cada ámbito médico, recogidos de MedMCQA. Google AI Studio facilitó la tarea del ajuste de los cinco agentes, aportando a cada uno conocimientos específicos de su especialidad.

La interacción entre los agentes definidos fue diseñada con la ayuda de LangGraph, que nos permitió crear debates y conversaciones estructuradas antes de generar un diagnóstico final. Así, dada una consulta médica, cada experto artificial individualmente propone un diagnóstico y debe justificar su decisión, y seguidamente se empieza un debate en qué cada agente discute los argumentos de los anteriores, hasta llegar a una decisión consensuada.

Para medir la calidad de las respuestas de forma objetiva, utilizamos el dataset MedQA, que incluye preguntas cerradas tipo test en formato de consulta médica. Con el fin de utilizar los especialistas más adecuados para cada consulta, incorporamos un sistema de clasificación de agentes basado en embeddings semánticos, evitando la asignación aleatoria de especialidades y reduciendo posibles sesgos en la comparación entre sistemas homogéneos y heterogéneos.

Los resultados muestran que la combinación de agentes mediante consenso supera sistemáticamente el rendimiento de agentes individuales. Además, la precisión mejora ligeramente a medida que se añaden más agentes al proceso, incluso cuando alguno de ellos (como el de ciencias básicas) parece poco relevante según la similitud semántica. Esto sugiere que la diversidad funcional puede favorecer el razonamiento colectivo.

Estos resultados nos permiten finalmente validar –al menos en este escenario experimental– la hipótesis inicial propuesta por Telefónica: la colaboración entre agentes diversos de inteligencia artificial puede generar decisiones más precisas y fundamentadas que el razonamiento homogéneo.

## **2. Motivación y Descripción del Reto**

Telefónica nos ha planteado investigar cómo influye la diversidad de agentes en la calidad de sus decisiones. Con el apoyo de Lars Stalling (Telefónica Discovery) y Ramon Sangüesa, accedimos a bibliografía y casos prácticos en inteligencia colectiva y sistemas multiagente.

Los sistemas generativos de IA suelen emitir una respuesta de manera unitaria, lo que a veces desemboca en errores o sesgos difíciles de detectar. Nuestra hipótesis es que una interacción iterativa entre varios agentes, cada uno experto en un campo distinto, permitiría contrastar y depurar esas respuestas, reduciendo sesgos e inconsistencias.

Esto es lo que se nos propuso con este proyecto. ¿Podría la diversidad de puntos de vista en los agentes mejorar los sistemas de tomas de decisiones con inteligencia artificial? Para responder esta pregunta, nuestro objetivo principal es simular agentes diversos y analizar si su colaboración tiene un impacto en la calidad de la toma de decisiones, en comparación con agentes homogéneos. Así, queremos validar la hipótesis de que la diversidad de los agentes es importante en la toma de decisiones con inteligencia artificial, así como lo es en otros ámbitos.

Además, queremos identificar las combinaciones óptimas de agentes que producen unas mejores soluciones para cada caso. Esto permitiría maximizar la calidad de las respuestas para cada query, personalizando los agentes involucrados en función de la consulta y ofreciendo unos resultados más adecuados al problema.

Como nuestro proyecto aborda un concepto más bien abstracto, como sería la importancia de la diversidad en los agentes de decisión, nos decantamos por utilizar un ejemplo más concreto con el que poder trabajar. Decidimos centrarnos en la toma de decisiones en el ámbito médico, puesto a que es un campo donde una buena resolución de problemas es vital, y a la vez es sencillo establecer unas métricas para evaluar la calidad de las decisiones tomadas, lo cual nos servirá para evaluar el impacto de los diferentes grupos. Así, enfocamos la toma de decisiones al diagnóstico médico de un paciente, y se tienen en cuenta diferentes tipos de agentes que dominan distintos campos de conocimiento médico. Consideramos que la interacción entre agentes inteligentes sería clave en el campo de la medicina a la hora de diagnosticar un paciente con precisión y agilizar el proceso de identificación de la enfermedad, lo que permitiría actuar más rápidamente.

### **3. Objetivos y alcance del proyecto**

#### **3.1 Objetivos**

Este proyecto tiene como objetivo investigar el impacto de la diversidad de agentes en la precisión y calidad de sus decisiones. Esto lo haremos mediante el desarrollo de un sistema de Inteligencia Artificial para el diagnóstico médico.

Diseñaremos el sistema para simular la colaboración entre agentes de IA que representan a diferentes especialidades médicas (e.g., cardiología, dermatología). Compararemos la interacción entre estos agentes diversos con sistemas que utilizan agentes homogéneos.

El alcance del proyecto incluye diversas áreas. Entre ellas se encuentra la creación de agentes de IA que simulan el conocimiento y la experiencia de diferentes especialidades médicas. También se implementará la interacción entre los agentes, mediante el diseño y programación de la lógica que les permita comunicarse, compartir información y colaborar en el proceso de diagnóstico. Otra de las áreas es la medición y el análisis de cómo la diversidad de los agentes afecta la precisión, la eficiencia y la equidad de los diagnósticos generados por el sistema. Por último, podemos incluir también la creación de una interfaz que permita a los profesionales médicos ingresar datos de pacientes y visualizar los diagnósticos y el razonamiento del sistema.

Este proyecto se centra en la investigación y el desarrollo de un prototipo, por lo tanto, la implementación clínica a gran escala queda fuera del alcance del presente trabajo.

## 3.2. Diseño del sistema

### 3.2.1 Investigación previa

En primer lugar, realizamos una revisión exhaustiva de los fundamentos teóricos que sustentan nuestro enfoque de inteligencia colectiva. Partimos de *El libro de la inteligencia colectiva* de Amalio A. Rey, donde se define la capacidad de un grupo para razonar, aprender y resolver problemas como algo más que la mera suma de sus componentes. Rey subraya que “saber elegir” es el verdadero síntoma de la inteligencia y plantea la paradoja de si, en todos los casos, un conjunto de individuos supera a los individuos por separado. El famoso duelo de Garry Kasparov contra “el ajedrez del mundo” en 1999 ejemplifica este dilema, y servirá de referencia cuando Magnus Carlsen repita la experiencia en abril de 2025.

A continuación, exploramos estudios académicos sobre diversidad y toma de decisiones grupales, que evidencian cómo la heterogeneidad de opiniones mitiga problemas como el fenómeno del free-rider y potencia la calidad de las conclusiones. En particular, trabajos en sistemas multiagente aplicados al ámbito médico muestran que la colaboración entre agentes mejora la fiabilidad de los diagnósticos .

Para traducir estos conceptos a la práctica, definimos primero qué entendemos por “agente” de IA. Según SuperAnnotate, un agente combina análisis de datos, planificación estratégica, recuperación de información y aprendizaje de experiencias previas, e incluso puede invocar herramientas externas (calculadoras, llamadas a redes neuronales o ejecución de código). Como prototipo multiagente, Lars Stalling nos acercó a STORM y su extensión Co-STORM, desarrolladas en Stanford: un sistema diseñado para reducir las “alucinaciones” de los LLM mediante la consulta a fuentes de internet y la generación de texto de estilo enciclopédico, lo que resultó inspirador para nuestro propio diseño de agentes especializados.

### 3.2.2 Design Thinking

El Design Thinking es una metodología de resolución de problemas que se centra en desarrollar soluciones innovadoras poniendo a las personas en el centro del proceso. Su principal objetivo es entender las necesidades reales y, en ocasiones, desconocidas de los usuarios para generar ideas y prototipos que respondan a problemas complejos de manera creativa y práctica.

Esta metodología se caracteriza por su enfoque iterativo y colaborativo, donde se fomenta la participación activa de equipos multidisciplinares. Se inicia con la fase de empatía, que consiste en observar y comprender profundamente a los usuarios, dejando a un lado supuestos y sesgos. Posteriormente, se define el problema de manera precisa, lo que permite orientar el proceso hacia soluciones realmente significativas. Las etapas siguientes ideación, prototipado y prueba implican la generación de múltiples ideas, la creación de versiones simplificadas de posibles soluciones y la experimentación para recoger feedback y mejorar continuamente el producto o servicio.

Hemos aplicado esta estrategia en el desarrollo de nuestro proyecto. Procederemos a explicar cada uno de los métodos que hemos utilizado, y si nos han sido útiles o no para el proceso de ‘Design Thinking’.

#### Lluvia de ideas

La lluvia de ideas fomenta la generación libre de propuestas, sin críticas tempranas, creando un ambiente dinámico en el que cada sugerencia queda registrada y visible. Esta fase impulsa el pensamiento divergente; después, las ideas se refinan y seleccionan para convertirlas en soluciones concretas.

Aunque se trata de una etapa muy inicial (muchas ideas pueden perder relevancia más adelante), nos ayudó a explorar diversas perspectivas del reto. Entre las propuestas surgieron:

- **“Tinder” para empresas:** plataforma estilo app de citas que conecta perfiles profesionales según afinidades y necesidades.
- **Simulador de decisiones:** herramienta para comparar cómo grupos con distintos grados de diversidad resuelven un mismo problema.
- **Método de evaluación de diversidad:** sistema de métricas que cuantifique la diversidad organizacional y su impacto en resultados de negocio.
- **Diagnóstico de políticas D&I:** soluciones para medir la implementación real de iniciativas de inclusión, más allá de los textos corporativos.
- **Prototipado con feedback real:** pruebas piloto que ofrezcan datos sobre el rendimiento de equipos diversos y permitan ajustar prácticas.
- **Equipos de telepresencia:** reclutamiento de profesionales remotos para enriquecer la diversidad geográfica y cultural.
- **Decisiones según tamaño de equipo:** estudio de cómo varía la eficacia y sinergia en función de la composición y el número de integrantes.



En retrospectiva, la lluvia de ideas no generó un único concepto definitivo, pero cada propuesta aportó elementos clave para nuestro marco de trabajo. Desde la conexión de perfiles hasta la medición de diversidad, todas las aportaciones han enriquecido la visión global del proyecto. Gracias a este ejercicio, pudimos identificar herramientas y enfoques que más adelante combinaríamos en un modelo unificado: diagnóstico médico colaborativo, evaluación de resultados, prototipado y optimización de equipos diversos.

### Matrix

Una matriz es una herramienta de organización que estructura la información en filas y columnas para facilitar la comparación y evaluación de distintos elementos o variables. Permite visualizar de forma clara los pros y contras de cada opción, apoyando una toma de decisiones más fundamentada y colaborativa.

En el contexto de los agentes en la medicina, una matriz de TP/TN/FP/FN:

		Machine Prediction	
		Positive	Negative
User reality	Positive	- Hacer un diagnóstico correcto (e.g. Dice que tienes la gripe, y sí que la tienes)	- Hacer un diagnóstico incorrecto en negativo (e.g. Dice que no tienes la gripe, cuando sí que la tienes )
	Negative	- Dar un diagnóstico positivo cuando realmente no lo es (e.g. Dice que tienes la gripe, pero no la tienes)	- Decir que no hay ninguna enfermedad, siendo esta información cierta

La matriz de confusión sirve para comparar lo que predice el sistema (positivo o negativo) frente a la realidad del usuario (positivo o negativo), ilustrando así las distintas consecuencias de un diagnóstico correcto o incorrecto. En el ejemplo de la gripe, si el sistema indica que tienes la enfermedad y en efecto la tienes, hablamos de un diagnóstico correcto (predicción positiva con realidad positiva). Sin embargo, si el sistema asegura que no estás enfermo cuando sí lo estás, se produce un error grave (predicción negativa con realidad positiva) que puede retrasar el tratamiento y afectar tu salud. Por otro lado, si te diagnostican la gripe pero en realidad no la tienes (predicción positiva con realidad negativa), es un falso positivo que puede generar tratamientos innecesarios. Finalmente, cuando el sistema confirma que no hay enfermedad y en verdad no la hay (predicción negativa con realidad negativa), se considera un diagnóstico acertado. De esta forma, la matriz de confusión ayuda a entender el impacto de cada uno de estos aciertos y errores, tanto para el usuario como para los demás involucrados en el proceso.

Hemos encontrado la matriz extremadamente útil. Nos ha permitido visualizar de forma clara la clasificación de las predicciones del sistema con la realidad, facilitando la identificación de

aciertos y errores en los diagnósticos. Este enfoque ha ayudado a comprender mejor las implicaciones de cada resultado, ayudando a ver que errores son más críticos y cuales nos conviene más evitar.

En este caso, pensamos que el error que es mejor evitar, es el falso negativo, ya que no tratar a una persona con urgencia puede empeorar su estado de salud a largo plazo.

Lean Canvas

El Lean Canvas es una herramienta de planificación estratégica que ayuda a crear un plan de negocios de manera rápida y sencilla. Consta de nueve secciones que permiten estructurar de manera clara y concisa los elementos clave de un modelo de negocio: problema, solución, indicadores clave, propuesta de valor, ventaja competitiva, segmentos de clientes, canales, estructura de costos y fuentes de ingresos. Utilizamos esta herramienta como guía para definir nuestro proyecto. Además, al ir la actualizando podemos observar como evolucionan nuestras ideas gracias a la implementación de mejoras o nuevas ideas.

The Lean Canvas

Telefónica Discovery

Designed by: Carla Atienza, Nils Duran, Alba Figueras y Lola Monroy

Date: 1/04/2025

Version: 2

<b>Problem</b> <div>1. Diferentes problemas requieren conocimientos especializados variantes.</div> <div>2. Agentes homogéneos pueden generar menor creatividad e innovación. La falta de perspectivas diversas afecta a la resolución de problemas complejos.</div>	<b>Solution</b> <div>La solución implementará un enfoque híbrido, combinando APIs con la ejecución local de modelos open-weights utilizando tecnologías como LangChain y el OpenAI Agents SDK, con el objetivo de optimizar la toma de decisiones.</div>	<b>Unique Value Prop.</b> <div>Nuestra solución no solo mide la diversidad, sino que la convierte en una ventaja competitiva tangible mediante simulaciones interactivas y análisis de impacto, permitiendo optimizar dinámicamente la composición de "equipos de agentes" para maximizar su creatividad y rendimiento.</div>	<b>Unfair Advantage</b> <div>La diversidad de nuestros agentes tiene en cuenta características de todo tipo, no sólo estudios o conocimientos, sino también el sesgo intrínseco de los datos con los que se entrenan los LLMs o sesgos culturales.</div>	<b>Customer Segments</b> <div>Target Customers<ul style="list-style-type: none"><li>- Telefónica (interno)</li><li>- Equipo de trabajadores de Telefónica</li></ul></div>
<b>Existing Alternatives</b> <div>1. Un solo agente</div> <div>2. Método de Data Science o modelos de ML tradicionales</div> <div>3. ChatGPT</div>	<b>Key Metrics</b> <div>1. Diversidad de los perfiles (con un valor numérico o niveles categóricos, baja, media o alta diversidad).</div> <div>2. ¿Consenso o no? ¿Los distintos agentes están de acuerdo con las decisiones? ¿A quién hacemos caso entonces?</div> <div>3. Benchmarks de medicina para evaluar objetivamente la toma de decisiones (AgentClinic, MMMU, MMLU, MedMCQA, MedQA-USMLE, PubMedQA).</div>	<b>High-Level Concept</b> <div>Unión en la Diversidad</div>	<b>Channels</b> <div>La solución incluirá una UI personalizada integrada en el sistema de Telefónica, diseñada para interactuar con la infraestructura interna existente. También se ofrecerá como una solución independiente para empresas que prefieran mantener los modelos locales en su propio entorno.</div>	<b>Early Adopters</b> <div>Investigadores de Telefónica Discovery o directivos, para convencerles de la importancia de tener múltiples agentes especializados en la toma de decisiones, en lugar de depender sólo de un solo LLM general.</div>
<b>Cost Structure</b> <div>- Infraestructura local: Servidores, almacenamiento y mantenimiento de hardware para la ejecución de modelos open-weights.</div> <div>-Licencias: Herramientas y plataformas necesarias como LangChain, OpenAI Agents SDK.</div> <div>-Personal especializado: Ingenieros de IA, científicos de datos, desarrolladores de software, y expertos en infraestructura para asegurar el correcto funcionamiento y actualización de los modelos.</div> <div>-Mantenimiento: Costos asociados a la actualización de modelos y soporte continuo.</div> <div>-Consultoría: Enfoques personalizados para clientes corporativos, adaptando la solución a sus necesidades específicas.</div>			<b>Revenue Streams</b> <div>-Consultoría personalizada: Ingresos por proyectos o tarifas por hora.</div> <div>-Licencias de uso: Ingresos recurrentes por licencias mensuales/anuales.</div> <div>-Suscripción: Ingresos por suscripción mensual/anual.</div> <div>-Suscripción Premium: Ingresos por suscripción con acceso a características avanzadas.</div> <div>-Anuncios: Ingresos por publicidad en la versión gratuita de la plataforma.</div>	

© 2025 Telefónica Discovery. All rights reserved. | Última actualización: 1/04/2025

Este documento es una herramienta de planificación estratégica. No constituye una oferta de inversión ni asesoramiento financiero. PowerPoint implementation by: Néstor Chirinos Limited

Imágenes de fondo: Pexels.com, Unsplash.com, Freepress.com, CC BY-SA 3.0

Después de nuestra reunión con Lars y las diversas actividades de Design Thinking, con una idea mucho más clara de cómo sería nuestro proyecto actualizamos el Lean Canvas. Las modificaciones clave que realizamos son varias, especificamos las tecnologías clave que usaremos en las soluciones. La estructura de costos ahora refleja la apuesta por ejecutar los modelos de manera local, minimizando la necesidad de infraestructura en la nube, además mencionamos el costo de personal, el equipo necesario para desarrollar, mantener y escalar la solución. En los flujos de ingreso mantuvimos las suscripciones y añadimos consultorías personalizadas y licencias de uso de la plataforma para empresas. Conseguimos resumir

mejor nuestro reto y presentación con el eslogan “Unión en la Diversidad”. A medida que avanzamos, seguimos enfocados en mejorar y adaptar nuestra propuesta para ofrecer soluciones innovadoras y eficientes. Siempre con el objetivo de poner la diversidad como un motor fundamental en la toma de decisiones.

## Stakeholders

Los “Stakeholders” son todas las personas, grupos u organizaciones que pueden influir o verse afectadas por el desarrollo de un proyecto. Este concepto abarca desde usuarios finales y clientes hasta colaboradores, proveedores, inversores y comunidades. Entender quiénes son los stakeholders, sus expectativas y necesidades es esencial para asegurar que las soluciones diseñadas sean relevantes, sostenibles y cuenten con el apoyo necesario para su implementación exitosa.

En nuestro proyecto, los stakeholders son elementos clave, ya que su involucramiento y retroalimentación aseguran que la solución responda a necesidades reales y cumpla con altos estándares de calidad. Esto incluye:

- **Pacientes:** Usuarios finales que dependen de diagnósticos precisos para su salud.
- **Profesionales Médicos:** Médicos y personal sanitario que utilizarán el sistema como herramienta de apoyo y que asumen responsabilidad en el proceso de atención.
- **Instituciones Sanitarias:** Hospitales y clínicas cuyo prestigio y reputación están vinculados a la calidad de la atención médica ofrecida.
- **Equipo de Desarrollo:** Profesionales de TI y diseñadores encargados de construir, mantener y mejorar el sistema.
- **Inversores y Socios Comerciales:** Quienes aportan recursos y esperan un retorno de la inversión, siendo fundamentales para la viabilidad económica del proyecto.
- **Organismos Reguladores y Aseguradoras:** Entidades que velan por el cumplimiento de normativas y la seguridad del sistema, garantizando que se respeten estándares como el GDPR y la EU AI Act.
- **Comunidades de Pacientes:** Grupos que pueden influir a través de sus experiencias y feedback, ayudando a afinar la herramienta en función de las expectativas y necesidades reales.

## Benchmarking

El benchmarking es una técnica de análisis comparativo que se utiliza para evaluar el desempeño, procesos o resultados de una organización, producto o servicio frente a los de la competencia o referentes del sector. La idea es identificar las mejores prácticas y áreas de mejora para implementar cambios que optimicen el rendimiento propio.

Hemos desarrollado un documento en el que representamos un eje. En este eje se representa el porcentaje de aciertos (del 0% al 100%) y, a medida que se avanza, se van anotando diferentes niveles de exactitud: desde un diagnóstico perfecto (sin errores) hasta diagnósticos con algún margen de equivocación.

El objetivo principal es determinar un estándar o “baseline” que sirva de referencia para evaluar el rendimiento de un sistema o modelo de IA, comparándolo con la precisión humana u otros métodos. También se discuten los niveles mínimos de confianza aceptables para que el usuario final confíe en el diagnóstico, teniendo en cuenta el impacto que puede tener un error (por ejemplo, un falso negativo o un falso positivo) en ámbitos sensibles como el de la salud. De este modo, se busca establecer en qué punto de la escala el sistema ofrece una precisión suficiente para su uso real, considerando tanto la seguridad del paciente como las expectativas de fiabilidad.

- **Diagnóstico Perfecto (100% acierto):** Indica que el sistema logra identificar correctamente la condición médica en todos los casos, sin incurrir en errores.
- **Errores en el Diagnóstico:** Se diferencian dos tipos principales (explicados también en la matriz de confusión):
  - **Falsos Positivos:** Cuando el sistema diagnostica una condición (por ejemplo, la gripe) en un paciente que en realidad no la padece. Esto puede llevar a tratamientos innecesarios o alarmas infundadas.
  - **Falsos Negativos:** Cuando el sistema no detecta una condición presente en el paciente, lo cual es especialmente crítico en el ámbito médico, ya que puede retrasar intervenciones urgentes.

Como los umbrales mínimos de confianza representan el nivel mínimo de precisión que el sistema debe alcanzar para que se considere seguro y fiable, debemos hacer que estos sean muy altos porque el campo de la medicina no se puede permitir casi margen de error.. Es decir, antes de emitir un diagnóstico, nuestro sistema debe demostrar, por ejemplo, que tiene un 90-95% de certeza en sus predicciones. Esto garantiza que los diagnósticos emitidos minimicen los riesgos de errores críticos, como falsos negativos o falsos positivos, que podrían tener consecuencias graves para el paciente. Estos umbrales actúan como una medida de control de calidad, asegurando que solo se presenten resultados cuando la confianza del sistema es suficientemente alta, lo que refuerza la seguridad y la fiabilidad de la solución en un entorno clínico.

¿Nos ha sido útil?

Sí, ha sido muy útil. Nos ha permitido establecer claramente los parámetros de seguridad y fiabilidad necesarios para que el sistema pueda ofrecer diagnósticos precisos, lo cual es fundamental en el contexto clínico. Este enfoque nos ayuda a identificar cuándo el sistema está listo para su uso y a priorizar mejoras en aquellos casos en que los umbrales mínimos de confianza aún no se cumplen.

## User Journey - UX of AI

El “User Journey” es el recorrido que experimenta un usuario al interactuar con un producto o servicio, desde el primer contacto hasta la etapa final de uso. Esta herramienta mapea cada fase de la experiencia del usuario, identificando puntos de interacción, momentos clave, posibles obstáculos y oportunidades de mejora. Analizar el user journey permite optimizar la experiencia del usuario, garantizando que cada paso sea intuitivo, satisfactorio y alineado con sus expectativas y necesidades.

Cada una de estas herramientas contribuye de manera complementaria al proceso de innovación, ofreciendo perspectivas y metodologías específicas para abordar los desafíos de diseño de forma integral y centrada en las personas.

<b>Explicabilidad</b> - <i>¿Cómo ayudaremos a nuestros usuarios a entender ciertos resultados?</i> Desglose de decisión/opinión de cada agente.  Recursos/documentos/experiencias usadas por el agente para tomar la decisión.	<b>Gestión de Expectativas</b> <i>¿Cómo estableceremos expectativas realistas?</i>  Mensajes de advertencia claros.	<b>Fallo Elegante y Responsabilidad</b> <i>¿Cómo gestionaremos la confianza en caso de error?</i>  Mostrar pregunta y respuesta; explicar errores.
<b>Retroalimentación del Usuario</b> <i>¿Cómo podrá el usuario proporcionar feedback al sistema?</i>  Formularios, encuestas, sugerencias.	<b>Autonomía del Usuario</b> <i>¿Cómo podrá el usuario personalizar su experiencia?</i>  Opciones de personalización (p.ej., propuestas de equipo).	<b>Privacidad y Seguridad de los Datos</b> <i>¿Cómo protegemos los datos? (ej.: GDPR + EU AI Act)</i>  Cumplir GDPR y EU AI Act.
<b>Traducción Computacional</b> <i>¿Cómo transformaremos</i>	<b>Sesgo e Inclusividad</b> <i>¿Cómo preveniremos el sesgo y garantiremos la inclusión?</i>	<b>Ética y Consecuencias No Intencionadas</b> <i>¿Cómo vigilarémos el impacto negativo y positivo?</i>

<i>las necesidades en requisitos?</i>  Convertir necesidades en requisitos claros.	Selección de agentes diversos	Seguimiento de decisiones y propuestas de mitigación.
<b>Otros Desafíos (Diseño)</b> <i>¿Qué otros desafíos prevés?</i>  Integración tecnológica en personas reticentes, escalabilidad, interoperabilidad.		

Además de clarificar responsabilidades, este esquema también puede ser útil para evaluar y priorizar nuevas funcionalidades. También puede servir de referencia en futuras iteraciones del proyecto y facilitar la comunicación con stakeholders y clientes. Por último, sirve para identificar áreas de riesgo y oportunidades de mejora. Establecer un marco para la auditoría y seguimiento del sistema. Por lo que nos ha sido extremadamente útil para amueblar nuestro pensamiento.

## **4. Arquitectura del Sistema**

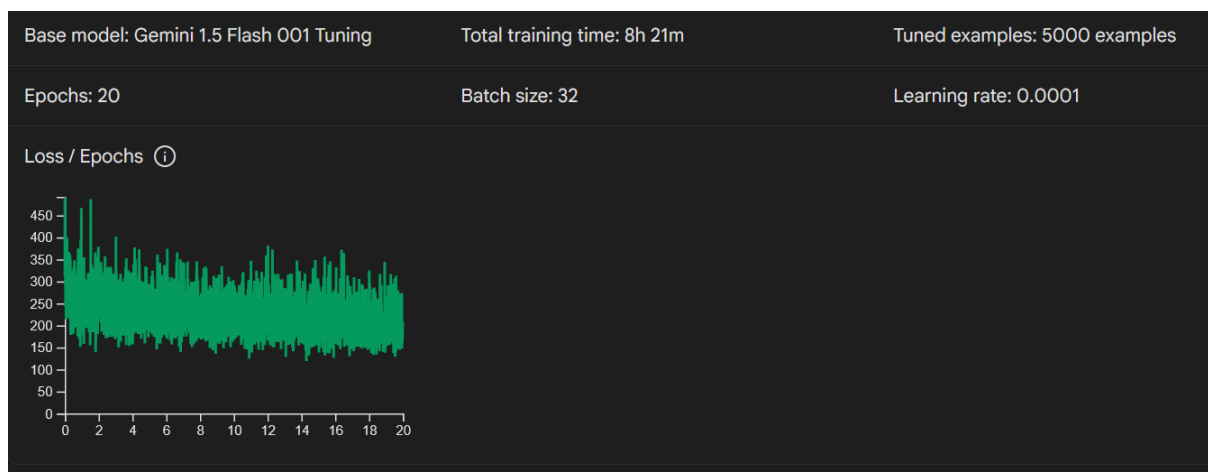
El sistema está diseñado como una arquitectura modular que simula un entorno colaborativo entre varios agentes con perfiles especializados. Cada módulo cumple una función específica dentro del flujo general, está implementado de tal forma que se puede reutilizar o experimentar con ello.

### **4.1. Módulo de agentes especializados**

Representamos a 5 agentes basados en modelos de lenguaje (LLMs), cada uno especializado en un campo de la medicina:

- Medicina general
- Cirurgia
- Pediatría i ginecología
- Patología i farmacología
- Ciències bàsiques (anatomia, fisiologia)

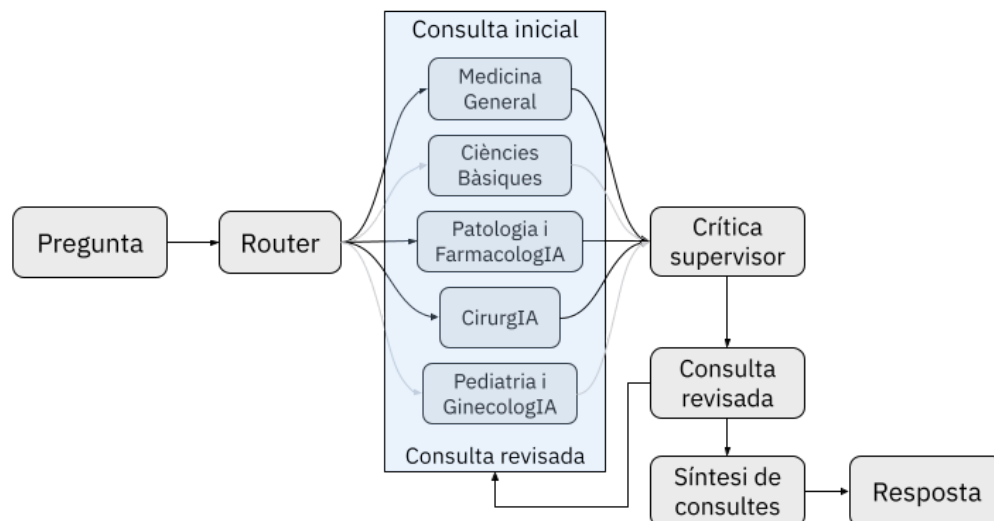
Para simular estos especialistas médicos utilizamos un fine-tuning con LoRA sobre el modelo base Gemini 1.5 Flash empleando los datos de MedMCQA, que contiene preguntas especializadas en distintos ámbitos de la medicina. Esto nos permitió ajustar un modelo de lenguaje preentrenado de manera eficiente y con un coste computacional bajo. Los parámetros usados en el entrenamiento son los siguientes:



### **4.2. Módulo de interacción y coordinación de agentes**

Para implementar el debate entre agentes, hemos utilizado LangGraph, que permite definir un flujo de trabajo secuencial y condicional. Este módulo tiene diferentes componentes:

- Router: Selecciona a los expertos más relevantes para la pregunta, basándose en la similitud semántica y la opción de diversidad especificada.
- Consulta al experto inicial: Recopila los primeros análisis y conclusiones de los expertos seleccionados sobre la pregunta.
- Crítica del supervisor: Un supervisor revisa las respuestas iniciales de los expertos y genera una crítica constructiva para ayudarles a refinar sus análisis.
- Consulta revisada: Los mismos expertos consultados inicialmente reevalúan la pregunta, teniendo en cuenta la crítica del supervisor para mejorar sus respuestas.
- Síntesis del supervisor: Un supervisor final integra las respuestas (revisadas o iniciales) de los expertos para producir un informe o respuesta consolidada y coherente.



## 4.3. Módulo de benchmarking y evaluación

Hemos evaluado nuestro sistema de dos formas distintas:

### 4.3.1. Evaluación objetiva con preguntas cerradas

En primer lugar, para una evaluación objetiva, seleccionamos preguntas del dataset MedQA, que contiene consultas médicas ficticias con sus correspondientes diagnósticos, en formato de opción múltiple. Este enfoque nos permite evaluar la precisión de las respuestas de forma objetiva sin tener conocimientos del dominio.



Comparamos el rendimiento de:

- Agentes individuales
- Grupos homogéneos de agentes
- Sistemas multiagente con diversidad baja, media o alta (controlada a través del router)

También experimentamos con parámetros como:

- Número de agentes consultados
- Grado de diversidad funcional (baja/media/alta)

La temperatura de los modelos, para observar el efecto de la variabilidad en las respuestas

#### 4.3.2. Evaluación cualitativa con casos clínicos abiertos:

Para casos más abiertos y con un componente ético, hemos implementado una variante que llamamos “modo batalla”. Este método busca eliminar los sesgos a la hora de juzgar las respuestas generadas. El funcionamiento es sencillo: Se seleccionan dos combinaciones aleatorias de agentes (puede ser un solo agente o un equipo de hasta 5, con más o menos diversidad), y se hace una misma consulta médica a ambos sistemas por separado.

Seguidamente, el sistema muestra solo las dos respuestas finales generadas por cada combinación, y el usuario vota cuál de las dos le parece mejor (o puede declarar empate). Solo cuando ya se ha registrado el voto, el sistema revela qué combinación de agentes había detrás de cada respuesta, evitando sesgos inconscientes en la decisión del usuario.

Aunque los que hemos evaluado los diagnósticos somos nosotros, que no somos expertos en el dominio, la valoración se realiza más en función de la prudencia y argumentación de las respuestas, además de que se trata de preguntas abiertas que no tienen una solución predeterminada.

Además, hemos puntuado también estas respuestas mediante el sistema de puntuación Elo [8] a partir de las votaciones de los usuarios, que evalúan la calidad de la salida de los LLMs.

Finalmente, hemos ordenado las configuraciones de agentes según estas métricas de rendimiento, lo cual nos ha ayudado a determinar si en casos clínicos con dilemas éticos, la diversidad genera respuestas más matizadas.

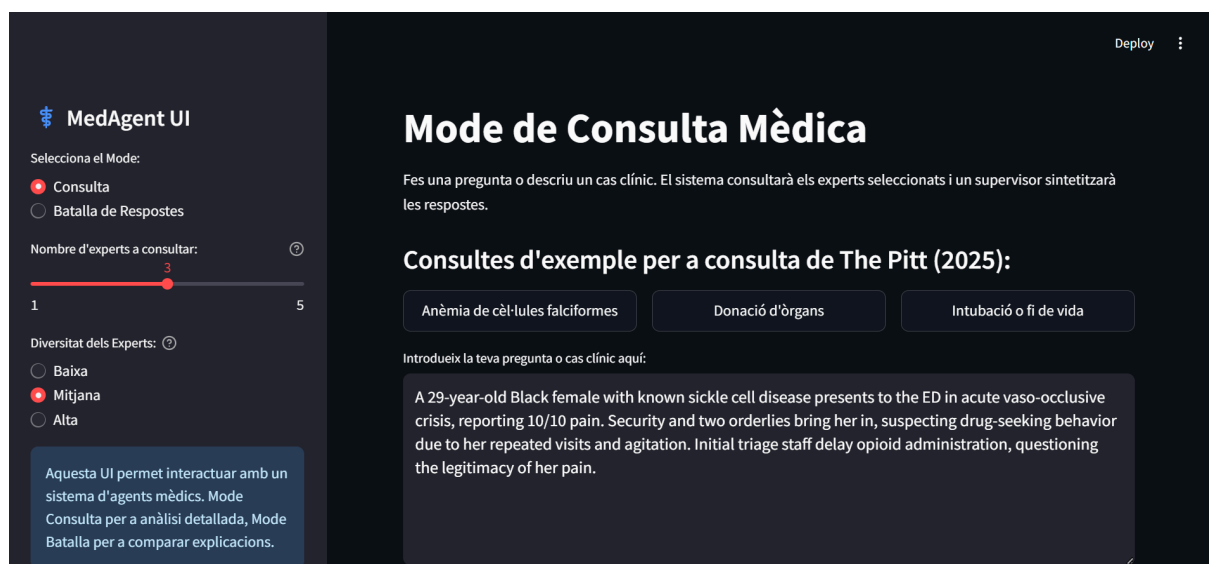
## 4.4. Módulo de interfaz de usuario

Construimos una interfaz web para la interacción del usuario con el sistema. En esta pueden configurar el Modo al que quieren acceder, ya sea “Consulta” o “Batalla de Respuestas”, también pueden regular el número de expertos (1-5) y el nivel de diversidad (Bajo, Medio, Alto)


En el Modo “Consulta” el usuario puede configurar el número de expertos que necesita y cuánta diversidad quiere que haya en el equipo. Después debe introducir un caso clínico. Así observará:

- Las respuestas individuales de cada agente
- La explicación del supervisor
- Las respuestas revisadas de cada agente
- La síntesis final que incluye los acuerdos y desacuerdos entre agentes y la conclusión general

El objetivo es hacer visible el razonamiento colaborativo entre IA y facilitar el análisis cualitativo del sistema.



En el Modo “Batalla de Respuestas” el usuario introduce un caso clínico y escoge la mejor respuesta entre dos grupos de agentes anónimos. Una vez haya escogido la mejor respuesta, se revelará qué nivel de diversidad corresponde a cada grupo. Con esto podemos evaluar, y hacer partícipes a los propios usuarios, si la diversidad afecta realmente a las respuestas.

 **MedAgent UI**

Selecciona el Mode:

☐ Consulta

☒ Batalla de Respostes

Mode Batalla

- Es generen automàticament dues respostes diferents
- Cada resposta té una configuració aleatòria de nombre d'experts i diversitat
- Pots comparar i triar la resposta que prefereixis

Aquesta UI permet interactuar amb un sistema d'agents mèdics. Mode Consulta per a anàlisi detallada, Mode Batalla per a comparar explicacions.

Deploy

Consultes d'exemple per a batalla de The Pitt (2025):

Anèmia de cèl·lules falciformes

Donació d'òrgans

Intubació o fi de vida

Introdueix la teva pregunta aquí per a la batalla:

A 29-year-old Black female with known sickle cell disease presents to the ED in acute vaso-occlusive crisis, reporting 10/10 pain. Security and two orderlies bring her in, suspecting drug-seeking behavior due to her repeated visits and agitation. Initial triage staff delay opioid administration, questioning the legitimacy of her pain.

Iniciar Batalla / Nova Pregunta

Opció A

Opció B

Opció A (# Experts: 2, Diversitat: Mitjana)

Empat

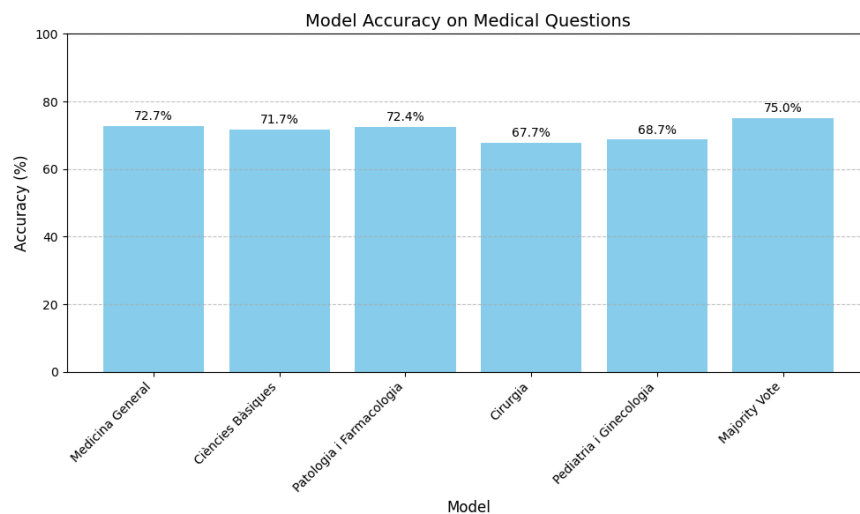
Opció B (# Experts: 4, Diversitat: Baixa)

Has votat per: Opció B. Les configuracions s'han revelat als botons.

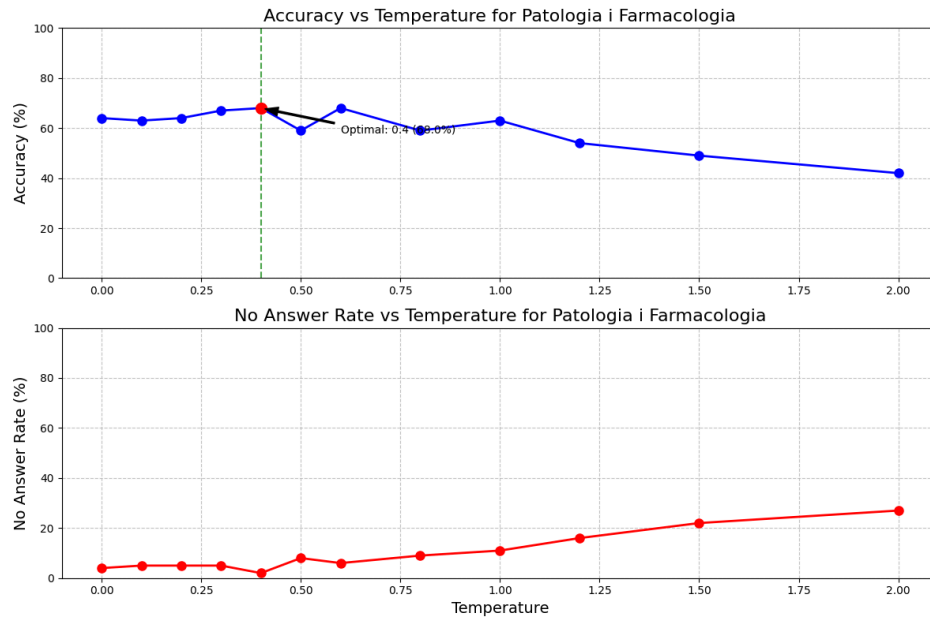
Per a una nova batalla, introdueix una nova pregunta i prem 'Iniciar Batalla / Nova Pregunta'.

## **5. Evaluación del Sistema**

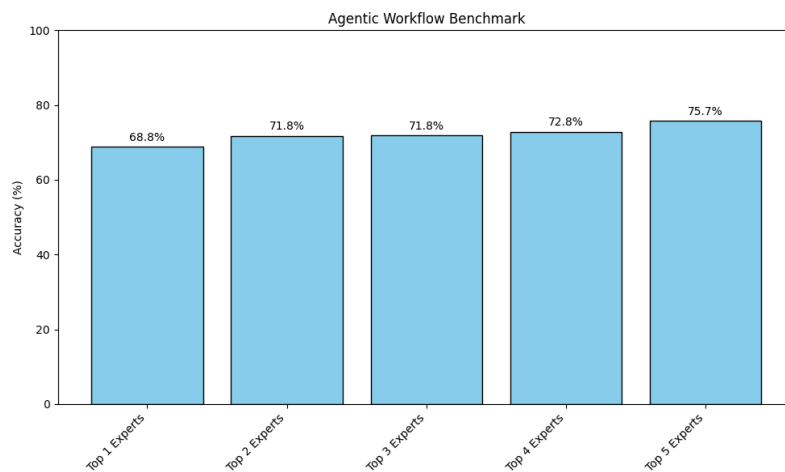
Primero de todo, quisimos evaluar los modelos adaptados de forma individual. Algunos tienen rendimientos ligeramente superiores a otros, pero todos se sitúan alrededor de un 70% de precisión. Los modelos SOTA (State of the Art) alcanzan una precisión del 96,5%, por lo que los nuestros quedan bastante lejos, aunque hay que tener en cuenta el tamaño de los modelos en sí. Los modelos individuales superan en precisión a los mejores modelos abiertos disponibles en enero de 2024 (Meditron 70B, con un 70,2%), pero están por debajo de Med-PaLM 2 (86,5% en mayo de 2023). La moda (es decir, el voto mayoritario) de los modelos individuales alcanza un 75% de precisión, comparable a GPT-4o mini.



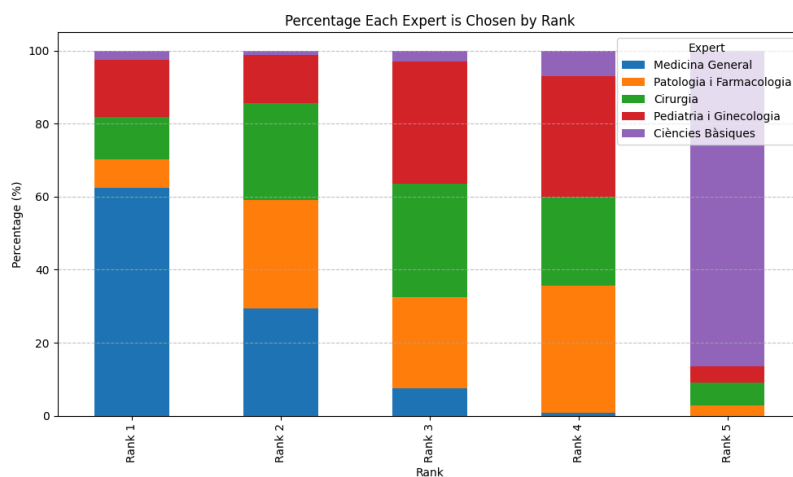
Después, quisimos evaluar el impacto de la temperatura sobre la precisión y la tasa de errores de tipo refusal (cuando el modelo no devuelve una respuesta en el formato esperado, incluso tras un reintento). Observamos que, a medida que aumenta la temperatura —especialmente con valores superiores a 1—, disminuye la precisión y aumenta el porcentaje de rechazos. Hay un punto intermedio entre 0 y 1 (entre 0,3 y 0,7) donde los modelos ofrecen mejores resultados. Un poco de aleatoriedad ayuda al modelo a no quedarse bloqueado.



Para finalizar la parte de evaluación del *benchmark*, quisimos estudiar la hipótesis del proyecto: si la diversidad mejora o no el rendimiento de los sistemas multiagente. Aunque el aumento es moderado, vemos que, efectivamente, cuantos más agentes participan, mayor precisión se obtiene, mejorando ligeramente también la precisión de la moda.



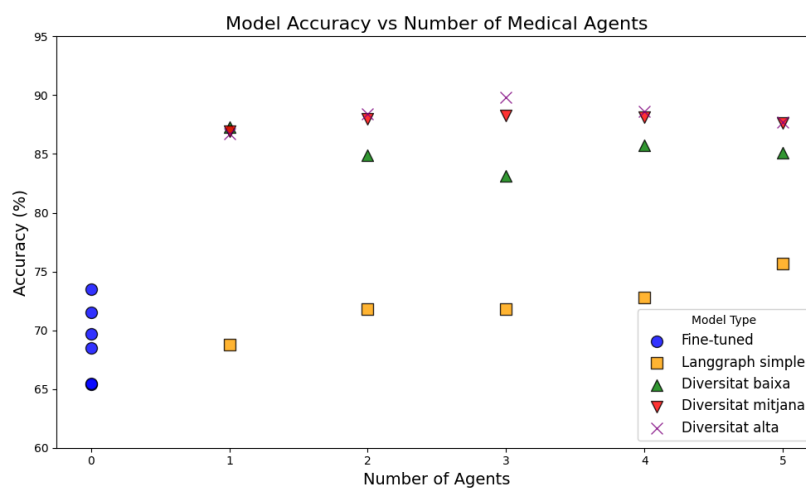
Dado que LangGraph decide a qué experto consultar, también nos pareció interesante analizar cuáles prioriza (elige primero) y cuáles no. Una arquitectura del tipo Mixture-of-Experts intenta equilibrar el uso de sus expertos para que todos resulten más o menos igual de útiles. En cambio, nuestros expertos tienen dominios específicos que no se aplican por igual en todos los casos.



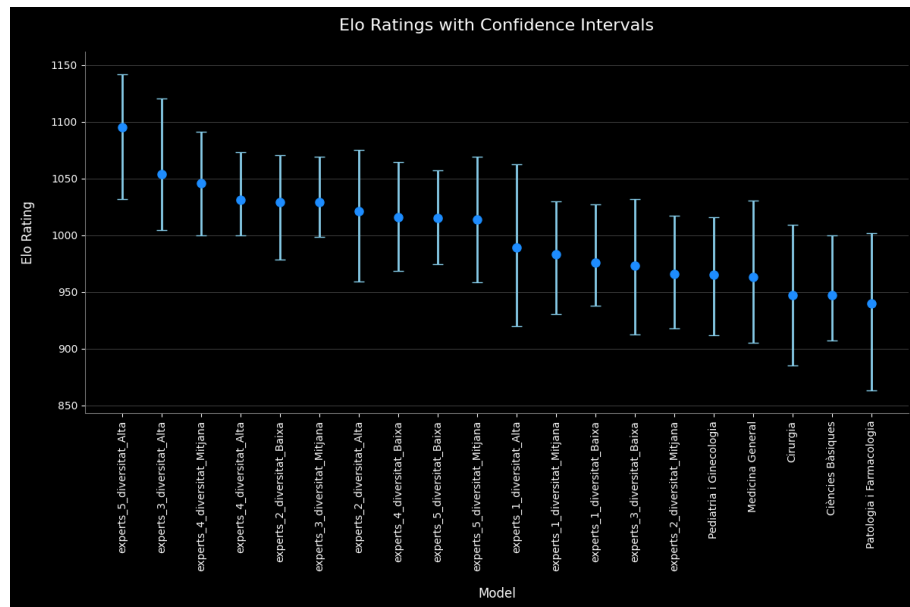
Posición media	
Medicina General	1.47
Cirurgía	2.87
Patología y Farmacología	2.95
Pediatría y Ginecología	2.98
Ciencias Básicas	4.74

El agente experto en medicina general es, con diferencia, el más relevante en la mayoría de los dominios, con una posición media de 1,47. Le siguen los expertos en cirugía, patología, farmacología, pediatría y ginecología, con distribuciones algo distintas pero medias similares, ligeramente por debajo del rango 3. Por último, los expertos en ciencias básicas tienen un dominio muy limitado y casi siempre son elegidos en último lugar.

Evaluamos los expertos en el benchmark de MedQA i los comparamos con los resultados previos:



En la pantalla de resultados mostramos una clasificación de los modelos según su puntuación Elo, calculada a partir de los votos de los usuarios y del *benchmark* de MedQA. Además, presentamos una serie de gráficos para entender el impacto del número de agentes y de la diversidad en la precisión y el comportamiento de los modelos.



Los intervalos de confianza son muy grandes porque como nosotros somos los únicos usuarios, tenemos pocos votos. Pero es un problema que se solucionaría solo con el uso de más personas.

## **6. Principales Dificultades Encontradas y Solución**

Uno de los retos principales que nos encontramos con este proyecto fue decidir cómo plantearlo, partiendo de una hipótesis tan abstracta y conceptual: evaluar el impacto de la diversidad entre agentes de inteligencia artificial. A diferencia de otros proyectos más técnicos o aplicados, aquí el foco estaba en una hipótesis de investigación, lo cual requería tomar decisiones de diseño que dieran sentido al experimento.

La primera pregunta que nos hicimos es, ¿Qué es la diversidad, en los agentes de inteligencia artificial? Podíamos imaginar el concepto de diversidad en términos humanos (género, origen, edad...), pero rápidamente vimos que eso no podía trasladarse directamente a agentes de IA. Por tanto, redefinimos el concepto en términos funcionales, y optamos por una diversidad técnica basada en la especialización de los agentes.

Además, para limitar el alcance de nuestro proyecto y hacerlo más concreto y evaluable, decidimos centrarnos en el ámbito médico. Así, cada agente representaría una especialidad médica distinta, simulando diferentes perspectivas basadas en el área de conocimiento.

Una vez definido el enfoque, nos enfrentamos a retos más técnicos. Por un lado, fue un desafío crear cinco agentes distintos basados en modelos de lenguaje especializados. En un inicio queríamos implementar un RAG para diseñar estos agentes expertos, pero nos dimos cuenta que, debido a la falta de recursos y la gran cantidad de datos médicos necesarios para esta tarea, esta opción no era viable. Finalmente, el uso de fine tuning sobre modelos base con la ayuda de Google AI Studio nos permitió obtener agentes especializados por un menor coste.

Por otro lado, diseñar una interfaz que permitiera coordinar las respuestas de los agentes, simular un debate iterativo y generar un informe conjunto también supuso una dificultad. La herramienta LangGraph nos ayudó a diseñar este flujo de forma sencilla y gratuita, estructurando la interacción de forma modular y escalable.



## **7. Organización del Equipo y Tareas**

El desarrollo del proyecto se organizó en torno a cuatro grandes bloques de trabajo que nos han servido para estructurar y repartir las tareas:

### **WP1. Planificación y gestión del proyecto**

Abarca todas las tareas de planificación y organización interna del proyecto: desde la lluvia de ideas, definición de objetivos, el alcance, la planificación temporal y la gestión de riesgos.

### **WP2. Desarrollo y configuración de los agentes**

En este bloque definimos la arquitectura multiagente del sistema. La identificación de los tipos de agentes necesarios para simular especialidades médicas. El desarrollo adaptando modelos de lenguaje con un fine-tuning específico a cada agente, y utilizando los datos médicos de cada especialidad de la base de datos MedMCQA. Y finalmente, programando la lógica de interacción entre agentes usando LangGraph, que permite controlar el flujo de turnos y coordinar el debate entre ellos de forma estructurada.

### **WP3. Diseño e implementación del sistema**

Este work package incluye el diseño y desarrollo de una interfaz de usuario interactiva mediante Streamlit, que permite introducir los casos clínicos y visualizar el proceso de razonamiento entre agentes.

También se definieron y ejecutaron casos de prueba para validar el comportamiento del sistema, y se documentaron todos los elementos clave del diseño.

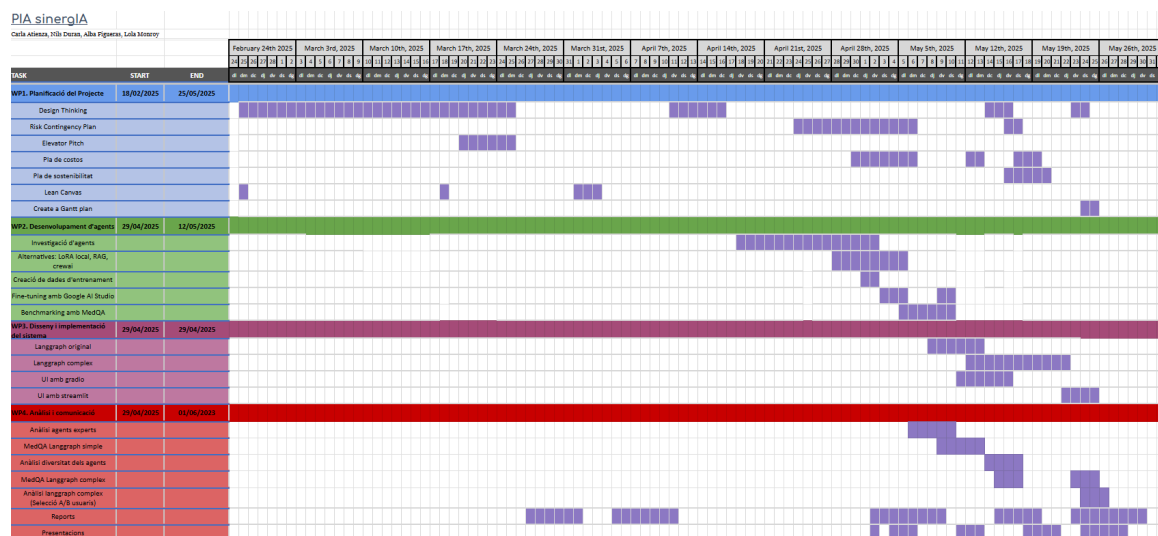
### **WP4. Análisis y comunicación**

El último bloque se ha centrado en el análisis de resultados y la comunicación del proyecto. Se han analizado las respuestas generadas por los agentes (individuales y colaborativas) y se han comparado con dos datasets distintos:

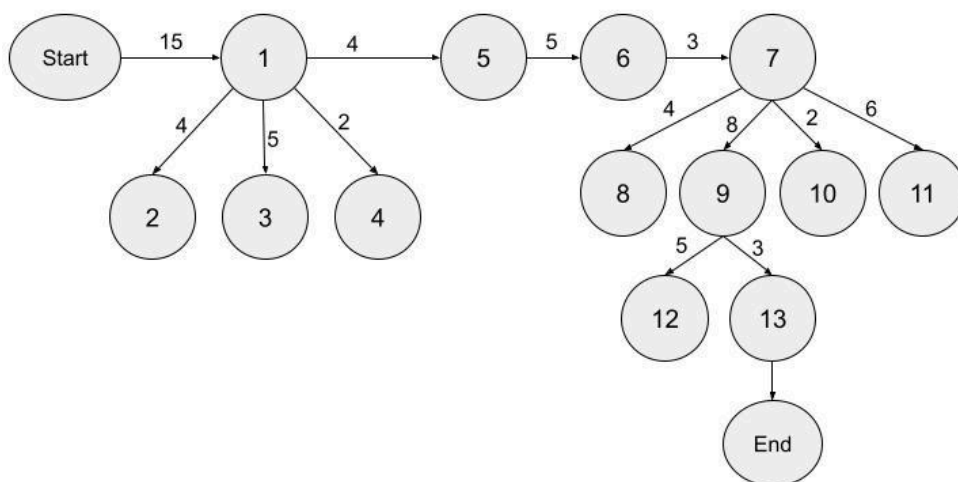
- el dataset MedQA para una comparación objetiva
- consultas médicas ficticias con dilemas éticos para evaluar si la diversidad lleva a soluciones más éticas, y si se logra alcanzar un acuerdo entre los diferentes agentes.

## 8. GANTT y PERT

Para tener una herramienta visual que nos permitiera observar el progreso de nuestro proyecto, hemos creado un gráfico de GANTT. Aunque no lo utilizamos para planificar las tareas, al hacerlo de forma retrospectiva nos muestra su utilidad, ya que vemos cómo la gran mayoría de las tareas las hemos acabado realizando en la segunda mitad o el último tercio del curso. Es cierto que ya sabíamos que el tema era bastante abstracto y que debíamos concretarlo de alguna manera, pero si hubiéramos hecho y seguido un GANTT, probablemente no habríamos ido tan estresados al final, ya que nos habríamos dado cuenta antes de que la planificación temporal no coincidía con el progreso real. Aun así, hemos alcanzado todos los objetivos y tareas con éxito.



## PERT



### Tareas

- 1) Investigación de agentes
- 2) LoRA local
- 3) RAG
- 4) CrewAI
- 5) Búsqueda de datos de entrenamiento
- 6) Fine-tuning con AI Studio
- 7) Benchmarking con MedQA
- 8) LangGraph original
- 9) LangGraph complejo
- 10) Interfaz de usuario con Gradio
- 11) Interfaz de usuario con Streamlit
- 12) Análisis del benchmark MedQA
- 13) Análisis cualitativo A/B

## **9. Presupuesto y Viabilidad Económica**

El plan de costes es un elemento crucial para la viabilidad del proyecto. Como startup conformada por un equipo de 4 trabajadores, lo detallaremos siguiendo el desarrollo de nuestro sistema. Definiremos los recursos necesarios y estableceremos un marco para la gestión eficiente de los mismos. Además, incluiremos la estimación de costes y la importancia del análisis de viabilidad financiera para asegurar el éxito del proyecto.

### **9.1 Estimación de Costes por Tarea**

Estimamos los costes para cada Work Package, incluyendo los costes de salario (tiempo estimado dedicado a cada tarea) y los recursos utilizados. Las cifras calculadas representan un valor simbólico del trabajo realizado. Para el cálculo del dinero necesario, tenemos que considerar que somos siempre los mismos 4 profesionales, y que cobraremos 30 euros la hora.

#### **WP1. Project Planning and Management**

Este Work Package abarca todas las actividades relacionadas con la iniciación, planificación, ejecución, seguimiento y cierre del proyecto. Incluye la definición del alcance del proyecto, la creación del cronograma del proyecto, la asignación de recursos, la gestión de riesgos, la realización de reuniones de seguimiento y la preparación del informe final del proyecto.

##### **Salario**

- Planificación de tareas y cronograma → 50 horas ( $50 \text{ h} \times 30 \text{ €/h} = 1\,500 \text{ €}$ )
- Gestión de riesgos → 10 horas ( $10 \text{ h} \times 30 \text{ €/h} = 300 \text{ €}$ )
- Reuniones de seguimiento y control de progreso → 50 horas ( $50 \text{ h} \times 30 \text{ €/h} = 1\,500 \text{ €}$ )
- Elaborar el reporte final del proyecto → 50 horas ( $50 \text{ h} \times 30 \text{ €/h} = 1\,500 \text{ €}$ )

Subtotal tareas:  $10 \text{ €} / \text{h} \times 4 \text{ personas} \rightarrow 4800 \text{ €}$

##### **Recursos**

- Google Meet → 0 €
- Google Drive → 0 €

Subtotal Recursos: 0 €

**Subtotal WP1:  $4800 \text{ €} + 0 \text{ €} = 4800 \text{ €}$**

#### **WP2. Agent Development and Configuration**

Este Work Package se centra en el diseño, desarrollo y configuración de los agentes inteligentes que participan en el proceso de toma de decisiones. Incluye la identificación de

los tipos de agentes necesarios, la definición de sus roles y capacidades, la adaptación de los modelos de lenguaje para representar diferentes especialidades, y la programación de la lógica de su interacción y colaboración.

### **Salario**

- Identificar tipos de agentes → 5 horas ( $5 \text{ h} \times 30 \text{ €/h} = 150 \text{ €}$ )
- Definir benchmarks → 20 horas ( $20 \text{ h} \times 30 \text{ €/h} = 600 \text{ €}$ )
- Adaptar los modelos de lenguaje para cada tipo de agente → 100 horas ( $100 \text{ h} \times 30 \text{ €/h} = 3\,000 \text{ €}$ )
- Programar la lógica de interacción entre agentes → 40 horas ( $40 \text{ h} \times 30 \text{ €/h} = 1\,200 \text{ €}$ )

Subtotal tareas:  $30 \text{ €} / \text{h} \times 4 \text{ personas} \rightarrow \mathbf{4\,950 \text{ €}}$

### **Recursos**

- LangChain/LangGraph → 0 €
- APIs para LLMs → 0.25 € / 1M tokens (Total 9 €)
- GitHub → 0 €
- Python, Visual Studio → 0 €

Subtotal Recursos: 9 €

**Subtotal WP2:  $4\,950 \text{ €} + 9 \text{ €} = 4\,959 \text{ €}$**

### **WP3. System Design and Implementation**

Este Work Package cubre el diseño e implementación de la arquitectura general del sistema, incluida la interfaz de usuario y la lógica que conecta a los agentes. Implica la creación de la interfaz para la entrada de datos, el diseño de casos de prueba para evaluar el rendimiento del sistema y la documentación de los detalles de diseño e implementación del sistema.

### **Salario**

- Implementar la interfaz para ingresar datos del paciente → 40 horas ( $40 \text{ h} \times 30 \text{ €/h} = 1\,200 \text{ €}$ )
- Diseñar casos de prueba para diferentes diagnósticos → 20 horas ( $20 \text{ h} \times 30 \text{ €/h} = 600 \text{ €}$ )
- Documentar el diseño del sistema y la implementación → 10 horas ( $10 \text{ h} \times 30 \text{ €/h} = 300 \text{ €}$ )

Subtotal tareas:  $30 \text{ €} / \text{h} \times 4 \text{ personas} \rightarrow \mathbf{2\,100 \text{ €}}$

### **Recursos**

- Streamlit → 0 €
- GitHub → 0 €

- Python, Visual Studio → 0 €
- Pandas → 0 €

Subtotal Recursos: 0 €

**Subtotal WP3: 2100€ + 0 € = 2100€**

#### WP4. Analysis and Communication

Este Work Package se centra en el análisis de los resultados obtenidos del sistema y la comunicación del progreso del proyecto a las partes interesadas relevantes. Incluye el análisis de los datos generados por el sistema, la evaluación de su rendimiento con respecto a los puntos de referencia definidos, la extracción de conclusiones y la preparación de informes y presentaciones para la comunicación con Telefónica.

#### Salario

- Analizar los resultados → 40 horas ( $40 \text{ h} \times 15 \text{ €/h} = 600 \text{ €}$ )
- Gestión de la comunicación con Telefónica/UPC → 40 horas ( $40 \text{ h} \times 10 \text{ €/h} = 400 \text{ €}$ )

Subtotal tareas: 15 € / h programadores; 10 € / h gestión → **1000 €**

#### Recursos

- Microsoft Teams → 0 €

Subtotal Recursos: 0 €

**Subtotal WP4: 1 000 € + 0 € = 1 000 €**

**Total estimado por tareas: 2400 € + 4959 € + 2100 € + 1000 € = 10459 € (valor simbólico del trabajo)**

## 9.2 Reservas

Este apartado contempla las reservas para cubrir posibles contingencias e imprevistos que puedan surgir durante la ejecución del proyecto.

- Reserva de Contingencia : Se destina un 10% del coste total estimado del proyecto para hacer frente a pequeños retrasos, problemas técnicos inesperados u otras eventualidades que puedan afectar al desarrollo.

- 1046 €

- Reserva de Project Manager: Se añade un 10% adicional sobre el total estimado (incluyendo la reserva de contingencia) para cubrir posibles costes asociados a la gestión del proyecto, como tiempo adicional de dedicación por parte del Project Manager o necesidades imprevistas de gestión.

- 1140,49 €

**Total reservas: 1046 € + 1150,49 € = 2 196,49 €.**

## 9.3 Presupuesto Total del Proyecto

Concepto	Total (€)
Costes estimados por tareas (tiempo estimado dedicado)	<b>10459</b>
Reservas (contingencia + project manager)	<b>2196,49</b>
Total estimado (valor simbólico)	<b>12655,49</b>

## 9.4 Análisis de Viabilidad Financiera

Analizamos la viabilidad económica del proyecto, considerando los costes y beneficios asociados a su desarrollo. Aunque el proyecto se enmarca dentro de un contexto académico, es importante evaluar su sostenibilidad financiera, especialmente si se llegase a contemplar una implementación a mayor escala. Para la realización de este proyecto, se cuenta con una financiación hipotética de 15 000 € proporcionada por Telefónica, que actúa como nuestro impulsor.

**Beneficio Neto**

Considerando la financiación de **15 000 €** y los costes a gastar de **12 644,50 €**, el cálculo simbólico del beneficio neto sería:

- **Ingresos:** 15 000 €
- **Costes:** 12 644,50 €
- **Beneficio Neto:** 15 000 € – 12 655,49 € = **2 344,51 €**

Este "beneficio neto" simbólico de 1275,75 € representa el excedente de la financiación de Telefónica una vez cubiertos los costes estimados del proyecto. En un escenario real, este excedente podría destinarse a:

- Inversión en mejoras y ampliación del sistema
- Desarrollo de nuevas funcionalidades
- Gastos de comercialización (si se decidiera lanzar el producto al mercado)

#### Retorno de la Inversión (ROI)

El Retorno de la Inversión (ROI) permite evaluar la rentabilidad o eficiencia de una inversión. Mide la ganancia obtenida por cada unidad de coste invertido. Indica cuánto beneficio se obtiene por cada euro gastado en el proyecto.

$$\text{ROI} = \frac{\text{Costes}}{\text{Beneficio Neto}} \times 100$$

$$\text{ROI} = \left( \frac{2355,50}{12644,50} \right) \times 100 \approx 18,63\%$$

Este ROI simbólico del 18,63 % indica que el valor del proyecto supera claramente el coste estimado del esfuerzo invertido, lo cual confirma su viabilidad y rentabilidad.

## 9.5 Seguimiento y Control del Presupuesto

El seguimiento del presupuesto permite no solo controlar el esfuerzo invertido, sino también desarrollar habilidades clave en gestión de proyectos. A pesar de que no hay un flujo de dinero real, aplicar metodologías como el *Earned Value Management (EVM)* y utilizar métricas de evaluación ayuda a detectar desviaciones, optimizar recursos y asegurar que se cumplen los objetivos dentro de los plazos establecidos.

#### Earned Value Management (EVM)

La *Earned Value Management* es una metodología que permite supervisar de manera integrada el alcance, el cronograma y los costes de un proyecto. A través de la comparación

entre lo planificado y lo realmente ejecutado, proporciona una fotografía cuantitativa del estado del proyecto en cada punto de control, facilitando la detección temprana de desviaciones y la adopción de acciones correctivas. Se aplicará de forma simbólica para evaluar el rendimiento del proyecto en función de tres variables clave:

- **Planned Value (PV):** Valor del trabajo planificado en cada fase (por ejemplo, desarrollo de agentes, pruebas, documentación), según el cronograma inicial.
- **Earned Value (EV):** Valor real del trabajo completado en cada momento, calculado en función del progreso real de cada tarea.
- **Actual Cost (AC):** Coste estimado en tiempo invertido por el equipo, comparado con lo inicialmente previsto.

Este enfoque permite responder a preguntas que plantean si estamos cumpliendo con lo planificado, si hemos invertido más tiempo del previsto o si el esfuerzo realizado refleja el avance esperado.

#### Métricas de Evaluación

Para evaluar si el proyecto se ha desarrollado de forma eficiente y dentro de lo previsto, se utilizarán las siguientes métricas:

- **Calidad del sistema:** Se analizará si simula correctamente la interacción de agentes diversos y si las decisiones médicas obtenidas son coherentes y justificadas. Tenemos que responder a la cuestión de si el agente responde bien.
- **Tiempo de desarrollo:** Se comparan las fechas previstas para cada paquete de trabajo con las fechas reales de finalización. Para que todo esté perfecto, se debería de cumplir el tiempo establecido.
- **Costes reales vs. planificados:** Aunque no hay costes económicos reales, se analizarán las horas de dedicación reales respecto a las estimadas. Una desviación significativa podría indicar sobrecarga, mala planificación o tareas no previstas.

Este control permitirá extraer conclusiones no solo sobre el cumplimiento del plan, sino también sobre la eficiencia del equipo y la validez del enfoque adoptado.



## 10. Gestión de Riesgos

La gestión de riesgos es un componente esencial para garantizar el éxito de cualquier proyecto, especialmente en entornos complejos y con recursos limitados como el nuestro. Este proceso nos permite anticipar, evaluar y responder a los posibles eventos que podrían afectar negativamente al desarrollo, calidad o resultados del sistema.

### 10.1 Asignación de Roles y Responsabilidades

Para garantizar la ejecución eficiente y la participación activa de todos los integrantes dentro del proyecto, es fundamental definir claramente los roles y responsabilidades tanto de los miembros del equipo como de los stakeholders clave. Por ello realizaremos esta asignación mediante la Matriz RACI, que es una herramienta que establece quién es el Responsable de realizar una tarea, quién es el Aprobador (responsable de la aprobación final), quién debe ser Consultado durante el proceso, y quién debe ser Informado sobre el progreso y los resultados.

La siguiente matriz proporciona una visión general de la asignación de roles y responsabilidades a lo largo del proyecto.

MATRIZ RACI

Tarea	Responsable	Aprobador	Consultor	Informado
Planificación de tareas y cronograma	Lola Monroy	Todos	Telefónica	Telefónica, Vicenç Fernandez
Gestión de riesgos	Carla Atienza	Carla Atienza	Vicenç Fernandez	Vicenç Fernandez
Identificar tipos de agentes	Lola Monroy	Lola Monroy	Telefónica	Telefónica
Definir benchmarks	Nils Duran	Nils Duran	Telefónica	Telefónica
Adaptar los modelos de lenguaje para cada tipo de agente	Alba Figueras	Alba Figueras	Telefónica, Ramon Sangüesa	Telefónica
Programar la lógica de interacción entre agentes	Alba Figueras	Alba Figueras	Ramon Sangüesa	Telefónica
Implementar la interfaz para ingresar datos del paciente	Nils Duran	Nils Duran	Ramon Sangüesa	Telefónica
Diseñar casos de prueba para diferentes diagnósticos	Carla Atienza	Carla Atienza	Telefónica	Telefónica
Analizar los resultados	Carla Atienza	Carla Atienza	Telefónica	Telefónica, Ramon Sangüesa

Documentar el diseño del sistema y la implementación	Lola Monroy	Lola Monroy	Telefónica	Telefónica
Elaborar el reporte final del proyecto	Todos	Lola Monroy	Telefónica	Telefónica
Reuniones de seguimiento y control de progreso	Nils Duran	Nils Duran	Telefónica	Telefónica
Gestión de la comunicación con Telefónica/UPC	Alba Figueras	Alba Figueras	Telefónica	Telefónica

*Matriz RACI*

## 10.2 Definición de Umbrales de Riesgo

Dado el contexto de este proyecto universitario, los umbrales de riesgo deben ser definidos con respecto a las limitaciones de recursos, tiempo y calidad.

### **COSTE**

El presupuesto disponible para este proyecto es extremadamente limitado, ya que no disponemos de financiación. Priorizaremos la optimización de los recursos disponibles y la minimización de los gastos a través de herramientas gratuitas y open-source. Cualquier desviación significativa del presupuesto se considerará un riesgo que requiere una atención inmediata.

Aunque no contemplamos costes fijos directos, sí existen costes indirectos asociados al uso de recursos en la nube o APIs externas. Definimos un umbral de sobrecoste bajo, con un incremento máximo de 200€.

### **TIEMPO**

El proyecto está sujeto a un cronograma estricto, ya que tenemos fechas de entregas finales definidas. Identificaremos las tareas críticas que tienen un impacto directo en la fecha de finalización del proyecto. Estableceremos umbrales de tiempo para estas tareas críticas, definiendo el retraso máximo aceptable antes de que se requieran acciones de mitigación o sea necesario redistribuir las tareas o recortar el alcance.

### **CALIDAD**

Dado que el enfoque del proyecto es sobre la investigación del impacto de la diversidad en la toma de decisiones de los agentes, priorizaremos una funcionalidad que nos permita comparar agentes diversos con homogéneos. Estamos dispuestos a simplificar aspectos secundarios, como usar librerías externas o una interfaz más simple, siempre que eso nos ayude a optimizar tiempo y recursos. Sin embargo, estas decisiones no pueden afectar la

calidad de los resultados ni la validez de las conclusiones de la investigación. Por tanto, el umbral de calidad se supera si los resultados no permiten una comparación válida entre agentes diversos y homogéneos.

## 10.3 Identificación de Riesgos

La identificación de riesgos es esencial en la gestión de proyectos, nos permite anticipar posibles problemas que podrían afectar el éxito del desarrollo de nuestro sistema. Hemos identificado diversos riesgos clasificados por categoría para facilitar su análisis y mitigación:

### 10.3.1 RIESGOS TÉCNICOS

Dificultad en la integración de agentes y en la simulación de la diversidad real

Existe el riesgo de enfrentarnos a complicaciones al intentar crear y simular una interacción efectiva entre los agentes que representan diferentes especialidades médicas. Este proceso representa un desafío, sobre todo cuando tratamos de simular adecuadamente la complejidad de la toma de decisiones médicas en diferentes contextos.

Limitaciones de las herramientas de desarrollo

Existe la posibilidad de que las herramientas y tecnologías que seleccionamos para el desarrollo del sistema (LangChain, LM Studio, OpenAI Agents SDK, modelos open-weights) presenten limitaciones técnicas, problemas de compatibilidad o restricciones que dificulten la implementación de ciertas funcionalidades.

Dependencia de APIs externas

Nuestro proyecto depende del uso de APIs externas para acceder a modelos de lenguaje. Esto genera un riesgo de dependencia de terceros, que puede traducirse en problemas de disponibilidad, cambios en las condiciones de servicio, aumentos de costos o limitaciones en el control sobre el sistema.

Sesgos en los datos de entrenamiento

Existe la posibilidad de que los datos utilizados para entrenar a los agentes de IA contengan sesgos, lo que podría llevar a que el sistema genere diagnósticos discriminatorios o inexactos para ciertos grupos de pacientes.

### 10.3.2 RIESGOS DE ORGANIZACIÓN

Estimación incorrecta de tareas

Existe el riesgo de que las estimaciones de tiempo, esfuerzo o recursos necesarios para completar las tareas del proyecto sean inexactas, lo que podría generar retrasos en el cronograma o sobrecostos.

Falta de comunicación efectiva

La falta de comunicación entre los miembros del equipo de desarrollo o los stakeholders (Telefónica/UPC) podría dar lugar a malentendidos, errores, retrasos o falta de alineación con los objetivos del proyecto.

Cambios en los requerimientos del proyecto

Este riesgo puede derivar de la falta de comunicación entre los integrantes del proyecto, esto puede generar la necesidad de re trabajar componentes ya desarrollados, lo que afectaría al cronograma y los recursos.

Riesgos imprevistos

Este riesgo incluye todos los riesgos que no hemos incluido en el Action Plan Table, porque los hemos considerado irrelevantes o no los hemos llegado a tener en cuenta. Estos serán gestionados mediante la monitorización continua y la adaptación del registro de riesgos a lo largo del proyecto.

### 10.3.3 RIESGOS DE GESTIÓN

Disponibilidad de recursos

La falta de disponibilidad de recursos computacionales o acceso a datos relevantes podría obstaculizar el progreso del proyecto y afectar su calidad.

### 10.3.4 RIESGOS ÉTICOS Y LEGALES

Violación de la privacidad y seguridad de los datos

En principio, trabajaremos con datos abiertos. Sin embargo, a medida que avance el proyecto, podría surgir la necesidad de acceder a información de pacientes. Esto es altamente sensible, ya que podría resultar en violaciones de la privacidad y seguridad.

El agente toma una decisión equivocada

Existe el riesgo de que el sistema de IA proporcione un diagnóstico incorrecto, lo que podría tener graves consecuencias para la salud del paciente y generar problemas de responsabilidad legal.

### 10.3.5 RIESGOS FINANCIEROS

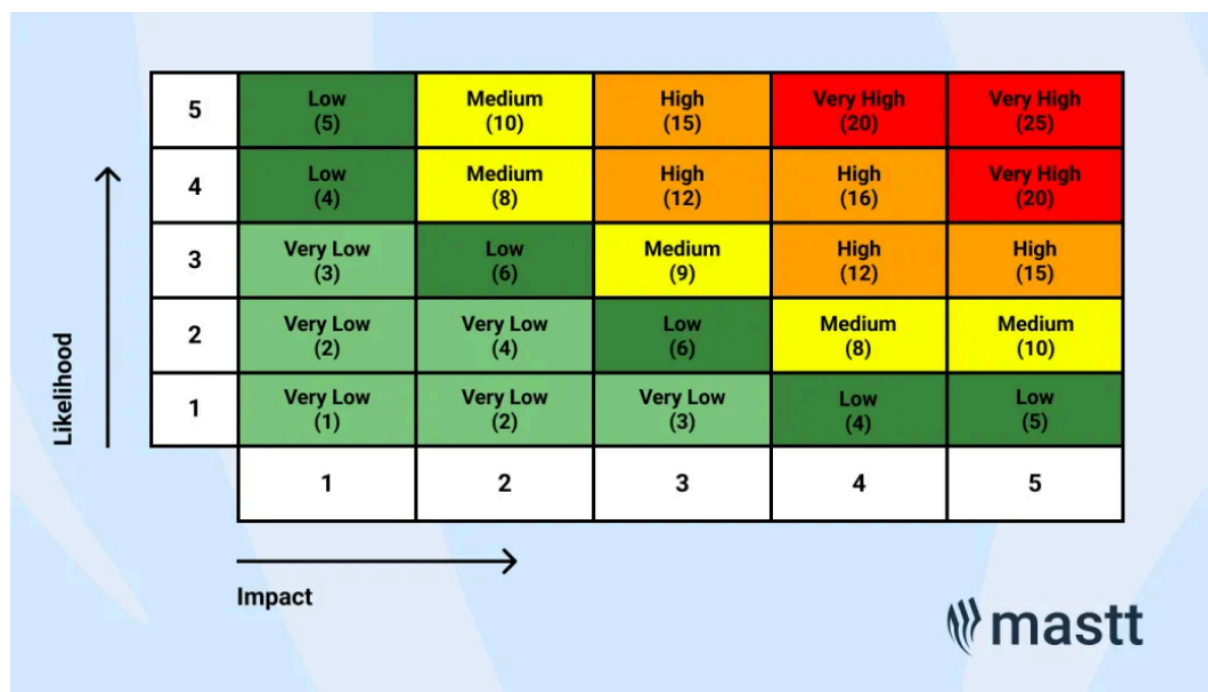
Costos superiores a los previstos

Los costos de desarrollo, incluyendo el uso de APIs, la infraestructura y otros gastos, podrían exceder el presupuesto asignado al proyecto, poniendo en peligro su viabilidad económica.

## 10.4 Valoración de Riesgos

Una vez identificados los riesgos potenciales que podrían afectar al proyecto, tenemos que valorarlos. Esta evaluación consiste en analizar qué tan probable es que ocurra cada riesgo y qué consecuencias o impacto tendría en el proyecto si llegara a suceder.

Para ello, utilizaremos una Matriz de Riesgos, que permite visualizar y clasificar los riesgos en función de su probabilidad e impacto, lo que facilitará la priorización de las futuras estrategias de mitigación. Como referencia utilizaremos la matriz de riesgos de Mastt para evaluar y priorizar los riesgos del proyecto.



Plantilla de Matriz de Riesgos:

<https://www.mastt.com/blogs/what-is-a-risk-matrix>

### RISK MATRIX

Riesgos	Probabilidad	Impacto	Riesgo
Dificultad en la integración de agentes y en la simulación de la diversidad real	2	3	Bajo
Limitaciones de las herramientas de	1	3	Muy bajo

desarrollo			
Dependencia de APIs externas	2	2	Muy bajo
Sesgos en los datos de entrenamiento	2	3	Bajo
Estimación incorrecta de tareas	4	2	Medio
Falta de comunicación efectiva	2	2	Muy bajo
Cambio en los requerimientos del proyecto	5	3	Alto
Riesgos imprevistos	5	3	Medio
Disponibilidad de recursos	1	4	Bajo
Violación de la privacidad y seguridad de los datos	1	4	Bajo
Agente toma una decisión equivocada	4	5	Muy alto
Costos superiores a los previstos	2	3	Bajo

*Risk Matrix*

## 10.5 Estrategias de Mitigación y Activadores de Riesgo

Para cada uno de los riesgos identificados, hemos definido estrategias de mitigación proactivas destinadas a reducir la probabilidad de ocurrencia del riesgo o a minimizar su impacto en caso de que ocurra. Además, hemos establecido "activadores" o "triggers", que son señales o condiciones específicas que indicarán que un riesgo está a punto de ocurrir o que ya ha ocurrido, y que por lo tanto se debe poner en marcha la estrategia de mitigación correspondiente.

Este paso es importante para garantizar que podamos identificar situaciones críticas a tiempo durante el desarrollo de nuestro proyecto, y a la vez para tener herramientas para lidiar con estas y sacar adelante el trabajo.

Las estrategias de mitigación y los activadores asociados a cada riesgo se detallan en la "Action Plan Table".

## 10.6 Monitorización del Registro de Riesgos y Actualización

Para asegurar que las estrategias de mitigación sigan siendo efectivas y para identificar nuevos riesgos que puedan surgir a lo largo del proyecto, es esencial establecer un proceso de monitorización del registro de riesgos.

La monitorización del registro de riesgos implicará lo siguiente:

- **Revisiones Periódicas:** Como equipo del proyecto revisaremos el registro de riesgos de forma regular (por ejemplo, semanalmente). Durante estas revisiones evaluaremos el estado de los riesgos identificados, la efectividad de las estrategias de mitigación implementadas y la aparición de nuevos riesgos.
- **Actualización del Registro:** Actualizaremos el registro de riesgos con la información más reciente. Esto incluirá cambios en la probabilidad o el impacto de los riesgos existentes, el progreso en la implementación de las estrategias de mitigación, la identificación de nuevos riesgos y el cierre de los riesgos que ya no sean relevantes.
- **Comunicación:** Comunicaremos a los stakeholders relevantes la información sobre el estado de los riesgos. Esto permitirá mantenerlos informados sobre los posibles problemas que podrían afectar al proyecto y sobre las acciones que se están tomando para gestionarlos.
- **Acciones Correctivas:** Si las estrategias de mitigación no están siendo efectivas o han surgido nuevos riesgos que requieren atención inmediata, tomaremos las acciones correctivas necesarias. Esto podría implicar la modificación de las estrategias de mitigación, la asignación de recursos adicionales o la replanificación de las actividades del proyecto.

## ACTION PLAN TABLE

Una vez que los riesgos han sido identificados, valorados y priorizados, el siguiente paso crítico es desarrollar un plan de acción para mitigar o minimizar su impacto potencial en el proyecto. La siguiente tabla, "Action Plan Table", detalla las estrategias de mitigación específicas que implementaremos para cada riesgo identificado. Servirá como una herramienta de referencia, permitiendo un seguimiento claro y efectivo de las acciones de mitigación. Se actualizará periódicamente para reflejar el progreso realizado y cualquier ajuste necesario en las estrategias de mitigación.

Descripción riesgo	Categoría riesgo	Efecto Tiempo	Efecto Coste	Efecto Calidad	Efecto Alcance	Probabilidad	Impacto	Rango	Tiempo	Coste	Δ Tiempo	Δ Coste	Estrategia	Descripción	Persona	Trigger
Dificultad en la integración de agentes y en la simulación de la diversidad real	Técnicos	✓	☐	✓	✓	2	3	Bajo	2 semanas	0 euros	3 semanas	0 euros	Mitigar	Implementar pruebas de integración frecuentes, un marco de trabajo para la interacción de agentes y un modelo de datos común.	Nita	Resultados inconsistentes en las pruebas de integración.
Limitaciones de las herramientas de desarrollo	Técnicos	✓	✓	✓	✓	1	3	Muy bajo	2 semana	50 euros	18 días	100 euros	Mitigar	Evaluar exhaustivamente las herramientas, establecer un plan de contingencia y mantenerlas actualizado sobre sus versiones.	Alba	Funcionalidades clave del sistema que no pueden implementarse con las herramientas seleccionadas.
Dependencia de APIs externas	Técnicos	✓	✓	✓	✓	2	2	Muy bajo	2 semanas	300 euros	18 días	400 euros	Evitar	Seleccionar APIs confiables con SLAs, implementar fallbacks, monitorizar uso/costos y considerar modelos locales para reducir dependencia.	Lola	Errores en el sistema que se correlacionan con problemas conocidos de las APIs.
Seguros en los datos de entrenamiento	Técnicos	✓	☐	✓	✓	2	3	Bajo	2 semanas	0 euros	3 semanas	0 euros	Evitar	Preprocesar datos para corregir sesgos y evaluar el rendimiento en subgrupos de pacientes.	Carla	Métricas de evaluación que indican un rendimiento inconsistente en diferentes subgrupos.
Estimación incorrecta de tareas	Organización	✓	☐	☐	☐	4	2	Medio	1 semana	0 euros	2 semanas	0 euros	Mitigar	Actualizar el RACI y Action Plan Table en cada cambio de la planificación.	Carla	Retrasos en la ejecución de las tareas y evolución inestable del proyecto.
Falta de comunicación efectiva	Organización	☐	☐	✓	✓	2	2	Muy bajo	1 semana	0 euros	2 semanas	0 euros	Evitar	Utilizar herramientas de comunicación efectiva como la escucha activa y soft skills.	Alba	Falta de comunicación y malentendidos a lo largo de la ejecución de tareas.
Cambio en los requerimientos del proyecto	Organización	✓	✓	✓	✓	5	3	Alto	2 semanas	0 euros	3 semanas	50 euros	Aceptar	Intentar reaprovechar tanto trabajo como podamos para no tener que iniciar desde cero.	Nita	Reuniones de seguimiento y control en las que no logramos cumplir con los requerimientos solicitados.
Riesgos imprevistos	Organización	✓	✓	✓	✓	5	3	Alto	2 semanas	0 euros	3 semanas	50 euros	Mitigar	Lidiar con los riesgos imprevistos a medida que nos los encontremos.	Carla	Gestión de riesgos
Disponibilidad de recursos	Gestión	✓	☐	✓	✓	1	4	Bajo	1 semana	0 euros	10 días	0 euros	Evitar	Definir desde un inicio los recursos disponibles y necesarios para el proyecto. Asegurarnos de su disponibilidad en el transcurso del desarrollo del sistema.	Alba	Falta de recursos computacionales o acceso a datos esenciales para el desarrollo.
Violación de la privacidad y seguridad de los datos	Éticos y Legales	☐	☐	☐	☐	1	4	Bajo	2 semanas	0 euros	3 semanas	0 euros	Evitar	Usar únicamente datos abiertos, públicos, que no contengan información personal o sensible. Verificar la licencia de los datos antes de utilizarlos.	Lola	Aparece información sensible en los logs o salidas del modelo.
Agente toma una decisión equivocada	Éticos y Legales	☐	☐	✓	✓	4	5	Muy alto	0 días	0 euros	0 días	0 euros	Aceptar	Advertir al cliente de la necesidad de verificar con un experto las decisiones tomadas por el programa, ya que nuestro sistema no está desarrollado para sustituir a un médico.	Lola	La respuesta del agente contradice datos conocidos, o los resultados inconsistentes entre agentes. El modelo propone una acción peligrosa o poco ética.
Costos superiores a los previstos	Financieros	☐	✓	☐	✓	2	3	Bajo	0 días	150 euros	0 días	200 euros	Evitar	Priorizar herramientas gratuitas o de código abierto.	Nita	El entorno se queda sin espacio o tarda demasiado, o superamos el límite de uso de la API.



## **11. Impacto ambiental, económico y social**

Realizamos un plan de sostenibilidad, con el objetivo de garantizar que el desarrollo y uso del sistema de colaboración entre agentes especializados en diagnóstico médico se realice de forma responsable, ética y sostenible, considerando las perspectivas económicas, medioambientales y sociales. Además, queremos garantizar la transparencia en la gestión de datos personales (como serían los datos médicos), minimizar el impacto ambiental y asegurar un sistema justo y equitativo. Para ello utilizaremos la herramienta de la Matriz de Sostenibilidad, que desglosaremos en diferentes ámbitos.

### **11.1 Matriz de Sostenibilidad**

#### 11.1.1 Ámbito Económico

##### Desarrollo del proyecto

Se ha priorizado usar APIs de bajo consumo y herramientas gratuitas. Además, se ha optimizado el número de pruebas y simulaciones para reducir costes de computación innecesarios.

En lugar de entrenar modelos de lenguaje desde cero, se ha optado por realizar un *fine-tuning* sobre modelos base ya existentes mediante Google AI Studio, lo que ha reducido de forma significativa el coste computacional.

La integración del flujo multiagente utilizando LangGraph ha facilitado una colaboración eficiente sin necesidad de mantener servidores persistentes ni procesos complejos, reduciendo así el coste de infraestructura.

Por otro lado, el uso de Streamlit como interfaz de usuario ha eliminado la necesidad de desarrollar desde cero una plataforma web personalizada, ahorrando costes adicionales.

En resumen, se ha priorizado el uso de herramientas accesibles, escalables y muchas de ellas de código abierto o gratuitas para entornos educativos y de investigación.

##### Ejecución del producto/servicio:

El sistema, una vez desplegado, presenta un coste moderado asociado principalmente a las llamadas a las APIs de generación de contenido de los modelos entrenados en Google AI Studio.

No obstante, el sistema no está funcionando continuamente en segundo plano, y solo se hacen llamadas a la API cuando el usuario introduce nuevos casos clínicos. Por tanto, el coste es proporcional a las llamadas al sistema, que en nuestro caso serán limitadas.

El uso de APIs pre entrenadas ha reducido de forma considerable el coste respecto a entrenamientos completos. Sin embargo, en un escenario de uso público a gran escala, se podría realizar un entrenamiento más exhaustivo como alternativa más rentable y personalizada a largo plazo.

Por otra parte, la arquitectura modular del sistema permite realizar actualizaciones o mejoras puntuales sin necesidad de reconstruir el sistema completo. Esto facilita, por ejemplo, el reentrenamiento de agentes concretos en caso de actualización de las guías clínicas o la aparición de nuevos protocolos médicos sin la necesidad de un coste adicional.

### Riesgos y limitaciones económicas

Existen algunos riesgos que podrían comprometer la viabilidad económica a largo plazo:

- Cambios en el modelo de precios de Google AI Studio o en el acceso gratuito a la API utilizada para generación de contenido, lo cual podría ser un problema por la dependencia de un único proveedor.
- Límites de uso en plataformas como Streamlit si se superan los planes gratuitos.
- En caso de una aplicación real, también se deberían considerar los costes ocultos derivados de la necesidad de asegurar confidencialidad o almacenamiento seguro de datos clínicos.

## 11.1.2. Medioambiental

### Desarrollo del proyecto

El desarrollo del proyecto implica el fine-tuning de cinco modelos de lenguaje especializados y la ejecución de unas 50.000 consultas (aproximadamente 50 millones de tokens procesados). Se estima un consumo total entre 15 y 50 kWh, con emisiones asociadas de entre 5 y 20 kg de CO<sub>2</sub>e, según la infraestructura. Para minimizar el impacto se utilizan modelos eficientes (Gemini 1.5 Flash) y servicios en la nube con compromiso de energía renovable (Google Cloud). No se han utilizado materiales físicos, siguiendo un enfoque completamente digital.

### Ejecución del producto/servicio

Durante la vida útil del sistema, el impacto ambiental proviene principalmente del consumo energético asociado a la inferencia de modelos de lenguaje y la infraestructura cloud. El impacto es relativamente bajo y no se generan residuos físicos. El sistema puede optimizar el

uso de recursos sanitarios al mejorar el proceso diagnóstico, lo que potencialmente reduce pruebas o consultas innecesarias.

### Riesgos y limitaciones económicas

Un incremento en el número de consultas o el uso de modelos más grandes podría aumentar significativamente la huella de carbono. Las estimaciones actuales son aproximadas y no se ha hecho un seguimiento en tiempo real del consumo energético. En fases posteriores se integrarán herramientas como CarbonTracker o MLCO2 Impact para un seguimiento más preciso.

### 11.1.3. Ámbito Social

#### Desarrollo del proyecto

La documentación y las interfaces del sistema han sido redactadas con un lenguaje inclusivo adaptado a diferentes perfiles de usuario, siguiendo las recomendaciones de accesibilidad digital.

Desde el inicio del proyecto hemos intentado garantizar la representatividad y diversidad de los datos utilizados para el entrenamiento de los modelos especializados. Las bases de datos médicas seleccionadas para el fine-tuning contienen información relevante procedente de múltiples áreas de la medicina y, en la medida de lo posible, incluyen casos clínicos anonimizados diversos en cuanto a edad y sexo.

El objetivo ha sido reducir los sesgos implícitos que suelen derivarse del uso de datos clínicos homogéneos o poco representativos. Aunque las fuentes empleadas no permiten un control absoluto sobre la diversidad (por ejemplo, no siempre está disponible el origen étnico o el contexto socioeconómico del paciente), se ha priorizado la inclusión de datasets validados por la comunidad médica y con amplio reconocimiento académico.

Además, el uso de modelos de lenguaje ajustados a especialidades concretas permite adaptar el razonamiento del sistema a distintos contextos clínicos, lo cual mejora su aplicabilidad en grupos diversos.

#### Ejecución del producto/servicio

Al facilitar la colaboración entre agentes virtuales con perfiles médicos diversos, se espera que el sistema pueda reducir los sesgos y dar respuestas que incluyan puntos de vista diversos. Además, el fine tuning con datos de perfiles diversos que ya hemos comentado debería mejorar la calidad de los diagnósticos.

El sistema también procura una alta explicabilidad y transparencia, justificando cada decisión tomada y argumentando los puntos a favor y en contra, lo cual se logra con el debate organizado entre agentes, visible para el usuario.

### Riesgos y limitaciones sociales

Las principales limitaciones del análisis social en esta fase son la falta de validación con usuarios reales y la necesidad de pruebas específicas con colectivos vulnerables o con necesidades especiales. Además, en una aplicación real, un mal uso o una interpretación errónea de los resultados podría impactar negativamente a los usuarios, por lo cual es necesario supervisar el sistema continuamente y revisar toda la información con profesionales médicos.

## **11.2 Análisis de Protección de los Datos**

Para este apartado hemos tenido en cuenta el Reglamento General de Protección de Datos (RGPD), que explica los principios que se deben tener en cuenta al tratar con datos personales. Los principios del reglamento, junto con cómo los hemos implementado, son los siguientes:

### Licitud, lealtad y transparencia

Los datos utilizados provienen de fuentes públicas abiertas y muy extendidas en la investigación médica. Los casos planteados son hipotéticos y generados por expertos en el dominio, de modo que no comprometen la privacidad de los usuarios.

### Limitación de la finalidad

Los datos son recogidos únicamente con la finalidad de realizar un diagnóstico en el entorno de simulación, pero no son persistentes. La arquitectura no permite el almacenamiento de estos datos, de modo que estos no se usan en ningún caso para transferir información, analizar usuarios o entrenar otros modelos.

### Minimización de los datos

El sistema no recoge nombres, identificadores ni historiales clínicos completos. Únicamente se pide como entrada una descripción del caso clínico, que es la información necesaria para la finalidad de generar el diagnóstico.

### Exactitud

En un caso real sería necesario implementar mecanismos para verificar que los casos utilizados estén actualizados, y podría ser necesario ir actualizando los modelos con la aparición de nuevos descubrimientos médicos.

#### Limitación del plazo de conservación

Actualmente, el sistema no conserva los inputs ni los diagnósticos de forma persistente. En un caso real también se deberían borrar, o al menos se podría aplicar un procedimiento de anonimización para garantizar que estos datos no puedan identificar a los usuarios, siempre y cuando esta retención pueda ser útil para la mejora de la precisión del sistema.

#### Seguridad

Los datos no se envían a terceros, y únicamente se transmiten a las APIs de Google AI Studio, que cumplen los estándares generales de seguridad. Además, en una implementación médica real se tendrían que encriptar las transmisiones de datos para garantizar su confidencialidad.

#### Responsabilidad Activa

Esto implica prever posibles riesgos (por ejemplo, introducción accidental de datos reales), y tomar medidas de prevención como mensajes en la interfaz y el diseño sin almacenamiento persistente.

En caso de aplicarse en un entorno clínico real, el sistema permitiría incluir registros de tratamiento, control de acceso, evaluación de impacto y supervisión por un delegado de protección de datos, cumpliendo así con la normativa vigente.

## 11.3 Consideraciones Éticas

### Principios Éticos Aplicados en el Desarrollo y Uso del Sistema

**Beneficencia:** El sistema ha sido diseñado con el objetivo de contribuir a una mejor toma de decisiones clínicas, apoyando a los profesionales sanitarios y reduciendo el riesgo de errores médicos mediante la integración de múltiples perspectivas especializadas.

**No maleficencia:** Para minimizar posibles daños, se ha implementado una estrategia de validación cruzada entre agentes con distintas especialidades, lo cual permite contrastar diagnósticos y evitar conclusiones erróneas basadas en un único punto de vista.

**Justicia:** se promueve el acceso abierto al sistema, su capacidad de adaptación a diferentes contextos médicos, y el uso de datos representativos para evitar sesgos poblacionales.

**Autonomía:** el sistema no sustituye la toma de decisiones humanas, sino que la complementa. Se proporciona transparencia, explicabilidad y trazabilidad de los razonamientos del sistema, y las respuestas deberían ser consultadas posteriormente por un médico.

## Compromiso con la Inclusión y la Accesibilidad

El diseño del sistema ha tenido en cuenta:

- Un lenguaje neutro e inclusivo, evitando estereotipos o expresiones sesgadas.
- Una interfaz accesible desarrollada con Streamlit, adaptable a distintos niveles de formación y experiencia médica.
- Posibilidad de adaptación a diferentes idiomas y contextos culturales mediante el ajuste de agentes entrenados en nuevos entornos.

## Evaluación y Mitigación de Sesgos

Dado que los modelos de lenguaje pueden reproducir sesgos presentes en los datos de entrenamiento, se han adoptado las siguientes medidas:

- Uso de datasets médicos reconocidos y lo más representativos posible (como MedMCQA y MedQA-USMLE).
- Fine-tuning separado por especialidades para reducir la dominancia de perspectivas homogéneas.
- Debate entre agentes para detectar y corregir recomendaciones poco precisas o incompletas.
- Pruebas internas con casos clínicos éticamente sensibles (por ejemplo, salud reproductiva, pacientes en situación de vulnerabilidad).

Consideramos que la eliminación completa del sesgo no es posible, por lo que el sistema está diseñado para ser supervisado y validado por profesionales médicos humanos.

## **12. Conclusiones**

Hemos podido observar que la diversidad efectivamente es un atributo importante que mejora el rendimiento de agentes inteligentes. De hecho, la adición de un agente que, a primera vista según los embeddings, no parece muy relevante (ciencias básicas), mejora notablemente el rendimiento (~3%). Esto puede ser una indicación de que usar los embeddings y la distancia coseno para determinar si un agente y una pregunta son compatibles no es la mejor opción. También hemos visto cómo el consenso (votación mayoritaria entre agentes) mejora ligeramente a los agentes individuales. Para finalizar, también hemos estudiado el efecto del parámetro de temperatura respecto a la precisión y la tasa de rechazo de respuestas.

Más allá, podríamos hacer muchísimas cosas, ya que, como hemos podido ver la diversidad de agentes es un campo inmenso. De entrada, podríamos usar modelos base más grandes para el fine-tuning, hacerlo con más épocas, más datos y de mayor calidad. Más allá de los modelos adaptados, también podríamos añadir nuevos agentes y hacerlos más granulares. Por ejemplo, en lugar de combinar ginecología y pediatría, tener un agente específico para cada área. Incluso podríamos desglosar los agentes en subdominios, y desarrollar un método de supervisión jerárquica. En este método, habría un agente de medicina general, y por ejemplo un especialista en oncología, y dentro de este especialista habría subespecialistas en cáncer de mama, de pulmón...

Más allá de los métodos que hemos implementado hasta ahora, también se podrían añadir herramientas de RAG (que hemos probado pero no funcionaron del todo bien), o búsqueda en internet.

Por tanto, los resultados de este estudio no solo confirman la importancia de la diversidad en la mejora de la inteligencia de los agentes, sino que también abren la puerta a una exploración más profunda y estructurada de cómo la especialización y la colaboración pueden dar forma a sistemas inteligentes aún más robustos y adaptables en el futuro.

### **13. Referencias**

- [1] Rey, A. A. (2022). *El libro de la Inteligencia Colectiva*.
- [2] Wikipedia. (n.d.). Kaspàrov contra el Món. *Wikipedia*. Recuperado de [https://ca.wikipedia.org/wiki/Kasp%C3%A0rov\\_contra\\_el\\_M%C3%B3n](https://ca.wikipedia.org/wiki/Kasp%C3%A0rov_contra_el_M%C3%B3n)
- [3] Chess.com. (n.d.). Announcing Magnus vs. The World. *Chess.com*. Recuperado de <https://www.chess.com/news/view/announcing-magnus-vs-the-world>
- [4] Dionne, S. D., Sayama, H., & Yammarino, F. J. (2019). Diversity and social network structure in collective decision making: Evolutionary perspectives with agent-based simulations. *Complexity*, 2019, 7591072. <https://doi.org/10.1155/2019/7591072>
- [5] Chris M. Stolle, Bartosz Gula, Rongjun Yu, and Yi Huang. (2024). The impact of diversity on group decision-making in the face of the free-rider problem. *Judgment and Decision Making*, 19.
- [6] Salem, (2015). A Survey of Multi-Agent Based Intelligent Decision Support Systems for Medical Classification Problems. *International Journal of Computer Applications*, 123(10).
- [7] SuperAnnotate. (2025). LLM agents: The ultimate guide. *SuperAnnotate*. Recuperado de <https://www.superannotate.com/blog/llm-agents>
- [8] Stanford. (n.d.). Genie. *Stanford University*. Recuperado de <https://storm.genie.stanford.edu/>
- [9] Design Thinking – Benchmarking. Recuperado de <https://design-toolkit.recursos.uoc.edu/benchmarking/>
- [10] Design Thinking – Graphical User Interface. Recuperado de <https://design-toolkit.recursos.uoc.edu/graphical-user-interface/>
- [11] Human Interface Guidelines, Apple. Recuperado de <https://developer.apple.com/design/human-interface-guidelines/foundations>
- [12] Sistema de puntuació elo, Viquipèdia. [https://ca.wikipedia.org/wiki/Sistema\\_de\\_puntuació\\_Elo](https://ca.wikipedia.org/wiki/Sistema_de_puntuació_Elo)
- [13] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, & Hae Won Park. (2024). MDAGents: An Adaptive Collaboration of LLMs for Medical Decision-Making.