Benchmark Comparativo: Aptitud de Modelos IA para Revisión de Completitud Ética Documental

Introducción

La selección de un Modelo de Lenguaje Grande (LLM) adecuado es fundamental para la tarea de revisión de completitud ética en documentos. En este contexto, "completitud" se define como la verificación de que un documento incorpora y satisface todos los elementos y estipulaciones requeridos por un conjunto predefinido de 18 normas éticas. Este análisis se basa en un benchmark específico (reflejado en tablas comparativas) para categorizar los modelos de IA según su capacidad para esta tarea.

Contexto de Evaluación

Para comprender la clasificación de los modelos en este benchmark, es esencial entender tanto las métricas de evaluación utilizadas como el enfoque de ponderación aplicado. Esta sección clarifica el significado y la clasificación de los benchmarks más relevantes y cómo se ha buscado un equilibrio entre la capacidad general de los modelos ("*potencia*") y su efectividad específica ("*precisión*") para la tarea de revisión de completitud ética.

Significado de los Benchmarks Clave

Los siguientes benchmarks son comúnmente utilizados para medir diversas capacidades de los LLMs y son pertinentes para la tarea de revisión de completitud ética:

- *MMLU (Massive Multi Task Language Understanding):* Evalúa la comprensión del lenguaje y el conocimiento general del modelo a través de 57 tareas diversas. Una puntuación alta sugiere una buena capacidad para entender el contenido general de un documento y, por extensión, los conceptos subyacentes a las normas éticas.
- *MMLU-Pro:* Una versión más desafiante del MMLU, diseñada para probar el razonamiento profundo. Es relevante para la revisión de la completitud ética, que a menudo requiere inferencias complejas sobre la aplicación de las normas.
- *GPQA (Graduate-level Physics, Philosophy, and Economics Questions; también variantes como GPQA Diamond):* Mide las capacidades de razonamiento avanzado. Es un indicador clave de la habilidad del modelo para detectar lagunas lógicas o información faltante en relación con los requisitos de las normas éticas.
- *MATH / GSM8K (Grade School Math 8K):* Prueban el razonamiento lógico-matemático. Aunque no todas las normas éticas son cuantitativas, la capacidad de razonamiento lógico que miden estos benchmarks es transferible a la interpretación de reglas y condiciones normativas.
- *Ventana de Contexto:* Se refiere a la cantidad de texto (medida en tokens) que un modelo puede procesar y recordar de una sola vez. Una ventana de contexto amplia es crucial para la

revisión de las 18 normas éticas en documentos extensos, permitiendo al modelo mantener una visión global y coherente.

- *IFEval (Instruction Following Evaluation):* Evalúa qué tan bien un modelo puede seguir instrucciones complejas. Esta capacidad es vital para guiar al modelo sobre qué aspectos específicos de las 18 normas éticas debe verificar.

Sistema de Ponderación: Equilibrio entre Potencia y Precisión para la Revisión Ética

La clasificación final de los modelos en este benchmark no se basa únicamente en una métrica aislada, sino en un sistema de ponderación que busca un equilibrio entre la "*potencia*" general de un modelo y su "*precisión*" o efectividad específica para la tarea de revisión de completitud de las 18 normas éticas.

- *Potencia del Modelo:* Se refiere a las capacidades fundamentales y generales del LLM, reflejadas en su rendimiento en benchmarks estandarizados como MMLU, GPQA, MATH, y el tamaño de su ventana de contexto. Modelos con alta "potencia" suelen tener una comprensión más profunda del lenguaje, mejores habilidades de razonamiento y la capacidad de manejar grandes cantidades de información. Estas son características deseables, ya que sugieren que el modelo tiene las herramientas cognitivas necesarias para abordar tareas complejas.
- *Precisión para la Tarea (Revisión Ética):* Se refiere a qué tan bien un modelo aplica sus capacidades para la tarea específica de verificar la completitud de las 18 normas éticas en un documento. Esto se evalúa directamente a través de los resultados del benchmark específico que se ha realizado (las tablas Q1, Q2, etc., que muestran "APROBADO" o "NO APROBADO" para cada norma). Un modelo puede ser muy potente en general, pero si no aplica esa potencia de manera efectiva para identificar correctamente el cumplimiento o incumplimiento de las normas éticas, su "*precisión*" para esta tarea será baja.

El enfoque de ponderación utilizado en la clasificación general de aptitud (Sección de "*Recomendaciones finales"* del informe) intenta reflejar esta dualidad. Si bien los modelos con mejor rendimiento en benchmarks de potencia tienden a clasificarse más alto, su desempeño específico en la tarea de revisión ética (evidenciado en las tablas de "APROBADO"/"NO APROBADO") es un factor crucial. Por ejemplo, un modelo con una ventana de contexto masiva (alta potencia en ese aspecto) y un buen MMLU es prometedor, pero sí en la práctica falla consistentemente en identificar elementos clave de las normas éticas (baja precisión en la tarea), su utilidad disminuye.

Por lo tanto, la clasificación de "Aptitud Alta", "Media", etc., considera que, si bien muchos modelos pueden tener buenos resultados en benchmarks generales, no todos poseen la combinación óptima de capacidad de procesamiento (potencia) y la habilidad para aplicar esa capacidad de forma certera (precisión) a los matices de la revisión de 18 normas éticas. Las recomendaciones finales se inclinan hacia aquellos modelos que no solo son potentes, sino que también demuestran ser precisos y fiables en el contexto específico de este benchmark ético.

Categorías de Aptitud y Modelos Recomendados para la Revisión de Normas Éticas

A continuación, se presentan las categorías de aptitud, desde la más alta hasta aquellas con ciertas limitaciones para la tarea de revisión de completitud ética documental.

1. Aptitud Alta: Modelos Sobresalientes para la Verificación Exhaustiva de Normas Éticas

Estos modelos demuestran un alto rendimiento en la comprensión, interpretación y verificación del cumplimiento de múltiples normas éticas dentro de los documentos. Su capacidad de razonamiento avanzado, manejo de contexto y precisión son cruciales para esta tarea.

- ***Características Clave de la Categoría:***
    - Excelente capacidad para interpretar y aplicar los criterios de las 18 normas éticas a través del contenido del documento.
    - Alto rendimiento en benchmarks de razonamiento (GPQA, MATH/GSM8K), indicativo de su habilidad para manejar la lógica inherente a las normativas.
    - Sólida comprensión del lenguaje (MMLU/MMLU-Pro) para entender los matices del texto y de las propias normas.
    - Ventanas de contexto amplias (idealmente 128K tokens o más) para asegurar la evaluación coherente de las normas a lo largo de documentos extensos.
    - Capacidad para identificar con precisión si los elementos requeridos por cada norma están presentes, ausentes o se presentan de forma ambigua.
- ***Modelos Recomendados en esta Categoría:***
    - Deepseek-r1-distill-llama-70b (Groq - DeepSeek)
        - *Razones para la Aptitud Ética:* Su fuerte razonamiento matemático y lógico (MATH-500: 94.5%, GPQA Diamond: 65.2%) es vital para desglosar y aplicar normas éticas complejas que pueden tener múltiples condiciones o interdependencias. La ventana de 128K Tokens permite un análisis integral.
        - *Consideraciones:* Modelo en "Preview" en Groq.
    - Gemini-2.0-flash-001 - Gemini-2.0-flash (Google)
        - *Razones para la Aptitud Ética:* La ventana de contexto de 1 millón de tokens es una ventaja significativa para revisar documentos extensos contra las 18 normas sin perder detalles. Su rendimiento competitivo en MMLU-Pro (77.4%), GPQA (65.2%) y Math500 (88.0%) respalda su capacidad para comprender y razonar sobre el contenido normativo.
        - *Consideraciones:* El uso intensivo de contextos grandes puede tener implicaciones de latencia/coste si se superan los límites gratuitos.
    - Meta-llama/llama-4-maverick-17b-128e-instruct (Groq - Meta)
        - *Razones para la Aptitud Ética:* Un fuerte GPQA Diamond (69.8%) y MMLU Pro (59.6% Groq / 80.5% Nvidia) junto con una ventana de 128K tokens lo hacen muy capaz de interpretar y verificar el cumplimiento normativo.
        - *Consideraciones:* Modelo en "Preview" en Groq.

Modelos muy Adecuados					
Preguntas	2.0-flash	deepseek-r1-distill-llama-70b	2.0-flash-001	meta-llama/llama-4-maverick-17b-128e-instruct	Resultado Esperado
Q1	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q2	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q3	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q4	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q5	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q6	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.1	APROBADO	NO APROBADO	APROBADO	APROBADO	APROBADO
Q7.2	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.3	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q7.4	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.5	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.6	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q7.7	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q8	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q9	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q10	NO APROBADO	NO APROBADO	NO APROBADO	APROBADO	NO APROBADO
Q11	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q12	NO APROBADO	APROBADO	NO APROBADO	APROBADO	APROBADO
% De exito	**94.44%**	**94.44%**	**94.44%**	**94.44%**	**100.00%**

2. Aptitud Alta-Media: Modelos Competentes con Gran Potencial para la Revisión Ética

Estos modelos ofrecen un muy buen rendimiento general y son capaces de manejar la revisión de normas éticas eficazmente, aunque pueden no igualar a los de "Aptitud Alta" en todas las facetas del razonamiento complejo o en la consistencia a través de todas las 18 normas.

- ***Características Clave de la Categoría:***
    - Buena capacidad para procesar y aplicar criterios de normas éticas.
    - Ventanas de contexto muy amplias o buen rendimiento en benchmarks de razonamiento.
    - Adecuados para la mayoría de las tareas de revisión de completitud ética, especialmente si se busca un equilibrio con la eficiencia.
- ***Modelos Recomendados en esta Categoría:***
    - Gemini-2.0-flash-lite-001 - 2.0-flash-little (Google)
        - *Razones para la Aptitud Ética:* Conserva la ventana de 1 millón de tokens, crucial para la revisión exhaustiva. Su buen rendimiento en MMLU-Pro (71.6%) y MATH (86.8%) apoya la comprensión y aplicación de las normas.
        - *Consideraciones:* Una opción eficiente para cuando se necesita un gran contexto sin la máxima potencia de razonamiento del *Gemini-2.0-flash-001*.
    - Gemini-1.5-flash-002 (Google)
        - *Razones para la Aptitud Ética:* La ventana de 1 millón de tokens es una gran ventaja. Su rendimiento en MATH (77.9%) es bueno.
        - *Consideraciones:* Puntuaciones más modestas en MMLU-Pro (67.3%) y GPQA (51.0%) pueden afectar la interpretación de normas complejas
    - Llama-3.3-70b-versatile (Groq - Meta)
        - *Razones para la Aptitud Ética:* Amplia ventana de contexto (128K tokens) que es beneficiosa para el análisis de documentos. Su inclusión en "Modelos Adecuados" según tu benchmark sugiere potencial.
        - *Consideraciones:* Modelo en "Preview" en Groq. El benchmark específico para las normas éticas indica "AUN NO" en la tabla, lo que significa que su rendimiento real para esta tarea específica necesita ser validado una vez se complete su evaluación detallada contra las 18 normas.

Modelos Adecuados					
Preguntas	2.0-flash-lite	1.5-flash-002	2.0-flash-lite-001	llama-3.3-70b-versatile	Resultado Esperado
Q1	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q2	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q3	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q4	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q5	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q6	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.1	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.2	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.3	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q7.4	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO

Q7.5	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.6	NO APROBADO	NO APROBADO	NO APROBADO	APROBADO	NO APROBADO
Q7.7	APROBADO	NO APROBADO	APROBADO	APROBADO	NO APROBADO
Q8	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q9	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q10	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q11	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q12	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	APROBADO
% De exito	**88.89%**	**94.44%**	**88.89%**	**83.33%**	**100.00%**

3. Aptitud Media: Modelos Sólidos con Algunas Limitaciones para la Revisión Ética Detallada

Estos modelos pueden ser efectivos para verificar la presencia de elementos relacionados con las normas éticas, pero podrían tener dificultades con la interpretación de normas más ambiguas o con la evaluación de la consistencia a través de múltiples normas en documentos muy largos si su ventana de contexto es limitada.

- ***Características Clave de la Categoría:***
  - Rendimiento competente en la comprensión general del lenguaje.
  - Pueden tener ventanas de contexto más restringidas o un rendimiento en razonamiento avanzado que no es de primer nivel.
  - Útiles para revisiones donde las normas éticas son claras y los documentos de tamaño moderado.
- ***Modelos Recomendados en esta Categoría:***
  - Mistral-saba-24b (Mistral-Small-24B-Instruct) (Groq - Mistral)
    - *Razones para la Aptitud Ética:* Buen MMLU-Pro (66.3%) y excelente IFEval (82.9%) para seguir instrucciones específicas sobre qué verificar de las normas.
    - *Consideraciones:* Ventana de 32,000 tokens y GPQA (45.3%) más bajo.
  - Gemini-1.5-flash / Gemini-1.5-flash-001 (Google)
    - *Razones para la Aptitud Ética:* Ventana de 1 millón de tokens. Buen rendimiento en MATH (77.9% para la variante 002, base de estas).
    - *Consideraciones:* Puntuaciones más modestas en MMLU-Pro (67.3% para 002) y GPQA (51.0% para 002) pueden afectar la interpretación de normas éticas complejas. Su desempeño en la tabla de normas éticas los sitúa como "Moderadamente Adecuados".
  - Gemini-1.5-flash-8b / Gemini-1.5-flash-8b-001 (Google)
    - *Razones para la Aptitud Ética:* Ventana de 1 millón de tokens.

- *Consideraciones:* Puntuaciones significativamente más bajas en MMLU-Pro (58.7%) y MATH (58.7%) en comparación con otros modelos Gemini, lo que afecta la capacidad de análisis normativo complejo. Su desempeño en la tabla de normas éticas los clasifica como "Moderadamente Adecuados".

Modelos Moderadamente adecuados						
Preguntas	1.5-flash	1.5-flash-001	1.5-flash-8b	1.5-flash-8b-001	mistral-saba-24b	Resultado Esperado
Q1	NO APROBADO	NO APROBADO	APROBADO	APROBADO	NO APROBADO	NO APROBADO
Q2	APROBADO	APROBADO	APROBADO	APROBADO	NO APROBADO	APROBADO
Q3	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q4	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q5	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q6	APROBADO	APROBADO	APROBADO	APROBADO	NO APROBADO	APROBADO
Q7.1	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.2	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.3	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	APROBADO	NO APROBADO
Q7.4	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.5	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q7.6	NO APROBADO	NO APROBADO	APROBADO	APROBADO	NO APROBADO	NO APROBADO
Q7.7	NO APROBADO	NO APROBADO	APROBADO	APROBADO	APROBADO	NO APROBADO
Q8	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q9	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q10	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q11	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO	APROBADO
Q12	NO APROBADO	NO APROBADO	APROBADO	APROBADO	NO APROBADO	APROBADO
% De éxito	**94.44%**	**94.44%**	**83.33%**	**83.33%**	**72.22%**	**100.00%**

4. Aptitud Media-Baja: Modelos Adecuados para Verificaciones Éticas Específicas o Menos Exigentes

Estos modelos pueden ser útiles para verificar aspectos puntuales de las normas éticas o cuando las normas son muy directas y no requieren una inferencia compleja.

- ***Características Clave de la Categoría:***
  - Rendimiento aceptable en comprensión general.

- Pueden ser una opción si la velocidad o los límites de uso gratuito son prioritarios y la tarea de revisión ética es acotada.

- ***Modelos Recomendados en esta Categoría:***
  - Llama-3.1-8b-instant (Groq - Meta)
    - *Razones para la Aptitud Ética:* Ventana de 128,000 tokens y optimización para baja latencia.
    - *Consideraciones:* Rendimiento más bajo en MMLU (69.4%) y MATH (51.9%), lo que puede limitar la profundidad del análisis ético.
  - Gemini-1.5-flash-001-tuning (Google)
    - *Razones para la Aptitud Ética:* Mantiene la ventana de 1 millón de tokens, heredada de su base 1.5-flash-001.
    - *Consideraciones:* Aunque comparte la base con 1.5-flash-001 (categorizado como Aptitud Media en esta evaluación para normas éticas), su rendimiento específico en el benchmark de normas éticas (según la tabla) lo clasifica como "Poco Adecuado". Esto podría deberse a variaciones introducidas por el proceso de "tuning" o a cómo interactuó con las preguntas específicas de las normas éticas. Las puntuaciones modestas en MMLU-Pro y GPQA del modelo base 1.5-flash podrían ser un factor que contribuye a esta clasificación inferior para tareas de revisión ética detallada.
  - Meta-llama/llama-4-scout-17b-16e-instruct (Groq - Meta)
    - *Razones para la Aptitud Ética:* Buen GPQA Diamond (57.2%) y ventana de 128K tokens.
    - *Consideraciones:* Menor rendimiento que Maverick en algunos benchmarks; modelo en "Preview".

Modelos Poco Adecuados				
Preguntas	llama-3.1-8b-instant	1.5-flash-001-tuning	meta-llama/llama-4-scout-17b-16e-instruct	Resultado Esperado
Q1	APROBADO	NO APROBADO	APROBADO	NO APROBADO
Q2	APROBADO	APROBADO	APROBADO	APROBADO
Q3	APROBADO	APROBADO	APROBADO	APROBADO
Q4	APROBADO	APROBADO	APROBADO	APROBADO
Q5	NO APROBADO	APROBADO	APROBADO	APROBADO
Q6	NO APROBADO	APROBADO	APROBADO	APROBADO
Q7.1	APROBADO	APROBADO	APROBADO	APROBADO
Q7.2	NO APROBADO	APROBADO	APROBADO	APROBADO
Q7.3	NO APROBADO	NO APROBADO	APROBADO	NO APROBADO
Q7.4	APROBADO	APROBADO	APROBADO	APROBADO
Q7.5	APROBADO	APROBADO	APROBADO	APROBADO
Q7.6	APROBADO	NO APROBADO	APROBADO	NO APROBADO
Q7.7	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q8	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO

Q9	APROBADO	NO APROBADO	APROBADO	NO APROBADO
Q10	NO APROBADO	NO APROBADO	NO APROBADO	NO APROBADO
Q11	APROBADO	APROBADO	APROBADO	APROBADO
Q12	APROBADO	NO APROBADO	APROBADO	APROBADO
% De exito	**66.67%**	**94.44%**	**77.78%**	**100.00%**

Recomendaciones Finales

Para la selección del modelo en la *Plataforma de Asistencia Ética para el Hospital Universitario San Ignacio* consideramos que los modelos más adecuados son los de la Sección 1: Aptitud alta:

- ***DeepSeek-r1-distill-llama-70b***
- ***Gemini-2.0-flash-001 / Gemini-2.0-flash***
- ***Meta-llama/llama-4-maverick-17b-128e-instruct***

Consideraciones Adicionales para la Plataforma del Hospital

Al tomar la decisión final para la Plataforma de Asistencia Ética del Hospital Universitario San Ignacio, es importante considerar los siguientes aspectos prácticos:

- *Límites de las Capas Gratuitas:* De los modelos recomendados en la categoría de "Aptitud Alta", aquellos accesibles a través de *Groq (Deepseek-r1-distill-llama-70b y Meta-llama/llama-4-maverick-17b-128e-instruct)* actualmente presentan una capa gratuita con limitaciones más estrictas, como un posible límite de alrededor de 20 peticiones al día, dependiendo de la longitud de tokens consumida por cada documento enviado. Esto podría ser un factor restrictivo si se anticipa un volumen de uso elevado o el procesamiento de documentos muy extensos de forma frecuente.
- *Flexibilidad de la Capa Gratuita de Gemini:* En contraste, los modelos *Gemini (Gemini-2.0-flash-001 / Gemini-2.0-flash),* accesibles a través de Google AI Studio, tienden a ofrecer una mayor flexibilidad en sus capas gratuitas en términos de volumen de peticiones, teniendo un límite de 15 peticiones por minuto.
- *Tendencias de Respuesta de los Modelos y Ajuste al Evaluador:* La elección entre estos modelos de alto rendimiento puede depender de las preferencias del equipo evaluador del Hospital. En las pruebas realizada identificamos estos patrones:
    - *Gemini***:** Suele ofrecer respuestas que tienden a la neutralidad y objetividad. En sus justificaciones, tiende a extraer y presentar el texto literal del documento que sustenta su evaluación, lo que puede ser útil para una verificación directa, aunque en ocasiones estas citas textuales pueden resultar extensas.
    - *DeepSeek***:** Puede percibirse como un modelo más estricto en sus evaluaciones, mostrando una tendencia a marcar más preguntas como "NO APROBADO" en comparación con otros. Sus justificaciones suelen ser rigurosas y directas, enfocándose en la ausencia o insuficiencia de evidencia explícita. Esto podría ser

beneficioso si se busca un escrutinio altamente conservador del cumplimiento normativo.

- *Llama (Maverick):* Tiende a ofrecer explicaciones o paráfrasis en sus justificaciones en lugar de siempre citar textualmente el documento. En contraste con DeepSeek, puede mostrar una tendencia a marcar más preguntas cómo "APROBADO", lo que podría interpretarse como un enfoque ligeramente más laxo o inferencial.

Por lo tanto, si se prefiere una evaluación más conservadora y estricta con justificaciones directas, DeepSeek podría ser una opción. Si se busca un balance con mayor flexibilidad, justificaciones basadas en citas textuales y una capa gratuita potencialmente más permisiva para un alto volumen, Gemini podría ser más adecuado. Llama (Maverick) ofrece otra alternativa potente con un perfil de respuesta más explicativo y una mayor propensión a la aprobación, lo que podría ser útil si se valora la interpretación contextual por encima de la evidencia literal estricta.

- *Precisión de los modelos recomendados:* A continuación se presenta un tabla con la estadística de precisión de los modelos recomendados, con respecto a las documentos ya evaluados por el Hospital

Precisión			
Modelos muy Adecuados	1er Documento	2do Documento	Total
2.0-flash	94.44%	100.00%	97.22%
2.0-flash-001	94.44%	100.00%	97.22%
deepseek-r1-distill-llama-70b	94.44%	83.33%	88.89%
meta-llama/llama-4-maverick-17b-128e-instruct	94.44%	100.00%	97.22%
Modelos Adecuados			
2.0-flash-lite	94.44%	100.00%	97.22%
1.5-flash-002	94.44%	94.44%	94.44%
2.0-flash-lite-001	88.89%	100.00%	94.44%
llama-3.3-70b-versatile	83.33%	94.44%	88.89%

Anexos

1. BenchMark IAs

Referencias

[1] Groq, "Supported Models," *Groq Console*. [Online]. Available: https://console.groq.com/docs/models. [Consultado: Mayo 2025].

[2] Google Cloud, "Modelos de IA generativa en Vertex AI," *Google Cloud Documentation*, Mayo 23, 2025. [Online]. Available: https://cloud.google.com/vertex-ai/generative-ai/docs/models.

[3] Meta Llama, "Llama3 MODEL_CARD.md," *GitHub*, 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[4] DeepSeek AI, "DeepSeek-R1-Distill-Llama-70B model card," *Hugging Face*, 2025. [Online]. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B (o URL específica de la tarjeta del modelo si es diferente).

[5] NVIDIA, "Llama 4 Maverick 17B 128E Instruct Model Card," *NVIDIA NGC*, Abr. 5, 2025. [Online]. Available: https://build.nvidia.com/meta/llama-4-maverick-17b-128e-instruct/modelcard.

[6] Mistral AI, "Mistral-Small-24B-Instruct-2501 model card," *Hugging Face*, Ene. 2025. [Online]. Available: https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501.

[7] Artificial Analysis, "Gemini 2.0 Flash (exp): Intelligence, Performance & Price Analysis," *Artificial Analysis*, Dic. 2024. [Online]. Available: https://artificialanalysis.ai/models/gemini-2-0-flash-experimental.

[8] Vals.ai, "google/gemini-2.0-flash-001," *Vals.ai*, Feb. 5, 2025. [Online]. Available: https://www.vals.ai/models/google_gemini-2.0-flash-001.

[9] Vellum AI, "LLM Benchmarks Overview, Limits, and Model Comparison," *Vellum AI Blog*, Sep. 8, 2024. [Online]. Available: https://www.vellum.ai/blog/llm-benchmarks-overview-limits-and-model-comparison.

[10] Google Cloud, "Gemini 1.5 Flash," *Google Cloud Documentation*, Sep. 24, 2024. [Online]. Available: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-flash.

[11] Google Cloud, "Gemini 2.0 Flash-Lite," *Google Cloud Documentation*, Feb. 25, 2025. [Online]. Available: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash-lite.

[12] Groq, "Llama 3.1 8B," *Groq Console*. [Online]. Available: https://console.groq.com/docs/model/llama-3.1-8b-instant. [Consultado: Mayo 2025].

[13] Groq, "Mistral Saba 24B," *Groq Console*. [Online]. Available: https://console.groq.com/docs/model/mistral-saba-24b. [Consultado: Mayo 2025].

[14] Groq, "Llama-3.3-70B-Versatile," *Groq Console*. [Online]. Available: https://console.groq.com/docs/model/llama-3.3-70b-versatile. [Consultado: Mayo 2025].