Cómo elegir, comparar y dominar modelos de IA

Título sugerido:
🔍 *"¿Y ahora qué IA uso? Cómo elegir sin perderse (ni perder el tiempo)"*

🎯 OBJETIVO:

Dar herramientas para que cada persona pueda **explorar, comparar y decidir** con confianza qué modelos y herramientas de IA le convienen, según sus necesidades reales.

□ CONTENIDO:

1. La pregunta del millón: ¿por dónde sigo?

- "Ya vi muchas IA… ¿cómo sé cuál me sirve?"
- Esta sesión es una brújula: **te ayuda a no depender de modas ni de tutoriales en YouTube**, sino a tomar decisiones basadas en criterio.

---

2. Antes de elegir: hazte estas preguntas

- ¿Qué problema necesito resolver?
- ¿Qué tipo de contenido necesito (texto, imagen, video, análisis)?
- ¿Cuánta precisión necesito? ¿Tiempo real? ¿Privacidad?
- ¿Lo necesito una vez o será recurrente?

---

3. ¿Cómo saber qué modelos existen?

- Seguir expertos en LinkedIn, Substack o X (Ej: Ethan Mollick, All About AI, Salomón en español).
- Consultar bibliotecas como:
  - https://theresanaiforthat.com
  - https://futurepedia.io
  - GPT Store (Explorar GPTs)

- Estar atentos a retos o resúmenes tipo *"Top herramientas IA del mes"*.

---

4. Regla de oro: usa el mismo prompt en distintos modelos

- Ponles **la misma tarea a todos** (ej: crear un guion, redactar un post, generar una imagen).
- Evalúa:
  - ¿Quién entiende mejor lo que pido?
  - ¿Cuál da un resultado más útil?
  - ¿Cuál se ajusta mejor a mi tono o estilo?

---

5. Hoja de ruta práctica para evaluar herramientas de IA

Paso	Acción	Ejemplo
1. Explora	Encuentra 2 o 3 opciones para lo que necesitas.	"IA para hacer videos con voz humana"
2. Prueba	Usa el **mismo prompt o tarea real**.	"Crea un video de bienvenida para mi curso"
3. Compara	Evalúa resultados, interfaz, velocidad.	¿Cuál entiende mejor? ¿Cuál es más intuitiva?
4. Analiza costos	¿Es gratuita, de pago, por tokens? ¿Hay prueba gratis?	¿Qué tanto ofrece sin pagar?
5. Decide	Elige la que **resuelve mejor tu necesidad**, no la más famosa.	

---

6. Criterios para evaluar modelos y herramientas

Criterio	Qué revisar
Calidad del resultado	Claridad, relevancia, precisión
Facilidad de uso	¿Requiere configuración técnica o es intuitiva?
Personalización	¿Se adapta a mi estilo, tono o necesidades?
Costos	¿Cuánto cuesta al mes? ¿Tiene versión gratuita funcional?
Privacidad	¿Dónde quedan los datos que subo? ¿Usan mi info para entrenarse?
Integraciones	¿Se conecta con Google Docs, Notion, redes sociales, etc.?
Comunidad / soporte	¿Hay tutoriales, ejemplos, actualizaciones?

---

7. Dinámica en vivo (opcional si hay tiempo)

"El reto IA": da una tarea real (crear imagen, hacer resumen, generar guion) y pide a los asistentes que la prueben en dos herramientas distintas. Luego comparten cuál les gustó más y por qué.

---

8. Cierre con mensaje potente

La clave no es saber todo, sino saber **cómo explorar, comparar y decidir.**
No hay una IA perfecta: hay una IA para cada necesidad.
Y esa la eliges tú, con criterio.

Benchmark	Gemini 2.5 Flash Preview (05-20) Thinking	Gemini 2.0 Flash	OpenAI o4-mini	Claude Sonnet 3.7 64k Ext. Thinking	Grok 3 Beta Extended thinking	DeepSeek R1
Input price $/M tokens	$0.15	$0.10	$1.10	$3.00	$3.00	$0.55
Output price $/M tokens	$0.60 No reasoning $3.50 Reasoning	$0.40	$4.40	$15.00	$15.00	$2.19
Reasoning & knowledge Humanity's Last Exam (no tools)	11.0%	5.1%	14.3%	8.9%	—	8.6%*
Science GPQA diamond (single attempt (pass@1))	82.8%	60.1%	81.4%	78.2%	80.2%	71.5%
(multiple attempts)	—	—	—	84.8%	84.6%	—
Mathematics AIME 2025 (single attempt (pass@1))	72.0%	27.5%	92.7%	49.5%	77.3%	70.0%
(multiple attempts)	—	—	—	—	93.3%	—
Code generation LiveCodeBench v5 (single attempt (pass@1))	63.9%	34.5%	—	—	70.6%	64.3%
(multiple attempts)	—	—	—	—	79.4%	—
Code editing Aider Polyglot	61.9% / 56.7% whole / diff-fenced	22.2% whole	68.9% / 58.2% whole / diff	64.9% diff	53.3% diff	56.9% diff
Agentic coding SWE-bench Verified	60.4%	—	68.1%	70.3%	—	49.2%
Factuality SimpleQA	26.9%	29.9%	—	—	43.6%	30.1%
Factuality FACTS Grounding	85.3%	84.6%	62.1%	78.8%	74.8%	56.8%
Visual reasoning MMMU (single attempt (pass@1))	79.7%	71.7%	81.6%	75.0%	76.0%	no MM support
(multiple attempts)	—	—	—	—	78.0%	no MM support
Image understanding Vibe-Eval (Reka)	65.4%	56.4%	—	—	—	no MM support
Long context MRCR v2 (128k (average))	74.0%	36.0%	49.0%	—	54.0%	45.0%
(1M (pointwise))	32.0%	6.0%	—	—	—	—
Multilingual performance Global MMLU (Lite)	88.4%	83.4%	—	—	—	—

1. Razonamiento y conocimiento (Humanity's Last Exam)

- **¿Qué mide?** Qué tan bien "piensa" la IA sin buscar ayuda.
- **¿Cómo te afecta?** Si tiene bajo puntaje, puede equivocarse en temas complejos.
- **¿Cómo compensarlo?** Usa preguntas más simples o acompaña con contexto claro.

🔬 2. Ciencia (GPQA Diamond)

- **¿Qué mide?** Qué tanto sabe el modelo sobre ciencia avanzada.
- **¿Cómo te afecta?** Si trabajas en ciencia o salud, necesitas uno con alto puntaje.
- **¿Cómo compensarlo?** Verifica las respuestas con fuentes confiables.

➗ 3. Matemáticas (AIME 2025)

- **¿Qué mide?** Habilidad para resolver problemas matemáticos difíciles.
- **¿Cómo te afecta?** Si necesitas cálculos o lógica, busca uno con alto rendimiento.
- **¿Cómo compensarlo?** Usa calculadora para confirmar o dividir el problema paso a paso.

🤹‍♂️💻 4. Generación de código (LiveCodeBench)

- **¿Qué mide?** Qué tan bien escribe código desde cero.
- **¿Cómo te afecta?** Si programas o automatizas tareas, elige un modelo fuerte en esta área.
- **¿Cómo compensarlo?** Prueba el código y corrige errores con ayuda del mismo modelo.

□ 5. Veracidad (Facts Grounding)

- **¿Qué mide?** Si las respuestas que da son verdaderas y no inventadas.
- **¿Cómo te afecta?** Puedes recibir respuestas que suenan bien, pero son falsas.
- **¿Cómo compensarlo?** Verifica los datos importantes en Google o fuentes oficiales.

👁□□ 6. Razonamiento visual (MMMU)

- **¿Qué mide?** Si entiende imágenes y responde bien con base en lo que ve.
- **¿Cómo te afecta?** Si trabajas con fotos, gráficos o planos, es clave.
- **¿Cómo compensarlo?** Describe bien la imagen si el modelo no la puede ver.

📚 7. Long context

- **¿Qué mide?** Qué tanto texto puede recordar en una sola conversación.

- **¿Cómo te afecta?** Si pegas un documento largo, puede olvidar partes.
- **¿Cómo compensarlo?** Divide el contenido por partes y haz preguntas por segmentos.