

AI Agents for Economic Research

Korinek, A., 2025

Francisco Iñiqui Irustia, Federico Ariel Lopez, Martina Palazzini

7 de octubre de 2025

1. Evolución de la IA
2. Servicios y Costos
3. Tipos de Modelos
4. AI Agents para Investigación Económica
5. Under the Hood: Construyendo Agentes de IA
6. Protocolos

Progreso de las habilidades de las IA

- **LLM Tradicionales** (2022-2023)
- **Modelos de Razonamiento** (2024)
- **Agentic Chatbots** (2025)

Servicios Premium

Table 5
PRICING TIERS FOR CHATBOT SUBSCRIPTIONS OF LEADING AI LABS
(USD PER MONTH, JULY 2025)

Lab / Provider	Basic tier	\$/mo	Most expensive tier	\$/mo
OpenAI ChatGPT	Plus	\$20	Pro	\$200
Google DeepMind Gemini	AI Pro	\$20	AI Ultra	\$250
Anthropic Claude	Pro	\$20	Max 20× usage	\$200
xAI Grok	X Premium	\$8	SuperGrok Heavy	\$300
Microsoft Copilot	Copilot Pro	\$20		

Otros Modelos y Modelos Abiertos

- Mistral
- Kimi-K2
- Qwen
- DeepSeek
- Minimax Locales:
- Llama
- GPT-OSS (120B y 20B)
- Gemma
- Phi3
- Otra opción: T3chat ('agregador')

... fijarse en Ollama y HuggingFace

'System-1 thinking' (Kahneman)

- Sin acceso a información en tiempo real
- No puede seguir pasos o derivaciones matematicas
- Sigue siendo muy cercano al next-word predictor (GPT-1)

Comparación de LLMs: <https://lmarena.ai/leaderboard>

Recomendado: 'What Is ChatGPT Doing ... and Why Does It Work? (Stephen Wolfram)'

'System-2 thinking'

- Resolución deliberada y paso a paso de problemas
- Identificación y corrección de errores
- Siguen sin ser más que next word predictors

Table 3
TOP REASONING AND AGENTIC CAPABILITIES BY LAB, GPQA &
SWE-BENCH-V SCORES

Lab	Model	Last Updated	GPQA \diamond Score	SWE-Bench V
OpenAI	GPT-5	2025-08-07	89.4%	75%
xAI	Grok-4	2025-07-10	88.9%	72-75%*
Google DeepMind	Gemini 2.5 Pro	2025-06-17	84.0%	63.8%
Anthropic	Claude Opus 4.1	2025-08-05	83.3%	72.7%
DeepSeek	DeepSeek-R1	2025-03-24	71.5%	49.2%

Source: Compiled by author. Last update: Aug 20th, 2025; * marks preliminary data.

Realizan acciones + usan herramientas externas

Los modelos previos que generan respuestas 'con lo que ya tienen', los agentes tienen funcionalidad extra:

- Búsquedas web
- Interacción con bases de datos
- Ejecución de código
- Manipulación de archivos

Es 'agentic' porque deja a la IA 'tomar la decisión' de usar una herramienta a la que tiene acceso.

EXTRA: ¿Qué quieren decir con 'Agente'?

En computación los agentes perciben su ambiente mediante 'sensores' y actúan sobre el ambiente por 'actuators.'

Los agentes de IA combinan LLMs con herramientas externas (tools) y un orquestador que coordina el flujo: Think \rightarrow Act \rightarrow Observe.

Componentes:

- **Reasoning Engine:** LLM que decide qué hacer
- **Tools:** búsqueda web, ejecución de código, APIs
- **Memory:** contexto acumulado
- **Orchestrator:** coordina el ciclo

Alignment: Economía se centra en el problema de principal-agente; en computación el problema se le dice alignment (Hendrycks, 2025)

54

CHAPTER 3. FINITE MARKOV DECISION PROCESSES

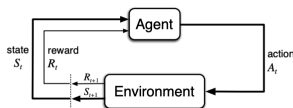


Figure 3.1: The agent-environment interaction in reinforcement learning.

Sutton & Barto

Arquitectura de un Agente de IA

Korinek: AI Agents for Economic Research

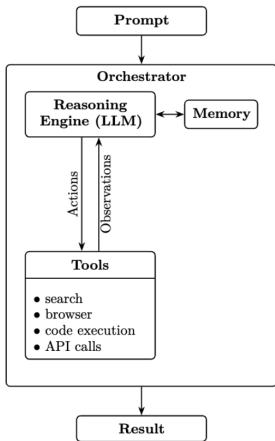


Figure 1. AI AGENT ARCHITECTURE

- Un *orchestrator* pasa el objetivo original (ej: prompt del usuario) y la lista de herramientas disponibles a un LLM de razonamiento.
- Este LLM decide qué herramientas externas llamar. Las herramientas proveen al sistema una interfaz con el mundo externo:
- Cada vez que el motor de razonamiento quiere llamar una herramienta, genera tokens que indican al orchestrator llamar la herramienta e insertar el resultado de vuelta antes de continuar.

Creado por Google DeepMind 2024

Table 6
OVERVIEW OF DEEP RESEARCH AGENTS

System	Availability & usage limits	Features	Time
Gemini	Free to try; better & higher allowance on paid plans	Proposes a research plan that user can confirm/ modify	5-10 min
Deep Research	Allowance of 25/200 reports per month for Plus/Pro plan	Asks follow-up questions to optimally target response	5-30 min
OpenAI	Available subject to limits for paid plans	Runs immediately; can connect with workspace	5-15 min
Claude	Free to try; unlimited under paid plan	Fast option covering lots of sources	2-4 min
Deep Research	Available under X premium plan	Real-time info through close integration with X platform	<1-10 min
Perplexity			
Deep Research			
xAI Grok			
DeepSearch			

- Un agente líder es el orchestrator que recibe el query
- Arma subtasks y cada una la realiza alguno de los subagentes especializados
- Síntesis final de resultados de múltiples fuentes

Terminal-based coding y 'Vibe Coding'

'Vibe coding' – crear proyectos de software completos basados en prompts en lenguaje natural. Ha hecho posible que usuarios sin experiencia en programación creen proyectos de software de principio a fin.

Full vibe coding:

- Claude Code
- Gemini CLI
- Codex + Open Codex

Herramientas intermedias:

- GitHub Copilot
- Cursor
- Windsurf
- Cline

Data Retrieval Agent (Primera Generación)

Lo que hace este código un 'agente' en lugar de un simple script es su comportamiento autónomo y dirigido a objetivos a través de múltiples pasos:

El Loop Think-Act-Observe:

1. **Think:** Piensa qué datos ayudarían a responder la pregunta
2. **Act:** Llama la herramienta externa apropiada (FRED API)
3. **Observe:** Analiza los datos retornados
4. **Respond:** Responde con una respuesta en lenguaje natural

Este loop Think-Act-Observe caracteriza sistemas agentic sofisticados.

¿Y pasa si necesitamos capacidades más complejas?

- ¿Y si el agente necesita obtener múltiples series de datos?
- ¿Y si queremos que el agente pueda retroceder cuando se da cuenta de que está siguiendo un camino improductivo?

LangGraph provee un framework para construir agentes con:

- Grafos de estado complejos
- Capacidad de backtracking
- Memoria persistente
- Coordinación multi-agente

The Model Context Protocol (MCP) y Agent2Agent Protocol

Anthropic, 2024

Para economistas, los servidores MCP de uso general proveen a los agentes de IA acceso automatizado a:

- Sistema de archivos del usuario
- Sistemas de email (Gmail, Outlook)
- Bases de datos financieras
- Bases de datos espaciales
- GitHub
- Slack
- Wrappers de FRED
- FMI
- más... en pulse

Agent2Agent: (Abril 2025) Protocolo para la interacción entre agentes

Si todo lo puede hacer la IA ¿para qué servimos los economistas? ¿la investigación va a quedar obsoleta?

Por ahora, no.

Mientras los agentes de IA sobresalen en síntesis e implementación,

- La identificación de preguntas de investigación novedosas
- La intuición conectando fenómenos dispares

siguen siendo contribuciones humanas.

Pero... hay esfuerzos en marcha que buscan automatizar estas contribuciones creativas usando sistemas de IA agentic — ver, por ejemplo, Google's AI Co-Scientist.



Benjamin Golub · 2°

Professor of Economics at Northwestern and Co-Founder at Refine

1 semana · Editado ·

✓ Siguiendo ...

I've been working on an AI tool, Refine, that reads research papers like a referee and finds issues with correctness, clarity, and consistency.

In my own papers, it regularly catches problems that my coauthors and I missed. Compared to a regular chatbot prompt, it's much more thorough and reliable.

We'd like researchers to test the beta product at <https://refine.ink> - if you create an account, you can try a preview to see how Refine performs on your own work.

Mostrar traducción

Session View

29/09/2025, 16:01:37

Share

New Document

History

questions about mothers' behavior during the last period, such as the use of serviettes. Attitudes includes questions about attitudes towards menstruation, where a higher value indicates more progressive attitudes. Girls' behavior includes mothers' reports about the psychosocial behavior of their daughters.

Table A17: Effects on bullying, disaggregated

VARIABLES | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |

| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |

| | | Self-report | Peer-report Combined (Z) | Peer-report Teasing only (Z) | Peer-report Intimidation/ harassment only (Z) | Teacher-report Combined (Z) |

Teacher-report Teasing only (Z) | Teacher-report Intimidation/ harassment only (Z) |

| | Self-report | Self-report | | | | | | |

| Combined (Z) | Teasing only (Z) | Intimidation/ harassment only (Z) | | | | | |

| Base Only | 0.029 | 0.108 | -0.083* | -0.074 | -0.029 | -0.081* | -0.067 | -0.023 | -0.074 |

| (0.059) | (0.066) | (0.046) | (0.052) | (0.061) | (0.046) | (0.056) | (0.067) | (0.049) |

| Base + YGL | -0.026 | 0.028 | -0.009* | -0.061 | -0.006 | -0.070 | -0.028 | 0.032 | -0.052 |

| (0.053) | (0.058) | (0.040) | (0.050) | (0.059) | (0.044) | (0.052) | (0.063) | (0.047) |

| Observations | 2,167 | 2,167 | 2,167 | 3,042 | 3,042 | 3,637 | 3,637 | 3,637 | 3,637 |

| Data source | Girls | Girls | Girls | Girls | Girls | Girls | Teachers | Teachers | Teachers |

| p: Treated = 0 | . 8933 | . 3858 | . 0624 | . 1459 | . 7864 | . 0652 | . 4343 | . 7762 | . 3917 |

| p: Base Only = Base + YGL | . 2927 | . 1093 | . 7349 | . 7788 | . 6944 | . 7869 | . 3908 | . 3266 | . 5793 |

Notes: Table shows girls' and teachers' reports of bullying. Columns (1)–(3) are girls' reports about bullying towards themselves. Columns (4)–(6) are girls' reports about bullying towards other girls, using a randomly selected subset of 2 or 3 girls in their gradelevel (and removing cases where the respondent did not know the other girl). Columns (7)–(9) are teachers' reports about randomly selected girls in their class. Each teacher reports on a randomly selected sample of 3 girls from their class; the observations are at the teacher's times girl level. Columns (1), (4) and (7) combine reports of both light teasing and severe intimidation/ harassment. Columns (2), (5), and (8) include only light teasing. Columns (3), (6) and (9) include only severe intimidation/ harassment. Combined indexes are constructed using loadings from exploratory factor analysis. The same relative loadings are used for teasing and intimidation/ harassment. Standard errors are clustered at the school level and are in parentheses. Sample includes all girls in online interview in person. Controls include stratum fixed effects and the baseline controls selected by double LASSO. All outcomes are in control group standard deviations.

Figure A18: Individual network questions

![[https://cdn.mathpix.com/cropped/2025_09_29_03059c90c2a2f033dfabg-51.jpg?height=966&width=1537&top_left_y=968&top_left_x=274]]

Notes: This figure shows the effects on each outcome that is used in the network index in Table 6. Only girls included in the baseline sample were used. All outcomes are normalized using the control mean and standard deviation from the endline. All regressions control for stratum fixed effects and confidence intervals are calculated based on standard errors clustered at the school level.

Table A19: Cleaning collaboration at school

| | | | |

| :--- | :--- | :--- | :--- | :--- |

| | (1) Cleaning collaboration index | (2) Days classroom cleaning (0–7) | (3) Days toilet cleaning (0–7) | (4) Days courtyard cleaning (0–7) |

Feedback 67

Overall ▾

Download

Priority Order

Paper Order

Dismissed

The bullying variables in Table 6 are described as “combining peer and self-reports.” However, the coefficients shown for teasing and for intimidation/ harassment match—after sign reversal—the self-report numbers in Appendix Table A17, while differing from the peer-report numbers. This indicates that the Table 6 outcomes are based almost entirely on self-reports, not on the advertised combination. Because peer- and self-reports tell different stories (e.g., self-reports suggest Base Only raises teasing, peer-reports suggest it lowers it), clarifying which data were used and, if necessary, recalculating the combined index is important for interpreting the program's effects on bullying and the broader mechanism discussion.

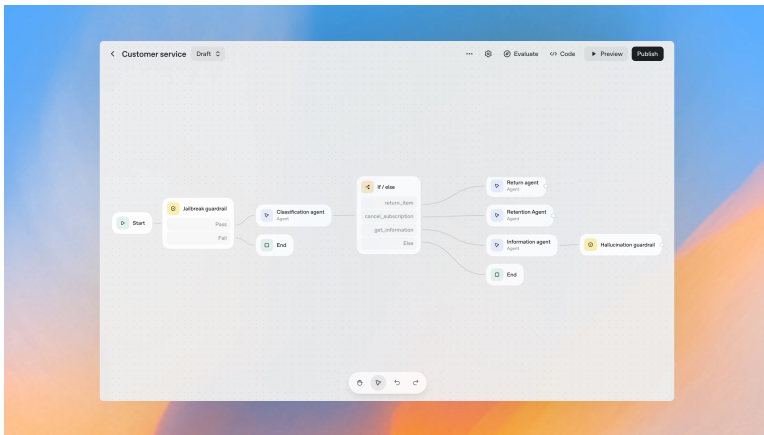
Reference to non-existent table column

As robustness tests, we drop or halve any observations above 175 bpm, and show that this still yields significant pooled effects (Table C3; $sp=0.095$ and 0.07 , columns 2 and 3). Results are also robust to using girl's mean heart rate during the interview rather than all the heart rates in each 30 second window ($sp=0.025$, column 1) and to controlling for age fixed effects ($sp=0.025$, column 4).

The paragraph cites “column 4” of Table C3 for the specification that adds age fixed effects, but Table C3 as printed displays only three columns. The age-FE regression—and the stated p -value of 0.02—cannot be located elsewhere in the appendix or main text. Please either reinstate the missing column or revise the text so that every robustness claim is documented in the table the reader is directed to.

AgentKit (Ayer)

AgentKit



limitaciones:

- Alucinaciones y errores que se propagan sin detección
- Fragilidad ante variaciones en prompts
- Dificultades con razonamiento económico genuino

"Tratar a los agentes de IA como un profesor trataría a un equipo de asistentes de investigación: requieren planificación cuidadosa, supervisión durante la ejecución, y verificación detallada de resultados."