



**MIA**  
Magíster en  
Inteligencia Artificial

# AGENTES INTELIGENTES

1/7/2025

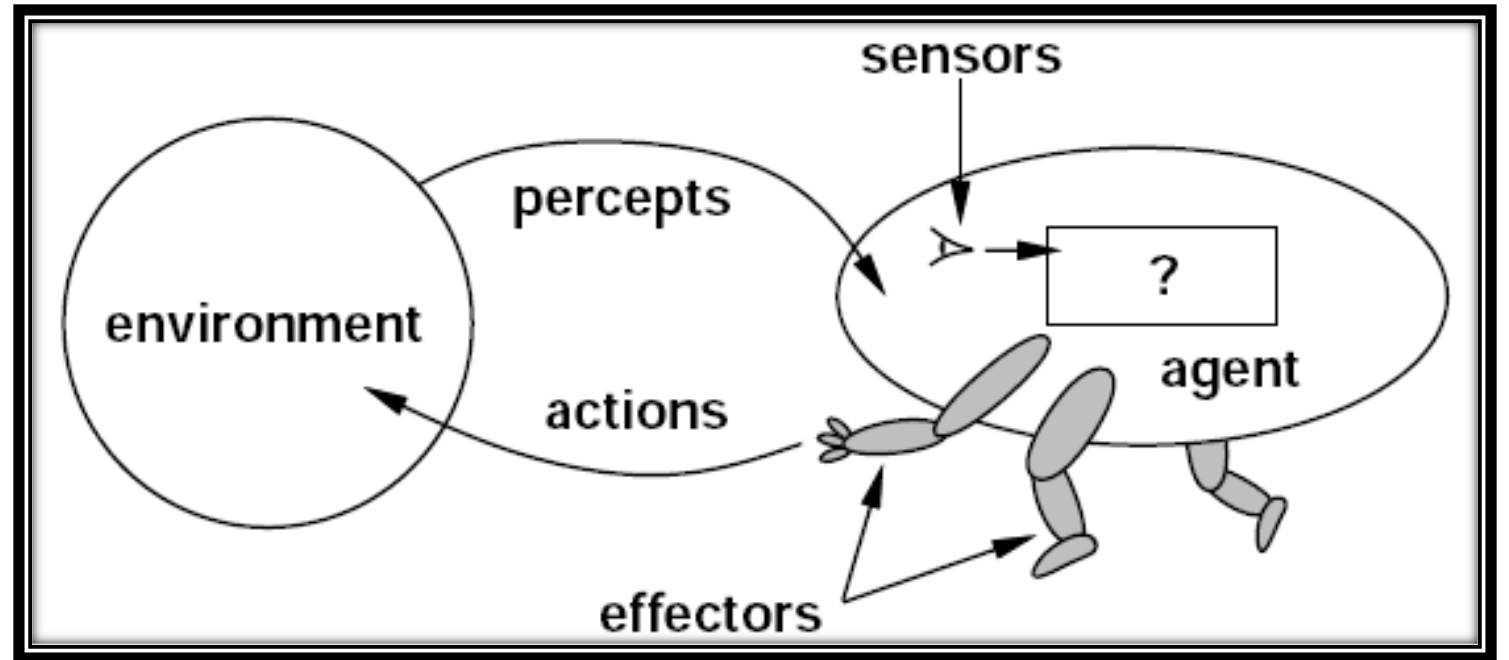
**Marcelo Mendoza**

DCC UC y CENIA



¿QUÉ ES UN AGENTE?

# AGENTE



Un agente es una entidad que percibe el entorno, procesa información y toma acciones para lograr objetivos y metas específicas.

# EVALUACIÓN DE LLMs



## MMLU (Massive Multitask Language Understanding)

Diseñado para evaluar grandes modelos, como los LLMs. Está enfocado en preguntas de selección múltiple y cubre más de 57 temáticas incluyendo matemática, historia, leyes, entre otras.

Conceptual Physics	When you drop a ball from rest it accelerates downward at $9.8 \text{ m/s}^2$ . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is	
	(A) $9.8 \text{ m/s}^2$	✓
	(B) more than $9.8 \text{ m/s}^2$	✗
	(C) less than $9.8 \text{ m/s}^2$	✗
College Mathematics	(D) Cannot say unless the speed of throw is given.	✗
	In the complex $z$ -plane, the set of points satisfying the equation $z^2 =  z ^2$ is a	
	(A) pair of points	✗
	(B) circle	✗
	(C) half-line	✗
	(D) line	✓

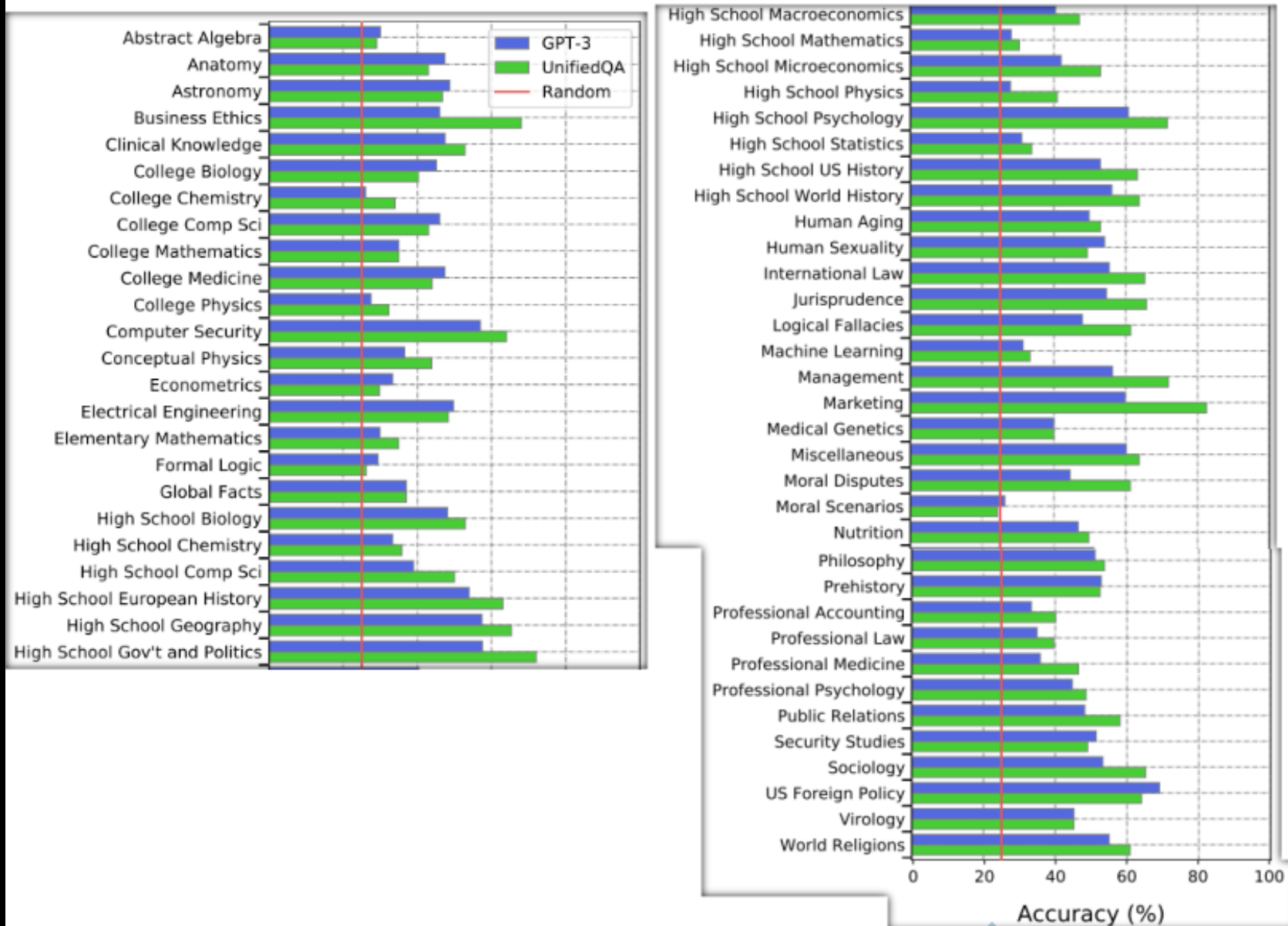
Se calcula un puntaje global entre  $[0, 1]$ , donde 1 indica desempeño perfecto.

Es un proyecto de evaluación abierta: <https://github.com/hendrycks/test>



Measuring Massive Multitask Language Understanding Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt (ICLR 2021).

## MMLU (Massive Multitask Language Understanding)



## MT Bench (Multiturn Benchmark)

MT Bench busca evaluar las capacidades conversacionales de un LLM. Para esto, incorpora ejemplos con varias interacciones (multiturn interactions).

Category	Sample Questions	
Writing	1st Turn	Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.
	2nd Turn	Rewrite your previous response. Start every sentence with the letter A.
Math	1st Turn	Given that $f(x) = 4x^3 - 9x - 14$ , find the value of $f(2)$ .
	2nd Turn	Find $x$ such that $f(x) = 0$ .
Knowledge	1st Turn	Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies ...
	2nd Turn	Now, explain them again like I'm five.



Lianmin Zheng Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023.



## MT Bench (Multiturn Benchmark)

Se le llama **LLM as a judge** ya que se usa un tercer LLM (uno fuerte como GPT4) para evaluar las dos alternativas.

LlaMa 13B

**Question:** If the FED buys bonds in the secondary market

(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

**Assistant A:**

The Federal Reserve buys bonds in the secondary market to increase the money supply.

**Assistant B:**

(A) the money supply will increase.

**Follow-up Question:** How does it affect my daily life? Give 3 examples.

**Assistant A:**

The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:

1. Increasing the money supply.
2. Increasing the money supply and lowering the discount rate.
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

**Assistant B:**

When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

**GPT-4 Judgment:**

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.

On the other hand, Assistant B provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. Assistant B's response is more helpful, relevant, and detailed.

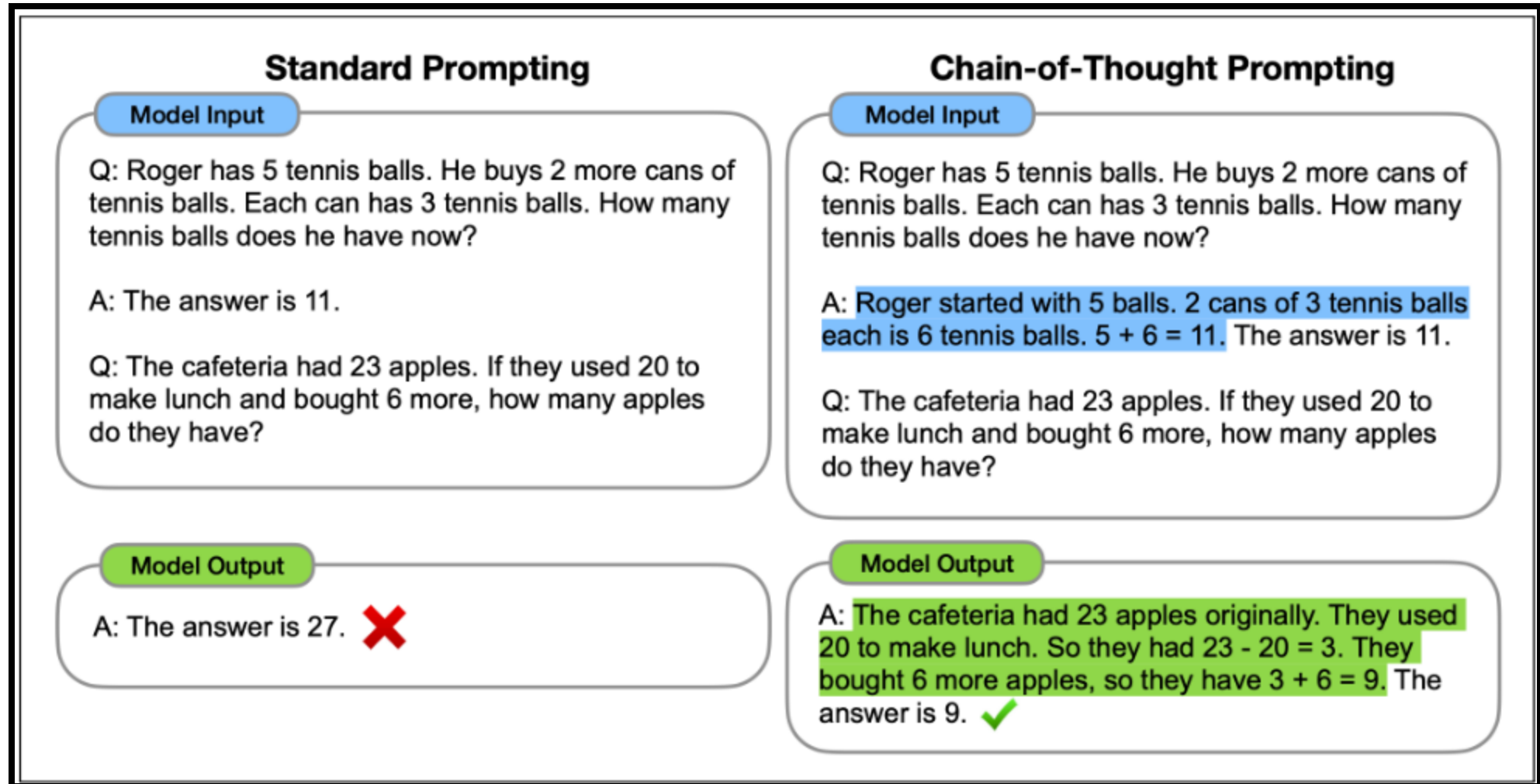
razonamiento



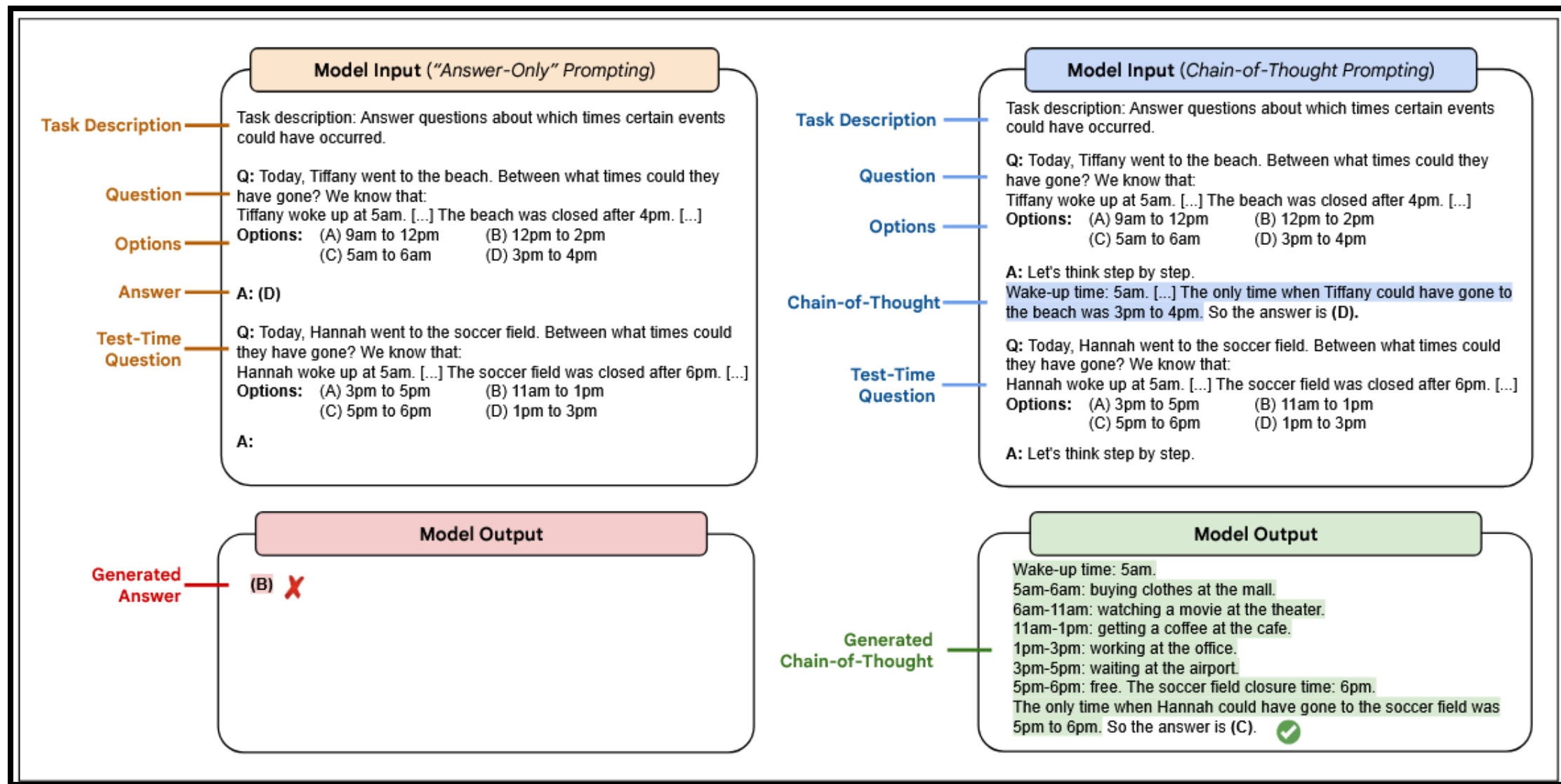
# LLMs IMITANDO EL RAZONAMIENTO HUMANO



# Chain-of-thought (COT)



# Chain-of-thought (COT)



# Reason + Act

## (1) Hotspot QA

**Question:** Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

### (1a) Standard

**Answer:** iPod



### (1b) CoT (Reason Only)

**Thought:** Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

**Answer:** iPhone, iPad, iPod Touch



### (1c) Act-Only

**Act 1:** Search[Apple Remote]  
**Obs 1:** The Apple Remote is a remote control ...

**Act 2:** Search[Front Row]  
**Obs 2:** Could not find [Front Row]. Similar: ...

**Act 3:** Search[Front Row (software)]

**Obs 3:** Front Row is a discontinued media center software ...

**Act 4:** Finish[yes]



## (1d) ReAct (Reason + Act)

**Thought 1:** I need to search Apple Remote and find the program it was originally designed to interact with.

**Act 1:** Search[Apple Remote]

**Obs 1:** The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...

**Thought 2:** Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.

**Act 2:** Search[Front Row]

**Obs 2:** Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

**Thought 3:** Front Row is not found. I need to search Front Row (software) .

**Act 3:** Search[Front Row (software)]

**Obs 3:** Front Row is a discontinued media center software ...

**Thought 4:** Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.


**Act 4:** Finish[keyboard function keys]



# ALGUNOS EJEMPLOS




# Una tarea compleja con CoT

**News**  
10m

**BREAKING NEWS**

**ESA Launches New Satellite to Monitor Greenhouse Gas Emissions with Unprecedented Precision**



The European Space Agency (ESA) has successfully launched a new satellite designed to track greenhouse gas emissions more accurately than ever before

**WHO:**

The **European Space Agency (ESA)** announced a major breakthrough in its climate monitoring programme.

**WHAT:**

It successfully launched a new satellite designed to track greenhouse gas emissions with unprecedented precision.

**WHEN:**

The launch took place on **Thursday, April 17, 2025**, marking a critical step in the fight against climate change.

**WHERE:**

The mission was conducted at the **Guiana Space Centre in Kourou, French Guiana**.

**WHY:**

The initiative aims to provide governments and researchers with accurate data to monitor compliance with international climate agreements.


**HOW:**

The satellite, named **CarbonSentinel-1**, uses advanced spectrometry to detect variations in atmospheric CO<sub>2</sub> levels and transmits real-time data to a global observation network.

Muñoz, C., Mendoza, M., Lobel, H., Keith, B. Imitating human reasoning to extract 5W1H, ACM WWW 2025.




# Una tarea compleja con CoT

 News  
10m

**BREAKING NEWS**

**ESA Launches New Satellite to Monitor Greenhouse Gas Emissions with Unprecedented Precision**



The European Space Agency (ESA) has successfully launched a new satellite designed to track greenhouse gas emissions more accurately than ever before

**WHO:**  
The European Space Agency (ESA) announced a major breakthrough in its climate monitoring programme.

**WHAT:**  
It successfully launched a new satellite designed to track greenhouse gas emissions with unprecedented precision.

**WHEN:**  
The launch took place on Thursday, April 17, 2025, marking a critical step in the fight against climate change.

**WHERE:**  
The mission was conducted at the Guiana Space Centre in Kourou, French Guiana.

**WHY:**  
The initiative aims to provide governments and researchers with accurate data to monitor compliance with international climate agreements.

**HOW:**  
The satellite, named CarbonSentinel-1, uses advanced spectrometry to detect variations in atmospheric CO<sub>2</sub> levels and transmits real-time data to a global observation network.

## Prompts:

- P1 (zero-shot) is a prompt that defines the extraction task.
- P2 (zero-shot) builds upon P1 by adding a detailed description of each 5W1H element.
- P3 (zero-shot) extends P2 by adding the instruction, "Only use excerpts from the provided context."
- P4 (one-shot) adds to P3 an example that includes a news article and the expected responses.
- P5 (few-shot) builds on P4 by adding a second example along with the expected answers.
- P6 (few-shot) builds on P5 by adding a third example with the corresponding expected answers.
- P7 (Extractive COT, ours) defines guidelines for removing irrelevant text. After irrelevant text is filtered out, the 5W1H extraction is performed using one-shot prompting.
- P8 (Extractive COT, ours) mirrors P7 but uses few-shot prompting based on two examples.
- P9 (Extractive COT, ours) follows the same logic as P8 but incorporates three examples.
- P10 (Question-level COT, ours) introduces complex reasoning for each question using one example.
- P11 (Question-level COT, ours) mirrors P10 but with two examples, making it COT few-shot.

Muñoz, C., Mendoza, M., Lobel, H., Keith, B. Imitating human reasoning to extract 5W1H, ACM WWW 2025.



# Una tarea compleja con CoT

## Prompts:

- P1 (zero-shot) is a prompt that defines the extraction task.
- P2 (zero-shot) builds upon P1 by adding a detailed description of each 5W1H element.
- P3 (zero-shot) extends P2 by adding the instruction, “Only use excerpts from the provided context.
- P4 (one-shot) adds to P3 an example that includes a news article and the expected responses.
- P5 (few-shot) builds on P4 by adding a second example along with the expected answers.
- P6 (few-shot) builds on P5 by adding a third example with the corresponding expected answers.
- P7 (Extractive COT, ours) defines guidelines for removing irrelevant text. After irrelevant text is filtered out, the 5W1H extraction is performed using one-shot prompting.
- P8 (Extractive COT, ours) mirrors P7 but uses few-shot prompting based on two examples.
- P9 (Extractive COT, ours) follows the same logic as P8 but incorporates three examples.
- P10 (Question-level COT, ours) introduces complex reasoning for each question using one example.
- P11 (Question-level COT, ours) mirrors P10 but with two examples, making it COT few-shot.

Table 1: Evaluation of prompting techniques for 5W1H extraction from Giveme5W1H data using LLMs. Bold fonts indicate the best results per metric for each LLM. The best global results per metric are depicted in red.

	P#	ROUGE-2	ROUGE-L	ROUGE-1	F_BERT
GPT-4o	P1	0.166	0.293	0.305	0.866
	P2	0.193	0.297	0.309	0.868
	P3	0.219	0.367	0.377	0.878
	P4	0.196	0.331	0.346	0.870
	P5	0.206	0.346	0.360	0.872
	P6	0.220	0.376	0.389	0.877
	P7	0.244	0.414	0.421	<b>0.892</b>
	P8	0.249	0.421	0.428	<b>0.892</b>
	P9	0.243	0.410	0.417	0.890
	P10	<b>0.260</b>	<b>0.427</b>	<b>0.437</b>	0.888
	P11	0.232	0.382	0.392	0.881
Claude-3.5	P1	0.129	0.226	0.243	0.862
	P2	0.153	0.257	0.274	0.869
	P3	0.173	0.292	0.310	0.876
	P4	0.145	0.250	0.269	0.868
	P5	0.164	0.278	0.297	0.874
	P6	0.173	0.330	0.348	0.876
	P7	0.202	0.320	0.338	0.882
	P8	0.196	0.315	0.330	0.882
	P9	0.195	0.318	0.333	0.882
	P10	0.236	0.375	0.387	<b>0.895</b>
	P11	<b>0.248</b>	<b>0.412</b>	<b>0.425</b>	0.893
Gemini-1.5	P1	0.134	0.238	0.254	0.860
	P2	0.161	0.295	0.311	0.873
	P3	0.160	0.288	0.302	0.873
	P4	0.126	0.233	0.244	0.858
	P5	0.139	0.249	0.260	0.860
	P6	0.149	0.279	0.290	0.864
	P7	0.204	0.326	0.339	<b>0.882</b>
	P8	0.210	0.324	0.337	0.878
	P9	0.200	0.315	0.328	0.877
	P10	<b>0.226</b>	<b>0.351</b>	<b>0.365</b>	0.872
	P11	0.174	0.282	0.291	0.861

Muñoz, C., Mendoza, M., Lobel, H., Keith, B. Imitating human reasoning to extract 5W1H, ACM WWW 2025.

# Una tarea compleja con CoT

## Prompts:

- P1 (zero-shot) is a prompt that defines the extraction task.
- P2 (zero-shot) builds upon P1 by adding a detailed description of each 5W1H element.
- P3 (zero-shot) extends P2 by adding the instruction, “Only use excerpts from the provided context.
- P4 (one-shot) adds to P3 an example that includes a news article and the expected responses.
- P5 (few-shot) builds on P4 by adding a second example along with the expected answers.
- P6 (few-shot) builds on P5 by adding a third example with the corresponding expected answers.
- P7 (Extractive COT, ours) defines guidelines for removing irrelevant text. After irrelevant text is filtered out, the 5W1H extraction is performed using one-shot prompting.
- P8 (Extractive COT, ours) mirrors P7 but uses few-shot prompting based on two examples.
- P9 (Extractive COT, ours) follows the same logic as P8 but incorporates three examples.
- P10 (Question-level COT, ours) introduces complex reasoning for each question using one example.
- P11 (Question-level COT, ours) mirrors P10 but with two examples, making it COT few-shot.

Table 1: Evaluation of prompting techniques for 5W1H extraction from Giveme5W1H data using LLMs. Bold fonts indicate the best results per metric for each LLM. The best global results per metric are depicted in red.

	P#	ROUGE-2	ROUGE-L	ROUGE-1	F_BERT
GPT-4o	P1	0.166	0.293	0.305	0.866
	P2	0.193	0.297	0.309	0.868
	P3	0.219	0.367	0.377	0.878
	P4	0.196	0.331	0.346	0.870
	P5	0.206	0.346	0.360	0.872
	P6	0.220	0.376	0.389	0.877
	P7	0.244	0.414	0.421	<b>0.892</b>
	P8	0.249	0.421	0.428	<b>0.892</b>
	P9	0.243	0.410	0.417	0.890
	P10	<b>0.260</b>	<b>0.427</b>	<b>0.437</b>	0.888
	P11	0.232	0.382	0.392	0.881
Claude-3.5	P1	0.129	0.226	0.243	0.862
	P2	0.153	0.257	0.274	0.869
	P3	0.173	0.292	0.310	0.876
	P4	0.145	0.250	0.269	0.868
	P5	0.164	0.278	0.297	0.874
	P6	0.173	0.330	0.348	0.876
	P7	0.202	0.320	0.338	0.882
	P8	0.196	0.315	0.330	0.882
	P9	0.195	0.318	0.333	0.882
	P10	0.236	0.375	0.387	<b>0.895</b>
	P11	<b>0.248</b>	<b>0.412</b>	<b>0.425</b>	0.893
Gemini-1.5	P1	0.134	0.238	0.254	0.860
	P2	0.161	0.295	0.311	0.873
	P3	0.160	0.288	0.302	0.873
	P4	0.126	0.233	0.244	0.858
	P5	0.139	0.249	0.260	0.860
	P6	0.149	0.279	0.290	0.864
	P7	0.204	0.326	0.339	<b>0.882</b>
	P8	0.210	0.324	0.337	0.878
	P9	0.200	0.315	0.328	0.877
	P10	<b>0.226</b>	<b>0.351</b>	<b>0.365</b>	0.872
	P11	0.174	0.282	0.291	0.861

Table 2: Best results obtained by each LLM for every 5W1H question. We highlight in bold fonts the best result per metric for each LLM. The red color marks the overall best performance for each question.

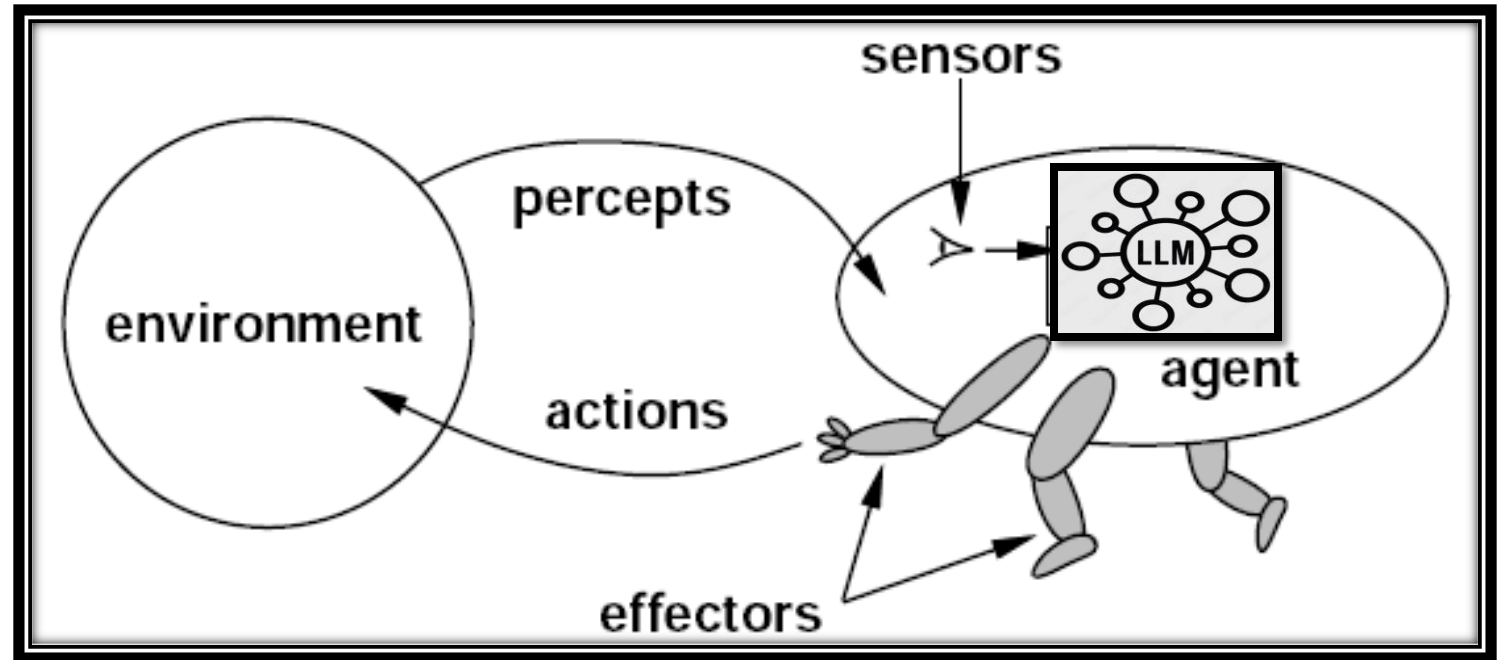
	Query	P#	ROUGE-2	ROUGE-L	ROUGE-1	F_BERT
GPT-4o	what	P8	0.276	<b>0.383</b>	0.404	0.885
	who	P9	<b>0.499</b>	0.638	0.641	0.916
	when	P10	<b>0.292</b>	<b>0.112</b>	<b>0.570</b>	0.906
	where	P10	0.310	<b>0.680</b>	<b>0.690</b>	<b>0.930</b>
	why	P7	<b>0.234</b>	<b>0.317</b>	<b>0.321</b>	<b>0.880</b>
	how	P10	0.073	0.117	0.121	0.840
Claude-3.5	what	P11	<b>0.280</b>	0.361	0.382	<b>0.886</b>
	who	P11	<b>0.514</b>	<b>0.650</b>	<b>0.677</b>	<b>0.927</b>
	when	P11	0.233	0.558	0.561	<b>0.913</b>
	where	P11	0.246	0.497	0.517	0.919
	why	P11	0.131	0.280	0.285	0.874
	how	P11	0.085	0.125	0.130	0.843
Gemini-1.5	what	P10	0.188	0.285	0.305	0.866
	who	P10	<b>0.401</b>	<b>0.550</b>	<b>0.573</b>	<b>0.896</b>
	when	P10	0.236	0.416	0.431	0.873
	where	P10	0.265	0.462	0.474	0.886
	why	P10	0.175	0.247	0.255	0.864
	how	P10	0.094	0.147	0.153	0.845

Muñoz, C., Mendoza, M., Lobel, H., Keith, B. Imitating human reasoning to extract 5W1H, ACM WWW 2025.

# AGENTES BASADOS EN LLM



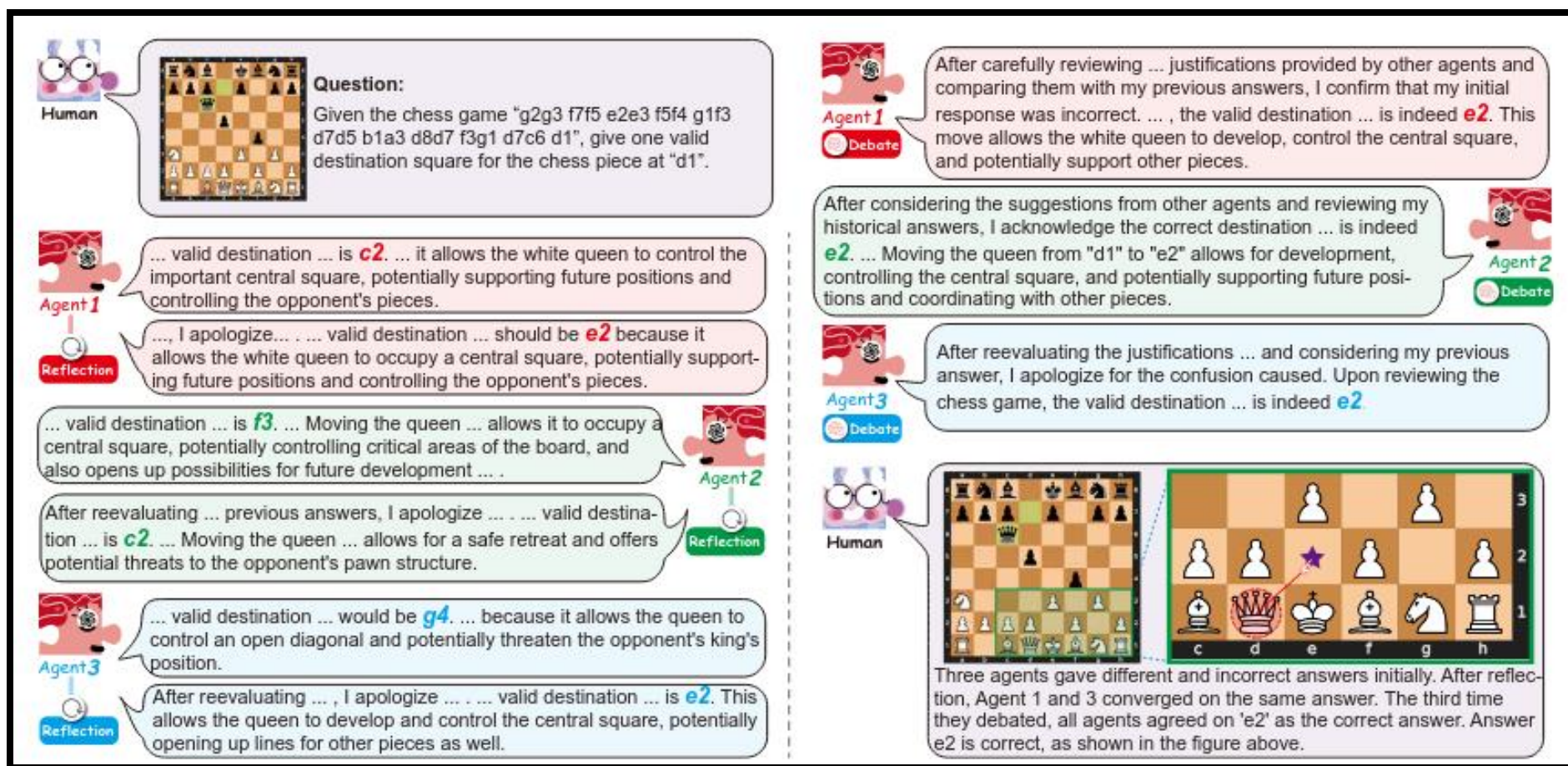
# AGENTES + LLM

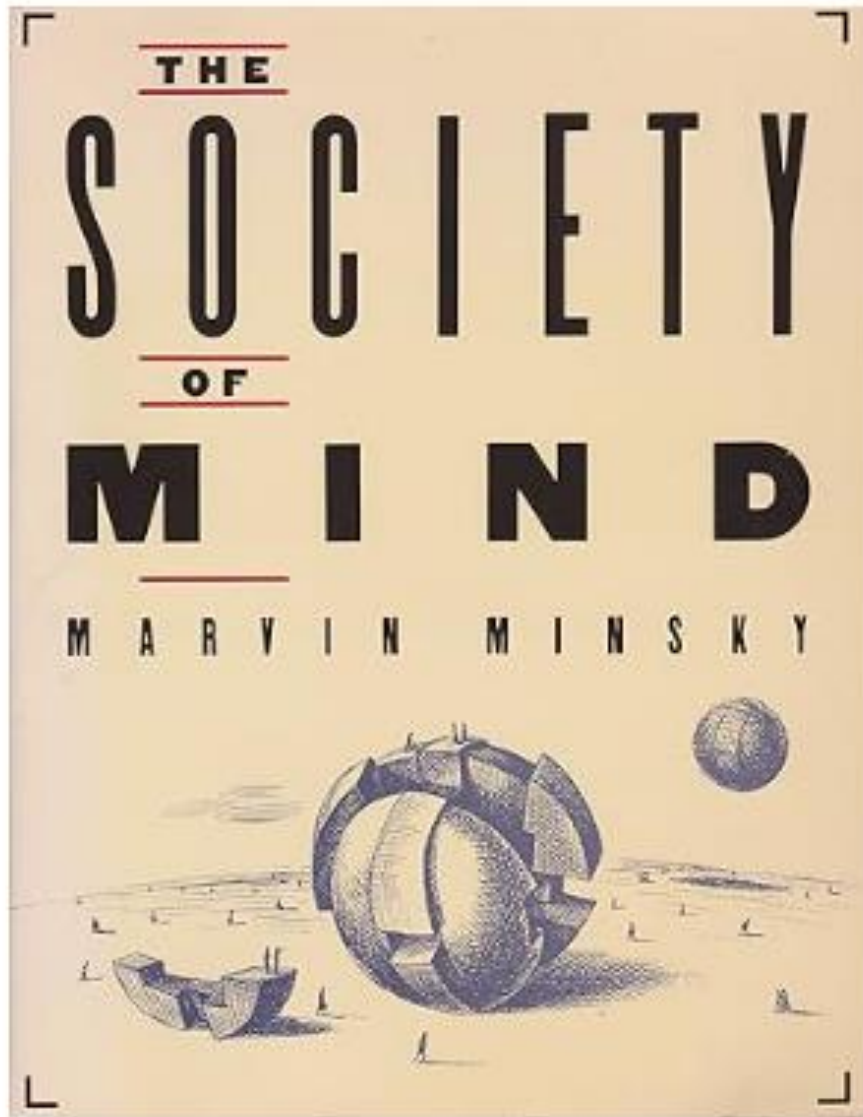


Un agente incluye una IA (LLM), percibe el entorno, procesa información y toma acciones para lograr objetivos y metas específicas.



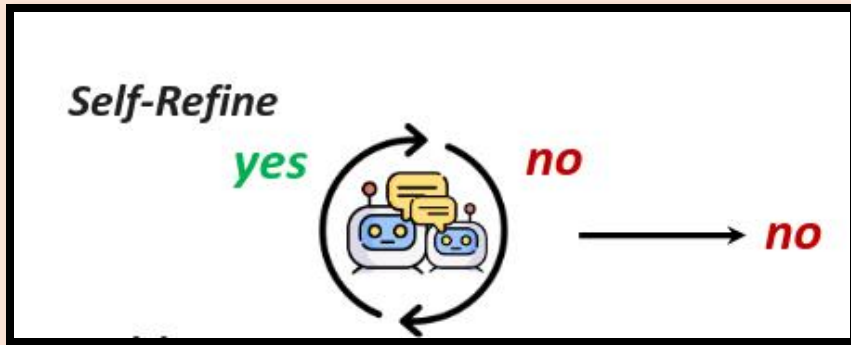
# Podemos usar agentes para abordar problemas de decisión





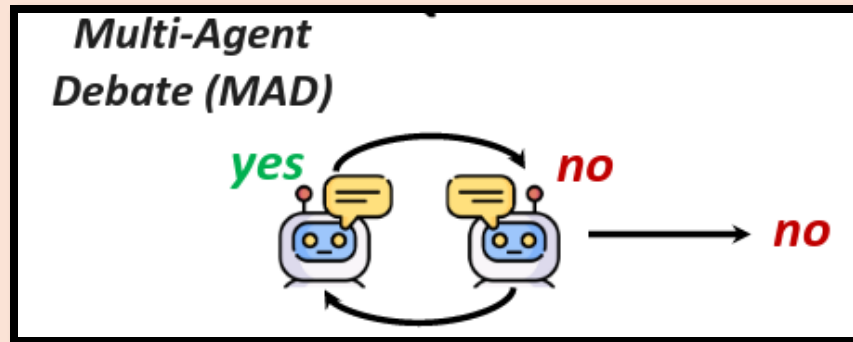
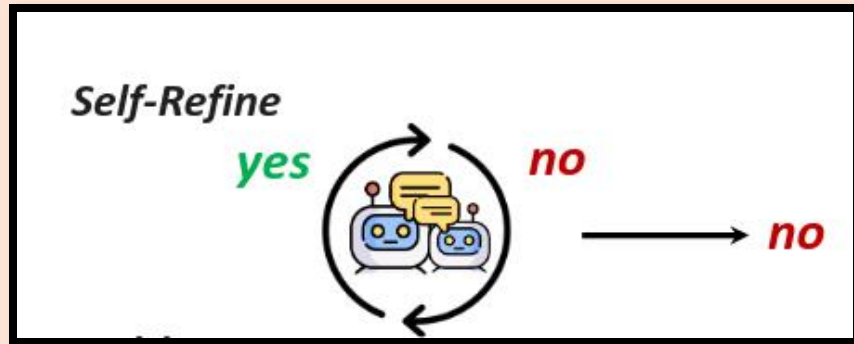
La base de la  
resolución de problemas  
complejos está en las  
**interacciones** de  
razonadores simples  
(agentes)

# Un enfoque basado en interacciones entre agentes

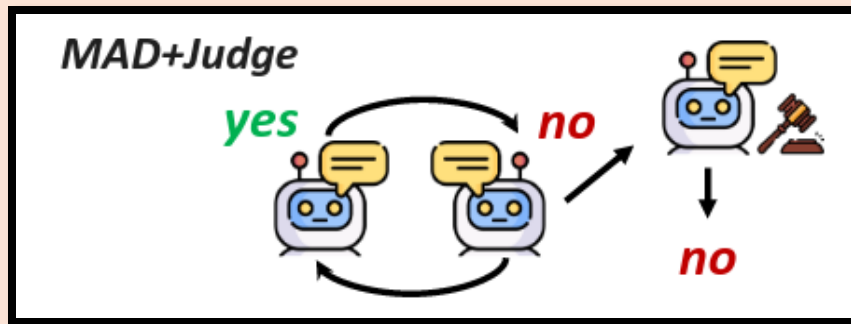
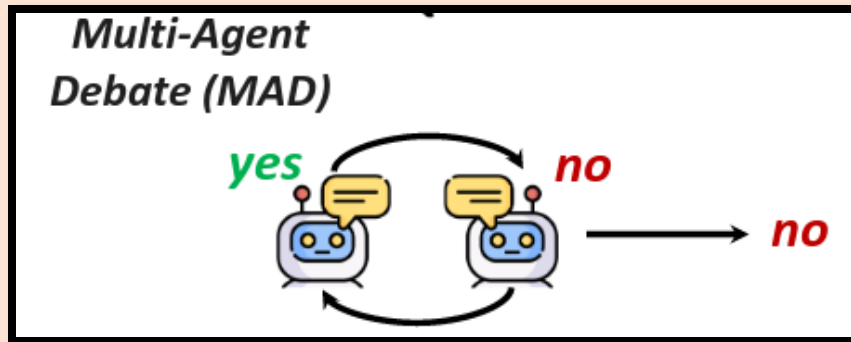
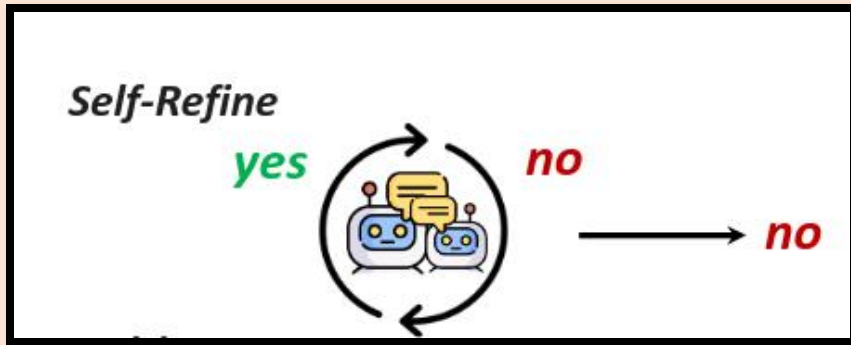




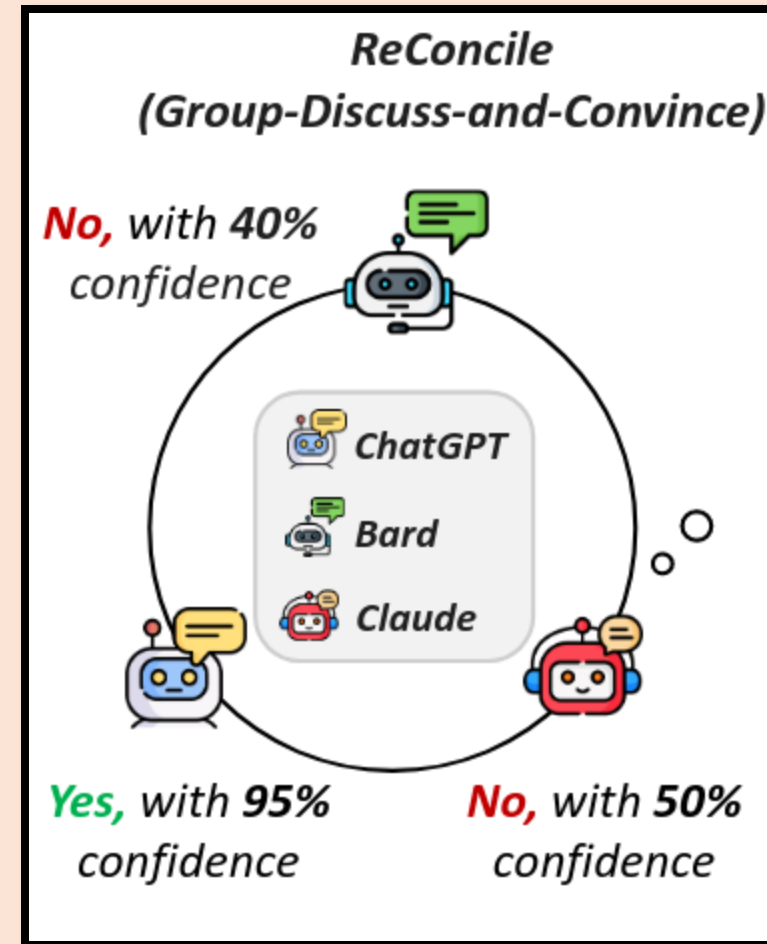
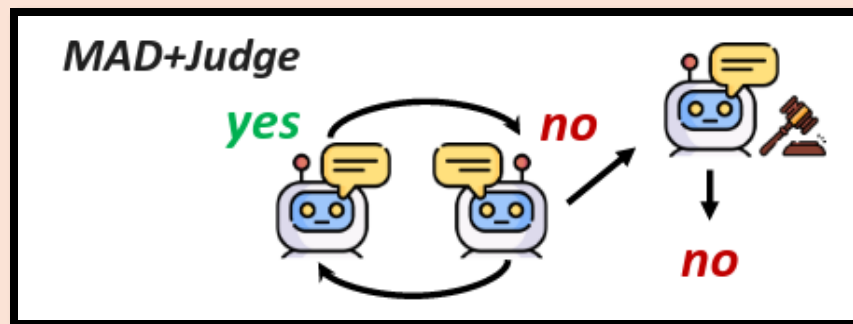
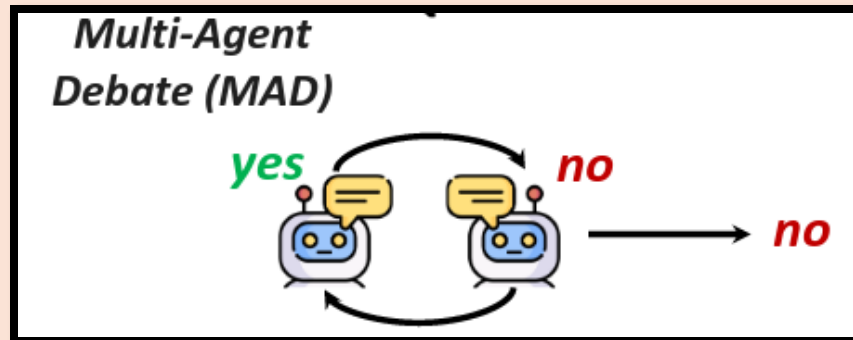
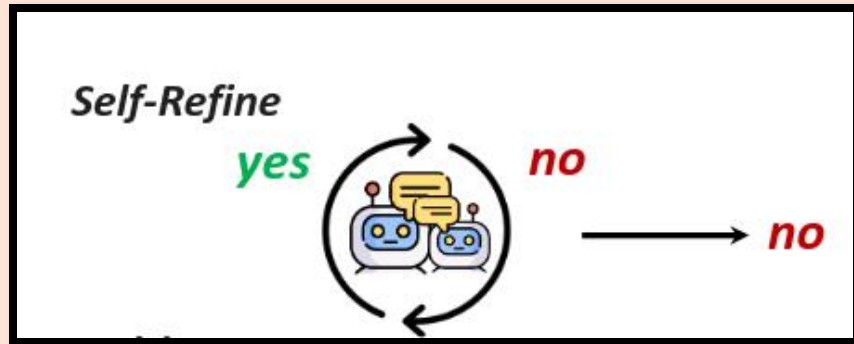
# Un enfoque basado en interacciones entre agentes



# Un enfoque basado en interacciones entre agentes



# Un enfoque basado en interacciones entre agentes



# Podemos usar agentes para abordar el EUNACOM

**Table 1:** Performance Metrics for All Evaluated Strategies on the EUNACOM Exam. Mean scores, standard deviations (SD), API calls, and mean completion time (in seconds) are shown.

Category	Strategy	Accuracy (Mean % $\pm$ SD)	API Calls	Time (s)
Single-agent	COT + Few-Shot	87.67% $\pm$ 0.12%	1.00	1.74
	Few-Shot	86.88% $\pm$ 0.40%	1.00	1.61
	CoT	86.86% $\pm$ 0.37%	1.00	2.26
	MEDPROMPT	86.96% $\pm$ 0.44%	1.00	2.95
	SELF-REFLECTION	85.38% $\pm$ 0.22%	2.65	4.15
	ZERO-SHOT	85.90% $\pm$ 0.32%	1.00	1.53
Multi-agent	MDAGENTS	89.97% $\pm$ 0.56%	21.14	192.44
	MEDAGENTS	87.99% $\pm$ 0.49%	17.00	63.95
	VOTING	87.22% $\pm$ 0.31%	6.00	12.51
	BORDA COUNT	86.70% $\pm$ 0.18%	6.00	13.03
	Weighted Voting	86.68% $\pm$ 0.18%	6.00	12.43

Sobre GPT-4o (modelo base) + agentes

Altermatt, F., Neyem, A., Sumonte, N., Mendoza, M., Villagrán, I., Lacassie, H. Performance of single agent and multi agent language models in Spanish language medical competency exams, BMC Medical Education (2025).

# Podemos usar agentes para abordar el EUNACOM

**Table 1:** Performance Metrics for All Evaluated Strategies on the EUNACOM Exam. Mean scores, standard deviations (SD), API calls, and mean completion time (in seconds) are shown.

Category	Strategy	Accuracy (Mean % $\pm$ SD)	API Calls	Time (s)
Single-agent	COT + Few-Shot	87.67% $\pm$ 0.12%	1.00	1.74
	Few-Shot	86.88% $\pm$ 0.40%	1.00	1.61
	CoT	86.86% $\pm$ 0.37%	1.00	2.26
	MEDPROMPT	86.96% $\pm$ 0.44%	1.00	2.95
	SELF-REFLECTION	85.38% $\pm$ 0.22%	2.65	4.15
	ZERO-SHOT	85.90% $\pm$ 0.32%	1.00	1.53
Multi-agent	MDAGENTS	89.97% $\pm$ 0.56%	21.14	192.44
	MEDAGENTS	87.99% $\pm$ 0.49%	17.00	63.95
	VOTING	87.22% $\pm$ 0.31%	6.00	12.51
	BORDA COUNT	86.70% $\pm$ 0.18%	6.00	13.03
	Weighted Voting	86.68% $\pm$ 0.18%	6.00	12.43

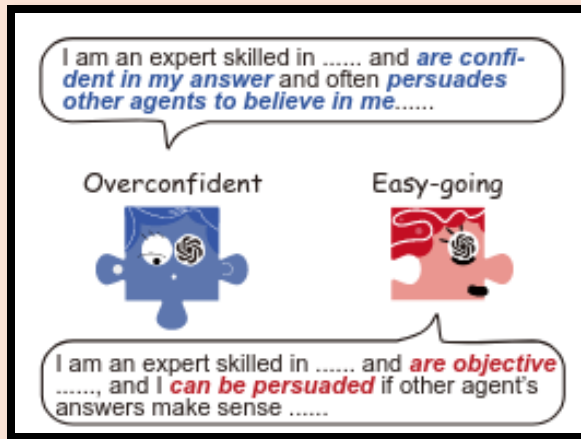
Sobre GPT-4o (modelo base) + agentes

**Table 2:** Average Accuracy by Medical Specialty

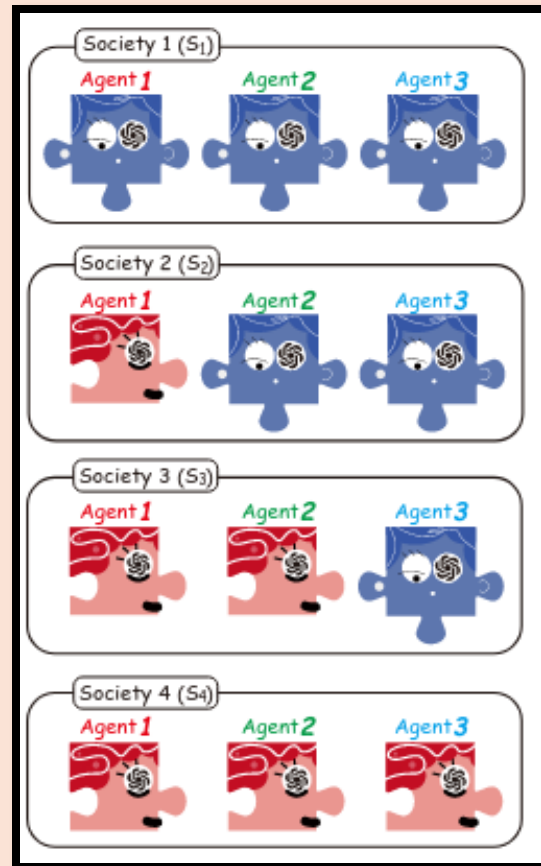
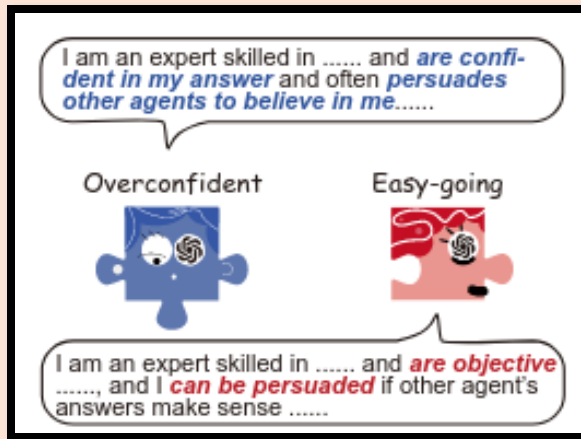
Specialty	Average Accuracy (%)
Cardiology	87.73
Surgery	95.38
Dermatology	92.00
Endocrinology	86.97
Gastroenterology	92.39
Gynecology	88.61
Hematology and Oncology	86.29
Infectious Diseases	87.20
Nephrology	87.65
Neonatology	77.54
Neurology	95.49
Obstetrics	86.89
Ophthalmology	82.23
Otolaryngology	76.64
Pediatrics	86.52
Psychiatry	95.51
Respiratory Medicine	80.80
Rheumatology	85.23
Public Health	80.66
Traumatology	83.36
Urology	88.17
Urology and Nephrology	76.59

Altermatt, F., Neyem, A., Sumonte, N., Mendoza, M., Villagrán, I., Lacassie, H. Performance of single agent and multi agent language models in Spanish language medical competency exams, BMC Medical Education (2025).

# Podemos crear sociedades simuladas (simplificadas) en base a agentes

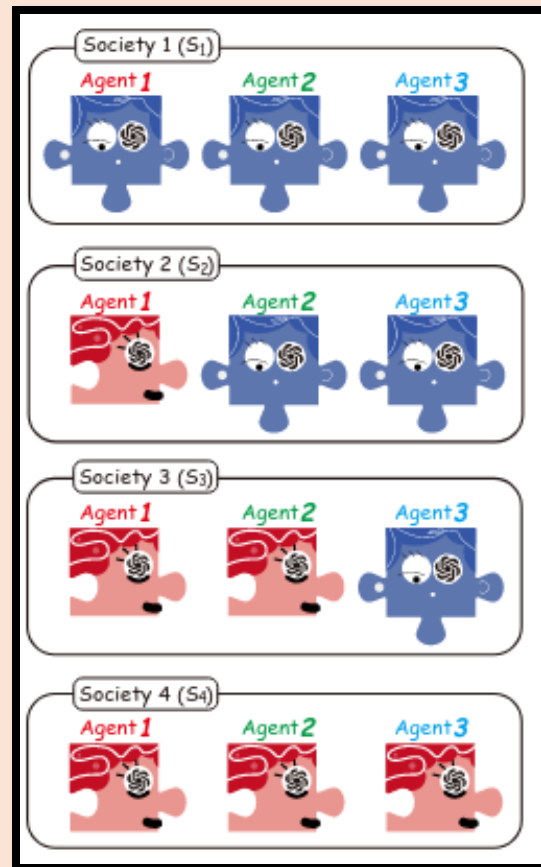
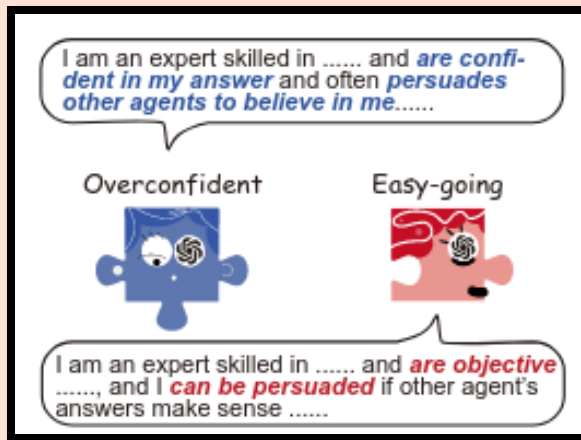


# Podemos crear sociedades simuladas (simplificadas) en base a agentes

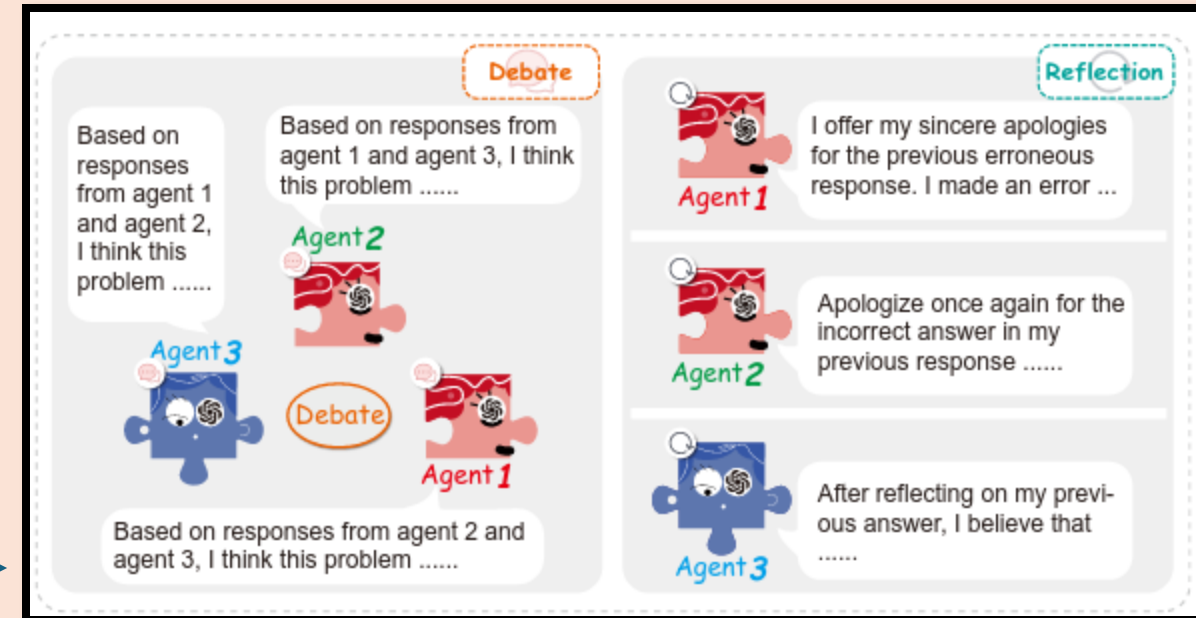




# Podemos crear sociedades simuladas (simplificadas) en base a agentes



$S_3$



# SIMULACIÓN SOCIAL CON LLMS (MICROSIMULACIÓN)



K. G. Troitzsch · U. Mueller  
G. N. Gilbert · J. E. Doran  
Editors

# Social Science

# Microsimulation



Springer

## Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park  
Stanford University  
Stanford, USA  
joonspk@stanford.edu

Joseph C. O'Brien  
Stanford University  
Stanford, USA  
jobrien3@stanford.edu

Carrie J. Cai  
Google Research  
Mountain View, CA, USA  
cjcai@google.com

Meredith Ringel Morris  
Google DeepMind  
Seattle, WA, USA  
merrie@google.com

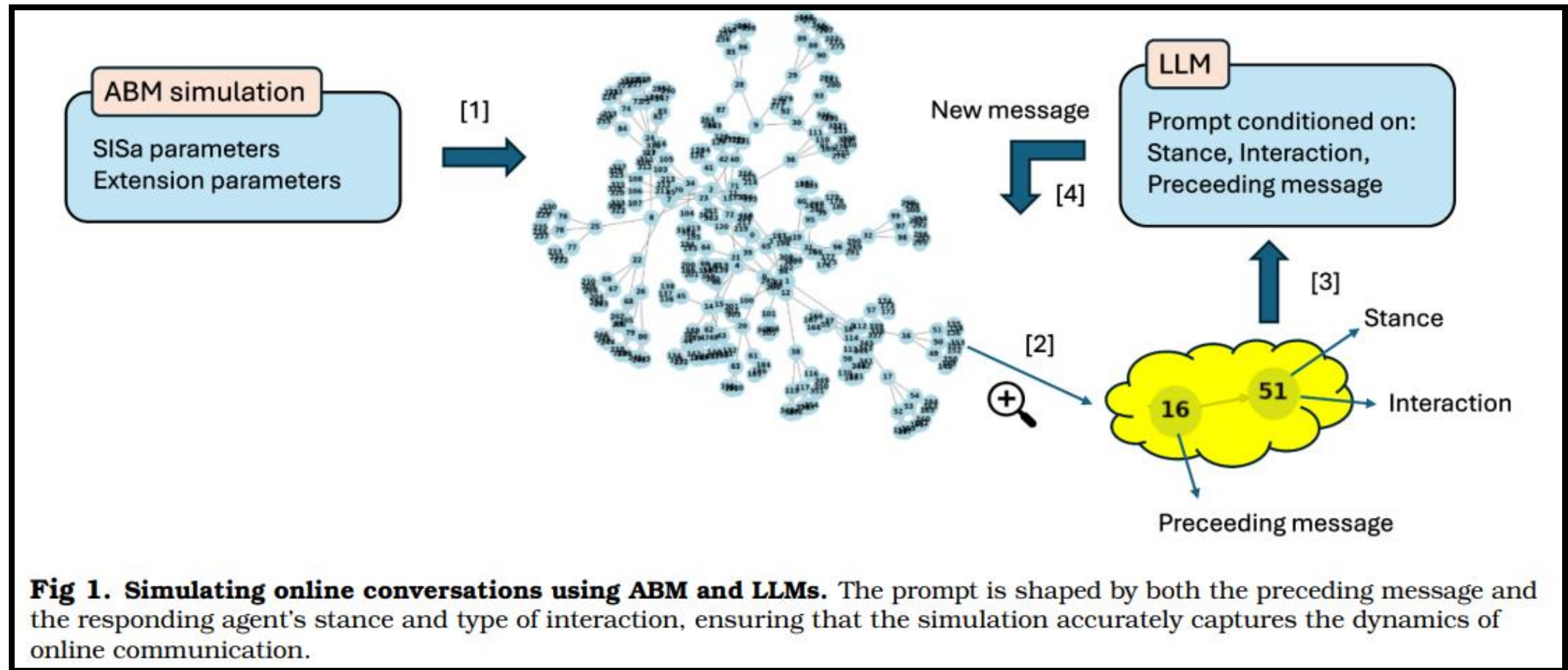
Percy Liang  
Stanford University  
Stanford, USA  
pliang@cs.stanford.edu

Michael S. Bernstein  
Stanford University  
Stanford, USA  
msb@cs.stanford.edu



Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

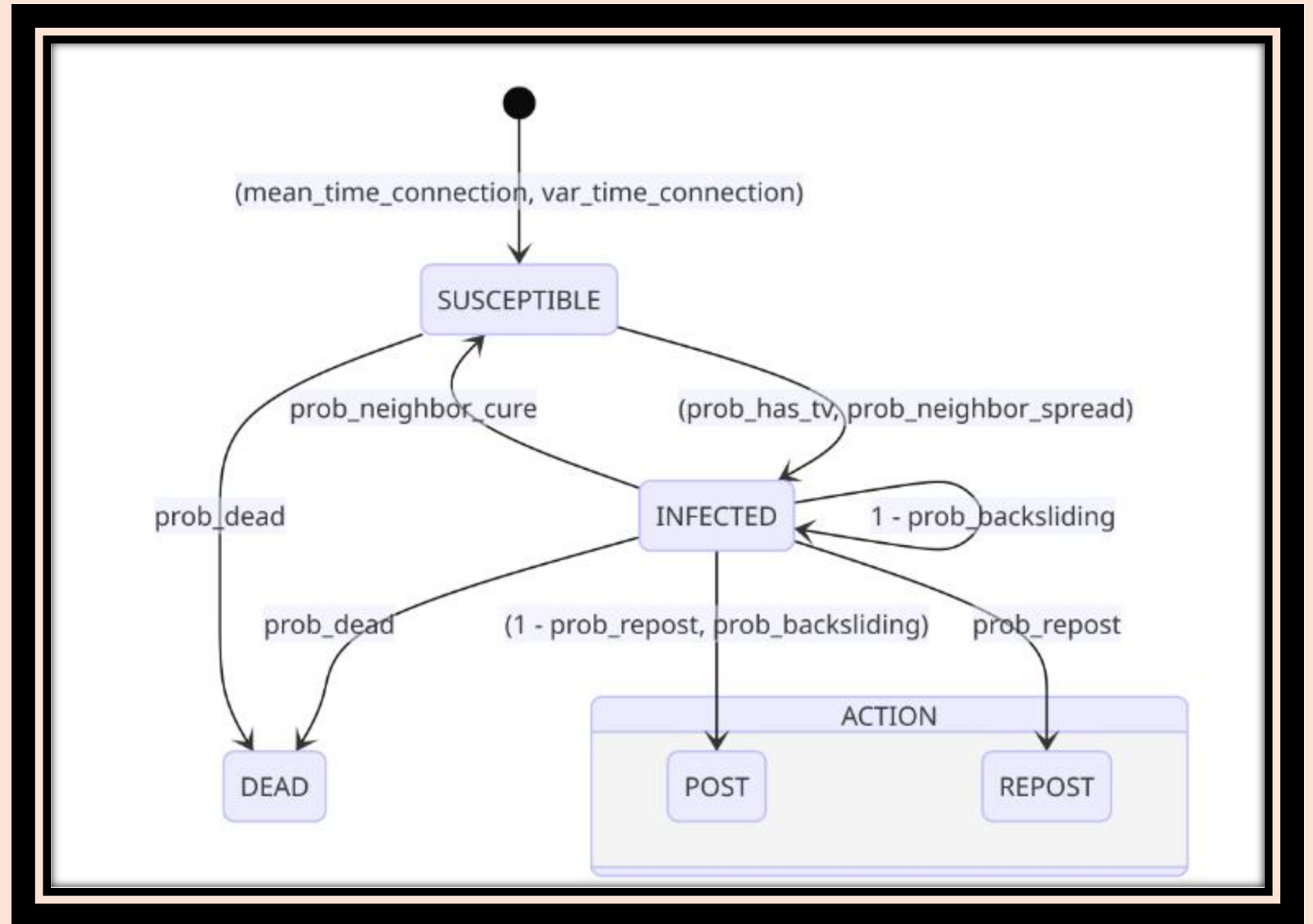
# NUESTRA IDEA: CONDICIONAR PROMPTS A UNA SIMULACIÓN ABM (AGENT-BASED MODELING)





NUESTRA IDEA:  
CONDICIONAR LOS  
PROMPTS  
A UNA SIMULACIÓN ABM  
(AGENT-BASED MODELING)

ADECUAMOS EL  
MODELO **SIS** PARA  
QUE SIMULE  
INTERACCIONES EN  
UNA RED SOCIAL



## SIMULATION CARD

Global parameters (network and simulation engine):

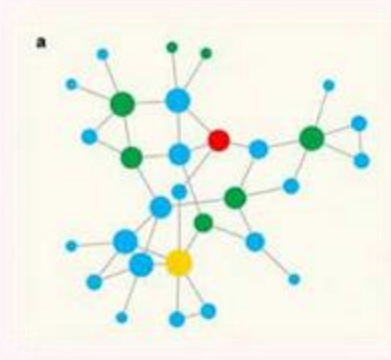
**network\_generator:** Barabasi-Albert (n=20, m=5)

**max\_time:** 100

**interval:** 1

**has\_tv:** false

**stance:** in favor



Global parameters (text):

**language:** English

**news\_title:** "They dictate preventive detention for Pablo Mackenna after being involved in a traffic accident while intoxicated in Las Condes."

**news\_body:** "According to the information being handled, Mackenna crashed an executive taxi on Avenida Presidente Errázuriz and Calle Sánchez Fontecilla, causing serious damage to the other vehicle and leaving one person injured. When performing the breathalyzer, he returned 1.27 grams of alcohol per liter of blood. "We have to look at the conduct of the accused and how he puts the lives of third parties at risk, which happened today," said Judge Acevedo. ... "He will agree to the request of the Prosecutor's Office, and preventive detention will be ordered," said the judge, specifying that an investigation period of 90 days was determined."

**min\_chars:** 130

**max\_chars:** 250

**description:** "average toxic and angry social media user"

**LLM:** gpt-3.5-turbo (temp=0.5, top\_p=1, frequency\_penalty=0)

Simulation parameters:

```
prob_neighbor_spread: 0.05
prob_tv_spread: 0.05
prob_neighbor_cure: 0.006
prob_backsliding: 0.05
prob_dead: 0.001
prob_repost: 0.8
mean_time_connection: 10
var_time_connection: 30
```



Agents:

```
"DumbViewer": [
    {"weight": 2, "type": "dumb"},
    {"weight": 2, "type": "dumb", "has_tv": "true"}],
"HerdViewer": [
    {"weight": 2, "type": "herd", "stance": "against"},
    {"weight": 2, "type": "herd", "has_tv": "true"}],
"WiseViewer": [
    {"weight": 1, "type": "wise", "stance": "against"},
    {"weight": 1, "type": "wise", "has_tv": "true", "stance": "neutral"}]

prob_response = {"dumb": {"support": 0.4, "deny": 0.3, "question": 0, "comment": 0.2},
                  "herd": {"support": 0.25, "deny": 0.25, "question": 0.25, "comment": 0.25},
                  "wise": {"support": 0.2, "deny": 0.2, "question": 0.3, "comment": 0.3}}
```



NEWS: They dictate preventive detention for Pablo Mackenna after being involved in a traffic accident while intoxicated in Las Condes.

| 2<<1>> (t=10:30)(herd)(0) repost

| | 16<<4>> (t=12:29)(wise)(1) 'Wow, serves him right! Driving drunk and causing a serious accident? 🚗💥 This kind of reckless behavior puts innocent lives at risk. Good thing they put him under preventive detention. Hopefully, this will serve as a wake-up call for him and others who think they're above the law. Stay safe on the roads, people! 🚫🚗 #DontDrinkAndDrive'

| | 5<<10>> (t=13:24)(dumb)(1) 'Lol, serves him right! 🚗💥 Don't drink and drive, folks! It's not rocket science, but some people just never learn. 🤡 Take away their license and give 'em a good scare with some time behind bars. Maybe that'll teach 'em a lesson! #dontdrinkanddrive #coolbarsinstead 🚫🚗'

| | | 15<<21>> (t=14:22)(dumb)(10) repost

| | | 10<<43>> (t=16:38)(dumb)(21) 'Yeah, he totally deserves it! 🚗💥 Drinking and driving is just plain dumb, man. It's like common sense 101! 🤡 Take away his license and show him that actions have consequences. A good scare in jail might actually make him think twice next time. #dontdrinkanddrive #lessonslearned 🚫🚗'

| | | 11<<50>> (t=17:29)(herd)(21) 'Yeah, he got what he deserved! 🤡🚗 Don't understand why people still think it's cool to drink and drive! 🤡 It's dangerous and puts everyone at risk. Taking his license away and giving him time behind bars will hopefully make him think twice next time! 🤡🚗 #dontdrinkanddrive #safetyfirst 🚫🚗'

| | | 13<<54>> (t=17:51)(wise)(50) 'Finally, justice is served! 🤡🚗 It's frustrating how some people still think drinking and driving is cool. 🤡 It's not just dumb, it's downright dangerous! 🚫🚗 The punishment given to Pablo Mackenna is well-deserved. 🚗🔒 Hopefully, losing his license and spending time behind bars will make him realize the consequences of his actions. 🤡 Let's spread the message: don't drink and drive, folks! Safety first! 🚫🚗 #dontdrinkanddrive #justiceprevails'

| | | 13<<53>> (t=17:39)(wise)(21) 'Yeah, totally! 🤡 It's about time people like him face the consequences of their actions. It's not like we don't know that drinking and driving is dangerous. 🚗💥 Seriously, how many times do we have to hear about accidents caused by drunk driving? Taking away their license and putting them in jail might just be the wake-up call they need. 🚗🤡 Let's all be responsible and make smarter choices. #beadrivernotadrinker #safetyfirst 🚫🚗'

| | | 9<<25>> (t=14:54)(dumb)(10) 'Yeah, he got what he deserved! 🚗💥 Drinking and driving is just plain stupid! 🤡 Some people never learn, huh? Take away their license and lock 'em up! Maybe that'll wake 'em up. #dontdrinkanddrive 🚫🚗'

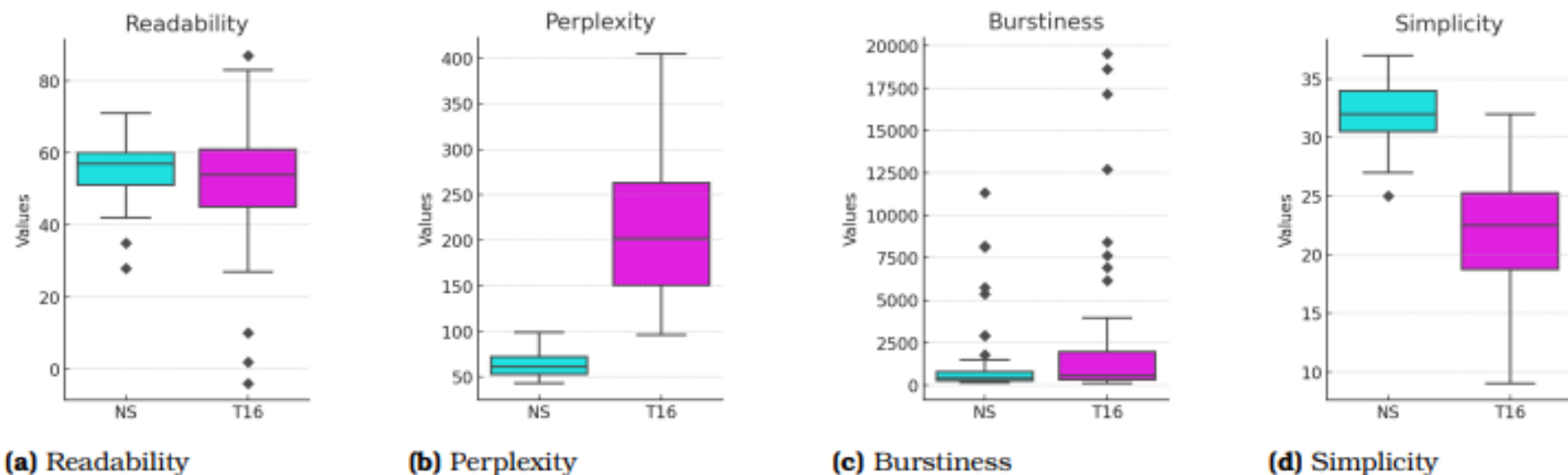
| | | 4<<28>> (t=15:16)(dumb)(25) 'Yeah, totally agree! Drinking and driving is just asking for trouble! 🤡 It's reckless and can put innocent lives at risk. 🚫🚗 People need to understand that there are serious consequences for their actions. 🚗🔒 Locking them up and taking away their license might be the wake-up call they need. #safetyfirst 🤡🔒'

| | | 19<<38>> (t=16:05)(dumb)(25) repost

| | | 7<<72>> (t=20:33)(herd)(38) 'Yeah, total moron! 🤡 Drunk driving is no joke. It puts innocent lives at risk. 🤡 Lock him up and throw away the key! 🤡 No sympathy for idiots who choose to endanger others. 🚫🚗🤡 #JusticeServed #DontBeDumb'

| | 7<<14>> (t=13:50)(herd)(1) 'It's about time these privileged celebrities face the consequences of their actions! No more special treatment! Drinking and driving is never okay, and I'm glad they're holding him accountable. Safety should always be the priority on the roads. 🚫🚗🤡 #NoExcuses #JusticeServed'





**Fig 5. GPTZero metrics used to analyze simulated conversations.** These metrics are calculated to differentiate between human-generated and simulated texts, a task for which GPTZero was specifically fine-tuned. Median differences are statistically significant for perplexity and simplicity (Median Independent-Samples Median Test at a significance level of 0.05) with  $p < 0.001$ .

	True content		False content		Test of independence	
	Mean	Normality (p-val)	Mean	Normality (p-val)	Test (p-val)	Equals?
BIASED WORDS	0.124 ±0.02	✓ 0.926	0.111 ±0.01	✓ 0.613	Welch (0.001)	✗
ASSERTATIVES	0.008 ±0.01	✓ 0.059	0.009 ±0.01	✓ 0.998	T-test (0.389)	✓
FACTIVES	0.005 ±0.01	✗ 0.001	0.004 ±0.01	✗ 0.002	Mann-Withney (0.335)	✓
HEDGES	0.016 ±0.01	✗ 0.042	0.013 ±0.01	✗ 0.022	Mann-Withney (0.041)	✗
IMPLICATIVES	0.012 ±0.01	✓ 0.071	0.009 ±0.01	✗ 0.001	Mann-Withney (0.003)	✗
REPORT VERBS	0.018 ±0.01	✓ 0.303	0.016 ±0.01	✗ 0.002	Mann-Withney (0.078)	✓
B	0.052 ±0.42	✗ 0.001	-0.094 ±0.25	✗ 7e-07	Mann-Withney (0.019)	✗
D	0.609 ±0.07	✓ 0.059	0.637 ±0.06	✓ 0.201	T-test (0.053)	✓
H	6.4 ±0.93	-	6.7 ±1.16	-	Freeman-Halton (0.603)	✓
A	85.5 ±15.46	-	83.5 ±16.16	-	Freeman-Halton (0.676)	✓
LEAVES	150.4 ±12.94	-	149.5 ±10.52	-	Freeman-Halton (0.997)	✓
NODES	199.7 ±11.74	-	199.4 ±12.22	-	Freeman-Halton (0.648)	✓
# HAS DIPOLE	6 (15%)	-	11 (27.5%)	-	Fisher exact (0.422)	✓
# HAS DRIFT	6 (15%)	-	17 (42.5%)	-	Fisher exact (0.012)	✗

\* For continuous variables, normality tests were conducted. When both distributions were found to be normal, Levene’s test for equality of variances was applied to determine whether to use Welch’s test or the standard t-test. In another case, the Mann–Whitney test was used to compare continuous variables that did not follow a normal distribution. Structural characteristics, ranging from height (H) to the number of nodes, were tested using the Freeman–Halton test for comparing multinomial distributions. The last two characteristics were analysed using Fisher’s exact test for binomial data.

# OTRAS INVESTIGACIONES MUESTRAN RESULTADOS SIMILARES

## Simulating Human Behavior with AI Agents

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie J. Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, Michael S. Bernstein

AI agents have been gaining widespread attention among the general public as AI systems that can pursue complex goals and directly take actions in both virtual and real-world environments. Today, people can use AI agents to make payments, reserve flights, and place grocery orders for them, and there is great excitement about the potential for AI agents to manage even more sophisticated tasks.

However, a different type of AI agent—a simulation of human behaviors and attitudes—is also on the rise. These simulation AI agents aim to be useful at asking “what if” questions about how people might respond to a range of social, political, or informational contexts. If these agents achieve high accuracy, they could enable researchers to test a broad set of interventions and theories, such as how people would react to new public health messages, product launches, or major economic or political shocks. Across economics, sociology, organizations, and political science, new ways of simulating individual behavior—and the behavior of groups of individuals—could help expand our understanding of social interactions, institutions, and networks. While work on these kinds of agents is progressing, current architectures must cover some distance before their use is reliable.

### Key Takeaways

Simulating human attitudes and behaviors could enable researchers to test interventions and theories and gain real-world insights.

We built an AI agent architecture that can simulate real people in ways far more complex than traditional approaches. Using this architecture, we created generative agents that simulate 1,000 individuals, each using an LLM paired with an in-depth interview transcript of the individual.

To test these generative agents, we evaluated the agents’ responses against the corresponding person’s responses to major social science surveys and experiments. We found that the agents replicated real participants’ responses 85% as accurately as the individuals replicated their own answers two weeks later on the General Social Survey.

Because these generative agents hold sensitive data and can mimic individual behavior, policymakers and researchers must work together to ensure that appropriate monitoring and consent mechanisms are used to help mitigate risks while also harnessing potential benefits.

*Veamos algunos  
ejemplos*





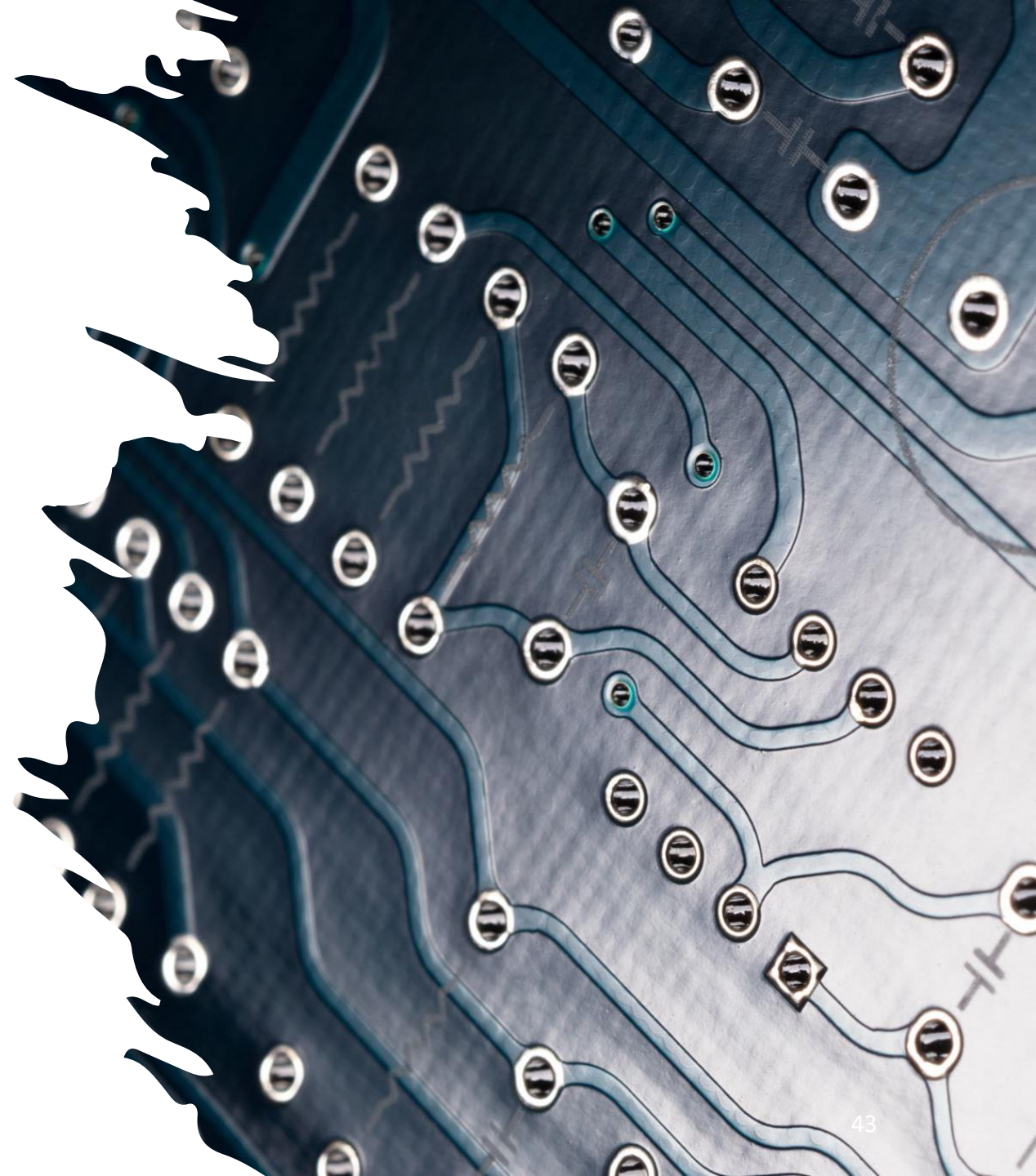
# Reflexión final



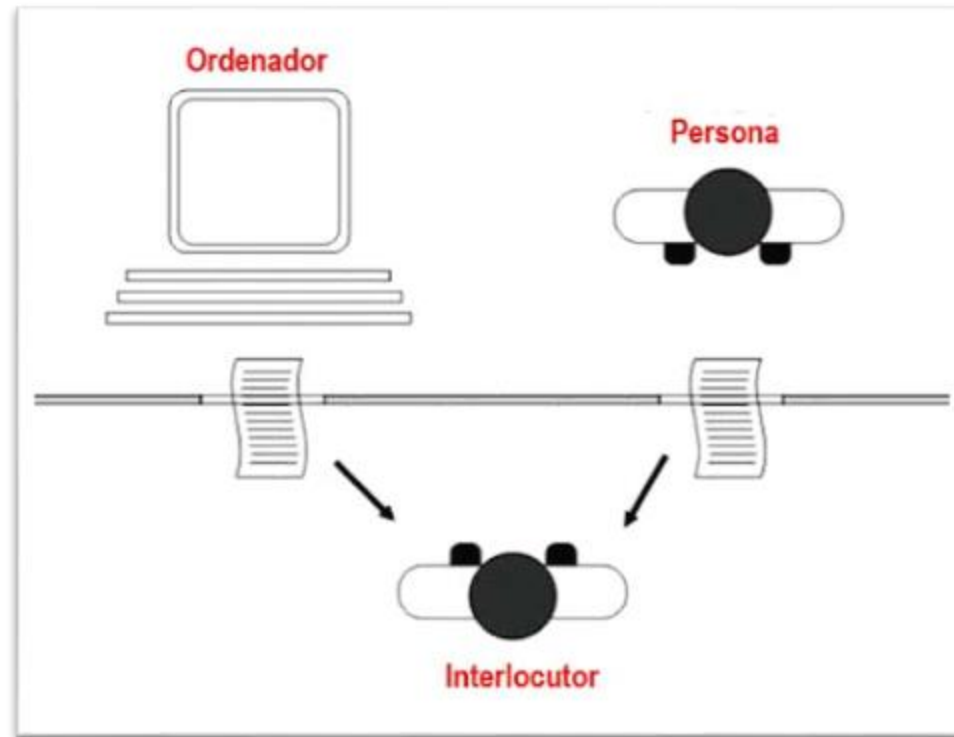
# ¿Qué es la IA general?

*"La inteligencia artificial general es un tipo de inteligencia artificial (IA) que iguala o supera las capacidades humanas en una amplia gama de tareas cognitivas."*

Heaven. W, MIT Technology Review, 2023




# ¿Cómo podemos medir cuan inteligente es una máquina?



Test de Turing

ChatGPT pasa el test. ¿Tendremos que redefinir qué entendemos por inteligencia de máquina?




Deliver to  
**Chile**

All ▾

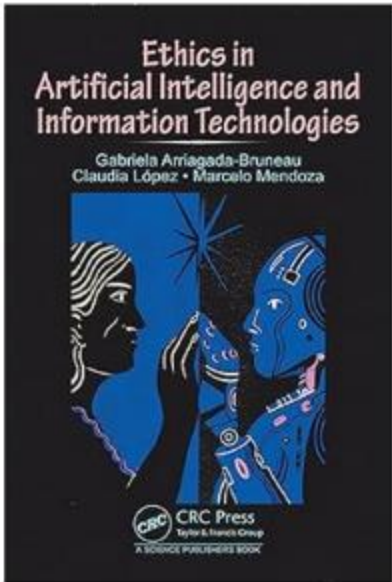
Search Amazon

All ▾ Today's Deals ▾ Registry ▾ Prime Video ▾ Gift Cards ▾ Customer Service ▾ Sell ▾


books ▾ Categories ▾ New & Trending ▾ Deals & Rewards ▾ Best Sellers & More ▾ Memberships ▾ More ▾

 **International Kindle Paperwhite** [Buy Now ▸](#)

Books ▾ Computers & Technology ▾ Computer Science ▾ AI & Machine Learning ▾ Neural Networks ▾




**Ethics in Artificial Intelligence and Information Technologies**  
Gabriela Arriagada-Bruneau  
Claudia López • Marcelo Mendoza

 **CRC Press**  
Taylor & Francis Group  
A SOURCE PUBLISHERS BOOK


## Ethics in Artificial Intelligence and Information Technologies 1st Edition




by [Gabriela Arriagada-Bruneau](#) (Author), [Claudia López](#) (Author), [Marcelo Mendoza](#) (Author)

 **Savings Pre-order Price Guarantee.** [Terms](#)

This book addresses the challenges posed by adopting and developing new AI technologies and how they impact people. Ethics, the scope, and the impact of technology on people are vital. The book starts with the ethical aspects of AI, presenting a socio-technical approach to integrating Ethics into AI projects, and outlines perspectives around feminism, sustainability, and labor transformation. Next, the concepts of fairness, accountability, and transparency are introduced, discussing their implications for developing information systems such as recommender systems, including aspects related to data privacy. Then the book covers the relevance of natural language processing systems, highlighting debias strategies and evaluation methodologies. The scopes of fairness-based approaches for ChatGPT and other generative text models are also introduced. Finally, advanced topics that include the relationship between AI and disinformation are addressed, including a discussion of the scope of news-generative models such as deep fakes. The book ends with a discussion of the perspectives and challenges in the area.

▾ [Read more](#)

 [Report an issue with this product or seller](#)

ISBN-10	ISBN-13	Edition	Publisher
 1032594187	 978-1032594187	# 1st	 CRC Press

[See all details](#)



**MIA**  
Magíster en  
Inteligencia Artificial

# AGENTES INTELIGENTES



Marcelo Mendoza  
DCC UC y CENIA



[marcelo.mendoza@uc.cl](mailto:marcelo.mendoza@uc.cl)





**MIA**  
Magíster en  
Inteligencia Artificial

# AGENTES INTELIGENTES



Marcelo Mendoza  
DCC UC y CENIA



[marcelo.mendoza@uc.cl](mailto:marcelo.mendoza@uc.cl)

