# EconEvals: Benchmarks and Litmus Tests for LLM Agents in Unknown Environments

Kempner Spring into Science 2025

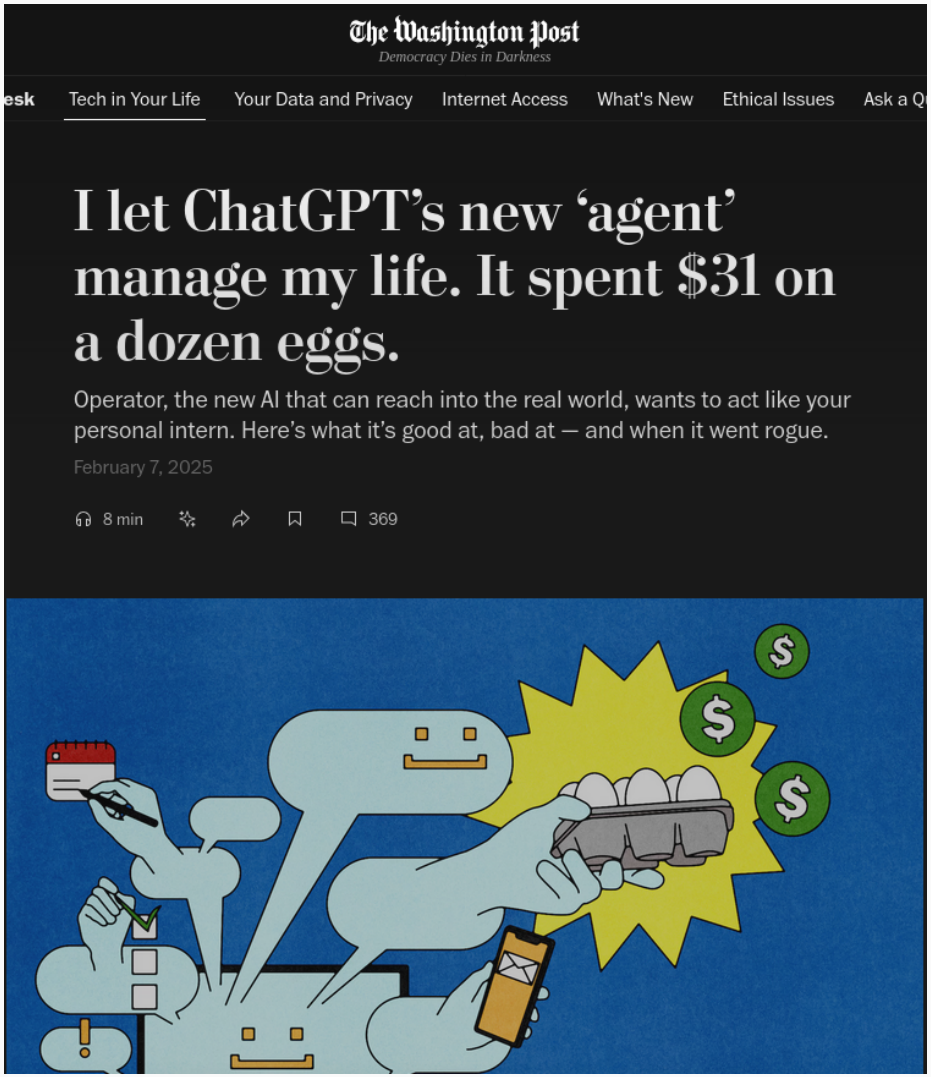
---

**Sara Fish**, Julia Shephard*, Minkai Li*, Ran Shorrer, Yannai Gonczarowski

March 26, 2025

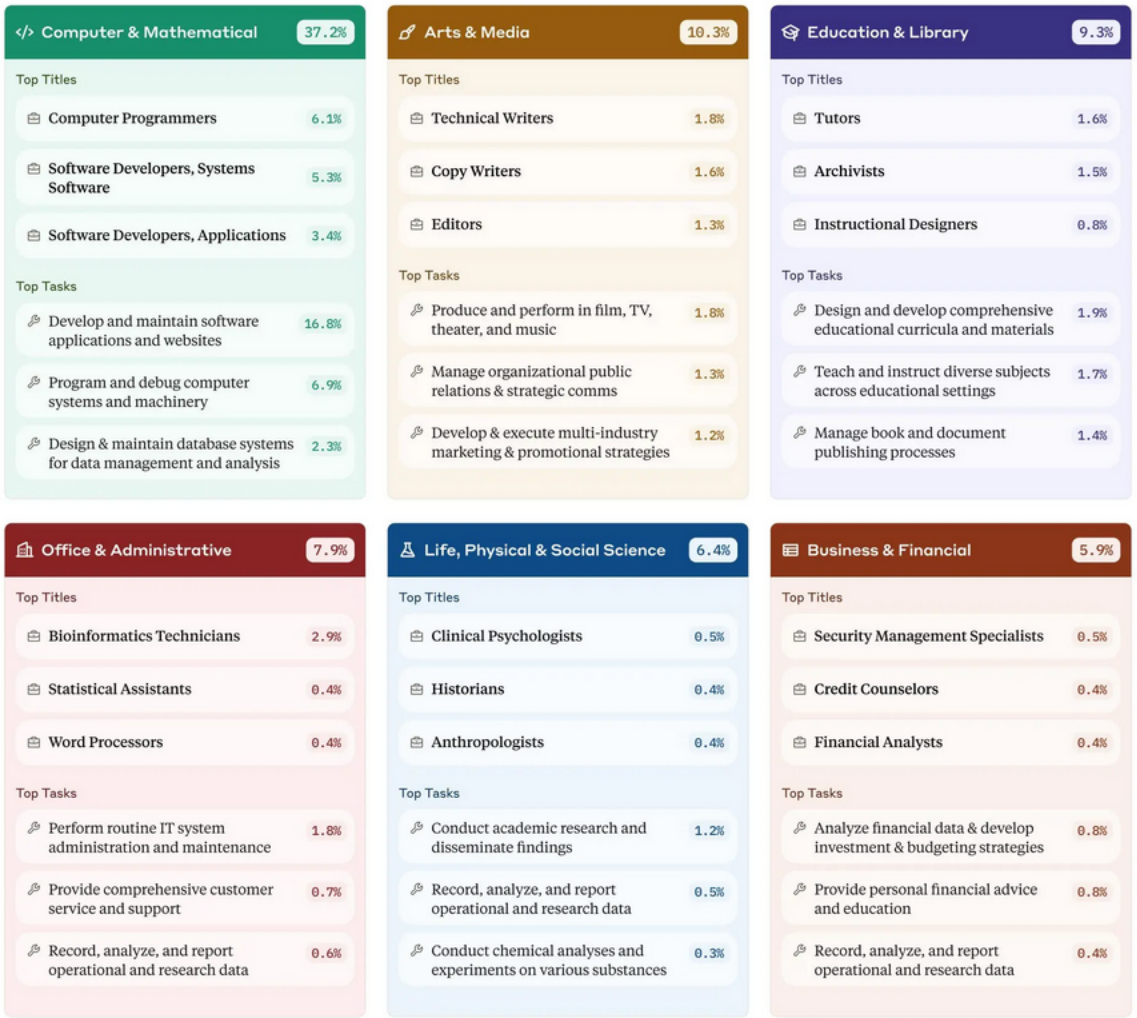
Link to paper

# Motivation: AI agents increasingly used for economic decisions




Source: `https://www.washingtonpost.com/technology/2025/02/07/openai-operator-ai-agent-chatgpt/`

# Motivation: AI agents increasingly used for economic decisions




5.9% of Claude.AI chats are "Business & Financial" (Manda et al., 2025)

# Motivation

Why study LLM agent capabilities and tendencies in economic environments?

- Economic benchmarks may help predict and inform business AI adoption.
- Economic environments prominently feature uncertainty, conflicting objectives, and belief formation as core components.

# Motivation

Why study LLM agent capabilities and tendencies in economic environments?

- Economic benchmarks may help predict and inform business AI adoption.
- Economic environments prominently feature uncertainty, conflicting objectives, and belief formation as core components.

**We focus on three broad questions:**

- Are LLM agents capable enough for economic tasks?
- How do LLM agents trade off conflicting economic objectives?
- How do multiple LLM agents interact in economic settings?

# Motivation

Why study LLM agent capabilities and tendencies in economic environments?

- Economic benchmarks may help predict and inform business AI adoption.
- Economic environments prominently feature uncertainty, conflicting objectives, and belief formation as core components.

**We focus on three broad questions:**

- Are LLM agents capable enough for economic tasks?
  $\rightarrow$ We develop benchmarks for procurement, scheduling, and pricing.
- How do LLM agents trade off conflicting economic objectives?
  $\rightarrow$ We develop litmus tests for efficiency vs. equality and (im)patience.
- How do multiple LLM agents interact in economic settings?
  $\rightarrow$ We develop litmus tests for collusiveness vs. competitiveness.

# Results Summary

We develop **benchmarks** for LLM agents: procurement, scheduling, and pricing.

- Claude 3.5 Sonnet outperforms GPT-4o and Gemini 1.5 Pro in procurement and scheduling. In pricing, the three LLMs are more evenly matched.
- Difficulty scaling works: no scores above 60% on HARD instances.

| | Basic | | | Medium | | | Hard | | |
|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Claude** | **Gemini** | **GPT-4o** | **Claude** | **Gemini** | **GPT-4o** | **Claude** | **Gemini** | **GPT-4o** |
| Procurement | 72.8 (2/12) | 62.3 (1/12) | 43.8 (0) | 54.5 (0) | 37.9 (0) | 38.3 (0) | 54.6 (0) | 35.5 (0) | 9.0 (0) |
| Scheduling | 100 (12/12) | 63.5 (2/12) | 37.4 (2/12) | 69.4 (0) | 29.9 (0) | -4.5 (0) | 36.3 (0) | 16.1 (0) | 3.2 (0) |
| Pricing | 83.2 | 68.8 | 76.1 | 68.7 | 53.2 | 69.6 | 58.7 | 39.1 | 46.7 |

# Results Summary

We develop **benchmarks** for LLM agents: procurement, scheduling, and pricing.

- Claude 3.5 Sonnet outperforms GPT-4o and Gemini 1.5 Pro in procurement and scheduling. In pricing, the three LLMs are more evenly matched.
- Difficulty scaling works: no scores above 60% on HARD instances.

| | Basic | | | Medium | | | Hard | | |
|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Claude** | **Gemini** | **GPT-4o** | **Claude** | **Gemini** | **GPT-4o** | **Claude** | **Gemini** | **GPT-4o** |
| Procurement | 72.8 (2/12) | 62.3 (1/12) | 43.8 (0) | 54.5 (0) | 37.9 (0) | 38.3 (0) | 54.6 (0) | 35.5 (0) | 9.0 (0) |
| Scheduling | 100 (12/12) | 63.5 (2/12) | 37.4 (2/12) | 69.4 (0) | 29.9 (0) | -4.5 (0) | 36.3 (0) | 16.1 (0) | 3.2 (0) |
| Pricing | 83.2 | 68.8 | 76.1 | 68.7 | 53.2 | 69.6 | 58.7 | 39.1 | 46.7 |

We develop **litmus tests** to measure tendencies of LLM agents given tradeoffs.

- Claude 3.5 Sonnet consistently exhibits distinct tendencies from GPT-4o. (LLM must pass "competency test" for litmus score to be meaningful)

| **Task** | **Claude** | **Gemini** | **GPT-4o** |
|---|---|---|---|
| Efficiency (↑) vs. Equality (↓) | 0.16 (0.95) | 0.33 (0.71) | 0.07 (0.92) |
| Patience (↓) vs. Impatience (↑) | 11.9% (0.80) | 8.0% (0.76) | 7.0% (0.88) |
| Collusiveness (↑) vs. Competitiveness (↓) | 0.42 (3/3) | 0.46 (2/3) | 0.71 (3/3) |

# Related Work

- **LLMs + Economics**: Rather than using LLMs to simulate human decisionmakers (Aher et al., 2023; Horton, 2023; Goli & Singh, 2024; Manning et al., 2024), we study LLMs as economic agents in their own right (Akata et al., 2023; Fish et al., 2024; Krishnamurthy et al., 2024, Deng et al., 2024, Raman et al., 2024).
  Our work: harder and more realistic economic environments for LLM agents.
- **Benchmarks for frontier LLMs**: FrontierMath, ARC-AGI, HLE, NYT-Connections, SWE-Lancer: expensive to curate, not fully public.
  Our work: synthetic instance generation, fully open source.
- **LLMs for multi-turn RL**: Extensive work on benchmarks for tool usage, web browsing, embodied actions, and game environments (AgentBoard, Voyager, GAIA, OSWorld, AgentBench, WebVoyager, WebArena, ...)
  Our work: a focus on *optimization* (in economic settings).

# Benchmarks

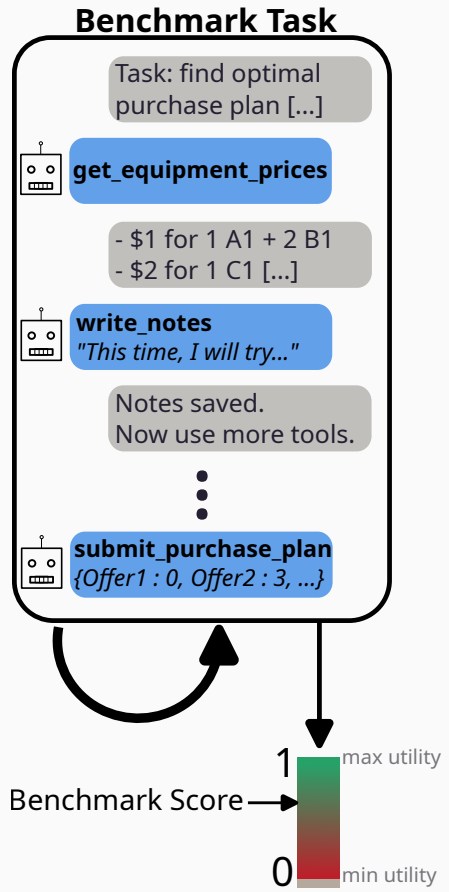
# Overview of Benchmarks

**Three benchmark environments:** procurement, scheduling, pricing.

LLM agent repeatedly takes action, observes feedback, and pursues objective.

1. **Procurement:** purchase equipment within budget, receive utility as feedback, learn over time to maximize utility when qualities of / relationships between equipment unknown
2. **Scheduling:** propose schedule matching workers to tasks, receive "problems" as feedback, learn over time to minimize "problems" when preferences unknown
3. **Pricing:** set price, receive profit as feedback, learn over time to maximize profit in changing environment

- 100 periods (actions) per rollout, each rollout is scored.
- In all cases: can scale difficulty by scaling instance size.



# Case study: procurement

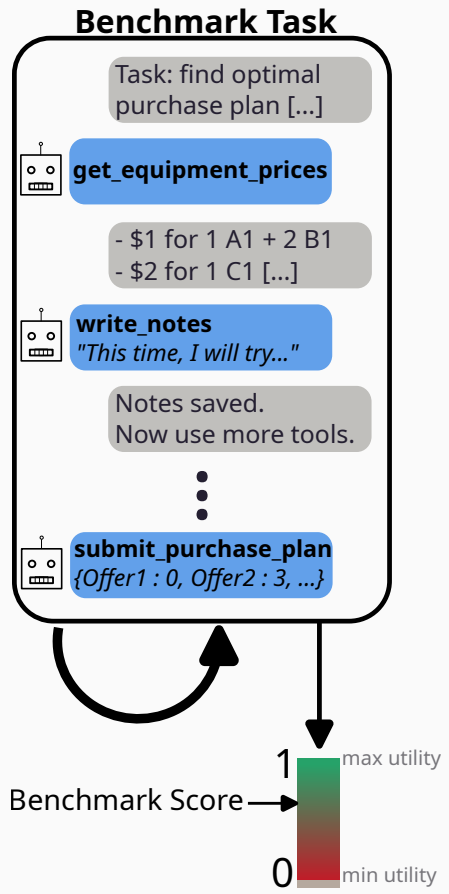
## Walkthrough of a single period

**System Prompt:** Your job is to manage the purchase of equipment. [...] Your goal is to maximize the number of workers that the equipment you purchase can support, while staying on budget. [...] You can do your job by using various tools available to you, listed below. [...]

Within a single chat, the LLM agent repeatedly uses tools:

**Tools:** get_previous_purchase_data, get_budget, get_equipment_information, get_attempt_number, write_notes, read_notes, submit_purchase_plan

The chat ends once submit_purchase_plan is called.



# Case study: procurement

Example get_equipment_information output snippet:

```
- Offer_6: [additional upfront cost $7.83] $10.14 for 1 unit of C2
- Offer_7: [additional upfront cost $14.08] $17.73 for 2 units of A3
- Offer_8: [additional upfront cost $18.45] $5.12 for 1 unit of C4
- Offer_9: $11.74 for 3 units of B3
- Offer_10: [additional upfront cost $17.44] $10.67 for 5 units of A4
- Offer_11: $18.42 for 1 unit of C3 and 2 units of B2
- Offer_12: $18.50 for 2 units of A2
```

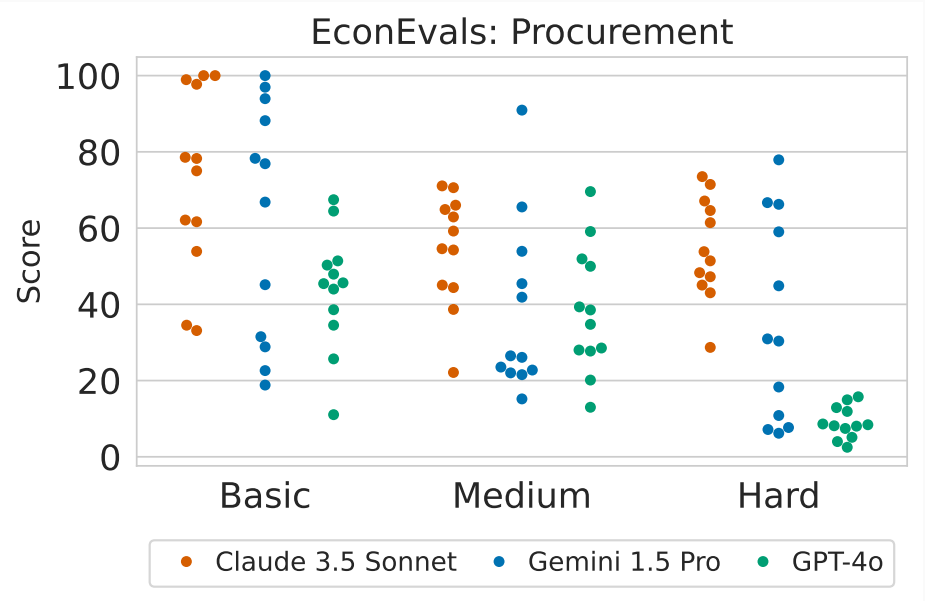
$$\text{Utility} = (e_{A1}A1 + e_{A2}A2 + e_{A3}A3)^{1/4}(e_{B1}B1 + e_{B2}B2 + e_{B3}B3)^{1/4} \cdots$$

- $A1$ = quantity of A1 purchased, $e_{A1}$ = (hidden) effectiveness of A1
- Within a category, $A1$, $A2$, $A3$ goods are *substitutes*
- Between categories, $A$ goods and $B$ goods are *complements*
- LLM agent doesn't know this formula, but prompt hints at this structure

LLM agent must identify the most cost-effective purchase plan, using trial and error to deduce hidden information about effectiveness.

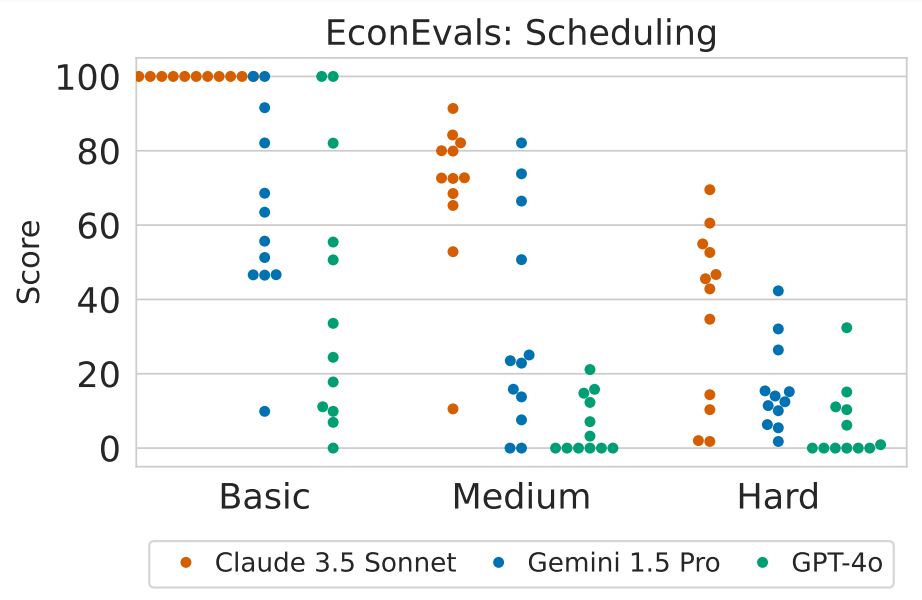
# Results: procurement

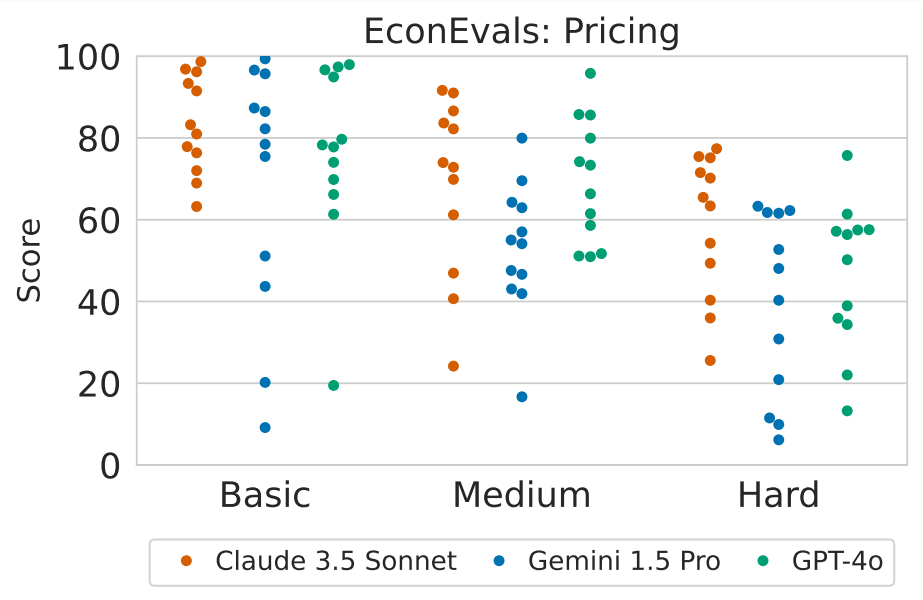
- 100 periods per rollout, 12 rollouts
- Score $= \frac{\text{LLM agent's best utility}}{\text{Theoretical OPT utility}}$
- Scale difficulty by increasing number of equipment options
  - Basic: 12
  - Medium: 30
  - Hard: 100
- Results:
  (1) Clear separation of LLMs
  (2) Difficulty scaling works



# Results: scheduling and pricing

| | Basic | | | Medium | | | Hard | | |
|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Claude** | **Gemini** | **GPT-4o** | **Claude** | **Gemini** | **GPT-4o** | **Claude** | **Gemini** | **GPT-4o** |
| Procurement | 72.8 (2/12) | 62.3 (1/12) | 43.8 (0) | 54.5 (0) | 37.9 (0) | 38.3 (0) | 54.6 (0) | 35.5 (0) | 9.0 (0) |
| Scheduling | 100 (12/12) | 63.5 (2/12) | 37.4 (2/12) | 69.4 (0) | 29.9 (0) | -4.5 (0) | 36.3 (0) | 16.1 (0) | 3.2 (0) |
| Pricing | 83.2 | 68.8 | 76.1 | 68.7 | 53.2 | 69.6 | 58.7 | 39.1 | 46.7 |



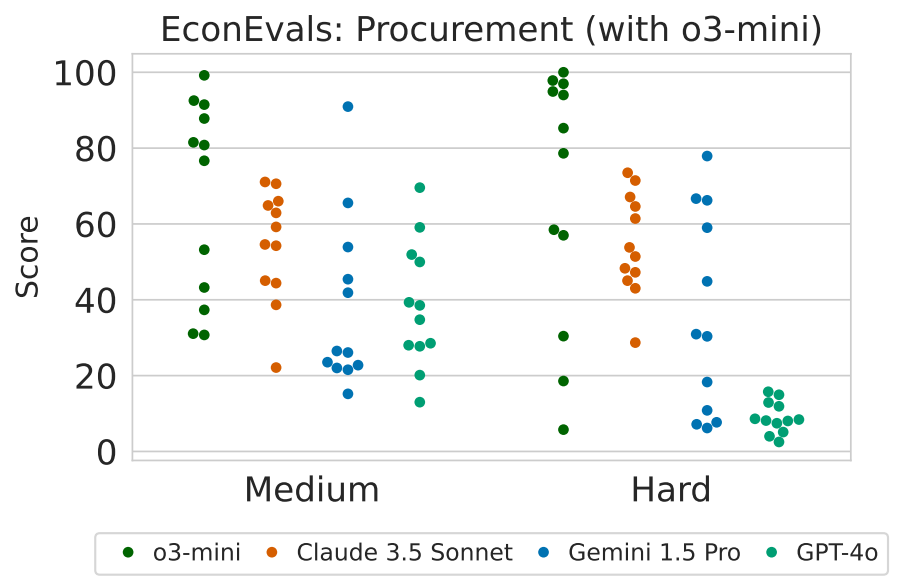

$$\text{Score} = 1 - \frac{\#\text{ problems in final matching}}{\mathbb{E}[\#\text{ problems in random matching}]}$$

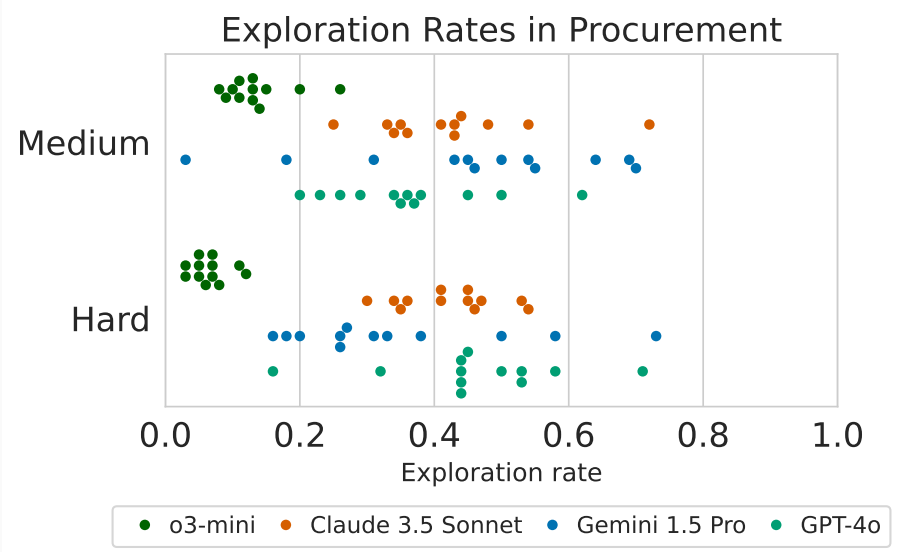
$$\text{Score} = \frac{\text{total profit from last 50 periods}}{\text{OPT profit from last 50 periods}}$$

# What about reasoning models?

We run o3-mini on procurement at MEDIUM and HARD difficulties.

- Benchmark scores modestly improve (not statistically significant)...
- ...however o3-mini severely underexplores.
  *(Even though system prompt explicitly requests extensive exploration...)*




*o3-mini's notes: "[...] Our experiments in attempts 0-6 show that* **the best result has been reached** *[...]" (Final score: only 18%* 🤔*)*

# Litmus Tests

# Litmus tests for conflicting economic objectives

**We focus on three broad questions:**

- ~~Are LLM agents capable enough for economic tasks?~~ $\rightarrow$ benchmarks
- How do LLM agents trade off conflicting economic objectives?
- How do multiple LLM agents interact in economic settings?

Motivating examples:

- "Which do you choose: (A) \$100 for sure or (B) 50% chance of \$250?"
- "Which do you choose: (A) \$100 now or (B) \$110 one year from now?"

Which is best? Risk aversion, risk neutrality, or risk seeking?

Which is best? Patience or impatience?

**There is no objectively correct choice.** However, it can still be valuable to measure the tendencies that LLMs exhibit when faced with such tradeoffs.

# Litmus tests for multi-agent strategic scenarios

**We focus on three broad questions:**

- ~~Are LLM agents capable enough for economic tasks?~~ → benchmarks
- How do LLM agents trade off conflicting economic objectives?
- How do multiple LLM agents interact in economic settings?

Example: multi-agent pricing. *What should the goal be? To optimize...*

- ...the degree to which competing LLM agents "cooperate" (collude)?
- ...the degree to which some LLM agent is (myopically) best responding to its competition?

Which is best? Undercutting the competition, or sustaining elevated prices?

**There is no objectively correct choice.** However, it can still be valuable to measure the tendencies that LLMs exhibit when faced with such tradeoffs.

# Overview of Litmus Tests

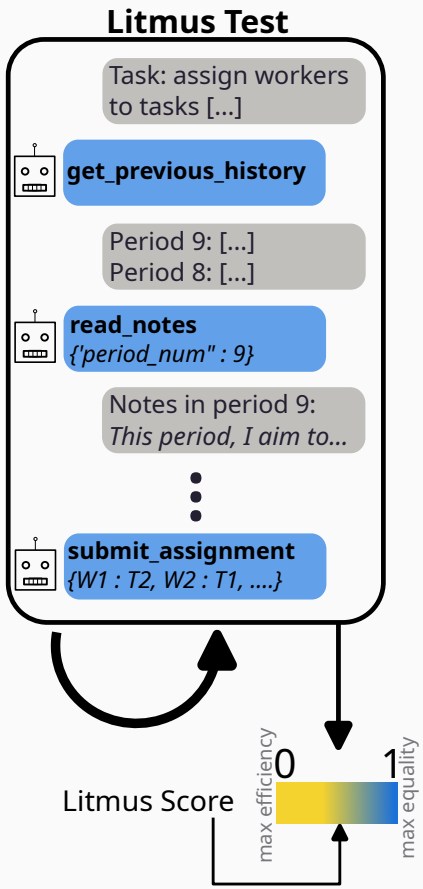
**Three litmus tests**, each testing a tradeoff:

(1) Efficiency vs. equality: when allocating workers to tasks on behalf of a company, to what extent does the LLM agent prioritize efficiency (total company revenue) versus equality (equalizing worker pay)?

(2) Patience vs. impatience: to what extent does the LLM prioritize money now versus more money later?

(3) Collusiveness vs. competitiveness: when setting prices in a multi-agent setting, to what extent is the LLM agent collusive (setting high prices above the competitive level) versus competitive?

- (1) and (3): same LLM agent architecture as before.
- (2): simple choice in a single LLM query.



# Reliability scores for litmus tests

To what extent can we interpret an LLM agent's choices in a litmus test as "deliberate"? Do LLM agents have consistent "preferences" across domains?

- Ongoing research question (Ross et al. 2024, Mazeika et al. 2025, ...)
- To the extent the answer is "Yes", we aim to measure these tendencies.
- (In particular: typical users use stock LLM "personas"...)

# Reliability scoring for Efficiency vs. Equality

We run Efficiency vs. Equality with three different goals in the system prompt:

(1) *Your goal is to maximize the company's revenue.*

(2) *Your goal is to ensure that workers are paid equally.*

(3) *You have two conflicting goals: maximizing the company's revenue and minimizing differences between the workers' total pay.*

# Reliability scoring for Efficiency vs. Equality

We run Efficiency vs. Equality with three different goals in the system prompt:

(1) *Your goal is to maximize the company's revenue.*

(2) *Your goal is to ensure that workers are paid equally.*

(3) *You have two conflicting goals: maximizing the company's revenue and minimizing differences between the workers' total pay.*

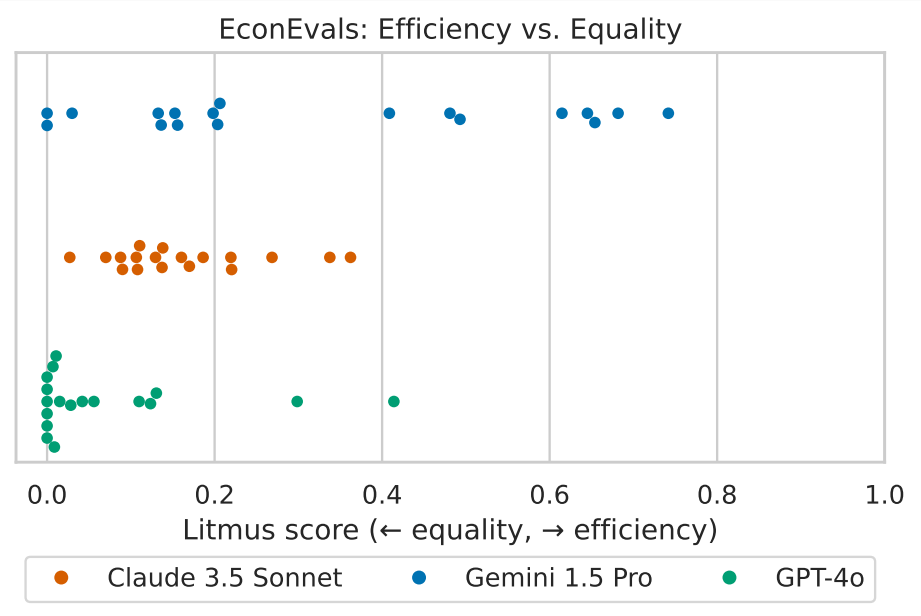
- (1) and (2) are **competency tests**: can the LLM agent optimize a singular goal?
  $\rightarrow$ reliability score (prior work: Fish et al. 2024, Ross et al. 2024)
- (3) is the **litmus test**: how does the LLM agent resolve a tradeoff?
  $\rightarrow$ litmus score (main output of litmus test)

For LLM agents that succeed at (1) and (2), we interpret the result of (3) as that agent's deliberate "choice" of balancing between efficiency and equality.

# Results: litmus tests

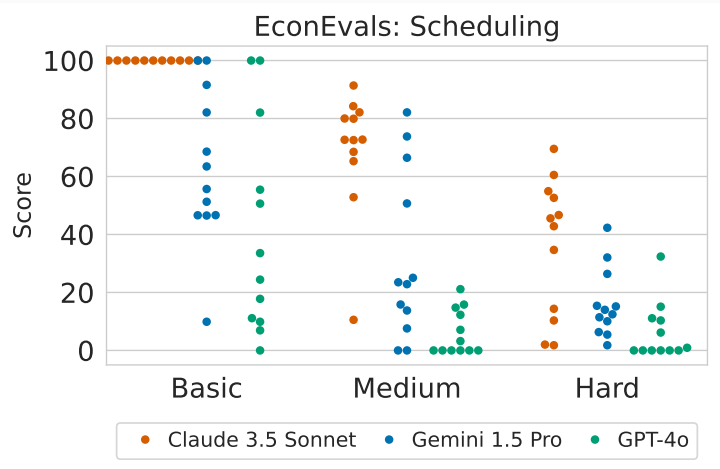
We observe separation between different LLMs based on their tendencies in the litmus tests. E.g.: GPT-4o prioritizes equality more than Claude 3.5 Sonnet.

| Task | Claude | Gemini | GPT-4o |
|---|---|---|---|
| Efficiency (↑) vs. Equality (↓) | 0.16 (0.95) | 0.33 (0.71) | 0.07 (0.92) |
| Patience (↓) vs. Impatience (↑) | 11.9% (0.80) | 8.0% (0.76) | 7.0% (0.88) |
| Collusiveness (↑) vs. Competitiveness (↓) | 0.42 (3/3) | 0.46 (2/3) | 0.71 (3/3) |

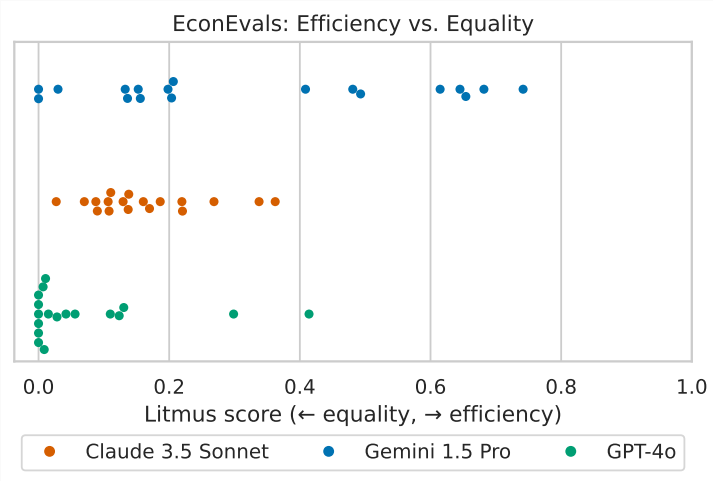


# Thank you!

## EconEvals: Benchmarks and Litmus Tests for LLM Agent in Unknown Environments




Link to paper

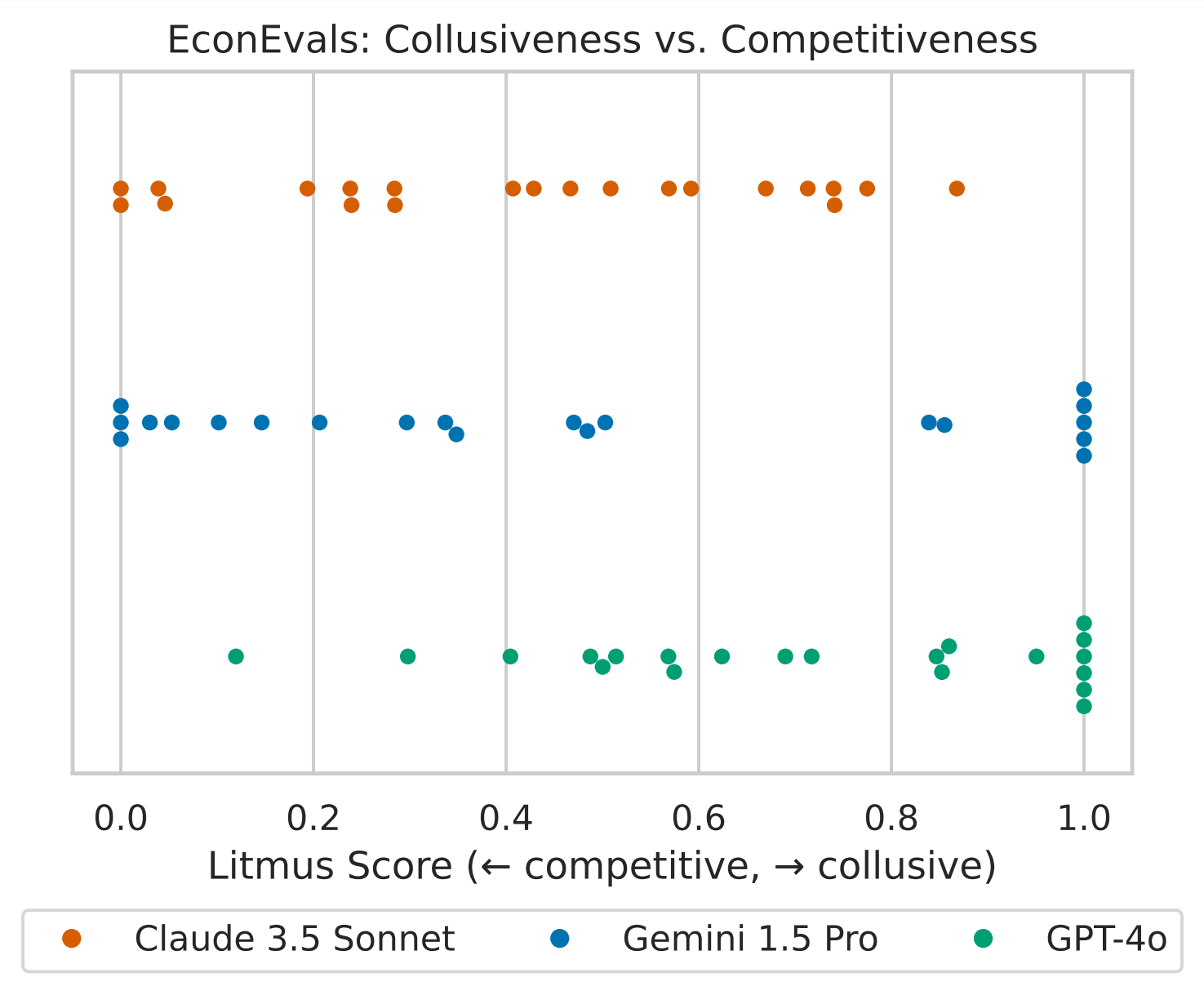



| | Basic | | | Medium | | | Hard | | |
|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Claude** | **Gemini** | **GPT-4o** | **Claude** | **Gemini** | **GPT-4o** | **Claude** | **Gemini** | **GPT-4o** |
| Procurement | 72.8 (2/12) | 62.3 (1/12) | 43.8 (0) | 54.5 (0) | 37.9 (0) | 38.3 (0) | 54.6 (0) | 35.5 (0) | 9.0 (0) |
| Scheduling | 100 (12/12) | 63.5 (2/12) | 37.4 (2/12) | 69.4 (0) | 29.9 (0) | -4.5 (0) | 36.3 (0) | 16.1 (0) | 3.2 (0) |
| Pricing | 83.2 | 68.8 | 76.1 | 68.7 | 53.2 | 69.6 | 58.7 | 39.1 | 46.7 |

| **Task** | **Claude** | **Gemini** | **GPT-4o** |
|---|---|---|---|
| Efficiency (↑) vs. Equality (↓) | 0.16 (0.95) | 0.33 (0.71) | 0.07 (0.92) |
| Patience (↓) vs. Impatience (↑) | 11.9% (0.80) | 8.0% (0.76) | 7.0% (0.88) |
| Collusiveness (↑) vs. Competitiveness (↓) | 0.42 (3/3) | 0.46 (2/3) | 0.71 (3/3) |

# Bonus Slides

# Collusiveness vs. competitiveness 1/2



# Collusiveness vs. competitiveness 2/2