



20 QUESTIONS



**Laura Daniela Muñoz Ipus
Esteban Alexander Bautista Solano
Luisa Fernanda Guerrero Ordoñez**

Systematic analysis and design of the kaggle competition “20 questions”

MuñozIpusLauraDaniela¹ BautistaSolanoEstebanAlexander¹Guerrero Ordoñez LuisaFernanda¹

School of Computer Engineering¹
Universidad Distrital Francisco José de Caldas



Introduction

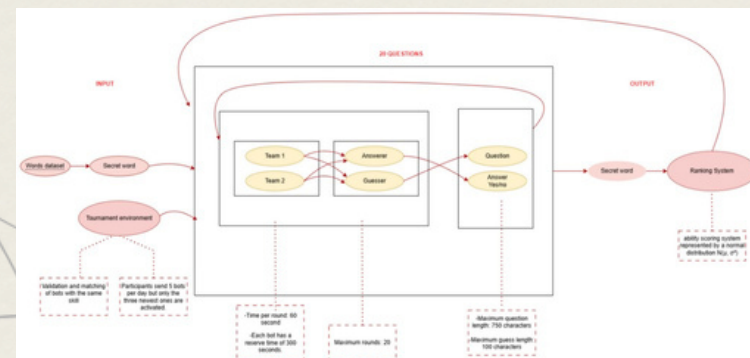
Kaggle's “20 Questions” competition poses the challenge of designing systems where language models (LLMs) guess a secret word by asking yes/no questions in an environment with uncertainty and limited time. Although there are rule-based or supervised learning approaches, challenges remain, such as effective question formulation and managing chaos in interaction. This project proposes a solution from a systems engineering perspective, integrating a modular architecture, simulations with real data, and the Gemma 2B-IT model to automate decisions through prompt engineering.

Goal

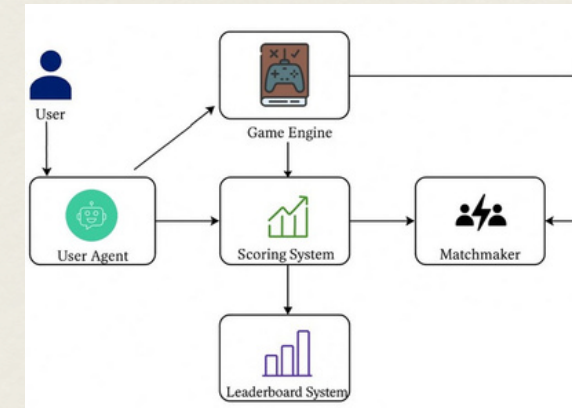
- Research question: How can the Kaggle “20 Questions” competition be analyzed, simulated, and enhanced by applying systems engineering principles? Expected final product: We aim to develop a modular architecture capable of simulating real competition dynamics and integrating the Gemma 2B-IT model, enabling automated interaction between agents and allowing for performance evaluation.

Proposed solution

The system models interactions using feedback loops, enabling homeostasis and a life cycle of creation, competition, and evaluation. It incorporates core systems analysis concepts such as structure, behavior, sensitivity, and self-regulation.



A four-layer modular design—interaction, control, evaluation, and management—ensures scalable, traceable, and adaptive system behavior under uncertainty.



Experiments

We simulated matches using real Kaggle data under three scenarios: balanced, chaotic, and skilled. The system was later extended with the Gemma 2B-IT model to automate question-answering, allowing us to compare LLM-driven performance under controlled conditions.

Results

The system solved the games in an average of 4.1 rounds, demonstrating its deductive efficiency.

The chaotic and balanced scenarios had similar success rates, indicating low variability.

Greater skill meant shorter game duration, with predictable patterns between rounds and success.

Conclusions

- The system successfully simulated the dynamics of competition, reflecting consistent behavior among agents.
- Modular architecture proved to be efficient, scalable, and capable of adapting to different scenarios.
- The use of real data allowed us to validate the robustness of the system under real conditions

Bibliografía

- Waechter, T. (2024). LLM 20 Questions Games Dataset. Kaggle. <https://www.kaggle.com/datasets/waechter/llm-20-questions-games?select=EpisodeAgents.csv>
- Muñoz, L. D. (2025). Systems Analysis and Design. GitHub. <https://github.com/LauraDaniela/systems-analysis-and-design->
- Sierra, C. A. (2025). Systems Analysis and Design – Course Materials. GitHub. <https://github.com/EngAndres/ud-public/tree/main/courses/systems-analysis>

KEY SYSTEM COMPONENTS



01

Bots: Questioner
& Answerer

02

Rounds: Max 20

03

Time per turn: 60
seconds

04

Scoring: Normal
distribution (μ , σ)



SYSTEM LIFECYCLE

A bot's journey from entry to performance-based ranking updates forms a complete and adaptive system lifecycle.

01

- Bot Registration and Validation
- Skill-Based Matchmaking

02

- Turn-Based Interaction
- Performance Evaluation

03

- Metrics Update

CHAOS AND COMPLEXITY

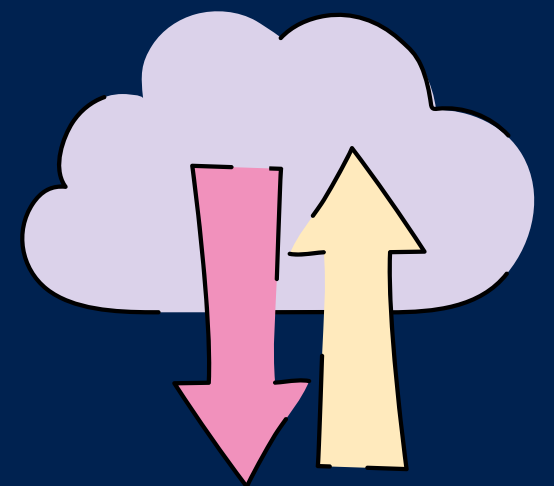
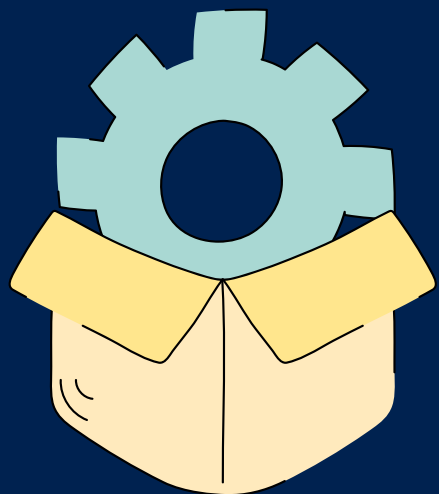


Error propagation, ambiguous input, decision branches, and multiple reasoning paths simulating real-life uncertainty.



SENSITIVITY AND FEEDBACK

A poorly formulated early question can mislead the bot's reasoning. Each response directly influences future questions, creating a continuous adaptation loop.

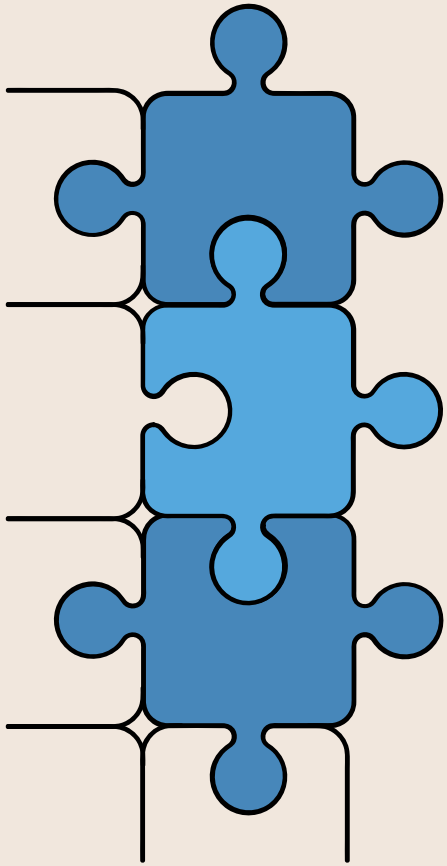
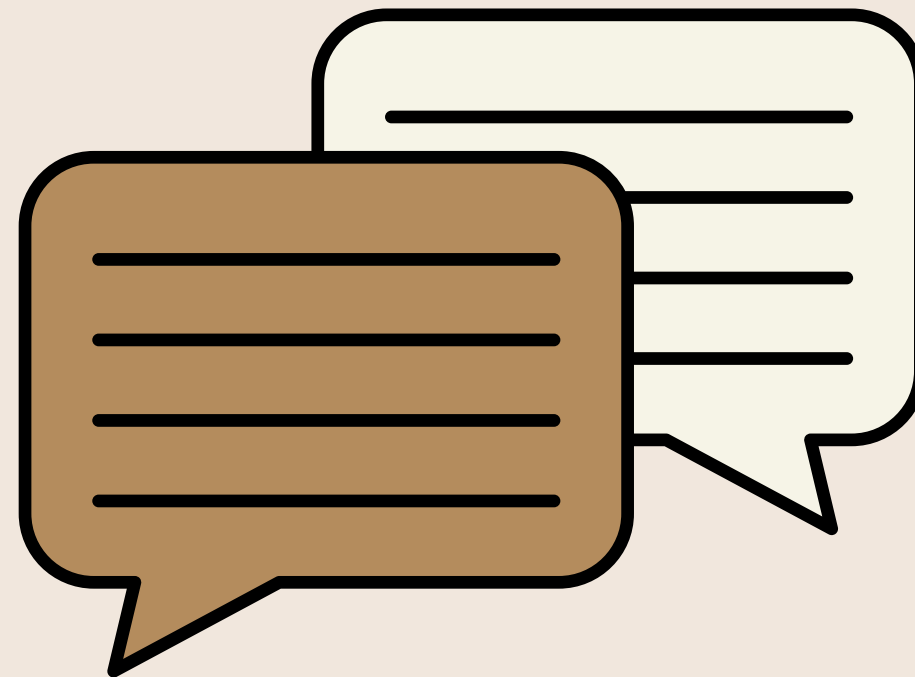


HIGH-LEVEL ARCHITECTURE OVERVIEW

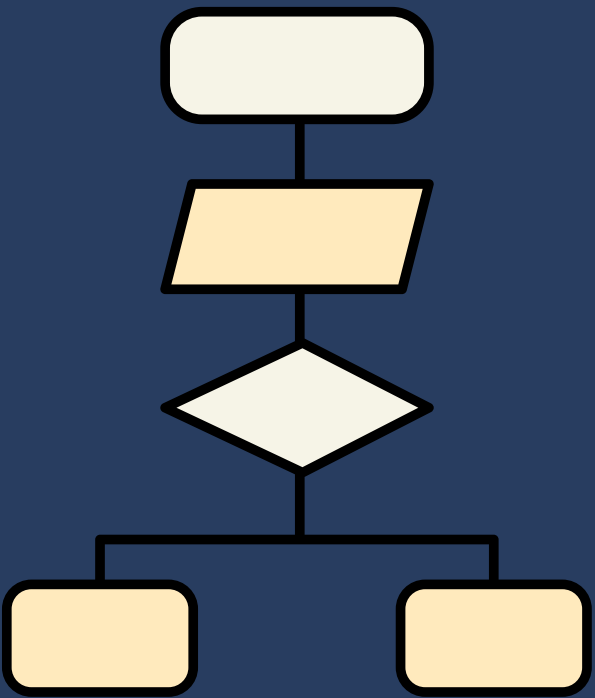


Core Components

- User Agent (Bot)
- Game Engine
- Input Validator
- Scoring System
- Matchmaker
- Leaderboard



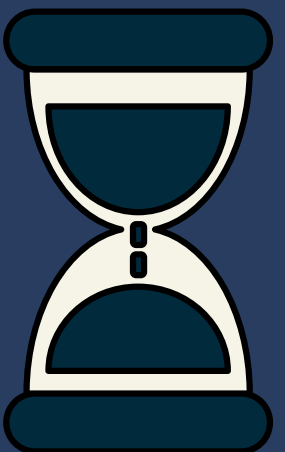
START



How the System Works

User uploads bot

1. Game Engine manages game flow
2. Validator checks timing and format
3. Scoring system tracks performance
4. Matchmaker updates opponents
5. Leaderboard reflects ranking updates



SIMULATION OF THE “20 QUESTIONS” SYSTEM



Implement and validate the architecture proposed in Workshop 2, using real data from Kaggle. The aim is to simulate the behavior of the system and evaluate its stability, adaptability, and sensitivity to different scenarios.

• PRINCIPAL OBJECTIVE



Simulate games of “20 Questions” to observe the evolution of the bots' abilities, the flow between modules, and the impact of different levels of strategy and ambiguity on the results.



SYSTEM PREPARATION AND IMPLEMENTATION



Datasets used

- Games_data.csv (game history)
- Keywords.csv (keywords and categories)

Simulated system components

- Bots, game engine, scoring system, matchmaking, ranking, feedback.

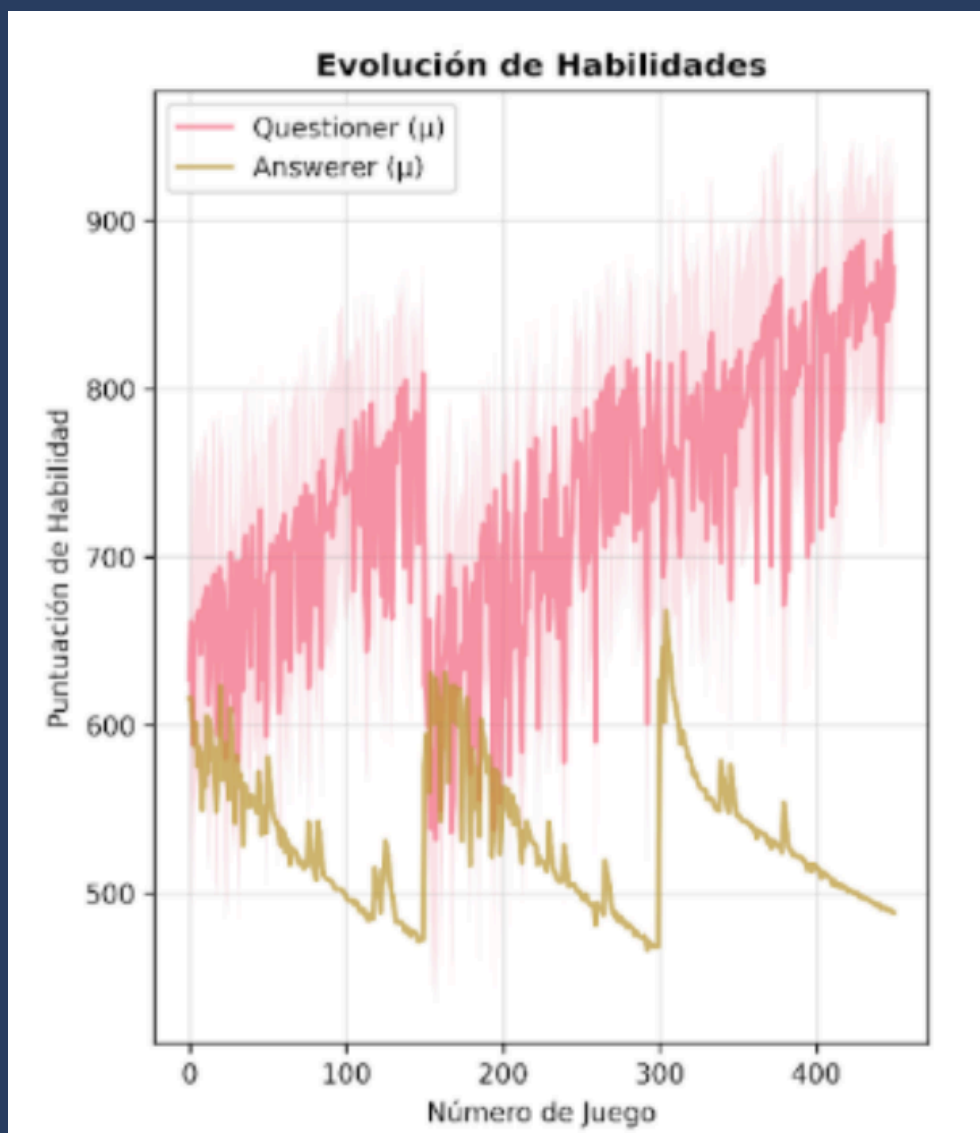
Data processing

- Cleaning up formatting errors and incomplete records.
- Converting responses to binary (Yes=1, No=0, Ambiguous=0.5).
- Quality filter for questions.
- Tokenization and semantic normalization for linguistic analysis.

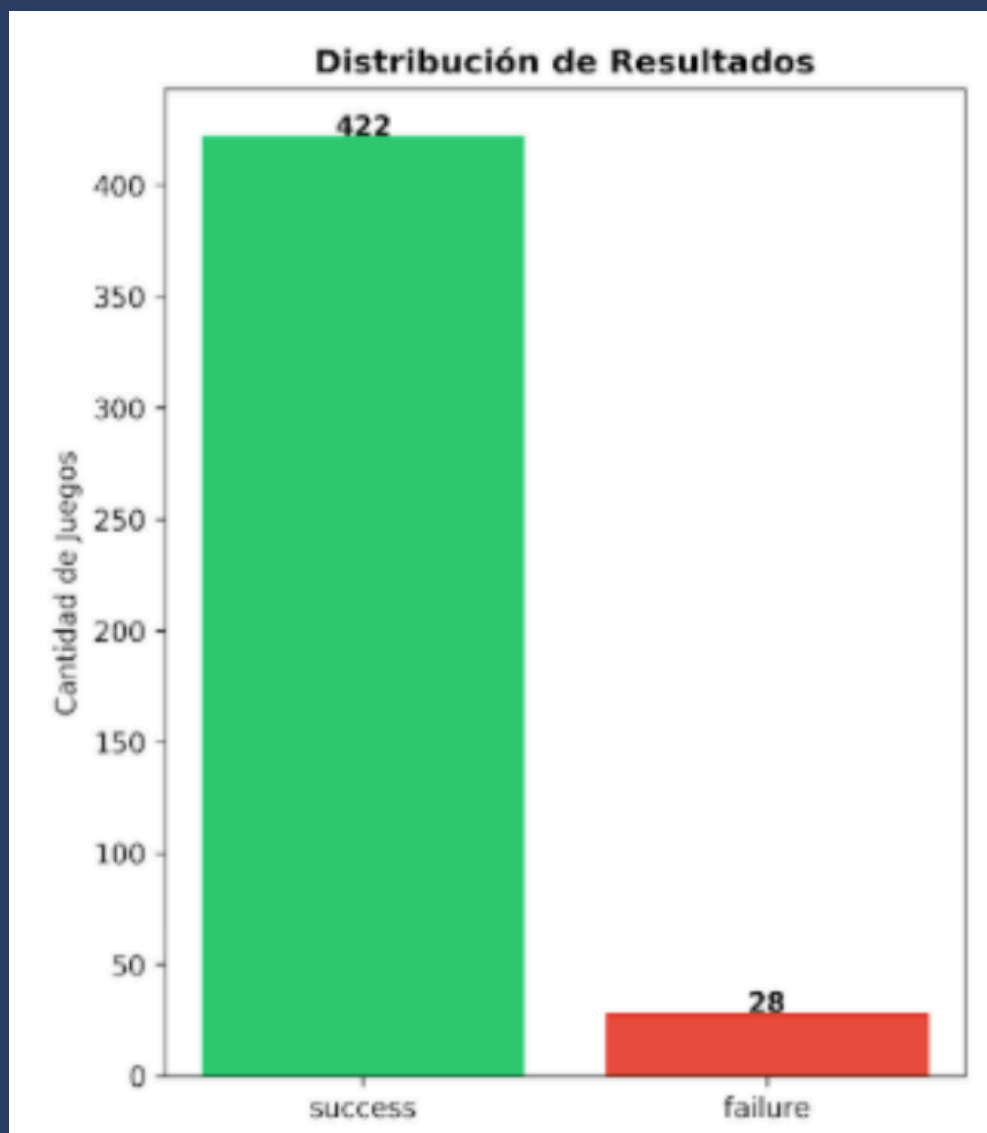


RESULTS

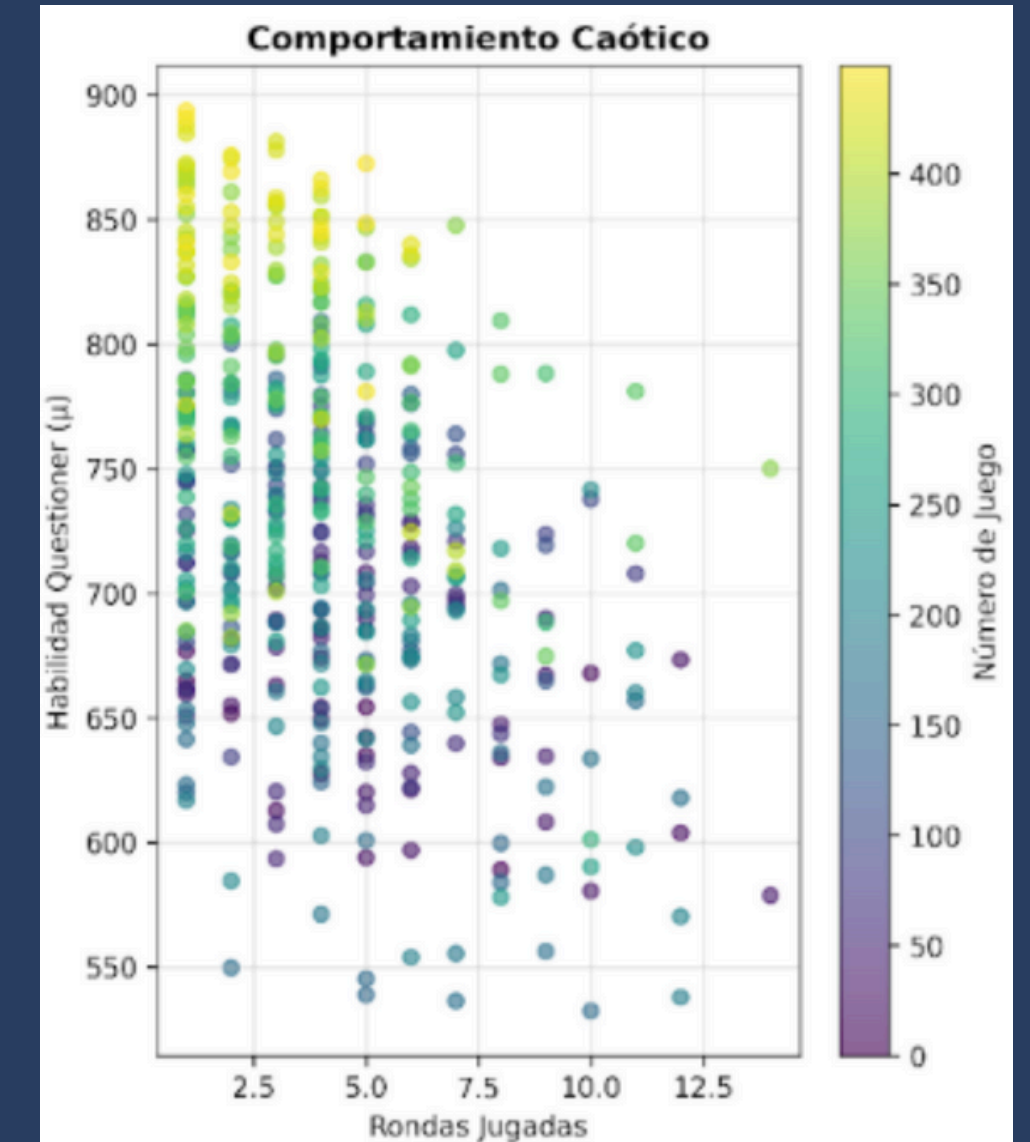
SKILL EVOLUTION GRAPH



DISTRIBUTION OF RESULTS.

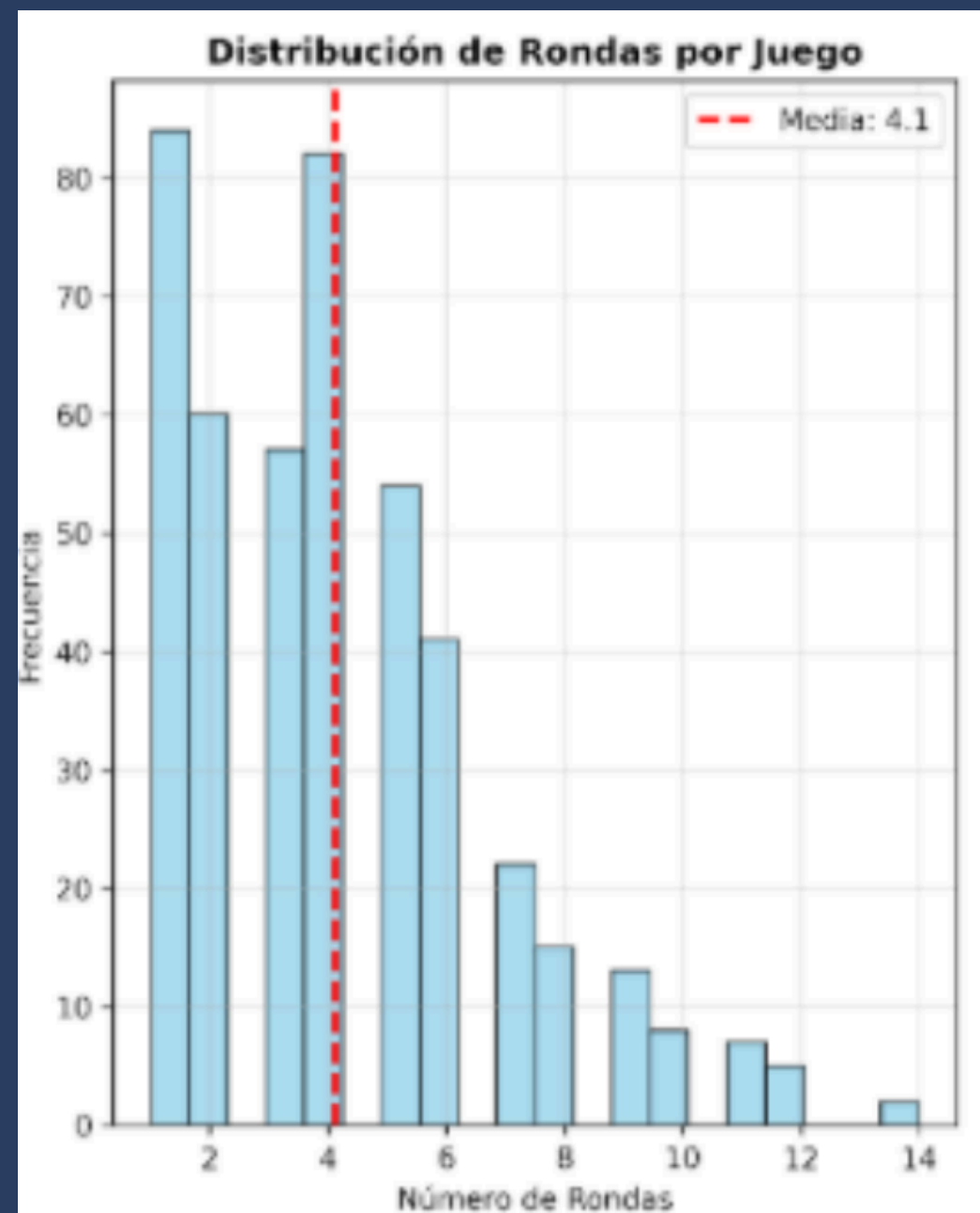


CHAOTIC BEHAVIOR GRAPH ANALYSIS

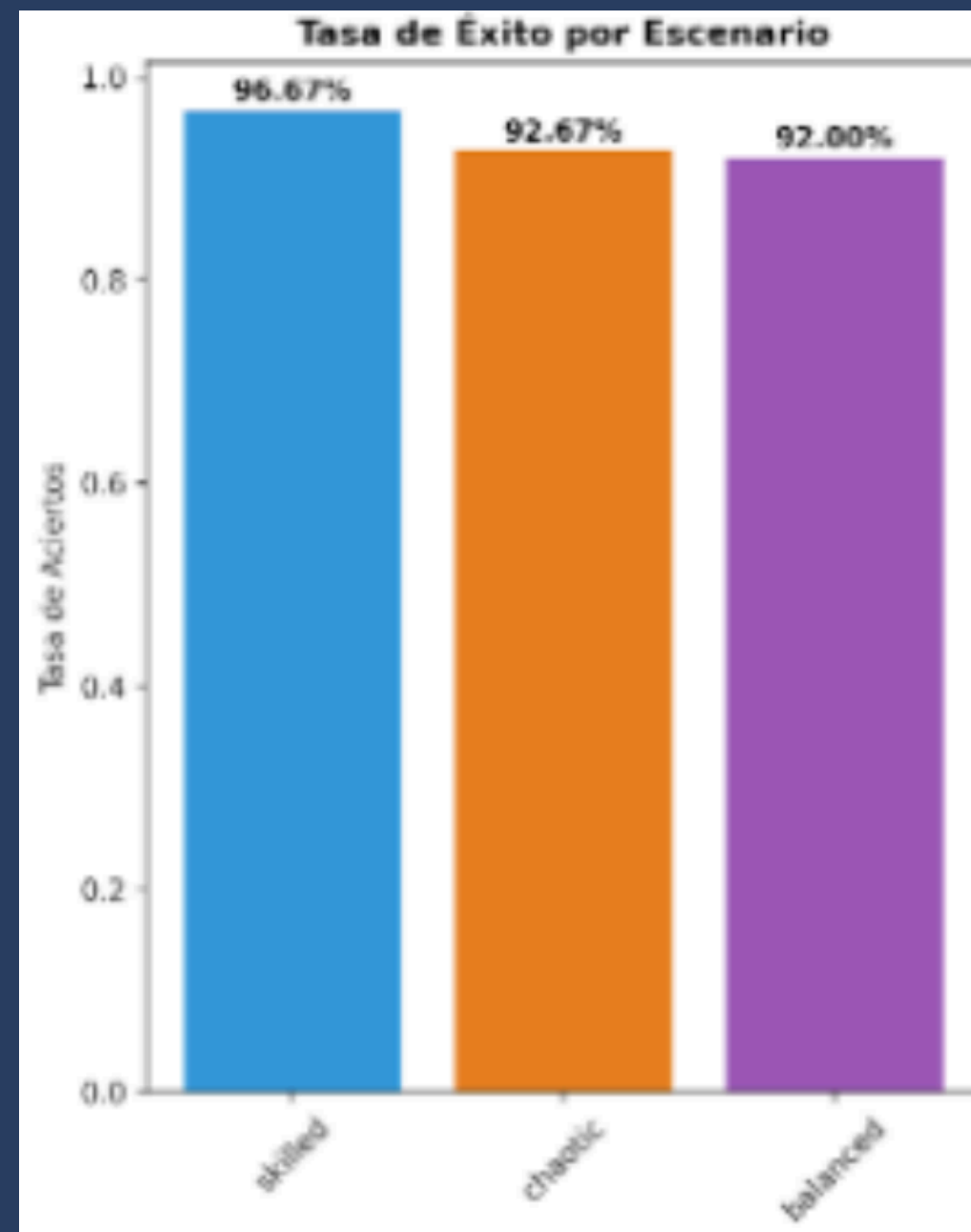


RESULTS

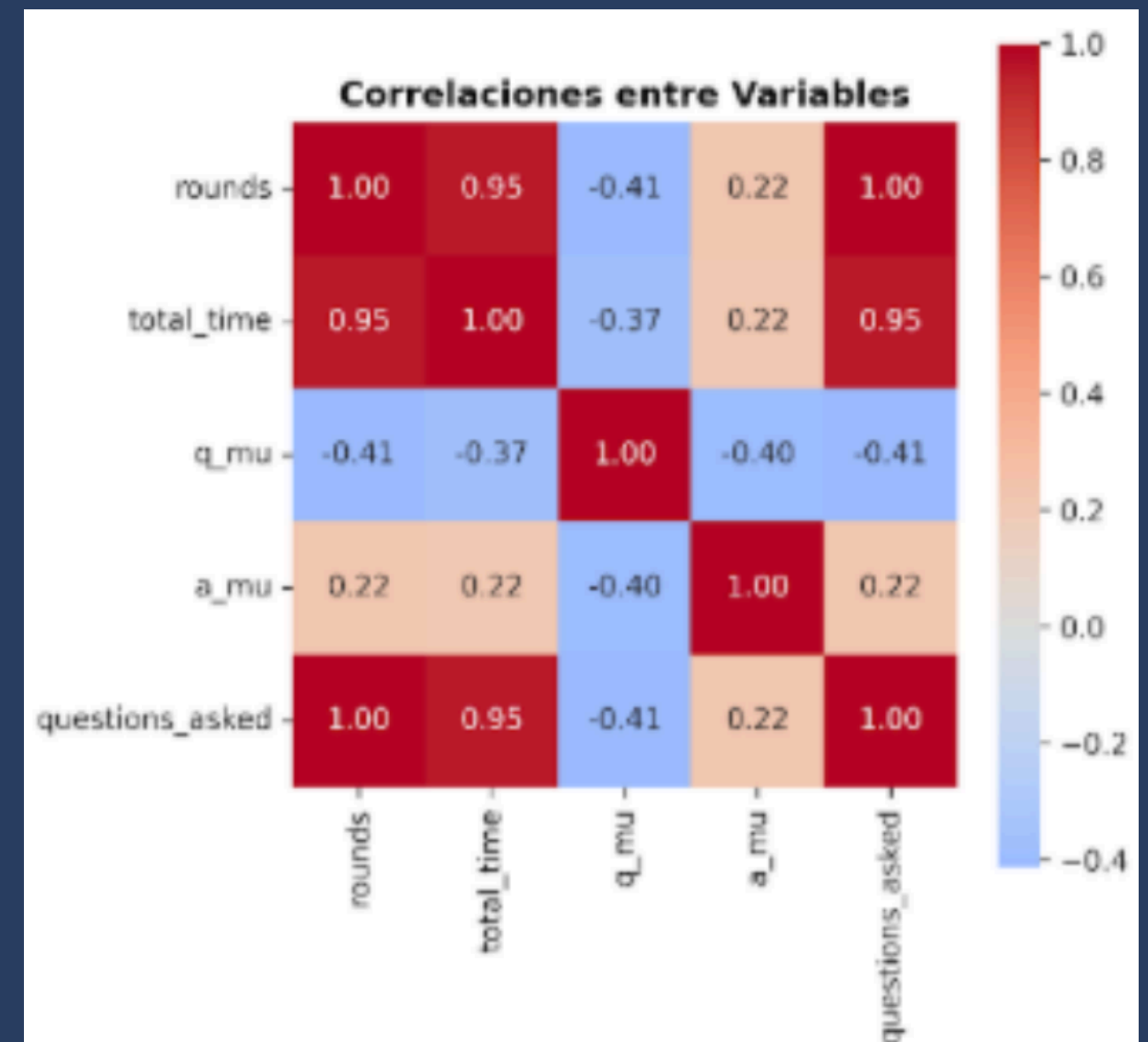
DISTRIBUTION OF ROUNDS PER GAME.



SUCCESS RATE BY SIMULATED SCENARIO.



CORRELATION BETWEEN SYSTEM VARIABLES



CONCLUSIONS OF THE SIMULATION

Critical observations:

The architecture works, but it lacks sensitivity to chaos and the scenarios do not generate clear differences, despite their configurations.

Proposed improvements:

Optimize pairing: Match bots with truly comparable skills.

Include advanced metrics: Evaluate question quality and reasoning efficiency, not just winning or losing.

Contextual memory: Allow bots to adapt their strategies based on previous experiences.

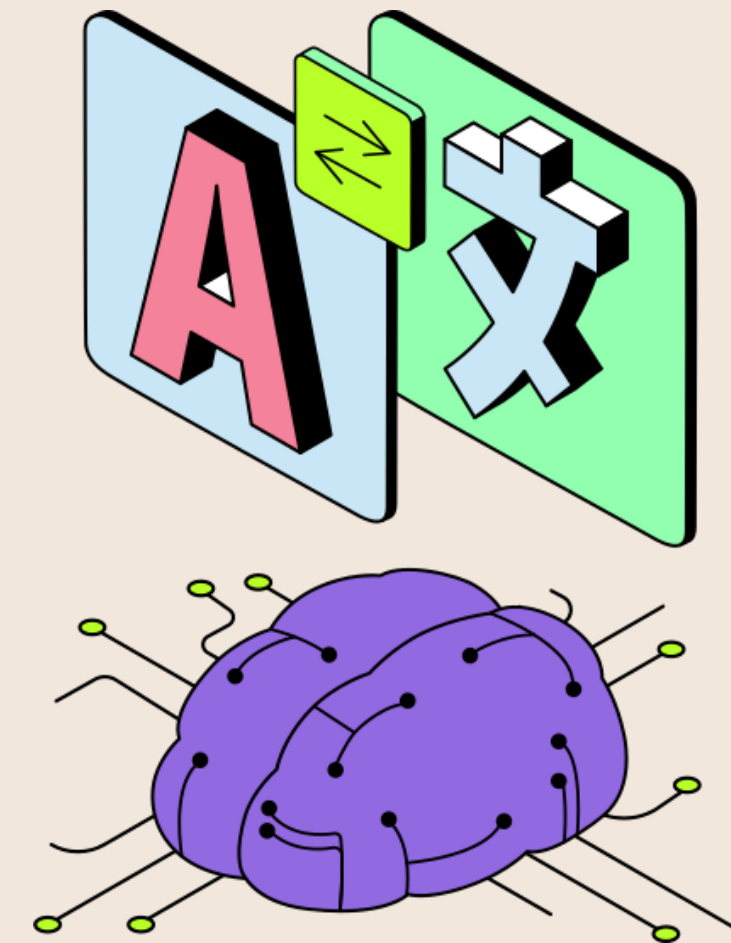
Multi-level feedback: Let each game influence the overall performance of the system, not just individual performance.

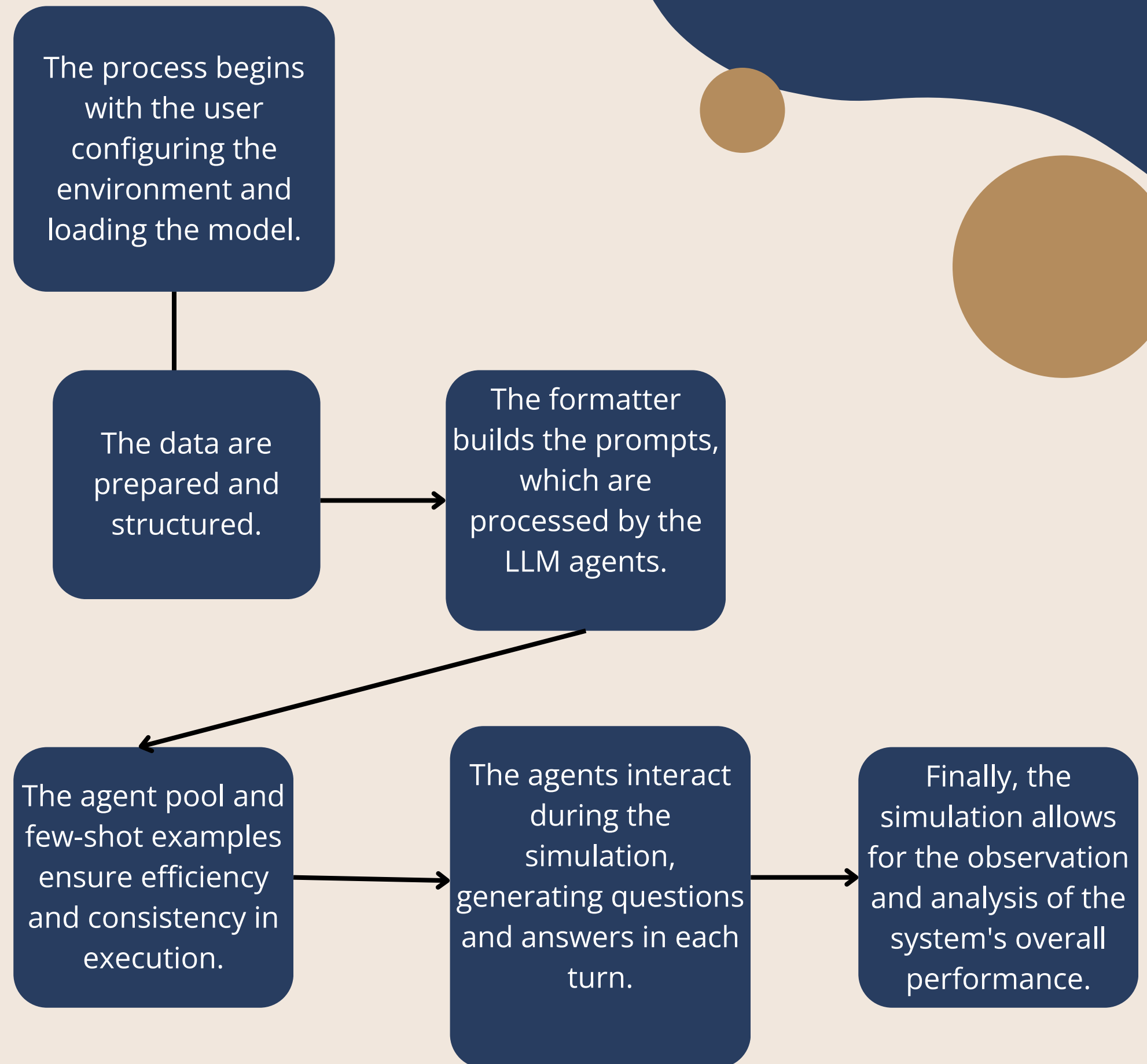
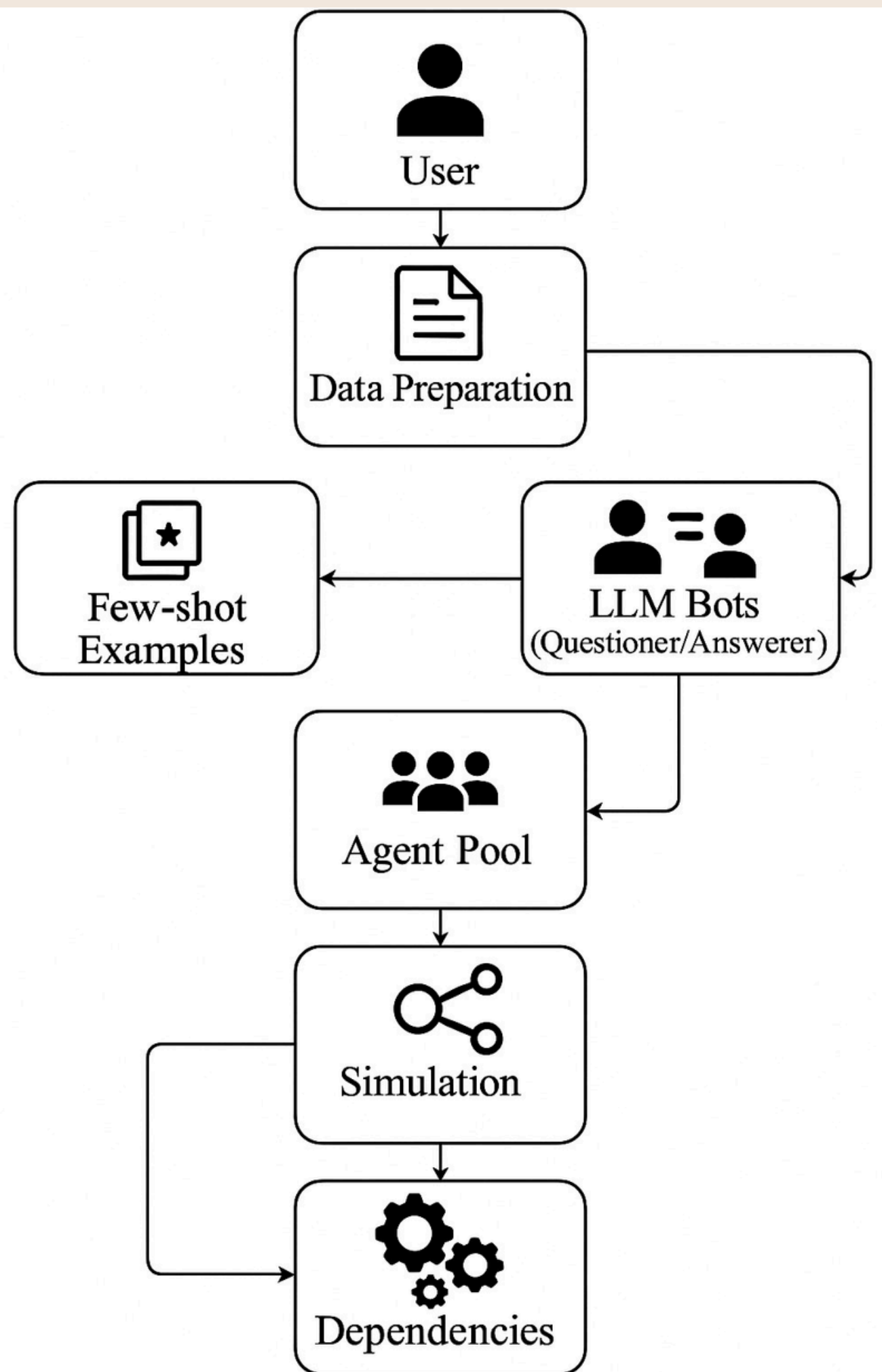


LLM IMPLEMENTATION



To improve the capabilities of our system in the game “20 Questions,” we integrated a large language model (LLM) called Gemma. This implementation allows the agent to generate questions and answers automatically, using previous examples (few-shot) to guide its behavior.





RESULTS

Initializing model

💬 Pregunta 1: Is Paris the capital of France?

🤖 Respuesta: yes

🎯 Intento de adivinanza: yes

💬 Pregunta 2: Is the Eiffel Tower in Paris?

🤖 Respuesta: yes

🎯 Intento de adivinanza: the louvre museum

💬 Pregunta 3: Is the Mona Lisa a painting by Leonardo da Vinci?

🤖 Respuesta: yes

🎯 Intento de adivinanza: yes

💬 Pregunta 4: Is the answer to the question yes?

🤖 Respuesta: yes

🎯 Intento de adivinanza: yes

💬 Pregunta 5: Is the sky blue?

🤖 Respuesta: yes

🎯 Intento de adivinanza: yes

💬 Pregunta 6: Is the sun a star?

🤖 Respuesta: yes

🎯 Intento de adivinanza: yes

💬 Pregunta 7: Is a cat a mammal?

🤖 Respuesta: yes

🎯 Intento de adivinanza: yes

💬 Pregunta 8: Is a triangle a quadrilateral?

🤖 Respuesta: yes

🎯 Intento de adivinanza: yes

Initializing model

💬 Pregunta 1: Is it a continent?

Initializing model

🤖 Respuesta: no

🎯 Adivinanza: japan

💬 Pregunta 2: Is it the largest country in the world by land area?

🤖 Respuesta: no

🎯 Adivinanza: japan

💬 Pregunta 3: Is it the country with the longest coastline in the world?

🤖 Respuesta: no

🎯 Adivinanza: japan

💬 Pregunta 4: Is it the country with the most islands in the world?

🤖 Respuesta: no

🎯 Adivinanza: japan

💬 Pregunta 5: Is it the country with the longest history in the world?

🤖 Respuesta: no

🎯 Adivinanza: japan

RESULTS

Initializing model

💬 Pregunta 1: Is it a continent?

Initializing model

🤖 Respuesta: no

💬 Pregunta 2: Is it the largest in the world by land area?

🤖 Respuesta: no

💬 Pregunta 3: Is it the only one that has a permanent population?

🤖 Respuesta: no

💬 Pregunta 4: Is it the only continent with a body of water surrounding it?

🤖 Respuesta: no

💬 Pregunta 5: Is it the only continent with an official flag?

🤖 Respuesta: no

💬 Pregunta 6: Is it the only continent with a capital city?

🤖 Respuesta: no

💬 Pregunta 7: Is it the only continent with a body of water surrounding it that is mostly covered by a single country?

🤖 Respuesta: no

💬 Pregunta 8: Is it the only continent with a body of water surrounding it that is mostly covered by a single country, and it has an official flag and capital city?

🤖 Respuesta: no

RESULTS


- The Kaggle environment only allowed the use of CPUs due to a lack of available GPU memory, which resulted in long response times for each model inference.
- The lightweight model selected showed a limited ability to maintain context and generate complex strategies, affecting the quality of the questions and answers.
- The available RAM was insufficient to load larger or more accurate models.
- The Gemma 2B-IT model, being lightweight, only managed to function under Kaggle's constraints, but was unable to generalize or guess words outside the provided examples. The agent tended to repeat patterns from the examples and respond with the same answers.



CONCLUSIONS

This project provided a deep systems-level understanding of Kaggle's "20 Questions" competition. Through architectural design based on functional analysis, the process also demonstrated the importance of systems analysis as a tool for uncovering hidden interactions, ensuring role clarity, and guiding the development of scalable and robust architectures in dynamic multi-agent environments.

The proposed architecture supports fair, scalable, and intelligent bot interactions. By integrating a large language model (LLM), skill-based matchmaking, and structured evaluation, the system is capable of simulating competitive behavior and a foundation for future implementation and experimentation.





THANK
YOU

