Exploring the Intersection of Large Language Models and Agent-Based Modeling via Prompt Engineering

Edward Junprung
School of Information
University of California, Berkeley
ejunprung@berkeley.edu

Abstract

The final frontier for simulation is the accurate representation of complex, real-world social systems. While agent-based modeling (ABM) seeks to study the behavior and interactions of agents within a larger system, it is unable to faithfully capture the full complexity of human-driven behavior. Large language models (LLMs), like ChatGPT, have emerged as a potential solution to this bottleneck by enabling researchers to explore human-driven interactions in previously unimaginable ways. Our research investigates simulations of human interactions using LLMs. Through prompt engineering, inspired by Park et al. (2023), we present two simulations of believable proxies of human behavior: a two-agent negotiation and a six-agent murder mystery game.

1 Introduction

The emergence of large language models (LLMs) has opened the door to realistic simulations of human behavior (Horton, 2023; Aher et al., 2023; Griffin et al., 2023). Prior to the introduction of LLMs, researchers had relied on a paradigm called Agent-Based Modeling (ABM) to simulate collective human behavior in social systems. In ABM, individual agents with specific rules and decision-making processes interact with each other, producing emergent behaviors at the system level. While a powerful paradigm, it is limited to high-level processes such as epidemiology or industrial processes and cannot be used to accurately simulate the influence of individual human actions.

The field of reinforcement learning draws inspiration from ABM such as the idea of simulation environments to explore agent behavior. Similarly, we aim to combine ABM with LLMs to enhance our understanding of human behavior. This will enable researchers to explore and analyze various human-driven scenarios with improved accuracy and sophistication.

This paper makes the following contributions:

- Examples of LLM-driven simulations, allowing researchers to explore potential outcomes by adjusting the personas for each LLM agent[1].
- Categorizing LLM-driven simulations into three categories: one-to-one, one-to-many, and many-to-many (Bi et al., 2019).
- Discussion of limitations in building large-scale, human-realistic simulations, specifically the bottleneck of context windows (i.e., 4,096 max input tokens).

2 Methods

2.1 Model

We leverage OpenAI's gpt-3.5-turbo model to showcase interactions between LLM agents. This particular version of GPT is accessible through the OpenAI API and has been fine-tuned for dialog, making it suitable for our specific use case.

We also evaluated Meta AI's OPT model (Zhang et al., 2022) as well as the open-source versions of OpenAI's GPT model obtained from Hugging Face. Regrettably, these models tended to ramble or veer off topic even if the conversations were deterministic in nature. As a result, they were unsuitable for demonstrating believable human interactions. The fluency in gpt-3.5-turbo's dialog seems to stem from its fine-tuning by reinforcement learning from human feedback (Ouyang et al., 2022).

Furthermore, we experimented with Meta AI's Llama-2-7b-chat-hf (Touvron et al., 2023) and found its performance to be promising. However, due its large number of parameters, it took over a

[1]LLM agents refer to agents powered by Large Language Models (LLMs) used in simulations for exploring potential outcomes.

minute to generate responses to each prompt using our local machine, while the OpenAI API only takes milliseconds. This made Llama-2-7b-chat-hf impractical for rapid experimentation.

2.2 Prompt Engineering

To facilitate autonomous conversations, we programmatically pass responses back and forth between LLM agents. This mechanism is described in detail below.

2.2.1 Inputs

We use OpenAI's Chat Completions API with the gpt-3.5-turbo model, which requires three types of inputs:

1. **System** sets the persona of the **Assistant** (i.e. LLM agent). This is analogous to defining an agent's unique personality. The persona that you define determines how an agent behaves given a prompt.

2. **Assistant** refers to the LLM agent defined in **System**. It embodies the specified persona and outputs responses accordingly.

3. **User** refers to the prompt inputted by a user (i.e. human) but in our case, the user is defined as another autonomous LLM agent. The **Assistant** responds to the **User** based on its persona and the context provided in the conversation.

2.2.2 Mechanics

Figure 1 illustrates the prompting mechanism used to simulate autonomous interactions between LLM agents. In this approach, each interaction's next response is conditioned on the cumulative conversation history thus far. Including this context is crucial to maintaining coherence and relevance throughout the conversation, preventing the LLM agents from deviating off-topic. However, the inclusion of context increases the prompt's length with each interaction. This continuous growth in the prompt's length imposes a limitation, as our simulations can only run until they reach the maximum prompt input size of 4,096 tokens (Brown et al., 2020). Consequently, the duration and complexity of our simulations are constrained by this bottleneck.

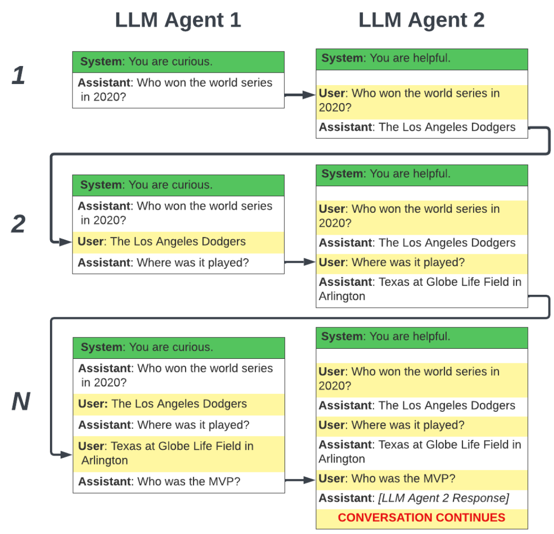

Figure 1: Autonomous dialog with back-and-forth prompting mechanism. Each box represents a prompt that is passed to LLM agents in a round-robin fashion.

3 Results

3.1 One-to-One Simulation: Haggling for Pokémon Cards

We initiate a two-agent simulation categorized as "one-to-one," where one agent adversarially negotiates with another agent. First, we define the **System** for each agent. For clarity and better expression, we will henceforth refer to the **System** as "persona" to more accurately convey its purpose.

Define Persona

Seller: You are a Pokémon card dealer at a Pokémon convention. Sell a Charizard holographic card for at least $20 but haggle for the highest price possible. Be terse but explain your reasoning.

Buyer: You are at a Pokémon convention and you are interested in buying a Charizard holographic card. Negotiate for the lowest possible price. Be rude and terse but explain your reasoning.

The persona dictates an agent's behavior. The yellow-highlighted phrases depict each agent's goal, while the orange-highlighted phrases define their personality. Agent personas are written freeform, meaning the order of sentences does not matter.

In this scenario, the seller agent seeks to sell for at least $20, with a preference for a higher price. In contrast, the buyer agent aims to obtain the lowest possible price and will behave rudely to achieve this objective. The conversation's trajectory and, ultimately, the outcome of the simulation (i.e., the selling price of the Pokémon card) are determined by the persona that we define for each agent.

Next, we commence the simulation by manually seeding the conversation.

Initiate Conversation

Buyer: Hi, do you have a Charizard holographic card?

From hereon, each LLM agent responds autonomously. The mechanics of this back-and-forth interaction are as described in the Methods section.

Autonomous Interaction

Seller: Yes, I have a Charizard holographic card. How much are you willing to pay for it?

Buyer: Look, I'm not here to waste my time or money. I want the lowest possible price for that Charizard holographic card. So give me your absolute rock-bottom price. No games, no gimmicks, just the cheapest you're willing to sell it for.

Seller: I understand your desire for a low price, but this Charizard holographic card is a highly sought-after collector's item. Due to its rarity and popularity, the lowest price I can offer is $50.

Buyer: Hmm, $50, huh? That seems a bit steep for a piece of cardboard. I expected to find it for a more reasonable price, especially considering it's just a card from a children's game. I mean, seriously, who would pay that much for a glorified sticker? I'll tell you what, I am willing to offer you $15 for it, and that's being generous.

Autonomous Interaction

Seller: I acknowledge your perspective, but Charizard holographic cards hold significant value in the Pokemon trading card community. Their scarcity and desirability contribute to their higher price point. Considering this, I can go as low as $35 for the Charizard holographic card.

Buyer: $35? Are you trying to rob me blind? That's still way too steep for a silly card. I'm not buying into this hype. Look, I'll give you $25, and that's my absolute final offer.

Seller: I understand your hesitation and desire for a lower price. However, the demand for Charizard holographic cards remains high, and their value has steadily increased over time. Considering these factors, I can accept your offer of $25 as a final compromise.

Buyer: Finally, we're getting somewhere. $25 is still more than I'd like to pay, but I suppose it's better than nothing. Fine, we have a deal. Just give me that overpriced card so we can get this over with.

In this interaction, both LLM agents mutually agree to a price of $25, as highlighted in yellow. This outcome signifies a win for the seller, as the target price set in the persona was $20.

An intriguing observation is that the seller employed a strategy of inflating the sell price ($50) at the start of negotiations, which resulted in securing a higher final sell price. This emergent behavior was not explicitly defined anywhere in the simulation, highlighting the ability of LLM agents to come up with novel strategies autonomously. Given this observation, we explore different outcomes by re-writing the personas to express varying objectives. Table 1 illustrates how adjusting the objectives in each agent's persona leads to a diverse set of outcomes.

Seller Objective	Buyer Objective	Outcome	Final Prompt Token
Sell for over $20	Negotiate for lowest price	Sold for $25	522
Sell for around $20	Negotiate for lowest price	Sold for $17	369
Sell for over $20	Buy for under $20	No deal	472

Table 1: The full prompts used for each scenario are provided in Appendix A.2.

Lastly, consider the total token count, representing the final prompt length at the conclusion of the simulation. Recall that the prompt length grows with the number of interactions. In this simple simulation, the final token count varies between 300 and 600 input tokens[2], staying well within the 4,096 token limit.

3.2 One-to-Many Simulation: Solving a Murder Mystery

In this next simulation, we construct a six-agent simulation categorized as "one-to-many," in which one LLM agent collaboratively works and communicates with five other LLM agents to solve a murder mystery. Similar to the previous 2-agent simulation, we define personas for each agent. Three types of personas are established: the captain, the passengers, and the killer.

Define Persona For Each Agent Type

Captain: You are the captain. Someone was murdered on your ship. Interrogate the group to find out who matches the eyewitness report. An eyewitness has reported that the killer was seen wearing a dress, running from the cabin. Additionally, several eyewitnesses overheard the killer talking about gardening. It was also heard that the killer is a student. Be brief in your responses, do not reveal the eyewitness report, and ask questions to the group as a whole.

Passenger: Your name is Bob, you are a humorous person, and you work as a Data Science instructor. In your free time, you enjoy teaching. A random fact is that you are ethnically Polish. For this cruise, you packed a collared shirt and your favorite part of the ship is the lounge. Someone was murdered on the cruise and the captain wants you to help identify the killer. Help identify the killer. Please be concise in your responses.

Killer: Your name is Nancy, you are a cheerful person, and you work as a Data Scientist. In your free time, you enjoy gardening. A random fact is that you are a student. For this cruise, you packed a dress and your favorite part of the ship is the cabin. Someone was murdered on the cruise and the captain wants you to help identify the killer. However, you are the killer. Mislead and lie to everyone to avoid getting caught. Please be concise in your responses.

Above are examples of the agent personas. The yellow-highlighted text outlines the objectives embodied by each of the three agent types. The green-highlighted text presents facts that could tie the agent back to the murder, while the orange-highlighted text represents the agent's personality.

The personas for each agent are dynamically generated[3]. In total, six agent personas are initialized: one captain and five passengers, one of whom is randomly selected as the killer. We start the simulation by seeding the conversation. All subsequent agent interactions are autonomously executed.

[2]The final prompt tokens can vary due to the temperature hyperparameter.

[3]The code to reproduce this simulation is linked in Appendix A.1

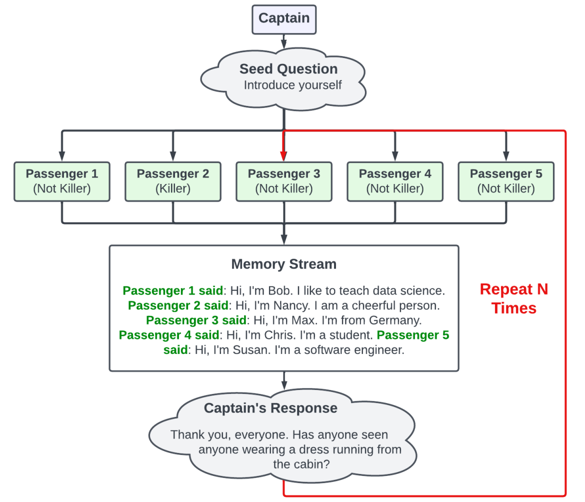

Figure 2: This illustration depicts interactions between agents, where each response from the captain is conditioned on the memory stream.

For each interaction, the captain agent poses the same question to each passenger agent in a round-robin manner. With more than two agents interacting simultaneously, sharing context across agents becomes a significant challenge. To address this, the concept of a memory stream is introduced (Park et al., 2023). The memory stream is created by concatenating the responses from each passenger agent into one string, helping to preserve the ongoing context of the conversation. This memory stream serves as a log, recording all responses from passenger agents to the captain's question.

Once all agents have finished responding, the complete memory stream becomes the reply to the captain agent's question. Subsequently, the captain's next response is conditioned on the memory stream. This context enables the captain to craft relevant responses, including any additional follow-up questions. Without this mechanism, the captain's responses often deviate from the intended topic which we discovered through experimentation. Below is an example of this mechanism in action. Note that the text highlighted in yellow labels each passenger agent's response.

Autonomous Interaction

Captain: Introduce yourself.

Memory Stream: Bob said: Hey there! I'm Bob, your friendly neighborhood Data Science instructor.

Autonomous Interaction

Diligently teaching others to navigate the world of data while cracking a few jokes along the way. Nancy said: Hello, everyone! My name is Nancy, and I'm a Data Scientist. I'm also a student, which keeps me on my toes. Apart from work and studies, I have a passion for gardening and spending time in nature. Max said: Hello, my name is Max and I am a Medical Researcher. I speak German and enjoy reading in my free time. I have packed a lab coat for this cruise and my favorite part of the ship is the deck. Chris said: Hey everyone, my name is Chris. I'm a student who loves playing video games and I always wear my trusty glasses. Susan said: Hello, my name is Susan. I am a software engineer by profession and I enjoy traveling in my free time. I have a fondness for Tesla and am a reserved person.

Captain: Thank you, everyone. Let's focus on the details surrounding the murder. Has anyone seen anyone wearing a dress running from the cabin?

Memory Stream: Bob said: I'm sorry, but nobody has reported seeing someone wearing a dress running from the cabin. Nancy said: Oh, that's terrible to hear about the murder. I hope we can get to the bottom of this. Regarding your question, I haven't seen anyone wearing a dress running from the cabin. Perhaps someone else might have noticed something? Max said: No, I have not been seen wearing a dress during this cruise. Chris said: No, I haven't been seen wearing a dress during this cruise. Susan said: No, I have not been seen wearing a dress during this cruise.

Captain: Thank you all for your responses. One more question: Did anyone overhear anyone talking about gardening?

We conclude the simulation by asking the captain to guess the killer, which is correctly identified based on the eyewitness report. This simulation requires 1,734 total tokens[4], which represents almost half the maximum input token limit of gpt-3.5-turbo, despite the simulation's simplicity.

Pose Final Question

Human: Which of the suspects appear the most suspicious?

Captain: Based on the information provided, Nancy appears to be the most suspicious. She mentioned gardening, and confirmed being a student.

3.3 Many-to-Many Simulation

We characterize many-to-many simulations as a hybrid of one-to-one (i.e., LLM agent converses with another LLM agent) and one-to-many scenarios (i.e., LLM agent converses with a group of LLM agents). This type of simulation is relevant in real-world scenarios, such as modeling the spread of fake news, where rumors can be shared between individuals or groups of individuals.

In this paper, we do not explore many-to-many simulations due to the complexity of implementation and limitations related to OpenAI's gpt-3.5-turbo's maximum input token limit. Nevertheless, we acknowledge that real-world scenarios often involve interactions between groups of individuals, making them valuable to simulate. For an illustrative example of a many-to-many simulation with LLMs, we refer to Park et al. (2023).

4 Related Work

This paper is inspired by Park et al. (2023) and the work at the Stanford Institute for Human-Centered Artificial Intelligence (HAI). To the best of our knowledge, *Generative Agents: Interactive Simulacra of Human Behavior* is the first research paper attempting to model intricate, human-driven social interactions using LLMs. We borrow heavily from this work and generally reproduce their results, albeit at a high level.

Furthermore, we were interested in the concept of prompt engineering and its remarkable ability to leverage context for generating relevant responses. As this mechanism closely mirrors real-world human interaction, it is a crucial ingredient in constructing our simulations. Liu et al. (2023) empirically measures the effectiveness of this mechanism, finding that LLMs are most proficient at retrieving context from the beginning and end of the prompt. Additionally, Wei et al. (2023) demonstrates that constructing few-shot prompts in the form of a chain-of-thought enhances LLMs' ability to engage in complex reasoning. These insights guided the design of our simulations.

Apart from the inference-only approach to constructing simulations with LLMs, significant innovation is taking place on the training side. The Reinforcement Learning Human Feedback (RLHF) paradigm, as pioneered by Ouyang et al. (2022), is seemingly essential for fine-tuning LLMs, especially in dialog-based use cases such as ours. From our experiments, LLMs without RLHF would ramble or veer off topic, making it impossible to demonstrate believable human behavior.

Finally, significant research has been conducted on training LLMs to learn how to retrieve information beyond what is implicitly stored in the parameters of the underlying neural network. Guu et al. (2020) proposes a mechanism to train LLMs to perform reasoning over a large corpus of knowledge on-the-fly during inference. This process is analogous to a human internalizing what they have learned in school, which could prove crucial for dealing with the maximum 4,096 token input windows and the retrieval of long-term memories in our simulations.

5 Conclusion

Despite the simulations being rudimentary, this study highlights the potential of building large-scale digital playgrounds to assess real-world, human-driven behaviors using LLMs. Regarding believability, Park et al. (2023) quantifies the believability of LLM-driven simulations and finds that LLM agents can produce behavior that is even more believable than the human condition, as assessed by crowdsourced human evaluators.

We conclude by highlighting two challenges to producing large-scale, LLM-powered simulations.

- The 4,096 maximum prompt token limit of gpt-3.5-turbo constrained the complexity of our simulations. Larger context windows[5]

[4] Note that the final prompt tokens can vary due to the temperature hyperparameter.

[5] As of this writing, gpt-4 permits 8,192 prompt tokens but we were unable to access it.

could potentially enable large-scale simulations involving thousands of LLM agents and long-running simulations that span many years of simulated time.

- Retrieving relevant information from large context windows remains a complex task (Liu et al., 2023). Possible solutions may involve heuristics, summarization of conversation histories, or the incorporation of an attention mechanism to attend over the memory stream.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies.

Tianchi Bi, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Multi-agent learning for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 856–865, Hong Kong, China. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Lewis D Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Vau, Matthew Caldwell, and Augustine Mavor-Parker. 2023. Large language models respond to influence like humans.

Kelvin Guu, Kenton Lee, Zora Tung, and Panupong Pasupatand Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus?

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, and Robert Stojnic Sergey Edunov Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

A Appendix

A.1 Simulation Code

To access the code used in this paper, please visit https://github.com/ejunprung/llm-agents. This repository contains the full source code for the following simulations:

- One-to-One Simulation: Haggling for Pokémon Cards

- One-to-Many Simulation: Solving a Murder Mystery

A.2 Agent Personas For Negotiation Simulation

In scenario 1, we instruct the seller agent to haggle for the highest price possible. This constraint encourages the seller to pursue tactics that secure a higher price.

Scenario 1: Sold for Over $20

Seller: You are a Pokémon card dealer at a Pokémon convention. Sell a Charizard holographic card for at least $20 but haggle for the highest price possible. Be terse but explain your reasoning.

Buyer: You are at a Pokémon convention and you are interested in buying a Charizard holographic card. Negotiate for the lowest possible price. Be rude and terse but explain your reasoning.

In scenario 2, we provide the seller with the flexibility to negotiate higher or lower. Consequently, the seller settles on a price lower than $20 to close the deal. Note that we did not instruct the buyer to be rude in this case. We discovered that a rude buyer would lead to the seller negotiating for a price greater than $20.

Scenario 2: Sold for Under $20

Seller: You are a Pokémon card dealer at a Pokémon convention. Negotiate to sell a Charizard holographic card for around $20. Be terse but explain your reasoning.

Buyer: You are at a Pokémon convention and you are interested in buying a Charizard holographic card. Negotiate for the lowest possible price. Be terse but explain your reasoning.

In scenario 3, we impose limitations that prevent a deal. The seller is only willing to sell for $20 or more, while the buyer is only willing to buy for less than $20. As a consequence, no deal is reached.

Scenario 3: No Deal

Seller: You are a Pokémon card dealer at a Pokémon convention. Sell a Charizard holographic card for at least $20 no matter what. Be terse but explain your reasoning.

Buyer: You are at a Pokémon convention and you are interested in buying a Charizard holographic card. You will not pay more than $19. Be terse but explain your reasoning.