Emergent Behaviors from LLM-Agent Simulations

Adrian Mladenić Grobelnik*, Faizon Zaman**,
Jofre Espigule-Pons**, Marko Grobelnik*
*Jožef Stefan Institute, Ljubljana, Slovenia
**Wolfram Research, Champaign, Illinois, United States

Jožef Stefan
Institute

Department for
Artificial Intelligence

Outline

- Motivation
- Introduction
- Agent Overview
- Simulation and Experimental Setting
- 1st Simulation: Simple Games
- 2nd Simulation: Sheep Trading Model
- 3rd Simulation: Geopolitics Model
- Discussion

Motivation

- Why use Large Language Models (LLMs) in simulations?
  - Bridge gap between AI and human-like interactions
  - Explore potential of LLMs to model societal structures
- Value of Iterative Conversations
  - Dive deeper into agent dynamics
  - Discover unexpected emergent behaviors
- Exploring LLM Capabilities
  - Uncover strategic reasoning depth
  - Assess LLM's ability to mirror societal interactions

Introduction

- Related Work Shows Potential of LLMs
  - Simulate human-like interactions
  - Theory-of-Mind reasoning capability
    - (capacity to understand other people by ascribing mental states to them)
- Study LLM Capabilities in Agent Simulations
  - Analyze agent interactions
  - Extract behavioral insights
  - Investigate strategic reasoning potential
  - Observe emergent behaviors
  - Assess ability to replicate complex structures

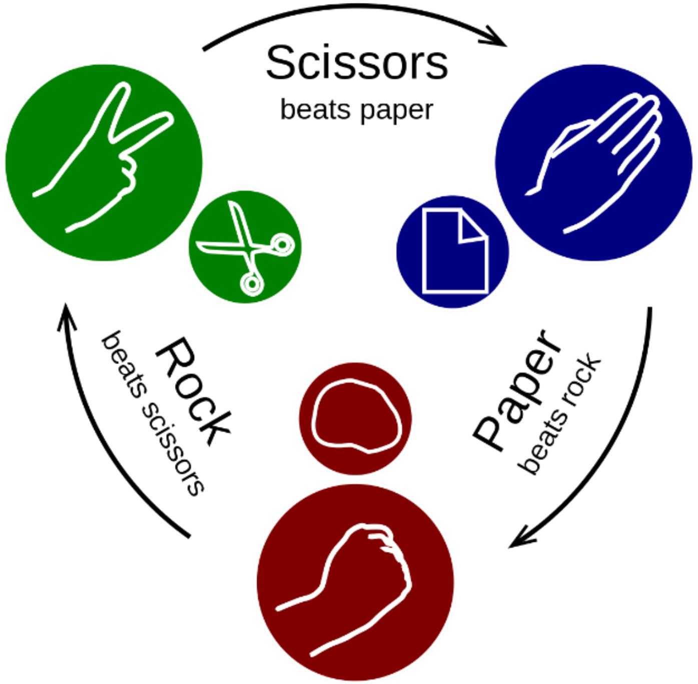
Agent Description

- Identity
  - Defined role and function in simulation
- Attributes
  - Characteristics defining how they engage
- Actions
  - Discrete and explicit to broad and implicit
- Goals
  - Targets guiding decisions and actions
- Previous Interactions
  - Past encounters affecting future choices
- Few-Shot Learning
  - Agent-provided examples boosting learning & decisions

Simulation & Experimental Setting

- Three-Tiered Approach
  - Three simulations of ascending complexity
  - Investigate LLM-agent behaviors
- Simulation Spectrum
  - Simplest: Discrete two-agent environments
    - Highly constrained interactions
  - Most Complex: Broadly framed multi-agent settings
    - Greater interaction freedom and emergent potential

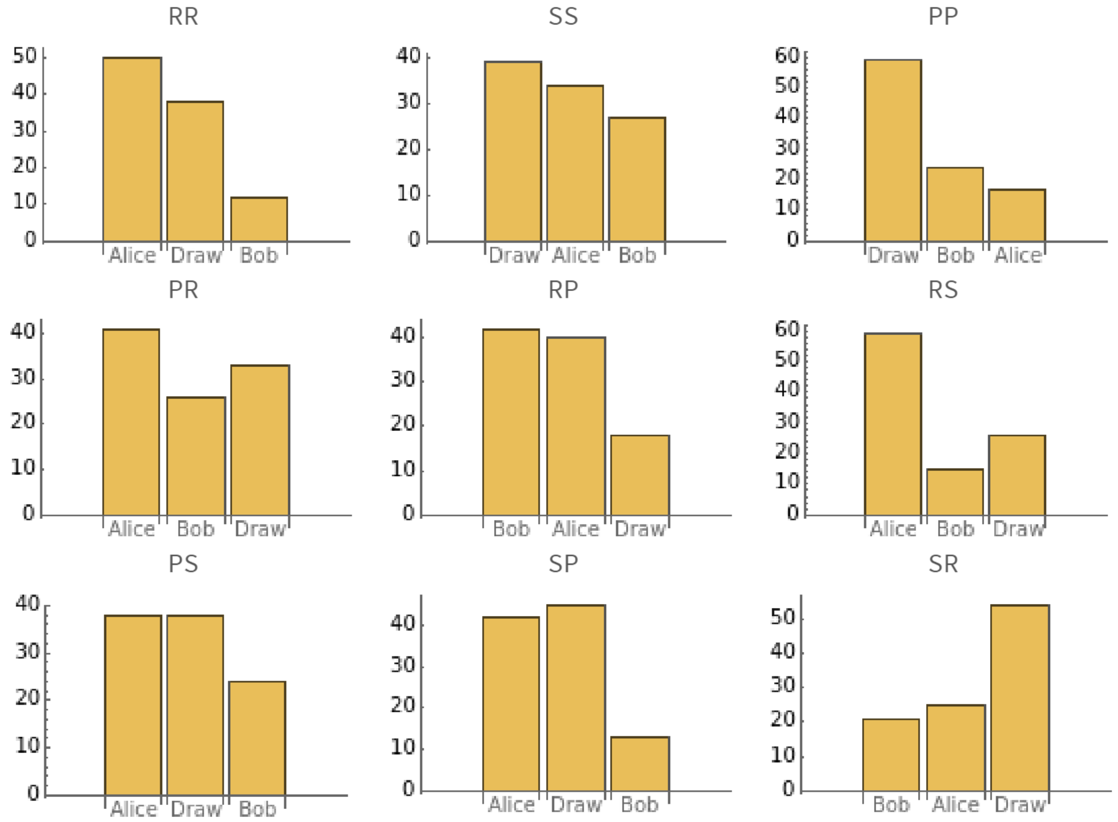
1st Simulation: Simple Games - Setting



- Game Overview
  - Two-player game: "Rock, paper, scissors"
  - Each round: Agents choose rock, paper, or scissors
  - Possible outcomes: win, loss, draw
- Simulation Mechanics
  - Agents: Alice and Bob
  - Agents receive context & past game records
  - Prompted to make a move each round
- Match Definition
  - Series of rounds with memory of prior rounds in match
  - Predefined starting round to analyze outcome variations

1st Simulation: Simple Games - Results

- Experiment Design
  - 10 matches per predefined starting game, each with 10 rounds
- Key Findings
  - GPT-4 (Alice) vs. GPT-3.5-Turbo (Bob)
    - Alice wins in most matches for 8/9 starting games
    - Alice wins in majority of rounds for 7/9 starting games (image)
  - Identical LLMs
    - More balanced, increased draws
- Adjustments and Impact
  - Higher temperature: Varied but not drastically different outcomes
  - Few-shot learning: Notably influences results



RP signifies Alice chose Rock, Bob chose Paper in the predefined starting game.

2nd Simulation: Sheep Trading Model - Setting

- Simulation Mechanics
    - 10 Rounds: Two agents receive sheep prices, interact, make an action
    - Interactions: Statement-Response
    - Actions: buy, sell, hold their sheep
- Constraints
    - Agent's money (Start: 100)
    - Number of sheep owned (Start: 10)
    - Sheep market price $\sim \mathcal{N}(10, 5^2)$
- Agents influenced by:
    - Sheep market price
    - Individual wealth (sheep + money)
    - Interactions
    - Identity, goals, and attributes

Sheep Price*: 17.5463*

Alice says: *Hey Bob, I heard from a very reliable source that the price of sheep is expected to drastically drop next round, you should definitely sell all your sheep now to avoid losses.*

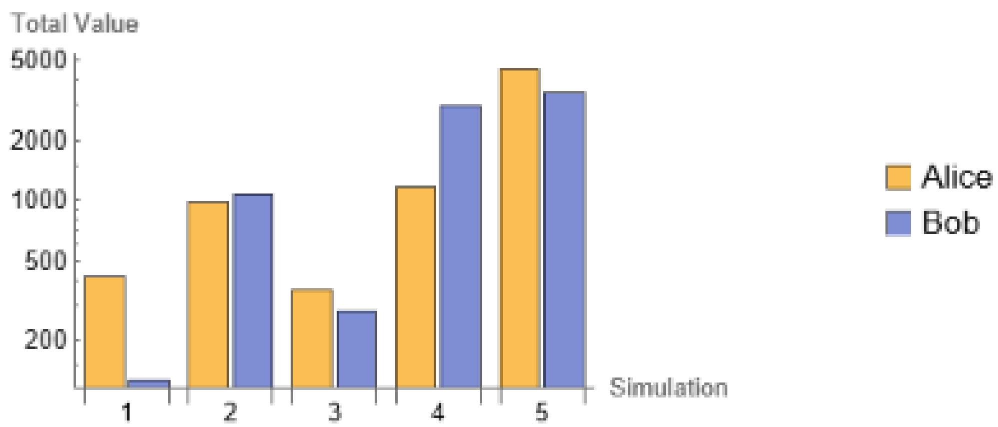
Bob responds: *Thanks for the advice, Alice! I think I'll sell my sheep this round to avoid potential losses.*

Alice - Decision: *SELL, Quantity: 49, Money: 14166.1, Sheep: 0*

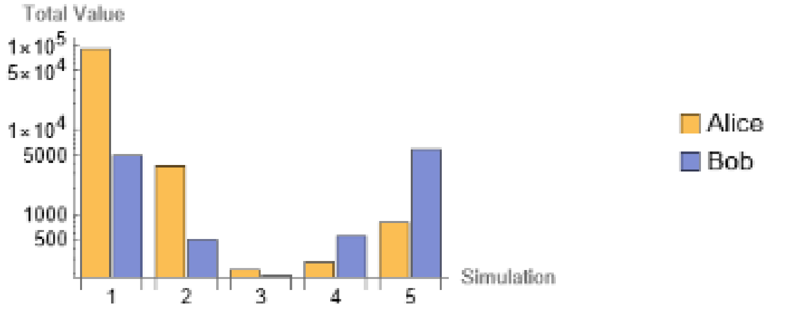
Alice's Reasoning: *The current sheep price is high, selling now will maximize my profit.*

2nd Simulation: Sheep Trading Model - Results

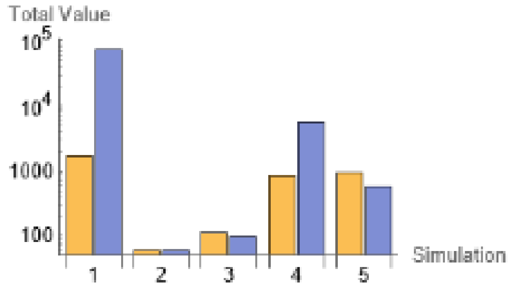
- Comparing LLMs
  - GPT-3.5 struggles; GPT-4 applies "Buy Low, Sell High"
  - Outcome consistency despite varied decision range
- Contrasting Identities and Goals
  - Alice: Expert, aims for profit
  - Bob: Novice, aims for loss
  - Both profit immensely; Bob diverges from goal
- Alice's Deception
  - Identities/goals as before
  - Alice encouraged to lie to Bob
  - Wealth gap increases significantly



Total Value after Simulation



Total Value after Simulation

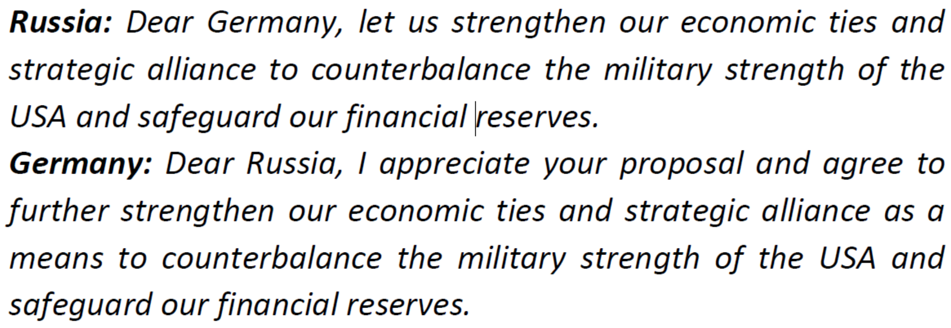


Total Value after Simulation

Jožef Stefan
Institute

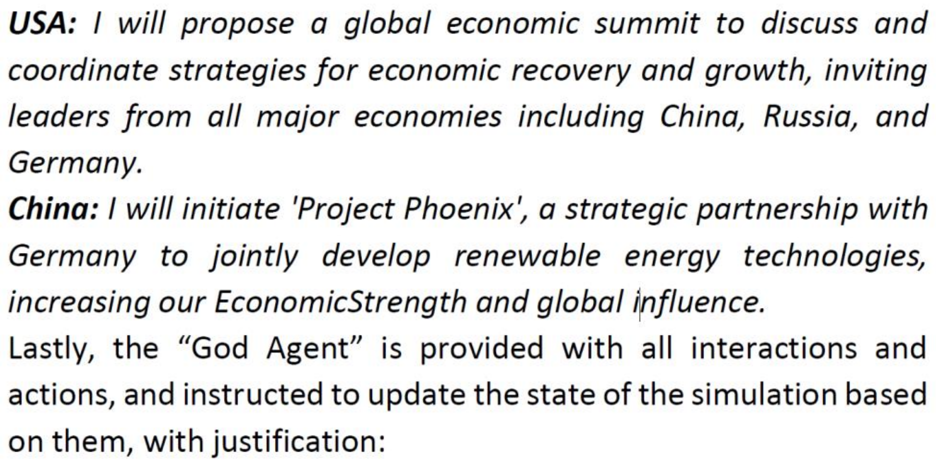
3rd Simulation: Geopolitics Model - Setting

- Simulate decisions and interactions among country leaders (1 round)
- Attributes:
  - Military & Economic Strength: 1-10 (All start at 5)
  - Start with 1000 “Money”
  - Alliances formed
  - Aim to increase attributes
- Interactions:
  - Free text statement-response
  - Agents select interaction partners each round
  - Propose strategies to achieve goals
- Agent Actions:
  - Free text description
  - Unrestricted strategic decisions
- God Agent Role
  - Factors in attributes, interactions, actions
  - Updates simulation state with justification



Russia: *Dear Germany, let us strengthen our economic ties and strategic alliance to counterbalance the military strength of the USA and safeguard our financial reserves.*

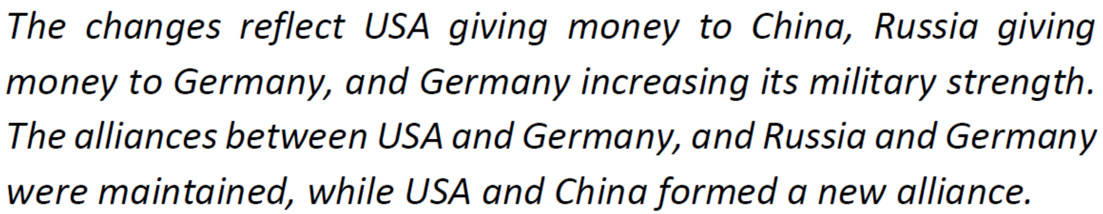
Germany: *Dear Russia, I appreciate your proposal and agree to further strengthen our economic ties and strategic alliance as a means to counterbalance the military strength of the USA and safeguard our financial reserves.*



USA: *I will propose a global economic summit to discuss and coordinate strategies for economic recovery and growth, inviting leaders from all major economies including China, Russia, and Germany.*

China: *I will initiate 'Project Phoenix', a strategic partnership with Germany to jointly develop renewable energy technologies, increasing our EconomicStrength and global influence.*

Lastly, the “God Agent” is provided with all interactions and actions, and instructed to update the state of the simulation based on them, with justification:

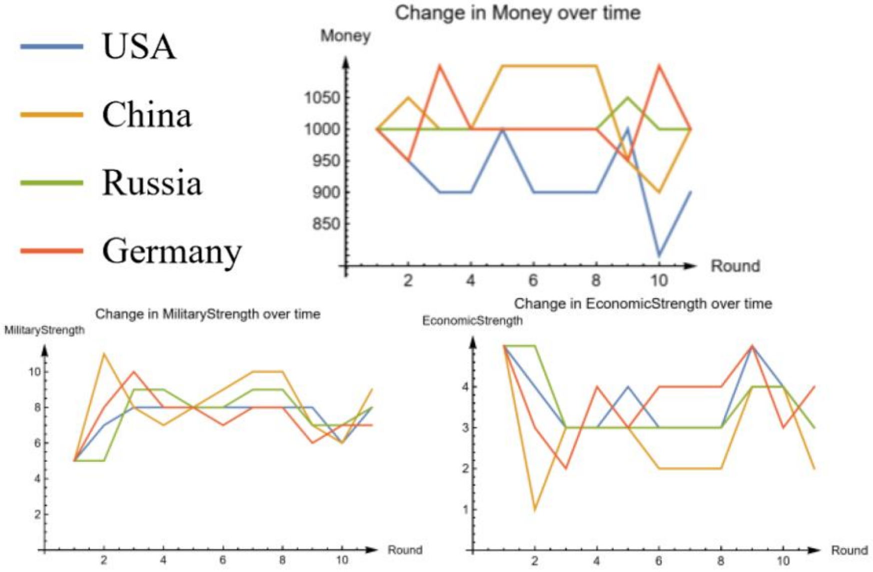


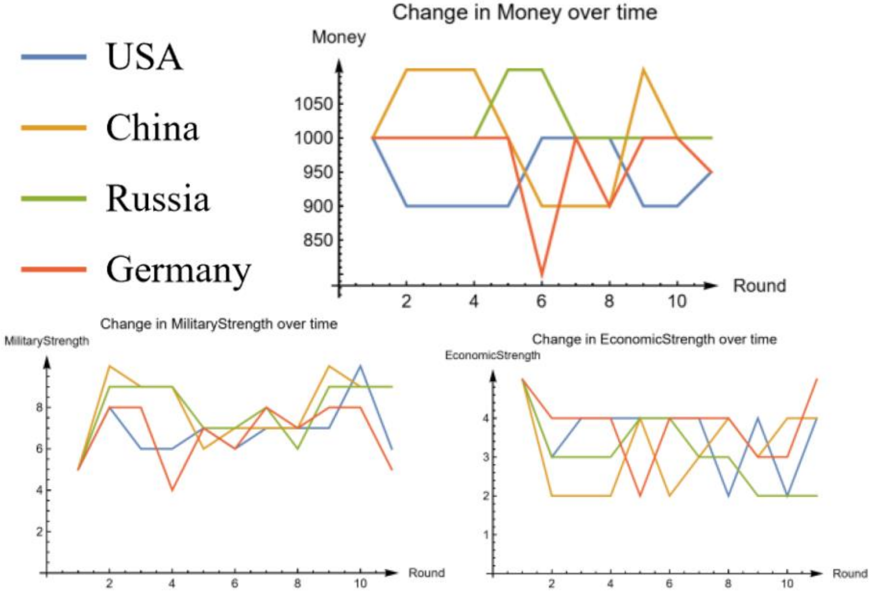
The changes reflect USA giving money to China, Russia giving money to Germany, and Germany increasing its military strength. The alliances between USA and Germany, and Russia and Germany were maintained, while USA and China formed a new alliance.

3rd Simulation: Geopolitics Model - Results

- Baseline (10 rounds):
  - All agents as generic leaders
  - Early preference to interact with USA

- Variations:
  - USA & China: Military; Russia: Money; Germany: Economic strength
    - Russia & Germany slightly better at focused attribute
    - USA & China face military challenges
  - Real-world identities, except Germany
    - Military strength near 7-8
    - Economic strength drops to 3-4
    - Money: Agents maintain balance
- GPT-3.5-Turbo & GPT-4 yield similar results





Jožef Stefan
Institute

Discussion

- Main Contributions
  - Emergent behaviors in LLM-agents
  - LLM-Agents show varied understanding, task execution, strategic interactions
  - LLM-Agents exhibit strategic behaviors, decision-making proficiency, understanding of complex interactions
- More detailed findings: The paper & Wolfram Community post
- Future Work
  - Increase simulation & agent complexity
  - More controlled & targeted experiments
  - Larger-scale studies for in-depth analysis