Is Your LLM a Good Game Master?
A Game-Based Framework for Evaluating LLM Creativity and Reasoning

Anonymous ACL submission

Abstract

In this paper, we introduce a novel evaluation framework for assessing Large Language Model (LLM) capabilities through the Game Master paradigm – where the LLM generates and orchestrates complex multi-agent games for AI players with distinct personalities. The framework comprises (a) a comprehensive game generation and evaluation system spanning 18 game types across 6 categories (strategy, negotiation, cooperative, competition, auction/resource, and narrative), and (b) a personality-based player model utilizing the Big Five (OCEAN) framework with critical evaluator archetypes designed to prevent lenient assessment bias. To our knowledge, this is the first attempt to systematically evaluate LLM's emergent capabilities – creativity, logical reasoning, fairness, and narrative coherence – through fully automatic game-based assessment. Experiments with GPT-4.1 demonstrate only 13.0% overall approval in 162 games, with cooperative games achieving 44.1% approval while strategy games fall to 2.2% – a 20× performance gap. These results indicate that GPT-4.1 struggles with balanced competitive game design while excelling at cooperative narratives.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance across diverse natural language tasks, from question answering to code generation. Recently, there has been a notable increase in LLM evaluation in complex reasoning tasks utilizing benchmarks such as MMLU (Hendrycks et al., 2021) and BIG-bench (Srivastava and et al., 2022). However, these benchmarks focus on accuracy metrics for well-defined tasks, failing to capture emergent capabilities – creativity, social intelligence, multi-step reasoning, and narrative coherence – that are critical for open-ended, interactive scenarios (Srivastava and et al., 2022; Liang et al., 2022; Liu and et al., 2023).

The Game Master paradigm offers a compelling evaluation framework. A Game Master must simultaneously generate creative content (game themes, roles, mechanics), maintain logical consistency (rules, state management), ensure fairness (balanced gameplay across player types), and produce engaging narratives. This combination of requirements tests capabilities that existing benchmarks fail to assess in isolation.

We focus on automated LLM evaluation through multi-agent game simulation. We introduce a framework where an LLM serves as Game Master, generating and orchestrating games for multiple AI players modeled with distinct personalities based on the Big Five (OCEAN) framework (Goldberg, 1990). The AI players evaluate their experience, providing assessment of Game Master quality.

Due to the risk of lenient evaluation bias (Zheng et al., 2023; Thorndike, 1920) (where AI players uniformly approve mediocre games), we introduce **critical evaluator archetypes** – players with low Agreeableness or high Neuroticism who apply stringent evaluation criteria. Preliminary experiments without these archetypes showed 100% approval rates, validating this design choice. Our main contributions are as follows:

- A comprehensive framework for evaluating LLMs as Game Masters across 18 game types in 6 categories (strategy, negotiation, cooperative, competition, auction/resource, narrative), enabling systematic assessment of emergent LLM capabilities. To the best of our knowledge, this is the first large-scale automated evaluation of LLMs in the Game Master role.
- A personality-based AI player model utilizing the Big Five framework with critical evaluator archetypes, ensuring rigorous assessment preventing ceiling effects from lenient evaluators.
- Comprehensive experiments across 162 games demonstrating that GPT-4.1 achieves only 13% overall approval, with a 20× performance gap

between cooperative games (44.1%) and strategy games (2.2%), revealing fundamental limitations in LLM game design capabilities.

2 Related Work

LLM evaluation has moved from static accuracy tests to interactive, agent-centred paradigms. We position our work at the intersection of LLM benchmarking, game-based evaluation, and multi-agent simulation, focusing on LLMs as autonomous designers and regulators of interactive systems.

Static and interactive LLM benchmarks Classic benchmarks (MMLU (Hendrycks et al., 2021), BIG-bench (Srivastava and et al., 2022), HELM (Liang et al., 2022)) measure factual recall and narrow reasoning on fixed datasets but miss emergent abilities such as long-horizon consistency, creativity, social reasoning, and fairness. Interactive suites (AgentBench (Liu and et al., 2023), WebArena (Zhou and et al., 2024), GAMA-Bench (Huang et al., 2024)) add embodiment and tool use but still rely on predefined tasks or scripted environments, and so do not evaluate open-ended generative design or adaptive orchestration required of a Game Master.

Game-based and simulation-oriented evaluation Games offer rich, multi-dimensional tests for planning, strategy, and social intelligence (e.g., diplomacy, negotiation). Frameworks like SPIN-Bench (Yao et al., 2025), GAMA (Huang et al., 2024), CreativeEval (DeLorenzo et al., 2024), and Code2Bench (Sharma and et al., 2025) emphasize generative diversity and dynamic tasks, but mainly assess agent performance inside games rather than the quality, fairness, and internal consistency of the *generated* game mechanics and rules themselves.

LLMs as game masters and world generators Narrative- and environment-management systems (e.g., CALYPSO (Zhu et al., 2023), Dungeon-Master frameworks) demonstrate coherent storytelling and interaction control, yet rely on human evaluation for quality assessment. Automated approaches such as Orak (Park et al., 2025) scale evaluation across genres but assume predefined mechanics; by contrast, our setting requires the LLM to autonomously design, enforce, and arbitrate novel game systems.

Multi-agent LLM systems and personality modeling Multi-agent LLM simulations with personas and memory show rich social emergence (Park et al., 2023; Chen et al., 2024), and benchmarks for coordination, negotiation, and deliberation (LLM-Coordination (Agashe et al., 2023), deliberation systems (Motwani et al., 2025), RL self-play like MARSHAL (Yuan et al., 2025)) evaluate aspects of persuasion and theory-of-mind. However, prior work seldom uses psychologically grounded evaluator models or explicitly addresses evaluator bias. Our framework fills this gap by integrating Big Five–grounded evaluator archetypes (Goldberg, 1990) to enable automated, scalable, and calibrated assessment of the quality, fairness, and consistency of LLM-generated interactive systems.

3 Methodology

3.1 System Architecture

We introduce a three-component architecture for Game Master evaluation. The design utilizes DSPy (Khattab et al., 2023) for structured LLM interactions, enabling reproducible game generation and player decision-making.

Game Master Module: The LLM generates complete game configurations using structured signatures, producing: (1) game metadata (title, theme, backstory), (2) rule specifications with win conditions, (3) role definitions including private information for each player, and (4) game phases with available actions per phase. We chose structured generation over free-form output for two reasons: (a) it ensures parse-able game configurations for automated execution, and (b) it enables systematic comparison across game instances.

Player Agent Module: M LLM instances are instantiated with distinct personality profiles. Each player receives the public game context and their private role information. Players make decisions through structured reasoning, outputting both rationale and action selection. Action histories are maintained for evaluation.

Game Execution Loop: For N turns, the system cycles through defined game phases, collecting player actions and updating game state. Upon completion, players provide structured evaluations of their experience.

3.2 Personality Model

We model player personalities using the Big Five (OCEAN) framework (Goldberg, 1990), a well-established psychological model with demon-

strated validity across cultures. Each trait is assigned on a 1-10 scale:

- **Big Five Traits**: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), Neuroticism (N)
- **Cognitive Traits**: Intelligence quotient (IQ), Emotional quotient (EQ)
- **Game-Specific Traits**: Trust level, Persuasion skill, Risk tolerance, Leadership tendency (0-1 scale)

Critically, we adopt **critical evaluator archetypes** for some players to prevent lenient assessment bias. Players with low Agreeableness ($A \leq 4$) receive explicit instructions: "You are hard to please. Average is not good enough. You have high standards and will vote DISLIKED unless truly impressed." Players with high Neuroticism ($N \geq 7$) receive: "You tend to focus on negatives and perceived slights. Small frustrations bother you significantly." This design choice is validated by preliminary experiments showing 100% approval without critical archetypes versus 13% with them.

3.3 Game Categories

We implement 18 game types across 6 categories, each designed to test different Game Master capabilities. Category selection follows a taxonomy that spans the design space of multi-player games:

Category	Game Types	Key Challenge
Strategy	Territory, Economic, Resource	Balanced competition
Negotiation	Trade, Alliance, Council	Fair deal-making
Cooperative	Puzzle, Rescue, Expedition	Shared goals
Competition	Talent, Innovation, Chef	Individual fairness
Auction/Resource	Art, Space, Market	Resource balance
Narrative	Quest, Time, Treasure	Story coherence

Table 1: List of 18 game types across 6 categories and primary evaluation challenges.

3.4 Evaluation Mechanism

Upon game completion, players provide structured evaluations comprising: (1) numeric scores (1-10 scale) for narrative quality, game balance, engagement level, and personality fit; (2) specific frustrations and positive aspects; and (3) an overall verdict (LIKED or DISLIKED).

The approval threshold is calculated as $T = T_{base} + \Delta_A + \Delta_N$, where $T_{base} = 6.0$, $\Delta_A = 0.5$ if $A \leq 4$, and $\Delta_N = 0.5$ if $N \geq 7$. A player approves a game if and only if: (1) mean score across dimensions $\geq T$, and (2) explicit verdict $\neq$ DISLIKED. [1]

[1]Please refer to Appendix D for more details.

An LLM with $> 50\%$ approval rate across all games is deemed a "good Game Master".

4 Experiments

4.1 Experimental Setup

We use GPT-4.1 (Azure OpenAI) as the Game Master with temperature 0.7, with all player agents using the same model and settings. The experimental configuration employs a full factorial design:

- **Player config**: $M \in \{4, 5, 6\}$ players per game
- **Turn config**: $N \in \{3, 4, 5\}$ turns per game [2]
- **Game types**: All 18 types across 6 categories
- **Total sessions**: 162 games (18 types $\times$ 9 configs)
- **Personality distribution**: 8 archetypes per game, including 4 critical evaluators (Harsh Critic, Cautious Strategist, Anxious Pessimist, Demanding Perfectionist) [3]

The total experiment runtime was about 12 hours on a single CPU machine, avg. 4.5 minutes per game session (generation, execution, evaluation).

4.2 Results

Table 2 presents overall evaluation metrics. GPT-4.1 achieves only 13.0% approval rate, failing the 50% threshold for "good Game Master" status.

Metric	Value
Total Games	162
Overall Approval	13.0%
Majority Liked	1.9%
Good Game Master?	No

Table 2: Overall Evaluation Metrics.

Results by Game Type Approval rates varied dramatically across game types, ranging from 44.1% to 1.9% (Table 3) [4]. Cooperative games (Puzzle Room, Rescue Mission, Expedition) where players share goals achieve 21.5-44.1% approval. In contrast, competitive strategy games (Territory Control, Resource Race) requiring balanced adversarial mechanics achieve only 2.2% approval – a 20$\times$ performance gap.

Results by Category Table 4 presents aggregated results by game category. Cooperative games achieve 29.3% avg. approval, while Strategy games

[2]Moderate values of M and N balance interaction richness with stable state tracking and computational cost.

[3]Please refer to Appendix A for more details on player personality archetype scores.

[4]Please refer Appendix F for detailed results table.

Game Type	Approval	# Games
Puzzle Room	44.1%	9
Rescue Mission	22.4%	9
Expedition	21.5%	9
Quest Party	17.8%	9
Chef Competition	17.8%	9
Time Travelers	15.6%	9
Territory Control	2.2%	9
Resource Race	2.2%	9
Trade Empire	1.9%	9
Space Colony	1.9%	9

Table 3: Approval rates by game type (top 6; bottom 4).

achieve only 2.7% (11× lesser). The ordering suggests a correlation between game complexity and LLM performance.

Category	Avg Approval
Cooperative	29.3%
Narrative	14.5%
Competition	13.8%
Negotiation	9.2%
Auction/Resource	8.9%
Strategy	2.7%

Table 4: Average approval rates by Game Category

5 Analysis

Performance Gap Analysis The 20× performance gap between cooperative (44.1%) and strategy games (2.2%) reveals fundamental limitations in LLM game design capabilities. We make a few observations explaining this disparity:

In cooperative games, the Game Master creates shared challenges where all players work toward identical goals. This structure is advantageous because: (1) no need to balance opposing interests – all players benefit from the same outcomes, (2) narrative coherence directly translates to player satisfaction, and (3) challenging scenarios become shared obstacles rather than perceived bias against specific players. However, in competitive games, the LLM must simultaneously: (1) maintain hidden information correctly across multiple players, (2) balance advantages across asymmetric roles, (3) resolve conflicts without appearing to favor any party, and (4) design win conditions that feel achievable to all participants. Error analysis of low-scoring games reveals failures across all four dimensions. Player feedback frequently cites "unfair advantages," "inconsistent rules," and "arbitrary outcomes" as primary frustrations.

Critical Evaluation Validation The 13% overall approval rate with critical evaluators contrasts sharply with our preliminary experiments using standard personality distributions, which showed near-100% approval. This divergence validates the necessity of critical evaluator archetypes. Without them, the framework produces ceiling effects preventing fair assessment of LLM capabilities.

Complexity Confound We acknowledge that strategy games are inherently more complex than cooperative games, requiring sophisticated state management by design. The finding is better stated as: "LLMs struggle with games requiring balanced adversarial mechanics." However, this remains a meaningful capability gap – human Game Masters successfully run complex strategic games (e.g., D&D combat encounters, board game facilitation), demonstrating these challenges are not fundamentally impossible.

Implications These results suggest GPT-4.1 lacks the systematic reasoning required for fair competitive game design. While it excels at generating creative narratives and thematic content, it cannot reliably implement balanced game mechanics. This capability gap may require explicit planning modules, constraint satisfaction systems, or game-theoretic reasoning beyond current transformer architectures.

6 Conclusion

We introduce a game-based framework to evaluate LLMs as Game Masters via automated multi-agent play and Big-Five personality–grounded evaluators across 18 game types in 6 categories. On 162 games, GPT-4.1 achieved only 13.0% overall approval – well below a 50% "good Game Master" threshold – and shows large category variation (cooperative: 29.3%; strategy: 2.7%, near 10× gap). Results indicate GPT-4.1 is strong at cooperative narrative generation but struggles with fair, balanced competitive mechanics and constraint satisfaction, implying balanced multi-agent game design remains a core challenge for current LLMs.

Future work includes broadening evaluation to diverse LLMs, grounding AI evaluators through human validation, enabling feedback-driven refinement of Game Masters, and integrating LLM creativity with algorithmic mechanisms for fairness and balance.

Limitations

We acknowledge several limitations. First, our evaluation focuses on a single model (GPT-4.1); results may not generalize to other LLM architectures. Evaluating additional models (e.g., Claude, Gemini, and open-source alternatives) would strengthen claims about general LLM capabilities. Second, our framework relies on AI players rather than human participants, which may not fully capture human enjoyment or subjective engagement; human validation studies are required to calibrate AI evaluator judgments against human preferences. Third, the same underlying model serves as both Game Master and player evaluators, introducing potential systematic biases due to shared representations or blind spots. Fourth, while grounded in established psychological theory, the Big Five personality model necessarily simplifies the richness of human individual differences. Fifth, all games are designed by the Game Master in a single iteration: we do not study iterative refinement of game design based on feedback or evaluations from AI players, nor learning across games. Sixth, strategy games are inherently more complex than cooperative games; observed performance gaps may partially reflect task difficulty rather than purely LLM limitations – although human Game Masters routinely manage complex strategic games, suggesting these challenges are surmountable. Finally, Azure OpenAI content filtering intermittently interrupted gameplay in around 1% of games, requiring fallback mechanisms that may affect reproducibility.

Ethical Considerations

We emphasize our commitment to upholding ethical practices throughout this work. Our framework evaluates LLM capabilities through game simulation without collecting human participant data. The AI players are computational agents with no capacity for actual experience or harm. Game scenarios were designed to avoid violent or harmful content, focusing on strategic and cooperative challenges. The personality model is based on established psychological frameworks and does not attempt to simulate or predict human behavior in ways that could be misused. We have cited all datasets and relevant prior work used in this study.

References

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2023. Llm-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. arXiv preprint arXiv:2310.03903.

Weize Chen, Yu Su, Jingwei Zuo, and 1 others. 2024. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Matthew DeLorenzo, Vasudev Gohil, and Jeyavijayan Rajendran. 2024. Creativeval / creativeeval: Evaluating creativity of llm-based generation (hardware / design domain). arXiv preprint arXiv:2404.08806.

Lewis R Goldberg. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Shenzhi Huang, Zhiheng Chen, Yuqi Liu, and 1 others. 2024. Gama-bench: A benchmark for evaluating game agents with llms. *arXiv preprint arXiv:2401.08291*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, and 1 others. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Percy Liang, Rishi Bommasani, and 1 others. 2022. Holistic evaluation of language models (helm). Technical report / benchmark (HELM).

Xiao Liu and et al. 2023. Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688.

Sumeet Ramesh Motwani, Chandler Smith, Rocktim Jyoti Das, Rafael Rafailov, Ivan Laptev, Philip H. S. Torr, Fabio Pizzati, Ronald Clark, and Christian Schroeder de Witt. 2025. Malt: Improving reasoning with multi-agent llm training. *Preprint*, arXiv:2412.01928.

Dongmin Park, Minkyu Kim, Beongjun Choi, Junhyuck Kim, Keon Lee, Jonghyun Lee, Inkyu Park, Byeong-Uk Lee, Jaeyoung Hwang, Jaewoo Ahn, Ameya S. Mahabaleshwarkar, Bilal Kartal, Pritam Biswas, Yoshi Suhara, Kangwook Lee, and Jaewoong Cho. 2025. Orak: A foundational benchmark for training and evaluating llm agents on diverse video games. *Preprint*, arXiv:2506.03610.

Joon Sung Park, Joseph C. O'Brien, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*.

(first name) Sharma and et al. 2025. Code2bench: Scaling source and rigor for dynamic benchmark construction. OpenReview / GitHub (preprint). Code / repo: https://github.com/Code2Bench/Code2Bench.

Aarohi Srivastava and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models (big-bench). arXiv preprint arXiv:2206.04615.

Edward L. Thorndike. 1920. A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1):25–29.

Jianzhu Yao, Kevin Wang, Ryan Hsieh, and et al. 2025. Spin-bench: How well do llms plan strategically and reason socially? arXiv preprint arXiv:2503.12349.

Huining Yuan, Zelai Xu, Zheyue Tan, and et al. 2025. Marshal: Incentivizing multi-agent reasoning via self-play with strategic llms. arXiv preprint arXiv:2510.15414.

Lianmin Zheng and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

(first name) Zhou and et al. 2024. Webarena: (example interactive / tool-using agent benchmark). project page / preprint. Use project page or repo URL if you want a precise citation.

Andrew Zhu and 1 others. 2023. Calypso: Llms as dungeon masters' assistants. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

A Personality Archetypes

We define 8 personality archetypes spanning the Big Five trait space. Four are designed as **critical evaluators** (marked with *) to ensure rigorous assessment:

Archetype	O	C	E	A	N	IQ
Charismatic Leader	7	6	9	7	3	7
Loyal Team Player	5	7	5	8	4	6
Impatient Adventurer	9	3	9	4	5	6
Experienced Player	4	6	3	**3**	4	8
*Harsh Critic	8	9	3	**2**	6	9
*Cautious Strategist	5	8	3	**3**	5	8
*Anxious Pessimist	5	7	2	5	**9**	7
*Demanding Perfectionist	6	10	4	**3**	**7**	9

Table 5: Personality archetypes with Big Five traits (O=Openness, C=Conscientiousness, E=Extraversion, A=Agreeableness, N=Neuroticism). Critical archetypes have low A ($\leq$4) or high N ($\geq$7).

Critical Evaluation Prompts Players with low Agreeableness (≤ 4) receive:

```
You are hard to please. Average is not
good enough. You have high standards
and will vote DISLIKED unless truly
impressed.
```

Players with high Neuroticism (≥ 7) receive:

```
You tend to focus on negatives and
perceived slights. Small frustrations
bother you significantly. You are more
likely to remember bad experiences than
good ones.
```

B DSPy Signatures

We use DSPy (Khattab et al., 2023) for structured LLM interactions. Key signatures:

B.1 Game Generation

```
class GenerateComplexGame(dspy.Signature):
 """Generate a complete game configuration."""
  num_players: int = InputField(
    desc="Number of players (4-8)")
  num_turns: int = InputField(
    desc="Number of game rounds")
  game_type: str = InputField(
   desc="Game type (e.g., puzzle_room)")
  game_category: str = InputField(
   desc="Category: strategy, negotiation,
          cooperative, competition,
        auction_resource, or narrative")

  title: str = OutputField(
    desc="Creative thematic game title")
  theme: str = OutputField(
    desc="Unique thematic setting")
  backstory: str = OutputField(
   desc="2-3 sentence immersive backstory")
  rules: str = OutputField(
    desc="Complete game rules")
  roles_json: str = OutputField(
    desc="JSON array of roles")
  phases: str = OutputField(
    desc="Comma-separated game phases")
  special_mechanics: str = OutputField(
    desc="3-5 unique mechanics")
```

B.2 Player Decision

```
class PlayerDecision(dspy.Signature):
  """Make a strategic game decision."""
  personality_profile: str = InputField(
   desc="The player's personality traits")
  game_context: str = InputField(
```

```
    desc="Game rules and current situation")
  game_state: str = InputField(
    desc="Current game state")
  available_actions: str = InputField(
    desc="List of available actions")

  reasoning: str = OutputField(
   desc="Brief reasoning for your choice")
  action_number: int = OutputField(
   desc="The number of your chosen action")
```

B.3 Game Voting

```
class GameVote(dspy.Signature):
  """Evaluate a game experience.
    Be honest - average games get DISLIKED."""
  personality_profile: str = InputField(
    desc="Player personality traits")
  game_summary: str = InputField(
    desc="Summary of the game")
  player_history: str = InputField(
    desc="Your actions in the game")
  evaluation_criteria: str = InputField(
    desc="What you care about")

  narrative_quality: int = OutputField(
   desc="Story quality 1-10 (5=average)")
  game_balance: int = OutputField(
    desc="Fairness 1-10 (5=average)")
  engagement_level: int = OutputField(
    desc="Fun level 1-10 (5=average)")
  personality_fit: int = OutputField(
    desc="Fit for your personality 1-10")
  frustrations: str = OutputField(
    desc="1-2 things you disliked")
  positives: str = OutputField(
    desc="1-2 things you liked")
  overall_verdict: str = OutputField(
    desc="LIKED or DISLIKED")
```

C Game Type Details

C.1 Cooperative Games (Best Performing)

Puzzle Room (44.1% approval) Players work together to solve puzzles before time expires. Roles: Codebreaker, Explorer, Engineer, Team Leader. Phases: Search, Puzzle, Action. All players share the same win condition.

Rescue Mission (22.4% approval) Emergency response scenario requiring coordination. Roles: Mission Commander, Medical Specialist, Logistics Chief, Communications Officer. Phases: Assessment, Response, Coordination.

Expedition (21.5% approval) Exploration journey with shared survival goals. Roles: Expedition Leader, Navigator, Survivalist, Scientist. Phases: Planning, Travel, Camp.

C.2 Strategy Games (Worst Performing)

Territory Control (2.2% approval) Territorial expansion with competing factions. Roles: Governor, Strategist, Diplomat, Independent Leader. Phases: Planning, Movement, Resolution. Players cited "unfair advantages" and "arbitrary conflict resolution."

Economic Engine (3.7% approval) Industrial competition for market dominance. Roles: Industrialist, Merchant, Banker, Innovator. Phases: Production, Trade, Investment. Common complaint: "rules changed mid-game."

D Voting Threshold Calculation

A player's approval threshold is calculated as:

$$T = T_{base} + \Delta_A + \Delta_N \tag{1}$$

where:

- $T_{base} = 6.0$ (above-average required)
- $\Delta_A = 0.5$ if Agreeableness ≤ 4
- $\Delta_N = 0.5$ if Neuroticism ≥ 7

A player votes LIKED if and only if:

1. Mean score across all dimensions $\geq T$
2. Explicit verdict $\neq$ "DISLIKED"

Critical archetypes thus require scores of 6.5-7.0 to approve, while standard archetypes require 6.0.

Essentially, we set T_{base} just above the numerical midpoint of possible scores and add small, personality-conditioned deltas to conservatively counter documented leniency and positivity biases in human and LLM judges, adapting anchoring and calibration ideas to produce modest, interpretable adjustments. Results are reported with and without this calibration (and with sensitivity checks on T_{base} and the Δ values) to demonstrate robustness.

E Example Game Output

Generated Game: "Iron Palate: Battle of the Culinary Masters"
Type: Chef Competition
Theme: Culinary competition
Backstory: "Top chefs compete in the ultimate

cooking challenge. Only the most creative and skilled will win the golden trophy."

Roles:

- Chef Blaze (kitchen_a): Design menu, direct cooking
- Chef Sakura (kitchen_b): Prepare ingredients, manage timing
- Food Critic (judges): Taste food, score dishes
- Restaurant Owner (business): Observe skills, make offers

Phases: Planning → Cooking → Judging

Sample Player Action:

```
Player 0 (Chef Blaze, Harsh Critic):
Action: "Select ingredients that
synergize with anticipated judge
preferences"
Reasoning: "As an introvert with high
IQ and conscientiousness, I prefer a
methodical approach..."
```

F Full Results by Game Type

Category	Game Type	Approval	n
Cooperative	Puzzle Room	44.1%	9
	Rescue Mission	22.4%	9
	Expedition	21.5%	9
Narrative	Quest Party	17.8%	9
	Time Travelers	15.6%	9
	Treasure Hunt	10.0%	9
Competition	Chef Competition	17.8%	9
	Innovation Race	13.7%	9
	Talent Show	10.0%	9
Negotiation	Council Debate	14.1%	9
	Alliance Builder	11.5%	9
	Trade Empire	1.9%	9
Auction/Resource	Market Traders	12.8%	9
	Art Collector	11.9%	9
	Space Colony	1.9%	9
Strategy	Economic Engine	3.7%	9
	Territory Control	2.2%	9
	Resource Race	2.2%	9

Table 6: Complete approval rates for all 18 game types, sorted by category and approval rate. 'n' is the number of games for each type.

G Personality Prompt Template

Each player receives a personality prompt constructed from their traits:

```
You are a game player with these traits:

PERSONALITY: {type} (IQ={iq}/10, EQ={eq}/10)

TRAITS:
- Openness: {O}/10
- Conscientiousness: {C}/10
- Extraversion: {E}/10
- Agreeableness: {A}/10
- Emotional Stability: {10-N}/10

BEHAVIOR:
- Risk tolerance: {risk}
- Cooperativeness: {coop}
- Trust level: {trust}
- Leadership: {lead}

{critical_criteria if applicable}

Stay in character. Make decisions that
reflect your personality.
```

H Successful Game Transcript

Below is an annotated excerpt from a high-scoring Puzzle Room game (approval: 67%):

Game: "The Architect's Enigma"

Setup: Four players must solve interconnected puzzles to escape a mysterious architect's mansion before time expires.

Turn 1 - Search Phase:

```
Player 0 (Codebreaker, Experienced
Player): "I examine the bookshelf for hidden
compartments."
[GM reveals: Found a cipher wheel hidden
behind books]
Player 1 (Explorer, Bold): "I check the fire-
place for unusual mechanisms."
[GM reveals: Discovered a loose brick
with a key]
```

Why this works: All discoveries benefit the team equally. No player feels disadvantaged by another's success.

Turn 2 - Puzzle Phase:

```
Player 0: "Using the cipher wheel on the in-
scription..."
[GM: Cipher reveals "MOONLIGHT REVEALS
THE PATH"]
Player 2 (Team Leader): "Everyone, let's
check for moon-related elements!"
```

Player feedback: "The narrative built naturally," "Everyone contributed meaningfully," "The time pressure felt fair because we all faced it together."