What to Ask Next? Probing the Imaginative Reasoning of LLMs with TurtleSoup Puzzles

Mengtao Zhou[1*], **Sifan Wu**[2,3*†], **Huan Zhang**[2,3], **Qi Sima**[1], **Bang Liu**[2,3]

[1]Huazhong University of Science and Technology
[2]University of Montreal
[3]Mila - Quebec AI Institute

Abstract

We investigate the capacity of Large Language Models (LLMs) for **imaginative reasoning**—the proactive construction, testing, and revision of hypotheses in information-sparse environments. Existing benchmarks, often static or focused on social deduction, fail to capture the dynamic, exploratory nature of this reasoning process. To address this gap, we introduce a comprehensive research framework based on the classic "Turtle Soup" game, integrating a benchmark, an agent, and an evaluation protocol. We present *TurtleSoup-Bench*, the first large-scale, bilingual, interactive benchmark for imaginative reasoning, comprising 800 turtle soup puzzles sourced from both the Internet and expert authors. We also propose *Mosaic-Agent*, a novel agent designed to assess LLMs' performance in this setting. To evaluate reasoning quality, we develop a multi-dimensional protocol measuring logical consistency, detail completion, and conclusion alignment. Experiments with leading LLMs reveal clear capability limits, common failure patterns, and a significant performance gap compared to humans. Our work offers new insights into LLMs' imaginative reasoning and establishes a foundation for future research on exploratory agent behavior.

Code — https://github.com/lin-ruo/TurtleSoup-Bench

1 Introduction

Large Language Models (LLMs) increasingly serve as the cognitive core of autonomous agents, enabling advanced reasoning, understanding, and decision-making (Bubeck et al. 2023; Xi et al. 2025; Hong et al. 2024; Park et al. 2023; Wu et al. 2025b,a, 2023). Yet these gains largely assume *information-complete* settings with fully specified rules, goals, and context. Many real-world tasks are *dynamic* and *information-scarce*—e.g., an archaeologist inferring daily life from a few pottery shards, or a police officer reconstructing a crime from sparse, ambiguous clues. In such cases, progress depends less on retrieving known facts than on constructing, testing, and revising speculative explanations of the missing pieces(Xu et al. 2025a; Schacter, Benoit, and Szpunar 2017). We refer to this advanced reasoning capability as **imaginative reasoning**.

Prior evaluation of imaginative reasoning in LLMs have relied on multi-agent social deception games such as Werewolf or Avalon (Xu et al. 2024; Lan et al. 2024). These works mainly emphasise role inference under hidden information. TurtleBench (Yu et al. 2024) employs an evaluation based on static question-answer pairs to assess a model's ability to answer deductive questions but fails to evaluate an agent's core decision-making capability to autonomously and strategically decide what to ask next. While valuable, these benchmarks do not directly measure the iterative process of hypothesis generation, testing, and belief updating that constitutes imaginative reasoning. Their focus remains on social deduction or static knowledge retrieval, not on the agent's ability to creatively explore an unknown problem space.

To bridge this gap and rigorously examine the imaginative reasoning potential of modern LLMs, we leverage the classic narrative-reasoning game **Turtle Soup**[1]. As illustrated in Figure 1, each puzzle reveals only a terse, enigmatic scenario as the *soup surface*. The solver's goal is to recover the complete latent story as the *soup bottom*—by asking a series of yes/no questions. Solving a puzzle, naturally, requires an iterative loop of abductive and deductive logic.Accordingly, we present *TurtleSoup-Bench*, a comprehensive benchmark grounded in turtle soup puzzles, crafted specifically to evaluate the imaginative reasoning ability of LLMs. It encompasses 800 stories drawn from online sources and expert creations. Table 1 presents the primary characteristics of our benchmark and contrasts it with prior benchmarks.

Building on TurtleSoup-Bench, we develop the **Mosaic-Agent** framework to model the iterative process of imaginative reasoning. The framework comprises a Questioner agent, a Responder agent, and a memory module. Nevertheless, evaluating such a creative, exploratory process presents a significant challenge, as conventional NLP metrics like BLEU or ROUGE collapse multiple dimensions of quality into a single, surface-level similarity score (Liu et al. 2023). As established in related fields like creative story generation, a comprehensive assessment requires a multi-faceted approach (Yang and Jin 2024). To address this, our evaluation protocol is designed to disentangle an agent's performance into two distinct perspectives: the quality of the

*Equal Contribution
†Correspondence: sifan.wu@umontreal.ca.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]Also known as *Situation Puzzles* or *Yes/No Puzzles*.

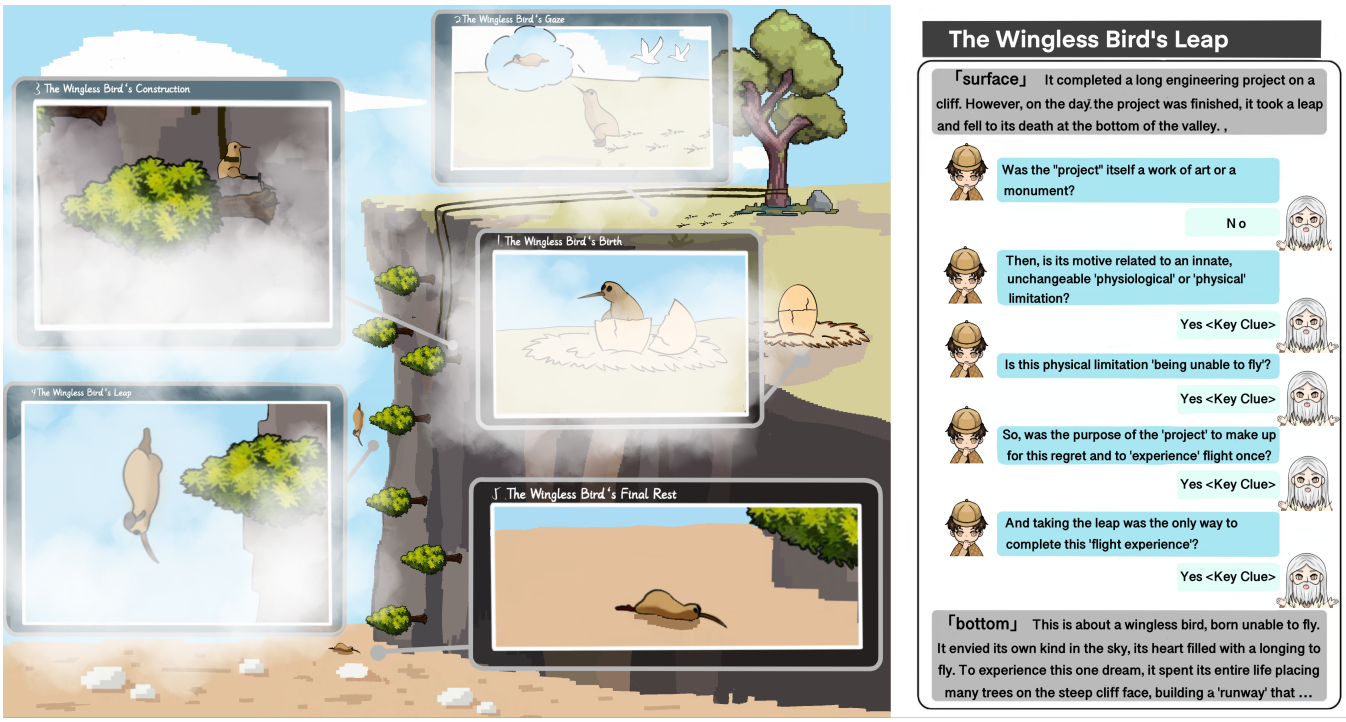

Figure 1: Left is the story of "The Wingless Bird's Leap" from TurtleSoup-Bench and right is the automatic evaluation through our Mosaic-Agent.

reasoning process and the accuracy of the final result. We evaluate the process using two metrics—Logic Accuracy to assess the coherence of the causal chain and Detail Fidelity to measure factual grounding—while the result is evaluated using Conclusion Match, which holistically compares the agent's final summary against the ground truth. Together, these metrics provide a more granular and meaningful assessment than a monolithic score, offering a finer-grained diagnostic lens for multi-step reasoning systems.

Overall, our main contributions are as follows: **(1)** We propose to use TurtleSoup Puzzles as the environment to assess the imaginative reasoning ability of LLMs, constructing *TurtleSoup-Bench* with 800 scenarios. **(2)** We design a multi-agent framework, Mosaic-Agent, aiming to solve puzzles fully automatically via two dynamic interaction agents. **(3)** Experiments on *TurtleSoup-Bench* demonstrate that state-of-the-art LLMs struggle with incomplete information and complex imaginative reasoning.

2 Related Work

LLM-based Autonomous Agents. Building on foundational frameworks like ReAct (Yao et al. 2023) and Reflexion (Shinn et al. 2023), recent work has increasingly tested LLM agents in complex multi-agent games. These studies span social deduction games like "Avalon" (Lan et al. 2024), "The Traitors" (Curvo 2025), and "Werewolf" (Sato, Ozaki, and Yokoyama 2024; Zhang et al. 2025; Xu et al. 2025b), to narrative mysteries like "Murder Mystery Games" (Zhu et al. 2024; Cai et al. 2025), "SpyGame" (Liang et al. 2023; Wei, Chen, and Xu 2025), and "Jubensha" (Wu et al. 2024). However, these works primarily focus on social strategy and inter-agent confrontation, rather than individual cognition in information-scarce environments. Our work takes a different path by avoiding social dynamics to focus on this fundamental dimension of cognition. We place the agent in a non-adversarial, puzzle-solving context, where the core challenge is to perform imaginative reasoning by constructing a coherent narrative from sparse clues.

Evaluating Reasoning in Interactive Environments. The evaluation of LLMs is shifting from static benchmarks like MMLU (Hendrycks et al. 2021) to dynamic, interactive environments, as static tests cannot assess the procedural capabilities required for active exploration (Hsia et al. 2024; Eriksson et al. 2025). Consequently, a new generation of interactive benchmarks has emerged, covering areas from open-world exploration (Wang et al. 2024a; Yang et al. 2025b) and multi-turn dialogue games (Ma et al. 2024; Bai et al. 2024; Li et al. 2025) to script-based role-playing (Wang et al. 2024b; Yu et al. 2025; Wang et al. 2025). The work most adjacent to ours is TurtleBench (Yu et al. 2024), which pioneered using turtle soup puzzles for reasoning assessment. However, its static question-answering protocol is verification-based, not exploratory, and thus cannot assess the core dynamic process of an agent formulating, testing, and revising beliefs through interaction. Our framework addresses this gap by shifting the evaluation focus from static outcome verification to measuring the entire dynamic process of imaginative reasoning.

Benchmark / Framework	Inter-action	Indi-vidual	Imagin-ative	Environ-ment	Evalu-ation
Werewolf	✓	✓	×	Adversarial	Outcome-based
Avalon	✓	✓	×	Adversarial	Outcome-based
Jubensha	✓	✓	×	Adversarial	Outcome-based
MIRAGE	✓	✓	×	Cooperative	Outcome-based
RoleLLM	✓	✓	×	Conversational	Output Fidelity
RPGBENCH	✓	✓	×	Simulative	Output Quality
Word Guess	✓	×	×	Adversarial	Outcome-based
TurtleBench	×	✓	×	Static Puzzle	Outcome-based
Ours	✓	✓	✓	Non-Advers.	Path & Fidelity

(a) Comparative Analysis

TurtleSoup-Bench Feature	Value
Crime Thriller	164
Mind Game	120
Supernatural	62
Constant Change	116
Clever Logic	138
Original (Expert-Authored)	200
Total	800
Surface Tokens (per scenario)	49.2
Bottom Tokens (per scenario)	143.3
Key Clues (per scenario)	5.7

(b) *TurtleSoup-Bench* Statistics

Table 1: Comparative analysis of TurtleSoup-Bench against other frameworks and detailed statistics of our benchmark.

3 *TurtleSoup-Bench*: A Benchmark for Imaginative Reasoning

Data Collection and Authoring. To balance the benchmark's breadth and novelty, we adopted a dual-source data collection strategy. First, we collected a large number of Chinese turtle soup puzzles from well-known online puzzle communities (e.g., (Tang 2025)). To select high-quality samples from this extensive pool, we implemented a two-stage filtering process. The first stage was a community-based pre-screening, where we retained only those stories with high upvote counts, top ratings, or a large number of positive comments, ensuring their popularity and recognized quality among players. This step narrowed down the candidates. The second stage was an expert-led final selection, where our expert team manually reviewed the pre-screened stories. Based on criteria such as logical soundness, narrative cleverness, and suitability for LLM evaluation, they selected the final 300 stories. To address the critical issue of data contamination and introduce more challenging scenarios, we then recruited a team of five experienced puzzle design experts to author 100 entirely new stories. The experts followed strict design principles during creation, including logical consistency (the bottom fully explains the premise), non-obviousness (the bottom is clever and non-intuitive), and self-containment (solvable only via "Yes/No/Unknown" questions). The 40 to 60 minutes required to craft a single story ensures the originality and high quality of this dataset.

Data Curation and Annotation. After compiling the 400 stories, we conducted a rigorous curation and annotation process for all data. All flawed or ambiguous samples identified during the selection process were corrected and re-edited by experts. Subsequently, to support a more nuanced analysis of agent capabilities, our expert team classified the collected stories into five core narrative genres, with our original stories treated as a distinct sixth category. Furthermore, a core contribution of our benchmark is the manual annotation of a *Key Clue Library* K_{lib} for every story. These expert-defined clues represent pivotal turning points in the reasoning process and provide effective guidance signals for the agent. To extend the benchmark's utility to the global research community, the complete corpus of 400 Chinese stories was professionally translated and culturally adapted into an equivalent set of 400 English stories.Ultimately, our *TurtleSoup-Bench* comprises 800 scenarios, with the detailed distribution shown in Table 1b. Some of the scenarios are shown in the Appendix.

4 Mosaic-Agent Framework: Interactive Environment for TurtleSoup-Bench

This section outlines our framework, Mosaic-Agent, designed to simulate a TurtleSoup puzzle solution. The goal is to find the *soup bottom* behind the *soup surface* by modeling multi-turn interaction between a questioner and a responder agent grounded in the real TurtleSoup situation (Wikipedia contributors 2024). As illustrated in Figure 2, the framework consists of: the questioner agent, the responder agent, and memory module. The questioner agent aims to act like the player to propose imaginative questions. The responder agent is like god to response the question and hint about key clue. And the memory module acts like a detective's notebook, recording the full conversation and pivotal clues.

4.1 Questioner: A Deliberative Cognitive Architecture

Efficient human problem-solving follows a logic progressing from analysis to decision-making. The process begins by analyzing information to form hypotheses (Binz and Schulz 2023), then uses strategic foresight to simulate potential outcomes (Schacter, Addis, and Buckner 2007). Ultimately, this analysis leads to a key decision—such as asking a highly informative question—aimed at most effectively reducing uncertainty (Kidd and Hayden 2015). We leverage this cognitive model by decoupling our questioner agent's question-generation mechanism into three corresponding processes: deliberation, meta-cognition, and action generation, to replicate the efficiency of human thought.

Deliberation Agent. This agent serves as the core analytical engine. At the beginning of each decision cycle, it processes existing information to form a comprehensive understanding of the situation and identify key directions for

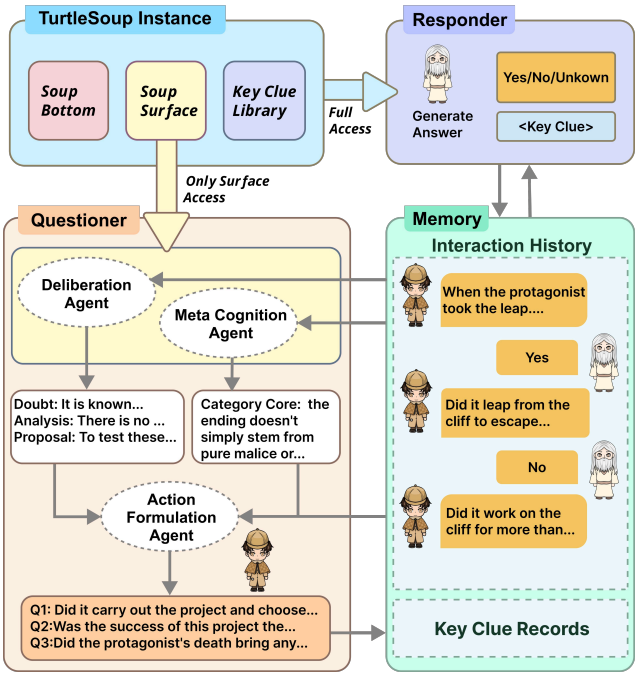

Figure 2: The Mosaic-Agent Framework

the next exploration. The workflow is designed as a hierarchical cognitive process that integrates two distinct but complementary levels of cognition: one for local analysis for immediate reaction, and another for global synthesis to build a macroscopic understanding. This progressive, local-to-global design is manifested in:

First, the agent performs a quick analysis of the most recent question-answer pair, (q_{t-1}, a_{t-1}), where q denotes the question, a the answer, and t the current turn. The purpose of this step is to quickly parse the newly acquired local information from the last interaction, establishing a clear starting point for the next line of reasoning.

To form a more macroscopic and in-depth understanding, the agent periodically (at a fixed interval of k turns) conducts a global deliberation to examine the global picture and prevent cognitive myopia. The core of this process is updating its internal Belief State, B_t, a structured json object of the agent's understanding of the story's core logic L_t, key details D_t, and overall conclusion C_t.This update is handled by the function f_{sum}, which directs the agent to retrieve the Interaction History H_{t-1}, Key Clue Records K_{rec} from the shared Memory module, and the belief state from the previous cycle B_{t-k} into the new belief state.

$$B_t = (L_t, D_t, C_t) = f_{\text{sum}}(H_{t-1}, K_{\text{rec}}, B_{t-k}) \quad (1)$$

Then the agent performs a self-diagnosis on the updated belief state B_t to identify logical gaps or missing information. The function f_{advice} performs this diagnosis, instructing the agent to generate a structured "Analysis and Proposal Set" (APS) based on the current belief state B_t and the *soup surface* S_{surf}:

$$APS_t = f_{\text{advice}}(B_t, S_{\text{surf}}) = \{(d_j, an_j, p_j)\}_{j=1}^{m} \quad (2)$$

This set contains m tuples, where each tuple represents a specific doubt d_j, its corresponding analysis an_j, and a question proposal p_j designed to resolve the doubt. Detailed illustration for are in Appendix.

Meta-cognition Agent. This agent dynamically adjusts the agent's macro-level strategy by periodically classifying the puzzle's narrative genre.

A static strategy is brittle in complex exploration. The agent initiates a judgment whenever it acquires several new clues or after several turns with no progress. This judgment involves a three-vote majority classification based on the *soup surface* S_{surf}, Interaction History H_{t-1}, and the Key Clue Records K_{rec}, to assign the puzzle to a narrative genre. An occasional misclassification can cause severe strategy oscillation. Imagine the agent misclassifies a crime story as supernatural; it would waste turns asking about ghosts.

To prevent this, we use a Smoothed Confidence mechanism. It maintains a policy confidence, c, representing the agent's confidence in its current assessment, initialized to a neutral 0.5 at the start of the game. When a new vote yields a confidence score v_c, we update it via the Exponential Moving Average formula:

$$c_t^{\text{smooth}} = \alpha \cdot c_{t-1} + (1 - \alpha) \cdot v_c \quad (3)$$

where c_{t-1} is the confidence from the previous cycle. A key stability feature is that the agent switches its strategy only if the new smoothed confidence, c_t^{smooth}, exceeds the sum of the old confidence and a predefined threshold, τ_{switch}. In our study, α is set to 0.7, and τ_{switch} is set to 0.1.

A successful genre switch makes the agent adopt a new questioning strategy, We designed unique questioning strategies for each narrative genre in collaboration with human experts, following the principles of compositional reasoning (Press et al. 2023). These strategies target the typical logical structures of each genre to make subsequent questioning more focused.

Action Formulation Agent. In this agent, we aim to integrate the various cognitive outputs to formulate the final single question action q_t.

First, in the candidate generation stage, the agent combines the Analysis and Proposal Set APS_t from the Deliberation agent with the chosen questioning strategy from the Meta-cognition agent to generate three candidate questions, Q_{cand}. Next, in the optimal screening stage, we let the agent act as its own decision-making critic. The agent considers the full Interaction History H_{t-1}, its own analysis APS_t, and a blacklist, $\mathbb{B}$. This blacklist is created at the start of the game and is dynamically updated with questions that are identified as ineffective or redundant during gameplay. Based on this complete context, the agent selects from the candidates the single best question, q_t, that is most likely to yield new information while avoiding redundancy:

$$q_t = \underset{q' \in Q_{\text{cand}}}{\text{argmax}} \ \text{Score}(q' | H_{t-1}, APS_t, \mathbb{B}) \quad (4)$$

This complete Deliberation-MetaCognition-Action process ensures that every question posed by the agent is well-considered and most likely to lead to the truth. Finally, q_t is sent to the Memory module and passed to the Responder.

4.2 Responder: Simulating an Interactive Environment

Using an LLM to simulate the Responder is validated by prior work(Yu et al. 2024). The Responder is thus designed as a faithful interactive environment; while inherently non-deterministic, we maximize its response fidelity via low temperature, fixed seed, and a constrained answer space. Its core function, f_{respond}, maps the Questioner's question at turn t, q_t, to a feedback tuple (a_t, f_t), based on the *soup bottom* S_{bot} and the *Key Clue Library* K_{lib}:

$$(a_t, f_t) = f_{\text{respond}}(q_t, S_{\text{sol}}, K_{\text{tips}}) \tag{5}$$

This tuple consists of two components, the answer a_t and the key clue flag f_t, which are detailed as follows:

Answer Generation. Given the Questioner's query q_t and the *soup bottom* S_{bot}, the agent first makes a logical judgment as to whether the content of q_t is consistent with S_{bot}. Based on this judgment, the agent provides a standardized answer, a_t, strictly confined to one of three categories:

- **Yes**: Indicates the statement in the question is true according to the solution. For instance, if the solution is "The killer wore a red coat", the answer to "Did the killer wear a red coat?" is "Yes".
- **No**: Indicates false statement. For the same example, the answer to "Did the killer wear a blue coat?" is "No".
- **Unknown**: This response encompasses cases where the information is either genuinely absent from the solution or simply irrelevant to the core mystery. For simplicity, we group these two cases into a single "Unknown" label, as both represent information that does not advance the puzzle's solution. This simplification has minimal impact on solving the core mystery, as the "Key Clue" mechanism safeguards the discovery of pivotal information.

Key Clue Identification. To better guide the Questioner's reasoning and help it recognize breakthroughs, we introduce the boolean flag f_t. That is, not all "Yes/No" answers are equally important, and the questions that hit upon the core of the puzzle will be highlighted by this flag.

The agent determines if the semantics of the question q_t are directly relevant to any of the predefined key clues in K_{lib}. If they are, f_t is set to true, and a `<Key Clue>` marker is appended to the answer string a_t. And they are passed to the Questioner via the memory module. For example, if the core of a solution is "The man was a dwarf and couldn't reach the button," a key clue in K_{lib} might be "The man's height is a critical factor." When the Questioner asks, "Was the man short?", the answer would be "Yes" and f_t would be true, so the final complete answer returned to the questioner would be the string "Yes`<Key Clue>`".

4.3 Memory Module

The memory module acts as a central hub in the agent's cognitive loop, recording information that facilitates the questioner's reasoning process. The module is partitioned into two key components: the complete interaction history and a curated record of key clues.

Interaction History. The Interaction History H_t is a complete, chronological log of every question posed by the Questioner and every answer given by the Responder. Its primary function is to provide complete context to the questioner, allowing agents to reflect on the entire conversational flow and prevent reasoning gaps due to forgotten information.

Key Clue Records. The Key Clue Records K_{rec} is a filtered, high-value memory pool. When a response from the Responder is flagged with `<Key Clue>`, the corresponding question-answer pair (q_t, a_t) is stored here. The purpose of this design is to allow the Questioner to quickly locate and access the information that is pivotal to solving the puzzle.

4.4 Automated Evaluation Protocol

To objectively evaluate Mosaic-Agent's open-ended summary, our protocol first decomposes the *soup bottom* S_{bot} into two sets of structured evaluation points using an LLM: Core Logic Points L_{true} and Key Details D_{true}. The number of points elicited is determined adaptively by a set of heuristic rules based on the solution's length and complexity—for instance, the number of logic points ranges from 2 to 5 depending on text length—to ensure comprehensive coverage.

The Questioner's summary, $B_{\text{final}} = (L_{\text{pred}}, D_{\text{pred}}, C_{\text{pred}})$, is then assessed across three dimensions. The motivation for these dimensions is to disentangle the agent's capabilities from the dual perspectives of the reasoning process and the final result. To evaluate the process, Logic Accuracy S_{logic} measures the coherence of the causal chain by matching L_{pred} against L_{true}, while Detail Fidelity S_{details} measures the factual grounding by matching D_{pred} against D_{true}. To evaluate the result, Conclusion Match $S_{\text{conclusion}}$ provides a holistic assessment by comparing the final summary text C_{pred} with the *soup bottom* S_{bot}.

In implementation, we use the latest LLM to perform semantic matching. To align with human judgment, the logic and detail scores undergo a Two-Threshold Calibration: a Validity Threshold of 0.5 ensures rigor by filtering out weakly-related matches, while a High-Confidence Threshold of 0.8 normalizes any strong match to a full score of 1.0, making the evaluation robust to paraphrasing. The final Overall Score is a weighted sum:

$$S_{\text{overall}} = w_l \cdot S_{\text{logic}} + w_d \cdot S_{\text{details}} + w_c \cdot S_{\text{conclusion}} \tag{6}$$

where the weights w_l, w_d, and w_c are set to 0.3, 0.3, and 0.4 in our study, respectively. Furthermore, we report all permetric results, enabling straightforward recomputation under any alternative weighting scheme.

4.5 Human Performance Baseline

To illustrate the difficulty of TurtleBench, we recruited 4 human players with extensive experience in the Turtle Soup game. These experts played all 400 Chinese scenarios in *TurtleSoup-Bench*. We meticulously recorded the gameplay of each human player, including the complete sequence of questions, the number of turns taken to solve each puzzle, and their final summary. The human baseline was established on the Chinese portion of the benchmark, which we posit sufficiently demonstrates human-level performance on this task. Critically, we do not subjectively score the human

Model	Lang	Crime Thrill.		Mind Game		Supernat.		Const. Change		Clev. Logic		Orig. Data.	
		Score	Reduce	Score	Reduce	Score	Reduce	Score	Reduce	Score	Reduce	Score	Reduce
claude-3.7-sonnet	Ch	**54.54**	(-15.34)	**54.66**	(-18.39)	**58.67**	(-16.39)	**63.19**	(-8.93)	**61.51**	(-11.82)	56.82	(-11.05)
	En	28.18	(-41.70)	30.63	(-42.42)	**37.00**	(-38.06)	39.93	(-32.19)	39.97	(-33.36)	31.56	(-36.31)
gemini-2.5-flash	Ch	52.39	(-17.49)	47.93	(-25.12)	57.42	(-17.64)	62.90	(-9.22)	45.45	(-27.88)	51.74	(-16.13)
	En	**33.59**	(-36.29)	**30.73**	(-42.32)	35.00	(-40.06)	40.95	(-31.17)	**42.17**	(-31.16)	**34.15**	(-33.72)
deepseek-r1	Ch	46.94	(-22.94)	46.88	(-26.17)	53.61	(-21.45)	59.36	(-12.76)	57.51	(-15.82)	**57.14**	(-10.73)
	En	29.82	(-40.06)	24.35	(-48.70)	36.77	(-38.29)	41.40	(-30.72)	36.86	(-36.47)	29.76	(-38.11)
gpt-4o	Ch	28.79	(-41.09)	28.57	(-44.48)	32.26	(-42.80)	38.07	(-34.05)	34.38	(-38.95)	33.85	(-34.02)
	En	31.46	(-38.42)	26.78	(-46.27)	32.19	(-42.87)	**41.76**	(-30.36)	39.23	(-34.10)	30.68	(-37.19)
qwen3-32b	Ch	42.07	(-27.81)	39.62	(-33.43)	52.19	(-22.87)	54.84	(-17.28)	46.58	(-26.75)	48.55	(-19.32)
	En	15.66	(-54.22)	14.67	(-58.38)	17.45	(-57.61)	22.95	(-49.17)	21.38	(-51.95)	20.25	(-47.62)
llama3-8b-instruct	Ch	10.51	(-59.37)	8.95	(-64.10)	6.74	(-68.32)	13.22	(-58.90)	12.57	(-60.76)	7.70	(-60.17)
	En	6.57	(-63.31)	5.12	(-67.93)	7.84	(-67.22)	8.67	(-63.45)	11.52	(-61.81)	6.88	(-60.99)
Human Baseline		**69.88**		**73.05**		**75.06**		**72.12**		**73.33**		**67.87**	

Table 2: Overall scores S_{overall} of different models and the human baseline, grouped by language, across all six genres from *TurtleSoup-Bench*. Scores are presented as percentages, with the reduction from the human baseline noted. The best-performing model in each column for each language group is highlighted in bold.

players. Instead, their generated final summaries are fed into the same Automated Evaluation Protocol described in the Section 4.4. This ensures that the performances of both the agent and the human players are compared under an identical, objective standard, guaranteeing a fair comparison.

5 Experiments

5.1 Experimental Setup

Environment and Models. All experiments are conducted on our *TurtleSoup-Bench*. We select a representative suite of state-of-the-art LLMs for evaluation, including claude-3.7-sonnet(Anthropic 2025), gemini-2.5-flash(Comanici et al. 2025), deepseek-r1(Guo et al. 2025), gpt-4o(Hurst et al. 2024), qwen3-32b(Yang et al. 2025a), and llama3-8b-instruct(Dubey et al. 2024). In our symmetric design, the Questioner and Responder employ the same model to simultaneously assess its reasoning and comprehension fidelity. To ensure fairness, we use deepseek-r1 for evaluation.

Evaluation Protocol. We employ the automated protocol from Section 4.4, with primary metrics being S_{logic}, S_{details}, $S_{\text{conclusion}}$, and S_{overall}. Furthermore, we introduce the Human Baseline established in Section 4.5 as a reference to measure the performance gap between LLM agents and human experts.

Implementation Details. In all experiments, the periodic deliberation interval k is set to 5, and the maximum number of question turns N_{max} is 30. Due to cost, all scenarios in this study are run only once.

5.2 Quantitative Analysis

Table 2 presents the results of our quantitative evaluation on *TurtleSoup-Bench*, providing concrete evidence for analyzing the imaginative reasoning capabilities of LLMs.

The results reveal a clear performance stratification among models that **top-tier proprietary models form a leading group, while open-source models, even the larger-parameter qwen3-32b (48.1%), exhibit a significant performance gap compared to the former**, such as claude-3.7-sonnet (58.8%). While the human performance achieved higher score as 67.87% for least. We believe that the performance bottleneck arises not only from the Questioner's deficiency in generating effective exploratory hypotheses but also from the Responder's difficulty in accurately understanding questions and faithfully providing correct answers. A flawed Responder introduces environmental noise that systematically derails the reasoning process.

Model performance correlates strongly with the narrative paradigm, revealing capability biases. For example, gemini-2.5-flash excels on "Constant Change" (62.9%) but drops sharply on the non-intuitive "Clever Logic" (45.5%). And when a task requires modeling complex human intent (as in "Crime Thriller") or non-linear logical reasoning, the performance of most models is severely challenged. This reveals that current LLM imagination is not a general but a collection of specialized skills optimized for specific tasks, with limited generalization.

Furthermore, **a systematic performance decline is observed across almost all models on the English dataset**. For example, the average score of deepseek-r1 drops by nearly 40%. Although we have done some cultural adaptations, the subtleties of many puzzles are rooted in their cultural and linguistic origins. The introduction of ambiguity and semantic loss during cross-language conversion also increase the difficulty of reasoning.

Finally, **a significant chasm persists between the best-performing agents and the human baseline.** The top model, claude-3.7-sonnet, still lags behind human experts by approximately 13 percentage points. The highly effective intuition, creative hypothesis generation, and the ability to efficiently eliminate irrelevant options from a vast space of possibilities that human players exhibit are core capa-

bilities that current models, reliant on probabilistic pattern-matching, have yet to replicate.

5.3 Qualitative and Error Analysis

Our qualitative analysis categorizes the failures of LLM in imaginative reasoning into four distinct levels, progressing from the micro to the macro. Figure 3, through a specific case analysis, demonstrates two of these modes.

Repaying Kindness with Enmity

Puzzle Context: The Soup Surface describes "I" repaying kindness with enmity for failing a promise. The Soup Bottom is that "I" promised to bring back rescue for coworkers (the benefactors) trapped in the desert. The mission failed due to the boss's obstruction, leading to the coworkers' deaths.

Turn 3-6: **Was the promise related to safeguarding / using / returning / transferring an object? (All denied)**

Turn 13: **Did the action involve... delivering or destroying an object? / Did the promise involve acquiring or creating an object?**

Turn 12:

Q: **Did... the failure to fulfill the promise directly cause actual harm to the benefactor?**

R: **No<Key Clue>**

Path Analysis: The **Questioner**'s failure is a typical case of **Deductive Pruning Failure**. After multiple "No" answers ruled out the possibility of the promise being related to an object, the agent failed to perform effective deductive reasoning to abandon this line of inquiry.The **Responder**'s failure is a **Context Construction Failure**, as it failed to understand the direct causal link between the 'failed promise' and the 'coworkers' deaths' within the context it was given, leading to a factual error.

Figure 3: Case Studies of Four Typical Failure Paradigms

The most foundational failure is **Semantic Fixation**, which occurs at the level of word meaning interpretation. This stems from the model's reliance on the statistical inertia of its training data, causing it to rigidly default to a word's most high-frequency, literal meaning while ignoring contradictory contextual clues. This causes the entire reasoning process to start from a flawed premise.

This micro-level error then precipitates a more macroscopic **Context Construction Failure** at the scene level. The core deficit here is in integration and updating. Even when the model understands all individual clues, it fails to effectively splice these fragmented and sometimes contradictory pieces of information into a coherent global context. Then the imaginative process stalls

Logic Blind Spots shows a higher-order bottleneck, a failure to reason about "why". Even with a correct factual context, the model struggles to conceive of the atypical causality that drives the scenario. Its reasoning paths are constrained by the common patterns within its training data, preventing it from proactively generating truly out-of-distribution hypotheses. This sharply defines the boundary of current LLM imagination: it excels at high-fidelity inference within its experiential space but lacks the creative leap required for genuinely novel solutions.

Finally, **Deductive Pruning Failure** is a fundamental deficit at the process and methodology level. This is less about content comprehension and more about the rigor of the reasoning process itself. The model ineffectively uses negative feedback to systematically eliminate falsified branches of the possibility space. Instead, it pursues redundant exploration down disproven paths. This demonstrates that model lacks the ability to adjust its own line of inquiry which renders its imaginative process both disordered and inefficient.

Agent Config	S_{logic}	S_{details}	$S_{\text{conclusion}}$	S_{overall}
Mosaic-Agent	54.72	56.68	59.45	57.14
w/o Deliberation	55.91	54.10	47.10	51.76
w/o Meta-Cog.	56.14	50.51	47.70	51.07
w/o Pruning	51.86	52.94	45.95	49.80
w/o Key Clue	46.84	45.85	47.45	46.73
w/o All agents	46.93	51.91	41.55	46.24

Table 3: Ablation study of the Mosaic-Agent framework driven by the deepseek-r1 model, evaluated on the original dataset from *TurtleSoup-Bench*.

5.4 Ablation study

To verify the necessity of each agent in our cognitive architecture, we conducted ablation studies by selectively removing key components. All variants use deepseek-r1 as the base model and were tested on the original dataset from *TurtleSoup-Bench*. The results are presented in Table 3.

Deliberation Agent. Removing the Deliberation Agent (57.14 $\rightarrow$ 51.76) harms performance, as the agent loses its ability to integrate scattered clues. This highlights that simple generative models are insufficient for such tasks.

Meta-cognition Agent. Disabling the Meta-cognition Agent (57.14 $\rightarrow$ 51.07) removes strategic adaptation, proving that advanced planning requires explicit architectural support.

Optimal Pruning Removing Optimal Pruning (57.14 $\rightarrow$ 49.80) bypasses self-correction, leading to less optimal decisions. This confirms the need to treat LLM outputs as proposals to be evaluated.

Key Clue Mechanism Removing the Key Clue mechanism (57.14 $\rightarrow$ 46.73) causes the most severe collapse. Without this high signal-to-noise feedback, exploration degrades into a near-random walk, suggesting feedback quality is a primary driver of efficiency.

Synergistic Effect of the Architecture Finally, the baseline agent (w/o All agents) scored 46.24, nearly identical to the score when only the Key Clue mechanism was removed (46.73). This reveals a crucial architectural synergy: sophisticated modules like Deliberation are ineffective without the high-quality information stream provided by the Key Clue mechanism. The ability to identify high-value information is thus a prerequisite for higher-order reasoning.

6 Conclusion

We introduce a framework to probe LLM imaginative reasoning, a critical capability for information-scarce environments. Our framework includes three components: *TurtleSoup-Bench*, the first large-scale interactive benchmark; *Mosaic-Agent*, featuring a novel deliberative architecture; and a multi-dimensional evaluation protocol. Unlike prior work on static outcomes or social dynamics, we pioneer evaluating the exploratory reasoning process itself. Experiments reveal LLM limitations and validate our path-analysis approach. This work establishes a standard for evaluating imaginative reasoning, shifting focus from final outcomes to the dynamic inquiry process.

References

Anthropic, C. 2025. 3.7 sonnet and claude code.

Bai, G.; Liu, J.; Bu, X.; He, Y.; Liu, J.; Zhou, Z.; Lin, Z.; Su, W.; Ge, T.; Zheng, B.; and Ouyang, W. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7421–7454. Bangkok, Thailand: Association for Computational Linguistics.

Binz, M.; and Schulz, E. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6): e2218523120.

Bubeck, S.; Chadrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Cai, Y.; Gu, Z.; Du, Z.; Ye, Z.; Cao, S.; Xu, Y.; Feng, H.; and Chen, P. 2025. MIRAGE: Exploring How Large Language Models Perform in Complex Social Interactive Environments. arXiv:2501.01652.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Curvo, P. M. P. 2025. The Traitors: Deception and Trust in Multi-Agent Language Model Simulations. arXiv:2505.12923.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models.

Eriksson, M.; Purificato, E.; Noroozian, A.; Vinagre, J.; Chaslot, G.; Gomez, E.; and Fernandez-Llorca, D. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. arXiv:2502.06559.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; Ran, C.; Xiao, L.; Wu, C.; and Schmidhuber, J. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations*.

Hsia, J.; Pruthi, D.; Singh, A.; and Lipton, Z. 2024. Goodhart's Law Applies to NLP's Explanation Benchmarks. In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 1322–1335. St. Julian's, Malta: Association for Computational Linguistics.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card.

Kidd, C.; and Hayden, B. Y. 2015. The psychology and neuroscience of curiosity. *Neuron*, 88(3): 449–460.

Lan, Y.; Hu, Z.; Wang, L.; Wang, Y.; Ye, D.; Zhao, P.; Lim, E.-P.; Xiong, H.; and Wang, H. 2024. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 128–145. Miami, Florida, USA: Association for Computational Linguistics.

Li, X.; Bao, K.; Ma, Y.; Li, M.; Wang, W.; Men, R.; Zhang, Y.; Feng, F.; Liu, D.; and Lin, J. 2025. MTR-Bench: A Comprehensive Benchmark for Multi-Turn Reasoning Evaluation. arXiv:2505.17123.

Liang, T.; He, Z.; tse Huang, J.; Wang, W.; Jiao, W.; Wang, R.; Yang, Y.; Tu, Z.; Shi, S.; and Wang, X. 2023. Leveraging Word Guessing Games to Assess the Intelligence of Large Language Models. arXiv:2310.20499.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.

Ma, C.; Zhang, J.; Zhu, Z.; Yang, C.; Yang, Y.; Jin, Y.; Lan, Z.; Kong, L.; and He, J. 2024. AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 74325–74362. Curran Associates, Inc.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N.; and Lewis, M. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5687–5711. Singapore: Association for Computational Linguistics.

Sato, T.; Ozaki, S.; and Yokoyama, D. 2024. An Implementation of Werewolf Agent That does not Truly Trust LLMs. In Kano, Y., ed., *Proceedings of the 2nd International AIWolfDial Workshop*, 58–67. Tokyo, Japan: Association for Computational Linguistics.

Schacter, D. L.; Addis, D. R.; and Buckner, R. L. 2007. Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, 8(9): 657–661.

Schacter, D. L.; Benoit, R. G.; and Szpunar, K. K. 2017. Episodic Future Thinking: Mechanisms and Functions. *Current opinion in behavioral sciences*, 17: 41–50.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K. R.; and Yao, S. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tang, H. 2025. Henre Tang's Personal Website. https://tanghenre.com/.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2024a. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research*.

Wang, L.; Lian, J.; Huang, Y.; Dai, Y.; Li, H.; Chen, X.; Xie, X.; and Wen, J.-R. 2025. CharacterBox: Evaluating the Role-Playing Capabilities of LLMs in Text-Based Virtual Worlds. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6372–6391. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Wang, N.; Peng, Z.; Que, H.; Liu, J.; Zhou, W.; Wu, Y.; Guo, H.; Gan, R.; Ni, Z.; Yang, J.; Zhang, M.; Zhang, Z.; Ouyang, W.; Xu, K.; Huang, W.; Fu, J.; and Peng, J. 2024b. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 14743–14777. Bangkok, Thailand: Association for Computational Linguistics.

Wei, C.; Chen, J.; and Xu, J. 2025. Exploring Large Language Models for Word Games:Who is the Spy? arXiv:2503.15235.

Wikipedia contributors. 2024. Situation puzzle — Wikipedia, The Free Encyclopedia. [Online; accessed 25-July-2024].

Wu, D.; Shi, H.; Sun, Z.; and Liu, B. 2024. Deciphering Digital Detectives: Understanding LLM Behaviors and Capabilities in Multi-Agent Mystery Games. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 8225–8291. Bangkok, Thailand: Association for Computational Linguistics.

Wu, S.; Khasahmadi, A.; Katz, M.; Jayaraman, P. K.; Pu, Y.; Willis, K.; and Liu, B. 2023. Cad-llm: Large language model for cad generation. In *NeurIPS 2023 Workshop on Machine Learning for Creativity and Design*.

Wu, S.; Khasahmadi, A. H.; Katz, M.; Jayaraman, P. K.; Pu, Y.; Willis, K.; and Liu, B. 2025a. CadVLM: Bridging Language and Vision in the Generation of Parametric CAD Sketches. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 368–384. Cham: Springer Nature Switzerland.

Wu, S.; Zhang, H.; Li, Y.; Effaty, F.; Ataei, A.; and Liu, B. 2025b. Seeing Beyond Words: MatVQA for Challenging Visual-Scientific Reasoning in Materials Science. arXiv:2505.18319.

Xi, Z.; Chen, W.; Guo, X.; et al. 2025. The rise and potential of large language model based agents: a survey. *Science China Information Sciences*, 68: 121101.

Xu, X.; Lawrence, R.; Dubey, K.; Pandey, A.; Falck, F.; Ueno, R.; Nori, A.; Sharma, R.; Sharma, A.; and González, J. 2025a. Re-Imagine: Symbolic Benchmark Synthesis for Reasoning Evaluation. In *ICLR 2025 - Workshop on Reasoning and Planning for LLMs*.

Xu, Z.; Gu, W.; Yu, C.; Wu, Y.; and Wang, Y. 2025b. Learning Strategic Language Agents in the Werewolf Game with Iterative Latent Space Policy Optimization. In *Forty-second International Conference on Machine Learning*.

Xu, Z.; Yu, C.; Fang, F.; Wang, Y.; and Wu, Y. 2024. Language agents with reinforcement learning for strategic play in the Werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 technical report.

Yang, D.; and Jin, Q. 2024. What Makes a Good Story and How Can We Measure It? A Comprehensive Survey of Story Evaluation. arXiv:2408.14622.

Yang, R.; Chen, H.; Zhang, J.; Zhao, M.; Qian, C.; Wang, K.; Wang, Q.; Koripella, T. V.; Movahedi, M.; Li, M.; Ji, H.; Zhang, H.; and Zhang, T. 2025b. EmbodiedBench: Comprehensive Benchmarking Multi-modal Large Language Models for Vision-Driven Embodied Agents. In *Forty-second International Conference on Machine Learning*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*.

Yu, P.; Shen, D.; Meng, S.; Lee, J.; Yin, W.; Cui, A. Y.; Xu, Z.; Zhu, Y.; Shi, X.; Li, M.; and Smola, A. 2025. RPGBENCH: Evaluating Large Language Models as Role-Playing Game Engines. arXiv:2502.00595.

Yu, Q.; Song, S.; Fang, K.; Shi, Y.; Zheng, Z.; Wang, H.; Niu, S.; and Li, Z. 2024. TurtleBench: Evaluating Top Language Models via Real-World Yes/No Puzzles. arXiv:2410.05262.

Zhang, Z.; Lan, Y.; Chen, Y.; Wang, L.; Wang, X.; and Wang, H. 2025. DVM: Towards Controllable LLM Agents in Social Deduction Games. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Zhu, Q.; Zhao, R.; Du, J.; Gui, L.; and He, Y. 2024. PLAYER*: Enhancing LLM-based Multi-Agent Communication and Interaction in Murder Mystery Games. *CoRR*, abs/2404.17662.