Probe by Gaming: A Game-based Benchmark for Assessing Conceptual Knowledge in LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Concepts represent generalized abstractions that enable humans to categorize and reason efficiently, yet it is unclear to what extent Large Language Models (LLMs) comprehend these semantic relationships. Existing benchmarks typically focus on factual recall and isolated tasks, failing to evaluate the ability of LLMs to understand conceptual boundaries. To address this gap, we introduce CK-Arena, a multi-agent interaction game built upon the Undercover game, designed to evaluate the capacity of LLMs to reason with concepts in interactive settings. CK-Arena challenges models to describe, differentiate, and infer conceptual boundaries based on partial information, encouraging models to explore commonalities and distinctions between closely related concepts. By simulating real-world interaction, CK-Arena provides a scalable and realistic benchmark for assessing conceptual reasoning in dynamic environments. Experimental results show that LLMs' understanding of conceptual knowledge varies significantly across different categories and is not strictly aligned with parameter size or general model capabilities. The code is available at this URL: `https://anonymous.4open.science/r/CK-Arena/readme.md`.

1 Introduction

As Large Language Models (LLMs) become integral to complex reasoning tasks, the demand is shifting from mere sequence prediction to a deeper grasp of conceptual structures and their related characteristics in the real world [1, 2, 3, 4]. A concept represents a generalized abstraction that encapsulates shared properties of entities, enabling humans to categorize and reason efficiently [5, 6, 7, 8, 9, 10]. For example, the concept *Primates* groups animals like *monkeys* and *apes* based on shared characteristics such as opposable thumbs, forward-facing eyes, and high cognitive abilities. While human cognition naturally leverages such conceptual structures for reasoning and adaptation, it remains unclear to what extent LLMs capture and utilize these abstractions. Current evaluations primarily focus on surface-level predictions, offering limited insight into whether LLMs truly understand concepts as structured semantic entities.

Traditional benchmarks for LLMs evaluation have contributed to improvements in model performance [11, 12, 13, 14], but they exhibit significant limitations. These benchmarks primarily assess token-level accuracy and factual recall through static question-answer formats, often breaking down knowledge into isolated questions. This fragmented evaluation approach captures surface-level information retrieval but fails to probe the inherent connections and boundaries between concepts. For example, a model may correctly identify that *monkeys* and *apes* belong to *Primates*, yet this does not indicate any understanding of the structural relationships or distinctive features that separate these groups within the broader taxonomy. Furthermore, as LLMs evolve towards more autonomous and interactive roles, traditional methods such as multiple-choice and true/false questions struggle to reflect their capabilities in complex and dynamic environments. The reliance on fixed datasets

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

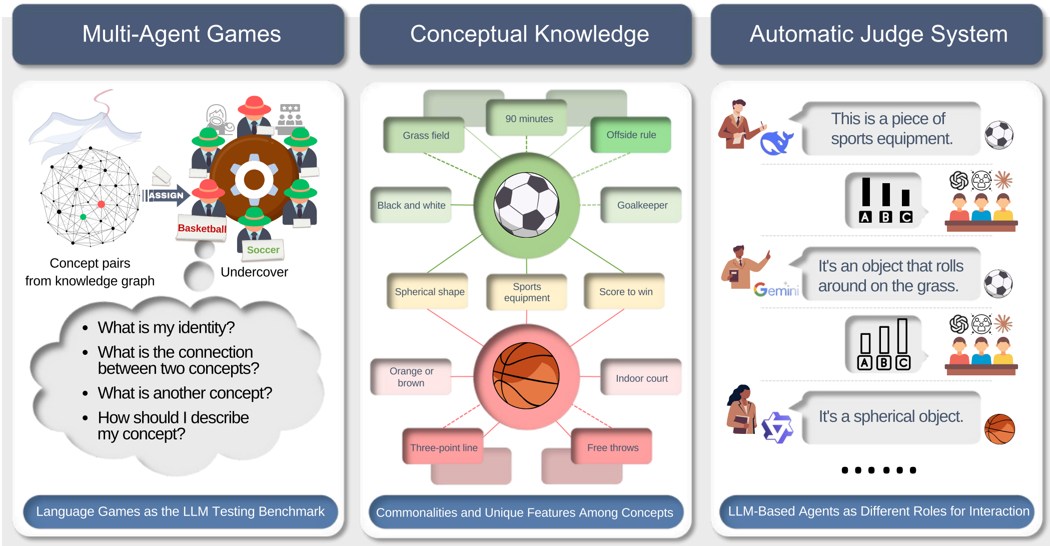

Figure 1: **Conceptual knowledge arena (CK-Arena).** A benchmark designed to evaluate the ability of Large Language Models (LLMs) to understand and reason with conceptual knowledge boundaries. Built upon the interactive game *Undercover*, CK-Arena challenges LLMs to take on roles as players and judges, navigating concept pairs that share both commonalities and unique distinctions. Through multi-agent interaction, LLMs generate descriptive statements, reason about semantic similarities and differences, and make strategic decisions based on partial information. Judges evaluate these interactions based on metrics such as novelty, relevance, and reasonableness, providing insights into the LLMs' conceptual reasoning capabilities in realistic, dynamic environments.

also limits scalability, as creating, maintaining, and updating these benchmarks is time-consuming and labor-intensive. This rigidity makes it difficult to adapt benchmarks to new concepts or evaluate models in evolving real-world scenarios.

In this context, recent work has explored concept-based processing in areas such as conceptual design generation [15], concept editing [9], and abstract concept understanding [16, 17]. However, despite these advances, there is still a lack of systematic benchmarks to evaluate conceptual processing capabilities. A well-designed benchmark is crucial to provide a standardized approach for evaluating LLMs in concept-based tasks, allowing effective measurement, comparison, and improvement of these models in concept comprehension and knowledge application. Simultaneously, interactive game-based environments have gained traction as novel evaluation paradigms to overcome the static nature of traditional benchmarks [18, 19, 20]. Unlike static question-answer formats, game-based evaluations create richer contexts for multi-step reasoning and decision-making. However, most game simulations primarily assess strategic reasoning, offering limited insight into the internal knowledge of models and their ability to convey structured concepts in dynamic multi-agent interactions.

To address the limitations of traditional benchmarks in evaluating conceptual understanding, we propose Conceptual Knowledge Arena (CK-Arena), a multi-agent interaction game benchmark inspired by *Undercover* [21]. Figure 1 illustrates the key aspects involved in this work. Unlike conventional methods that focus on isolated tasks, CK-Arena is designed to assess conceptual reasoning in interactive, multi-agent scenarios. In CK-Arena, participants (LLM-based agents) are assigned one of two similar concepts, representing different identities—*civilian* or *undercover agent*. Without knowing others' identities, agents engage in rounds of dialogue to describe their concepts, analyze others' statements, and attempt to identify undercover agents by discerning commonalities and distinctions. CK-Arena introduces structured evaluation mechanisms, including statement-level metrics for novelty, relevance, and reasonableness, as well as player-level metrics such as win rate and survival rate. To accommodate models with varying reasoning capabilities, CK-Arena also includes a game variant called *Undercover-Audience*, where players focus on shared attributes, and audience agents vote based on perceived inconsistencies. This design allows for scalable, flexible evaluation of conceptual reasoning in interactive settings, reflecting LLMs' ability to navigate semantic boundaries and engage in strategic communication.

Overall, our contributions are as follows: **1) A Game-based Conceptual Reasoning Benchmark:** We introduce CK-Arena, a benchmark built upon the *Undercover* game that effectively gauges LLMs' grasp of conceptual knowledge boundaries. This benchmark is designed to be easily expandable, closely mimics real-world interactive scenarios, and simultaneously evaluates reasoning and decision-making capabilities. **2) Comprehensive Game Variants and Metrics:** We develop multiple versions of the *Undercover* game, tailored to assess not only comprehensive reasoning but also the pure understanding of conceptual boundaries, independent of decision-making strategies. Additionally, we design robust evaluation metrics and an LLM-based automated process to support these assessments. **3) Baseline Evaluation of LLMs:** Using CK-Arena, we evaluate six popular LLMs as baselines for the assessment of conceptual knowledge reasoning. The results show that LLMs have their own strengths and weaknesses in different categories, and there are differences between traditional knowledge quantity assessment results and conceptual knowledge assessment results.

2 Related Work

Benchmarks for Conceptual Knowledge Reasoning. Commonsense reasoning benchmarks play an important role in assessing the capabilities of Large Language Models (LLMs). Widely used benchmarks such as Story Cloze Test [14], Choice of Plausible Alternatives (COPA) [22], and HellaSwag [12] largely rely on static formats like multiple-choice questions or binary judgments. While effective for evaluating factual recall and surface-level understanding, these static formats do not fully reflect real-world interactive scenarios. More recent benchmarks, including MMLU [11], CMMLU [23], BIG-Bench [24], and HELM [13], have introduced tasks such as logical reasoning, cloze tests, and multi-turn Q&A to expand the scope of evaluation. Although these efforts represent progress toward more interactive assessments, they still focus predominantly on factual recall and task-specific reasoning, offering limited insight into how well LLMs understand and manipulate conceptual knowledge boundaries in evolving contexts. In contrast, CK-Arena is designed to explicitly evaluate conceptual reasoning by immersing LLMs in interactive, multi-agent gameplay that requires real-time understanding of semantic boundaries.

Game-based Evaluation. Games provide a unique platform for evaluating AI capabilities, offering interactive and dynamic environments that differ from traditional benchmarks built on static datasets. They have been used to measure various skills, including environmental perception and planning in exploratory games [25, 20], strategic decision-making in competitive games [26, 27], team collaboration in cooperative games [28, 29], and social interaction and language comprehension in communication games [30, 31, 32]. Compared to static evaluations, game-based benchmarks offer more realistic interaction scenarios that better mimic real-world decision-making. However, existing game-based evaluations mainly focus on reasoning and decision-making without specifically addressing the acquisition and application of conceptual knowledge. CK-Arena fills this gap by integrating concept-based reasoning within multi-agent interactions, allowing LLMs to explore and articulate conceptual relationships dynamically, mirroring real-world cognitive processing.

3 Conceptual Knowledge Arena (CK-Arena)

Evaluating the understanding of conceptual knowledge boundaries in large language models (LLMs) is important for assessing their ability to reason with abstract semantic structures. An effective evaluation would consider three key aspects: (1) measuring how well LLMs can distinguish and relate different concepts, reflecting their understanding of semantic boundaries; (2) simulating interactive and dynamic environments that mirror real-world scenarios where concepts are applied flexibly; and (3) covering a diverse range of concepts across multiple domains to enable a comprehensive assessment. To address these needs, we introduce CK-Arena, a benchmark designed to evaluate LLMs' conceptual reasoning through interactive, multi-agent gameplay. CK-Arena provides a setting where models engage in concept-based interactions, encouraging a deeper examination of their ability to reason about similarities, differences, and boundaries between concepts in evolving scenarios.

3.1 The Undercover Game for Evaluation

Game Rule. CK-Arena is built on the multi-agent language game *Undercover* [21], which is originally designed to test the players' reasoning and strategic communication abilities. In the game, players are assigned either as “civilians" who are the majority of the players and know a common word, or as “undercover" who are given a different but related word. Note that each player is informed of their assigned concept word but remains unaware of their team identity or the concepts held by others. Through rounds of description, players must identify who the undercover agents are while undercover agents try to remain undetected by providing descriptions vague enough to seem plausible without revealing their ignorance of the civilians’ word. After each round, players participate in a voting process to eliminate the individual they suspect to be an undercover agent. The game concludes under one of two conditions: (1) if all undercover agents are eliminated, the civilians win; (2) if the number of civilians and undercover agents is equal, the undercover agents win.

Game Variants: Undercover-Audience. During the testing phase of the *Undercover* game in CK-Arena, we observed that certain LLMs with smaller parameter sizes or older architectures struggled with the reasoning and decision-making requirements necessary for effective participation. To accommodate these models, we introduced a variant game mode, called *Undercover-Audience*, designed to simplify the game’s cognitive demands while still evaluating conceptual understanding. In the *Undercover-Audience*, all players are directly informed of both concepts as well as their own identities (civilian or undercover agent). Rather than attempting to obscure their concept, players focus on describing the common features shared between the two concepts. This adjustment reduces the need for strategic reasoning, making the game accessible to models with more limited reasoning capabilities. To replace the traditional voting mechanism, we introduce an audience character. This audience agent is unaware of the two concepts and the identities of the players. After each round of descriptions, the audience agent selects the player whose statements appear most inconsistent or unsociable with the shared features. This modified setup still allows for effective evaluation of conceptual grasp, as successful players must articulate the overlapping characteristics convincingly while avoiding detection.

Why CK-Arena Works. To illustrate the effectiveness of the *Undercover* game used in CK-Arena, consider a concrete example: suppose the concepts *football* and *basketball* are assigned to the players, with *basketball* designated as the undercover concept. During the speaking phase, the undercover player must analyze the descriptions provided by others about *football*, identify shared attributes, and strategically describe *basketball* in a way that overlaps with common features, such as “This is a ball-shaped sports equipment” or “This sport is played by two teams.” This requires more than surface-level token predictions—it involves understanding conceptual commonalities and distinctions. If the player merely relies on token-based generation without grasping these relationships, they are more likely to expose their undercover role, leading to elimination. Thus, performance in CK-Arena reflects the model’s understanding of conceptual knowledge boundaries. Therefore, CK-Arena, with its high demands for conceptual knowledge understanding, active exploration through agent interactions, and extensive knowledge coverage enabled by its scalability, meets the requirements mentioned and can serve as a benchmark for evaluating the understanding of conceptual knowledge.

3.2 Overall Construction and Workflow

Pipeline. CK-Arena involves multiple LLMs as judges and LLM-based players for evaluation, with adjustable group sizes based on the experimental setup. In our experiments, the configuration includes 2 LLM judges and 6 LLM players, consisting of 4 civilians and 2 undercover agents. The game begins with an initialization phase, where players are randomly assigned their roles. Civilians receive a primary concept, while undercover agents are given a similar but distinct concept. During gameplay, each player takes turns making statements that describe their assigned concept while also attempting to identify potential undercover agents or civilians. After each statement, the LLM judges evaluate the description across three criteria: *novelty*, *relevance*, and *reasonableness*. If a player’s statement score falls below a predefined threshold, that player is automatically eliminated. This process continues for a predetermined number of rounds, after which an audience vote determines one additional player for elimination. The game progresses until one of three conditions is met: (1) all undercover agents are eliminated, resulting in a civilian victory; (2) the number of undercover agents matches the number of civilians, resulting in an undercover victory; or (3) the maximum number of rounds is reached. To maintain fairness and mitigate biases that may arise from LLM-based evaluations, multiple LLM judges with different strong base models are employed. The system records the mean and variance of their ratings for each statement. If the variance exceeds a predefined threshold, human reviewers consult a knowledge base to verify the judgment and adjust the final score if necessary.

This mechanism ensures more reliable and unbiased evaluation, enhancing the robustness of the assessment process. Implementation details are provided in Section 4.

Data Preparation The selection of concept pairs is crucial to the effectiveness of the *Undercover* game in CK-Arena. We manually constructed a dataset of semantically related concept pairs spanning a wide range of categories. The dataset underwent pilot screening to ensure two main properties: (1) Semantic proximity: concepts are sufficiently similar to create challenging gameplay yet distinct enough for meaningful differentiation; (2) Descriptive clarity: concepts are expressive enough to enable smooth interactions during the game. The final dataset consists of 529 English concept pairs, which include different parts of speech and semantic category emphases. For details, see the appendix. You can check the word statistics in the appendix E or view the source files in our project.

Prompt Design. To ensure effective communication and role-specific behavior, we designed tailored prompts for each type of LLM-based agent in CK-Arena, including players, judges, and audience members. (1) *Player*: the prompts include a comprehensive system prompt that provides game rules, input-output format guidelines, specific task instructions, basic strategic guidance, and example descriptions. In addition, each player receives a contextualized user prompt containing information about their assigned concept, historical statements, and analytical insights from previous rounds. (2) *Judge*: The prompts are built on a specialized evaluation framework that aligns with the assessment criteria outlined in Section 3.3. Each evaluation dimension—such as novelty, relevance, and reasonableness—is clearly defined with scoring guidelines and examples, ensuring consistent and transparent evaluations across game rounds. (3) *Audience*: The prompts are designed to summarize game history and player statements, enabling them to make informed elimination decisions based on accumulated evidence and analysis from each round.

Result Collection and Analysis. To ensure robust evaluation, CK-Arena integrates comprehensive data collection mechanisms throughout each gameplay session. Every game instance generates a structured JSON record that captures key information, including game metadata (game ID, timestamp, and selected concepts), player details (player IDs, LLM models, role assignments, and concept assignments), and judge information (judge IDs and model specifications). The system also logs the complete history of player statements with evaluation metrics for novelty, relevance, and reasonableness, along with records of vote decisions and elimination outcomes. Additionally, game summary statistics are collected to provide insights into overall performance and decision-making patterns. Game data is systematically organized by rounds, enabling multi-dimensional analysis of interactions and decision-making. We provide automated processing scripts that aggregate results across multiple game instances, producing statistical summaries and visualizations that highlight key indicators such as decision quality, elimination accuracy, and statement metrics.

3.3 Evaluation Metrics

Player-Level Metrics. To evaluate player agents in CK-Arena, we analyze both objective outcomes across multiple game instances and the statistical properties of their statements during gameplay. Three primary metrics are used to capture different aspects of player performance:

(1) *Win Rate (WR)*: This metric reflects the proportion of games won by the player, serving as a straightforward indicator of their effectiveness in fulfilling their assigned role. A higher win rate suggests stronger conceptual understanding and strategic decision-making during the game, whether as a civilian or undercover agent. (2) *Survival Rate (SR)*: It measures the number of rounds a player remains active before elimination. This metric evaluates a player's ability to navigate social dynamics and avoid suspicion, highlighting their skill in either blending in as an undercover agent or effectively communicating concept understanding as a civilian. (3) *Overall Statement Performance*: This metric aggregates the novelty, relevance, and reasonableness scores of a player's statements across all rounds. It provides a holistic view of the player's ability to generate creative, targeted, and logically consistent descriptions. High performance in this metric indicates a well-rounded capability to articulate concept features while maintaining strategic ambiguity or clarity as required. The formal definitions and scoring methods for these evaluation metrics are presented in Table 1.

Statement-Level Metrics. To assess the quality and effectiveness of individual statements during gameplay, CK-Arena employs three primary evaluation metrics, each quantified on a $0 - 1.0$ scale:

Table 1: **Evaluation metrics for CK-Arena.** Detailed breakdown of the metrics used to assess LLM performance in interactive gameplay.

Metric	Formula	Symbol Definitions
Win Rate (WR)	$\mathrm{WR} = \frac{G_{\mathrm{w}}}{G_{\mathrm{t}}}$	G_{w}: Number of games won by the player G_{t}: Total number of games played by the player
Survival Rate (SR)	$\mathrm{SR} = \frac{R_s}{R_t}$	R_s: Number of rounds the player survived R_t: Total number of rounds in all games
Novelty	$\mathrm{Nov}(s_i) \in [0, 1]$	s_i: Current statement $\mathrm{Nov}(s_i)$: Degree of new information in statement s_i compared to previous statements
Reasonableness	$\mathrm{Rea}(s_i, c) \in [0, 1]$	s_i: Current statement c: Target concept $\mathrm{Rea}(s_i, c)$: Logical coherence between statement s_i and concept c's properties
Relevance	$\mathrm{Rel}(s_i, c) \in [0, 1]$	s_i: Current statement c: Target concept $\mathrm{Rel}(s_i, c)$: Degree to which statement s_i specifically points to concept c

(1) *Novelty*: It measures the extent to which a statement introduces new information compared to previous descriptions in the game. High novelty scores indicate that the statement presents fresh insights or unique perspectives, while low scores suggest repetition or rephrasing of earlier descriptions. The purpose of this metric is to discourage simple repackaging of information and promote creative exploration of concept characteristics. Statements falling below the novelty threshold result in automatic elimination to maintain engagement and meaningful discourse. (2) *Reasonableness*: This metric assesses the logical coherence between the statement and the inherent properties of the assigned concept. High scores indicate that the statement logically matches the concept's attributes, while low scores suggest inconsistent or arbitrary descriptions. Ensuring reasonableness prevents players from making misleading or nonsensical claims during gameplay. Statements that fall below the reasonableness threshold trigger immediate elimination to preserve the integrity of the game. (3) *Relevance*: it evaluates how closely a statement aligns with the target concept. High relevance scores reflect descriptions that are specific and closely tied to the concept, making it easier for civilians to identify undercover agents. Conversely, low relevance indicates vague or overly broad descriptions that could apply to multiple concepts. This metric captures the strategic tension in the game—while civilians benefit from clear and targeted descriptions, undercover agents may intentionally opt for broader statements to avoid detection. Although relevance does not directly measure quality, it serves as a valuable scoring criterion for deeper analysis. All three metrics are generated by the Judge agent and reviewed for consistency and accuracy by human evaluators, ensuring fair and meaningful assessment throughout the game.

4 Experiments

In this section, we present a comprehensive evaluation of various language models within the CK-Arena framework. The experiments are designed to address several key research questions: (1) How do different large language models perform in understanding conceptual knowledge and executing strategic reasoning in interactive gameplay? (2) Is there a correlation between model size and the ability to grasp conceptual boundaries effectively? The experimental data consists of 261 game instances spanning eleven concept categories: *food, landforms, animals, artifacts, tools, people/social, plants, sports, stationery, electronics*, and *sundries*. A total of 3462 conceptual feature descriptions were generated during gameplay. Detailed data statistics can be found in the appendix E. Finally, the results reveal interesting patterns in model capabilities and challenge some common assumptions about model scaling.

4.1 Results of Different Models

Experimental Setting. We use the standard mode of CK-Arena to evaluate six widely adopted large language models, including *Claude-3-5-Haiku-20241022* [33], *GPT-4o-2024-11-20* [34], *Gemini-2.0-Pro-Exp* [35], *DeepSeek-V3* [36], *LLaMA-3.3-70B* [37], and *Qwen2.5-72B* [38]. To ensure consistent and fair evaluation, we selected *GPT-4.1-2025-04-14* [39] and *Claude-3-7-Sonnet-20250219* [40] as the LLM-based judges to score all statements across three key dimensions: novelty, relevance, and reasonableness. Following data collection, a human review panel examined statements with a score variance of 0.04or higher between the two LLM-based judges. A total of 163 statements met this criterion and were re-evaluated by human reviewers. For clarity and differentiation, human-assigned

Table 2: **Performance comparison of large language models in CK-Arena.** Analysis across *Civilian* and *Undercover* roles with multiple performance metrics. All values are normalized between 0 and 1. Higher values indicate stronger performance in WR, SR, Novelty, and Reasonableness. Relevance serves distinct strategic purposes—high relevance aids in concept revelation or verification, while low relevance is advantageous for concealment or misdirection.

LLM	Role	Performance Metrics				
		WR ↑	SR ↑	Novelty ↑	Reasonableness ↑	Relevance
DeepSeek-V3	Civilian	**0.6893**	0.6699	0.8285	0.9519	0.7515
	Undercover	**0.3902**	0.2927	0.8154	0.9449	0.7128
Qwen2.5-72B	Civilian	0.6796	**0.7184**	0.6664	0.9652	0.5988
	Undercover	0.3659	0.2927	0.7118	0.9774	0.6398
GPT-4o-2024-11-20	Civilian	0.6824	0.6588	0.6701	0.9717	0.6521
	Undercover	0.3448	0.2414	0.7405	0.9691	0.6206
Gemini-2.0-pro-exp	Civilian	0.6733	0.6832	0.8248	0.9681	0.7055
	Undercover	0.3488	**0.3256**	**0.8333**	0.9690	0.6575
LLaMA-3.3-70B-instruct	Civilian	0.6556	0.6333	0.8149	0.9716	0.6959
	Undercover	0.3148	0.1481	**0.8333**	**0.9775**	0.6539
Claude-3-5-haiku-20241022	Civilian	0.6237	0.6344	0.7588	0.9542	0.7102
	Undercover	0.2549	0.1765	0.8011	0.9274	0.6926

scores were recorded with four decimal places (*e.g.*, $0.X001$), distinguishing them from the original one-decimal-place LLM scores.

Qualitative Analysis. The performance of each model across different conceptual categories is summarized in Table 2. The results reflect the outcomes of multiple rounds of gameplay in CK-Arena, providing insights into the conceptual reasoning capabilities of each LLM. From the statistical results, the overall win rate (WR) for the civilian role is consistently higher than that for the undercover role. This suggests that playing as an undercover agent demands greater strategic thinking and effective concealment skills. Unlike civilians, undercover agents must deduce the other concept and identify shared attributes between two concepts without revealing their own identity. All LLMs performed well at reasonableness. This is partly attributed to the threshold elimination mechanism applied during the evaluation, which filters out low-scoring statements before final analysis. High reasonableness scores also indicate that current LLMs are capable of understanding tasks and generating structured language descriptions based on basic knowledge. However, differences in win rate and survival rate across models suggest that excelling in CK-Arena is influenced less by raw knowledge or task-tracking ability and more by the model's understanding of concepts and the grasp of relationships between them. Moreover, *DeepSeek-V3* achieved the highest performance across multiple indicators, demonstrating stronger conceptual reasoning and strategic adaptation.

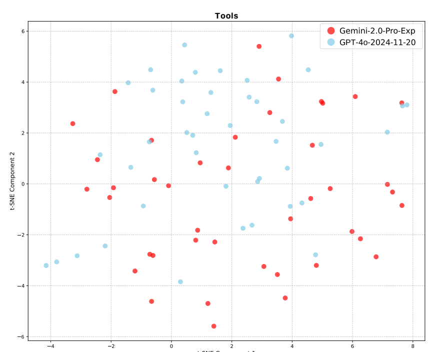

Figure 2: **The t-SNE visualization of all embedded statements in the Tools category for *GPT-4o* and *Gemini-2.0-pro-exp*.** It shows that the distribution of *Gemini-2.0-pro-exp*'s statements is more widespread, while *GPT-4o*'s distribution is more concentrated. This indicates that *Gemini-2.0-pro-exp* captures a broader range of conceptual knowledge, which indirectly reflects a deeper understanding of concepts.

We embedded the statements and compared them through dimensionality reduction and visualization. Assuming the same number of descriptions for the same concept, speakers with a shallow understanding and limited knowledge will inevitably have higher repetition in their descriptions, which will appear as clustered points in the t-SNE plot. In contrast, speakers with a rich knowledge base will exhibit more dispersed and evenly distributed points in the plot. Figure 2 shows the results for one category, and more categories can be found in Appendix E.

Figure 3 illustrates the win rate performance of various LLMs across different conceptual categories. The results highlight clear strengths and weaknesses for each model. For example, DeepSeek-V3

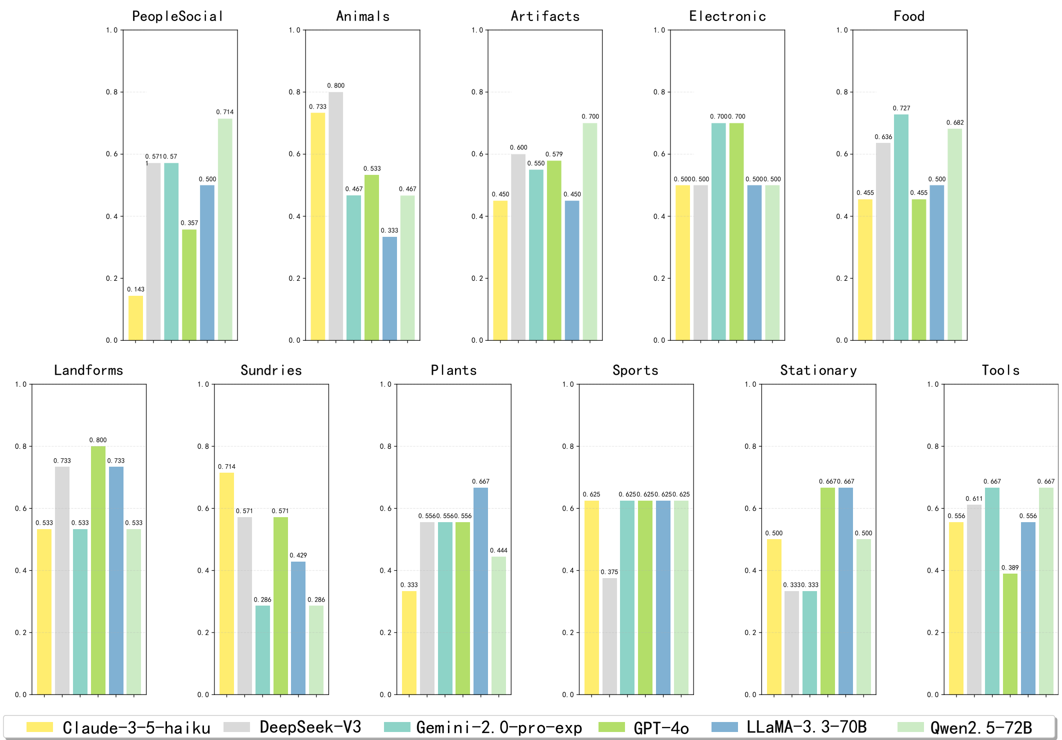

Figure 3: **The win rate performance of six LLMs across** 11 **categories.** A comparative analysis reveals that each model exhibits distinct strengths and weaknesses across different concept categories. These variations are likely influenced by differences in training data, architectural design, and optimization strategies specific to each model. The analysis reveals models' focus areas, knowledge gaps, and insights for improving conceptual reasoning.

achieves the highest win rate in the animal category, reaching 80%, indicating strong domain-specific understanding. Similarly, *GPT-4o-2024-11-20* excels in the landmark category with a win rate of 80%, reflecting its grasp of geographical concepts. In contrast, Claude-3-5-Haiku-20241022 demonstrates a notably low win rate of just 14.3% in the social category, suggesting limitations in handling social context. These performance differences are likely influenced by the models' training datasets and optimization strategies, highlighting domain-specific expertise and gaps in conceptual reasoning.

Figure 4 presents the relevance scores of different LLMs across various conceptual categories. The analysis shows that both the highest-scoring *DeepSeek-V3* and the lowest-scoring *Qwen2.5-72B* consistently rank high in win rate, suggesting that relevance does not directly correlate with win rate. This indicates that achieving high relevance alone is not sufficient for success in CK-Arena. The strategic decision-making likely plays an important role. In addition, comparisons across several categories do not reveal any significant imbalance, demonstrating that the manually selected concepts maintain a high degree of consistency in terms of “describability".

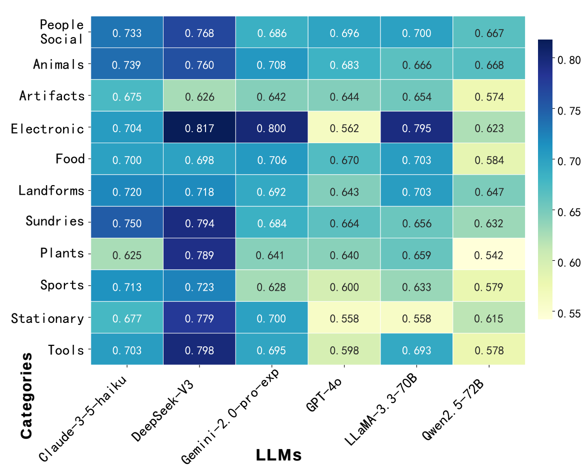

Figure 4: **Relevance scores of different LLMs across various categories**. In this heatmap, the darker the color, the higher the score, intuitively reflecting the association between the descriptions and concepts of each LLM in different categories.

4.2 Results of Models with Different Sizes

Experimental Setting. It is generally assumed that larger models within the same series possess greater capabilities than their smaller counterparts. To test this, we conduct experiments using the *Qwen2.5 series*, specifically comparing the performance of *Qwen2.5-72B*, *Qwen2.5-32B*, and *Qwen2.5-7B*. For these experiments, each model participates as two distinct players in CK-Arena. Given the limitations of *Qwen2.5-7B* in instruction following and decision-making, it is evaluated under the spectator mode described earlier. The referees are implemented using *GPT-4.1-2025-04-14* and *Claude-3-7-Sonnet-20250219*, while the spectator role is handled by *GPT-4.1-2025-04-14*. The experimental findings are interesting—*Qwen2.5-32B* consistently outperforms *Qwen2.5-72B* in both win rate and survival rate, which are comprehensive indicators of gameplay effectiveness. To investigate whether this result is merely due to chance, we conduct one-on-one experiments. In these settings, we design two configurations: (1) four *Qwen2.5-32B* models playing as civilians against two *Qwen2.5-72B* models playing as undercover agents, and (2) four *Qwen2.5-72B* models as civilians against two *Qwen2.5-32B* models as undercover agents. In both scenarios, *Qwen2.5-32B* consistently achieves a higher win rate, suggesting that its superior performance is not coincidental but indicative of better adaptability and strategic reasoning in CK-Arena.

Table 3: **Performance comparison of *Qwen2.5* models with different parameter sizes (14B, 32B, 72B) in CK-Arena.** While the 72B model achieves the highest Reasonableness, suggesting strong language quality, the 32B model demonstrates better performance in capturing conceptual commonalities, as reflected in its superior Win Rate (WR) and Survival Rate (SR).

LLM	Role	Performance Metrics				
		WR ↑	SR ↑	Novelty ↑	Reasonableness ↑	Relevance
Qwen2.5-14B	Civilian	0.3636	0.4935	0.6916	0.9539	0.5707
	Undercover	0.5385	<u>0.4872</u>	0.6512	0.9593	0.5512
Qwen2.5-32B	Civilian	**0.4368**	<u>0.5402</u>	0.7175	0.9495	0.5423
	Undercover	**0.7241**	**0.5862**	**0.7242**	0.9468	0.5145
Qwen2.5-72B	Civilian	<u>0.3824</u>	**0.5735**	<u>0.7205</u>	**0.9776**	0.6577
	Undercover	<u>0.5833</u>	0.4792	0.7108	<u>0.9774</u>	0.5893

Qualitative Analysis. The results in Table 3 and Table 4 show that model size does not always correlate with better performance in CK-Arena. Larger models, despite more training data, sometimes exhibit illusions and misunderstandings of concept relationships, affecting their strategic reasoning. Specifically, Qwen2.5-32B outperforms Qwen2.5-72B as an undercover agent, suggesting that effective conceptual understanding is not solely dependent on parameter size. These findings highlight the importance of model design and data alignment over raw scale for conceptual reasoning.

Table 4: **One-on-one performance comparison between *Qwen2.5-72B* and *Qwen2.5-32B* in CK-Arena.** The 32B model consistently achieves higher win rates than the 72B model in both civilian and undercover roles, suggesting that its advantage is not due to random variation.

LLM	Role	WR ↑	SR ↑
Qwen2.5-72B	Civilian	0.5000	0.5417
	Undercover	0.4483	0.3966
Qwen2.5-32B	Civilian	0.5517	0.6379
	Undercover	0.5000	0.3500

5 Conclusion

We introduce CK-Arena, a benchmark designed to evaluate the ability of Large Language Models (LLMs) to understand conceptual knowledge boundaries through interactive, multi-agent gameplay. Built upon the *Undercover* game, CK-Arena provides a scalable and dynamic environment where models reason about semantic similarities and distinctions—an aspect of human-like understanding that traditional static benchmarks often overlook. Our experimental results indicate that LLMs show varying degrees of conceptual understanding across different categories, and this understanding does not consistently align with model size or overall capabilities. This observation suggests that larger parameter counts do not necessarily lead to better conceptual reasoning, highlighting the need for targeted evaluation of semantic comprehension. CK-Arena serves as a step toward bridging this gap, encouraging further exploration into conceptual reasoning as an essential capability for advancing LLMs toward more robust and human-like understanding.

References

[1] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[2] Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. Benchmarking knowledge boundary for large language models: A different perspective on model evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2270–2286, 2024.

[3] Avinash Patil. Advancing reasoning in large language models: Promising methods and approaches. *arXiv preprint arXiv:2502.03671*, 2025.

[4] Sudarshan Nagavalli, Sundar Tiwari, and Writuraj Sarma. Large language models and nlp: Investigating challenges. *Opportunities, and the Path to Human-Like Language Understanding*, 11:571–584, 2024.

[5] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, pages 481–492, 2012.

[6] Yu Gong, Kaiqi Zhao, and Kenny Zhu. Representing verbs as argument concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[7] Lei Ji, Yujing Wang, Botian Shi, Dawei Zhang, Zhongyuan Wang, and Jun Yan. Microsoft concept graph: Mining semantic concepts for short text understanding. *Data Intelligence*, 1(3):238–270, 2019.

[8] Ningyu Zhang, Qianghuai Jia, Shumin Deng, Xiang Chen, Hongbin Ye, Hui Chen, Huaixiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, et al. Alicg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 3895–3905, 2021.

[9] Xiaohan Wang, Shengyu Mao, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, Huajun Chen, and Ningyu Zhang. Editing conceptual knowledge for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 706–724, 2024.

[10] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

[12] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, 2019.

[13] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

[14] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James F Allen. Lsdsem 2017 shared task: The story cloze test. In *2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51. Association for Computational Linguistics, 2017.

[15] Kevin Ma, Daniele Grandi, Christopher McComb, and Kosa Goucher-Lambert. Conceptual design generation using large language models. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 87349, page V006T06A021. American Society of Mechanical Engineers, 2023.

[16] Jiayi Liao, Xu Chen, and Lun Du. Concept understanding in large language models: An empirical study. 2023.

[17] Liuqing Chen, Duowei Xia, ZhaoJun Jiang, Xinyang Tan, Lingyun Sun, and Lin Zhang. A conceptual design method based on concept–knowledge theory and large language models. *Journal of Computing and Information Science in Engineering*, 25(2), 2025.

[18] Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. Agentsims: An open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026*, 2023.

[19] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

[20] Yue Wu, So Yeon Min, Shrimai Prabhumoye, Yonatan Bisk, Russ R Salakhutdinov, Amos Azaria, Tom M Mitchell, and Yuanzhi Li. Spring: Studying papers and reasoning to play games. In *Advances in Neural Information Processing Systems*, volume 36, pages 22383–22687, 2023.

[21] Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See-Kiong Ng, and Jiashi Feng. MAgIC: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7315–7332, 2024.

[22] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95, 2011.

[23] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11260–11285, 2024.

[24] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TRANSACTIONS ON MACHINE LEARNING RESEARCH*, 2022.

[25] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, Yitao Liang, and Team CraftJarvis. Describe, explain, plan and select: interactive planning with large language models enables open-world multi-task agents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34153–34189, 2023.

[26] Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. Chessgpt: Bridging policy learning and language modeling. In *Advances in Neural Information Processing Systems*, volume 36, pages 7216–7262, 2023.

[27] Weiyu Ma, Qirui Mi, Yongcheng Zeng, Xue Yan, Runji Lin, Yuqiao Wu, Jun Wang, and Haifeng Zhang. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. *Advances in Neural Information Processing Systems*, 37:133386–133442, 2024.

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

[29] Manuel Mosquera, Juan Sebastian Pinzon, Manuel Rios, Yesid Fonseca, Luis Felipe Giraldo, Nicanor Quijano, and Ruben Manrique. Can llm-augmented autonomous agents cooperate?, an evaluation of their cooperative capabilities through melting pot. *arXiv preprint arXiv:2403.11381*, 2024.

[30] Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

[31] Dan Qiao, Chenfei Wu, Yaobo Liang, Juntao Li, and Nan Duan. Gameeval: Evaluating llms on conversational games. *arXiv preprint arXiv:2308.10032*, 2023.

[32] Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. Deciphering digital detectives: Understanding llm behaviors and capabilities in multi-agent mystery games. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8225–8291, 2024.

[33] Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku, 2024.

[34] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[36] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[37] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[38] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[39] OpenAI. Introducing gpt-4.1 in the api, 2025.

[40] Anthropic. Claude 3.7 sonnet and claude code, 2025.

A Future Works

In the future, we plan to extend CK-Arena in several key directions: (1) Expanding the Concept Pair Dataset: We aim to increase the diversity of concept pairs by introducing more categories and refining the quality of selections, thereby building a more comprehensive knowledge network for evaluation. (2) Multilingual Extension: Adapting CK-Arena to support multiple languages holds significant potential. Different languages are deeply tied to unique cultural knowledge and conceptual representations, which can reveal cross-linguistic differences in conceptual understanding. (3) Diversifying Agent Forms: Beyond standard LLM-based agents, we intend to incorporate specialized language models trained in specific knowledge domains to serve as judges, and even explore scenarios where LLM-based agents interact and compete alongside human participants.

Furthermore, the rich set of statements generated during CK-Arena gameplay represents a valuable resource. These concept-driven descriptions can form a semantic norm, potentially serving as raw data for training concept-aware models, such as Large Concept Models (LCMs). Although the current dataset is functional, we aim to further enhance the automation process and evaluation system to transform this data into a high-quality, structured dataset. This would enable more effective training and evaluation of models designed for conceptual reasoning and knowledge-based tasks.

B Limitations

Despite its contributions, the CK-Arena benchmark has several limitations that are worth considering. First, CK-Arena is currently effective primarily for evaluating noun-based concepts, as nouns typically provide richer descriptive attributes for interactive reasoning. Extending the framework to assess verbs or abstract concepts remains an open challenge. Second, in the automated process of CK-Arena, the LLM (Large Language Model) serving as the judge needs to be a powerful and knowledgeable model (for example, in our experiments, we used the latest GPT-4.1 and Claude-3.7). During the process where the LLM acts as an agent, there can also be issues of hallucination. We need to set certain processing measures and manual screening to ensure the reliability of the final results. Third, all evaluations are conducted exclusively in English, which may introduce language-specific biases and limit cross-linguistic understanding of conceptual reasoning. Finally, the multi-agent interaction design of CK-Arena demands significant computational resources, which may restrict accessibility for researchers with limited computing power. Addressing these challenges is essential for enhancing the scalability and inclusiveness of the benchmark.

C Ethical Statement

This research was conducted following established ethical guidelines for AI research. Our benchmark CK-Arena evaluates AI systems' conceptual knowledge without collecting or processing any personally identifiable information. All concept pairs used in our experiments were carefully curated to ensure they do not contain harmful, offensive, or culturally insensitive content. The experiments involving multiple large language models were designed to analyze their capabilities in understanding conceptual boundaries without any deception or manipulation techniques. We have made our prompts, methodologies, and evaluation criteria transparent to ensure reproducibility and to facilitate critical assessment by the research community.

D Prompts

This section presents the detailed prompts used for different agent roles in CK-Arena. We provide the complete prompts to ensure reproducibility and transparency of our experimental setup.

```
You are an AI player participating in the "Undercover" game. You need
↪ to analyze the situation based on the information received,
↪ determine your identity, and devise appropriate speaking strategies
↪ and content.

# Game Rules

1. Each player receives a word. The majority of players receive the
↪ same word (civilians), while a minority (1-2 players) receive a
↪ different but related word (undercover agents).
2. The game proceeds in turns, with each player using one sentence to
↪ describe their word without directly saying it.
3. After each round of descriptions, all players vote for who they
↪ think is the undercover agent. The player with the most votes is
↪ eliminated.
4. If all undercover agents are eliminated, the civilians win; if the
↪ number of undercover agents equals or exceeds the number of
↪ civilians, the undercover agents win.

# Speaking Requirements

1. Your statement must be a brief descriptive sentence, not a lengthy
↪ exposition.
2. You cannot repeat statements made by other players in previous
↪ rounds.
3. Your description can be broad or specific, but must match the word
↪ you received. You cannot give descriptions that do not match your
↪ word.
4. Please adjust the level of detail in your description according to
↪ your strategic needs. Below are examples of different levels of
↪ detail.

# Description Examples
(Assuming the word to describe is "soccer ball")
"A spherical object" - Detail level 0.2 (too broad, many objects are
↪ spherical)
"A sports equipment" - Detail level 0.4 (more specific, but still
↪ covers a wide range)
"Mostly contacted by the lower body of athletes" - Detail level 0.6
↪ (more specific, stronger directional indication)
"Commonly depicted with a pattern of black and white pentagons and
↪ hexagons" - Detail level 0.8 (very specific, almost only soccer
↪ balls look like this)
"One of the most popular sports in the world, seen being kicked and
↪ headed by athletes on a green field" - Detail level 1.0 (completely
↪ points to soccer ball)

# Your Task

1. Based on the given word and other players' statements, analyze your
↪ possible identity (civilian or undercover agent)
2. With the goal of protecting yourself and accomplishing your game
↪ objective, provide your statement content.
3. Provide your analysis and decision-making process in JSON format
```

Listing 1: Player Agent System Prompt for Speaking Phase (I)

```
# Output Requirements

You must respond in JSON format, including the following fields:
{
"identity": "Analysis of your own and other players' identities",
"strategy": "Your thinking and decision-making process",
"statement": "Your final statement (you cannot include your analysis
↪ process in the statement field, and you cannot directly mention your
↪ word)"
}

# Strategy Tips

### At the beginning of the game or when identity is still undetermined:

start with very vague, broad characteristics or properties, then
↪ provide more detailed descriptions of the word after gradually
↪ determining your identity situation.

### As a civilian (you need to determine your civilian identity
↪ yourself):

Analyze other players' statements to find descriptions inconsistent
↪ with the majority
Gradually narrow down the word range to help identify the undercover
↪ agent
Ensure your description matches your word, don't say anything
↪ inconsistent with it

### As an undercover agent (you need to determine your undercover
↪ identity yourself):

Carefully analyze civilian statements to infer their word
Use vague descriptions that won't immediately arouse suspicion
Ensure your description can match both your word and the civilians'
↪ word (ambiguity)
Avoid descriptions obviously different from others, but don't
↪ completely follow them
```

Listing 2: Player Agent System Prompt for Speaking Phase (II)

```
Here is information related to this game round. Please analyze this
↪ information to complete your task.
    # Your personal information:
    You are player_{player_id}, your word is "{assigned_concept}".
    # Statement history for this game round:
    {statement_history}
    # Your identity analysis from the previous round:
    {last_analyze}
```

Listing 3: Player Agent User Prompt Template for Speaking Phase

```
You are an AI player participating in the "Undercover" game. You need
↪ to analyze the situation based on the information received,
↪ determine your identity, and decide to vote for a player to try to
↪ eliminate them.

# Game Rules

1. Each player receives a word. The majority of players receive the
↪ same word (civilians), while a minority (1-2 players) receive a
↪ different but related word (undercover agents).
2. The game proceeds in turns, with each player using one sentence to
↪ describe their word without directly saying it.
3. After each round of descriptions, all players vote for who they
↪ think is the undercover agent. The player with the most votes is
↪ eliminated.
4. If all undercover agents are eliminated, the civilians win; if the
↪ number of undercover agents equals or exceeds the number of
↪ civilians, the undercover agents win.

# Your Task

1. Based on the given word and other players' statements, analyze your
↪ possible identity (civilian or undercover agent)
2. With the goal of protecting yourself and accomplishing your game
↪ objective, provide your statement content.
3. Provide your analysis and decision-making process in JSON format

# Output Requirements

You must respond in JSON format, including the following fields:
{
"identity": "Analysis of your identity",
"strategy": "Thinking about your strategy",
"vote": "The player you want to vote for (just a number, no additional
↪ text explanation needed)"
}
```

Listing 4: Player Agent System Prompt for Voting Phase

```
Here is information related to this game round. Please analyze this
↪ information to complete your task.
    # Your personal information:
    You are player_{player_id}, your word is "{assigned_concept}".
    # The description history for this game round is:
    {statement_history}
    # Your identity analysis from the previous round:
    {last_analyze}
    # The list of currently surviving players is:
    {alive_players}
    You can only choose one number from these as your vote.
```

Listing 5: Player Agent User Prompt Template for Voting Phase

````
# Undercover Game Referee Guide

You are the referee for the "Undercover" game, responsible for analyzing each player's statement and scoring it according to the following criteria. You need to output your evaluation results in JSON format.

## Evaluation Dimensions

1. **Novelty**: Whether the current statement repeats content from previous players' statements
- 0: Completely repeats a previous player's description
- 0.2: Mostly repetitive, with only minimal new information
- 0.4: Partially repeats previous content, with some additional content
- 0.6: Mainly new content, but with some repetition
- 0.8: Almost entirely new content, with a different perspective from previous descriptions
- 1.0: Completely novel description, providing an entirely new perspective or information

2. **Relevance**: The degree of relevance and specificity between the statement and the word's characteristics
- 0: Completely irrelevant
- 0.2: Extremely broad description, applicable to a large number of objects/concepts
- 0.4: Broad but somewhat directional description
- 0.6: Clear clues, but could still point to multiple similar things
- 0.8: Highly specific description, basically only pointing to the target word or very few similar objects
- 1.0: Description that almost directly points to the word

3. **Reasonableness**: How reasonable the association between the description and the word is
- 0: Completely unreasonable, impossible to establish any association with the word
- 0.2: Extremely far-fetched association
- 0.4: Some association but rather far-fetched
- 0.6: Reasonable but not unique association
- 0.8: Highly reasonable association
- 1.0: Description completely matching the word's characteristics

## Output Format

```json
{
    "novelty": {
    "score": Value between 0 and 1 (limited to 0, 0.2, 0.4, 0.6, 0.8, 1),
    "explanation": "Explanation for why this score was given"
    },
    "relevance": {
    "score": Value between 0 and 1 (limited to 0, 0.2, 0.4, 0.6, 0.8, 1),
    "explanation": "Explanation for why this score was given"
    },
    "reasonableness": {
    "score": Value between 0 and 1 (limited to 0, 0.2, 0.4, 0.6, 0.8, 1),
    "explanation": "Explanation for why this score was given"
    }
}
```
````

Listing 6: Judge Agent System Prompt (I)

````
## Scoring Reference Examples

### Example 1: Soccer Ball

Assume the word is "soccer ball", player’s statement is "a spherical
↪ object", with no previous player statements:

```json
{
    "novelty": {
    "score": 1.0,
    "explanation": "This is the first statement, so it’s completely
↪ novel"
    },
    "relevance": {
    "score": 0.2,
    "explanation": "The description is very broad, applicable to any
↪ spherical object, doesn’t provide characteristics unique to a soccer
↪ ball"
    },
    "reasonableness": {
    "score": 1,
    "explanation": "The description is completely reasonable, a soccer
↪ ball is indeed a spherical object"
    }
}
```

### Example 2: Soccer Ball

Assume the word is "soccer ball", player’s statement is "one of the
↪ most popular sports in the world, can be seen being kicked by people
↪ on a green field", previous players have said "a spherical object"
↪ and "a black and white object":

```json
{
    "novelty": {
    "score": 1.0,
    "explanation": "The description provides completely new
↪ information, focusing on soccer ball as a sport attribute and usage
↪ scenario, completely different from previous descriptions focusing
↪ on appearance"
    },
    "relevance": {
    "score": 1.0,
    "explanation": "The description is highly relevant, ’being kicked
↪ by people on a green field’ directly points to a soccer ball, with
↪ almost no other possibilities"
    },
    "reasonableness": {
    "score": 1.0,
    "explanation": "The description is completely reasonably associated
↪ with a soccer ball, mentioning core features of soccer"
    }
}
```
````

Listing 7: Judge Agent System Prompt (II)

```
### Example 3: Soccer Ball

Assume the word is "soccer ball", player's statement is "it gives me a
↪ headache", previous players have said "a ball that can be kicked"
↪ and "used on a green field":

```json
{
    "novelty": {
    "score": 0.8,
    "explanation": "The description provides a new perspective (related
↪ to bodily sensation), completely different from previous
↪ descriptions focusing on physical characteristics and usage
↪ scenarios"
    },
    "relevance": {
    "score": 0.4,
    "explanation": "The description provides some clues (possibly
↪ alluding to headers), but is very vague, many things could cause
↪ headaches"
    },
    "reasonableness": {
    "score": 0.2,
    "explanation": "Although one could connect this to how heading a
↪ soccer ball might cause headaches, this association is quite
↪ far-fetched and not a typical or direct characteristic of soccer
↪ balls"
    }
}
```

Example 4: Soccer Ball

Assume the word is "soccer ball", current player's statement is "a ball
↪ kicked on grass", a previous player has said "a ball used on a green
↪ field":

```json
{
    "novelty": {
    "score": 0.4,
    "explanation": "The description largely repeats the previous 'green
↪ field' concept (grass), only adding the 'kicking' action detail"
    },
    "relevance": {
    "score": 0.8,
    "explanation": "The description is quite specific, 'a ball kicked
↪ on grass' largely points to a soccer ball, but could also be other
↪ ball sports"
    },
    "reasonableness": {
    "score": 1.0,
    "explanation": "The description is completely reasonably associated
↪ with a soccer ball, matching its basic characteristics"
    }
}
```
```

Listing 8: Judge Agent System Prompt

```
Please evaluate the following player's statement.
# Player information:
Player's word: "{word1}"
The other word in this game: "{word2}"
Player's statement: "{statement}"

# Historical statements:
{history}
```

Listing 9: Judge Agent User Prompt Template

E Implementation Details

Detailed data statistics The dataset we provided contains a total of 529 English pairs of concepts, including 220 concrete noun pairs, 100 abstract noun pairs, 109 adverb pairs, and 100 verb pairs. After initial experimental attempts, we concluded that concrete noun pairs are more suitable for our experimental setup and overall research questions. Therefore, for the specific experiments, we selected 11 different categories from the 220 concrete noun pairs. These categories consist of concrete noun pairs that are closest to our daily life and conversational contexts. All of those concepts can be considered with rich and clearly describable features. We believe that starting with these concept pairs can more reliably and steadily complete our experiments and yield preliminary results. In the future, we will further explore the other words.

Details in the experiment When requiring the LLMs to play as a player in the entire interaction process, the LLM's uncertainty can affect the game's progress. To address this, we have added some processing mechanisms in both the code and the game rules to ensure that a single game can proceed smoothly and that multiple games can be batched and continuously traversed through multiple pairs of concepts within a category, thereby improving the efficiency of the experiment. Specifically, the LLM is required to return in JSON format. The returned string will first be preprocessed at the code level to prevent the large model from generating content and symbols outside the specified format. If multiple layers of processing mechanisms fail to resolve the issue, a three-time retry mechanism will be adopted. If the problem still persists, the handling method will depend on the stage of the game. If it is the voting stage, the player will be considered to have forfeited the vote for this round. If it is the speaking stage, the player will be directly expelled for speaking in violation of the rules.

In our first experiment comparing six mainstream LLMs (Large Language Models), we used all 11 pairs of concepts mentioned earlier. Each pair of words was used in at least one game to evaluate their conceptual knowledge abilities on different themes. The evaluation results showed that *Qwen2.5-72B* performed well in the FOOD category, which also had the highest number of concept pairs among the 11 categories. More games in this category can reduce the randomness of the experiment; therefore, we chose this category for testing different sizes of the *Qwen-2.5* series. Due to resource limitations, we did not use the full set of 11 pairs of concepts in the second experiment. However, the additional results from the one-on-one tests with different sizes ensured the reliability of our conclusions.

The stability of the scoring process To verify the stability of the scoring process in our LLM-based evaluation framework (and thereby support the reliability and repeatability of evaluation results), we conducted three independent evaluations on the animal group. Based on the outcomes of these evaluations, we calculated key statistical indicators—mean, variance, and standard deviation—for each of the three core metrics (Novelty, Relevance, and Reasonableness). The specific statistical data are presented in Table 5. This table reflects the stability of the scoring process: all metrics exhibit small variances (ranging from 0.000042 to 0.000302) and standard deviations (ranging from 0.0065 to 0.0174), indicating that the LLM-based scoring results are consistent across repeated evaluations and have low random fluctuation—this provides direct support for the reliability of our evaluation framework.

One-on-one performance comparison between GPT-oss models To further verify the robustness of the conclusion (i.e., larger model size may not directly translate into better conceptual understanding) presented in the main text, we expanded the validation scope beyond the Qwen series—specifically

Table 5: Statistical indicators of three independent evaluations on the animal group.

Metric	Mean	Variance	Std Dev
Novelty	0.8150	0.000203	0.0142
Relevance	0.7034	0.000302	0.0174
Reasonableness	0.9672	0.000042	0.0065

incorporating the latest open-source GPT models (GPT-oss) with distinct parameter scales, and conducting head-to-head comparative tests to cross-validate the aforementioned result.

We selected two GPT-oss models with significant parameter differences: GPT-oss-120B (large-parameter model) and GPT-oss-20B (small-parameter model). Validation experiments were conducted across three representative categories (animals, sports, and food), with key performance metrics including Win Rate, Survival Rate, and average scores of the three core evaluation dimensions (Novelty, Relevance, and Reasonableness). The detailed comparative results of the two models are presented in Table 6.

Table 6: Performance comparison of GPT-oss models with different parameter scales.

LLM	WR	SR	Novelty	Relevance	Reasonableness
GPT-oss-120B	0.4058	0.4783	0.7813	0.5774	0.8890
GPT-oss-20B	0.4638	0.5362	0.7248	0.5810	0.8922

As shown in Table 6, the small-parameter GPT-oss-20B outperforms the large-parameter GPT-oss-120B in four key metrics: WinRate (0.4638 vs. 0.4058), SurvivalRate (0.5362 vs. 0.4783), AvgRelevanceScore (0.5810 vs. 0.5774), and AvgReasonablenessScore (0.8922 vs. 0.8890). This observation is consistent with the trend found in the Qwen series experiments (where larger models also did not show superior knowledge understanding). Collectively, the results from both the Qwen series and the latest open-source GPT models confirm that larger model size does not have a direct and positive correlation with better conceptual understanding, further supporting the reliability of the conclusion in the main text.

Quantitative analysis of embedding distributions We conducted a quantitative analysis of the embedding distributions of target models. Two key statistical metrics were adopted for this analysis: intra-cluster variance (to measure the dispersion of embeddings within clusters) and distribution entropy (to evaluate the semantic diversity of embeddings). The analysis focused on two representative models—Gemini-2.0-Pro-Exp and GPT-4o—and the detailed quantitative results are presented in Table 7.

Table 7: Intra-cluster variance and distribution entropy of model embedding distributions.

Metric	LLM	Score
Intra-cluster variance	Gemini-2.0-Pro-Exp	0.3549
	GPT-4o	0.3224
Distribution Entropy	Gemini-2.0-Pro-Exp	2.9886
	GPT-4o	2.9822

The quantitative results support our qualitative observation. Gemini 2.0 Pro Exp shows higher intra-cluster variance (0.3549 vs 0.3224), indicating approximately 10% greater dispersion in embedding space compared to GPT-4o. The distribution entropy values are nearly identical (2.9886 vs 2.9822), suggesting comparable semantic diversity, with Gemini's slightly higher entropy aligning with the variance findings to confirm marginally greater distributional spread.

More t-SNE plots In addition to the Tools category, we also embedded and visualized the other 10 categories, as shown in the following figure.

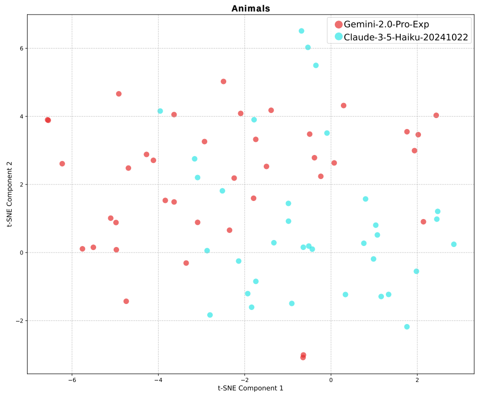

Figure 5: The t-SNE visualization of all embedded statements in the Animals category for *Gemini-2.0-Pro-Exp* and *Claude-3-5-Haiku-20241022*.

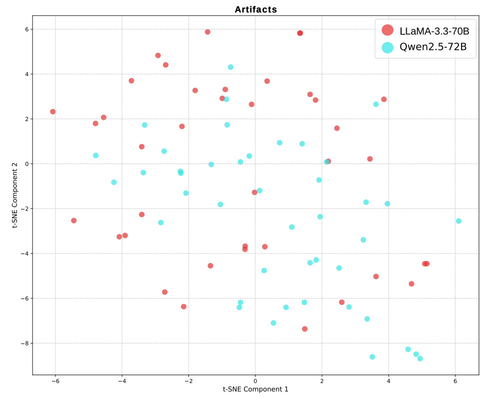

Figure 6: The t-SNE visualization of all embedded statements in the Artifacts category for *LLaMA-3.3-70B* and *Qwen2.5-72B*.

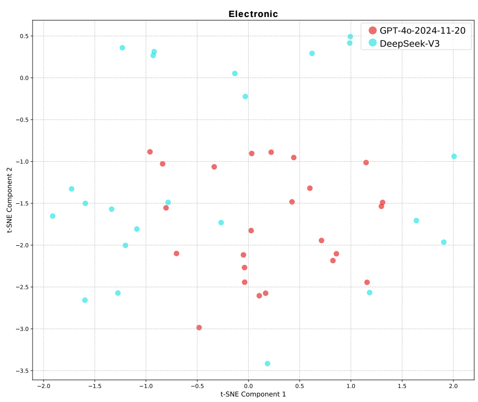

Figure 7: The t-SNE visualization of all embedded statements in the Electronic category for *GPT-4o-2024-11-20* and *DeepSeek-V3*.

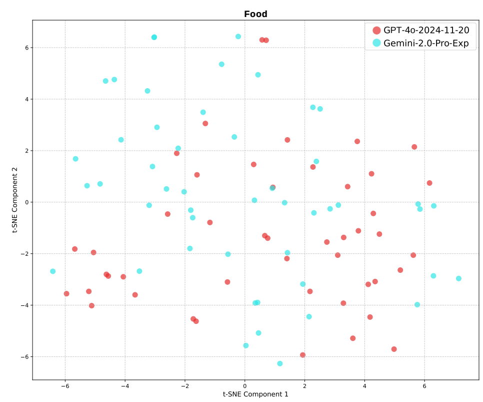

Figure 8: The t-SNE visualization of all embedded statements in the Food category for *GPT-4o-2024-11-20* and *Gemini-2.0-Pro-Exp*.

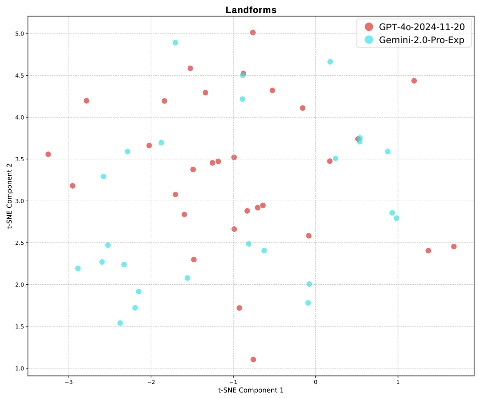

Figure 9: The t-SNE visualization of all embedded statements in the Landforms category for *GPT-4o-2024-11-20* and *Gemini-2.0-Pro-Exp*.

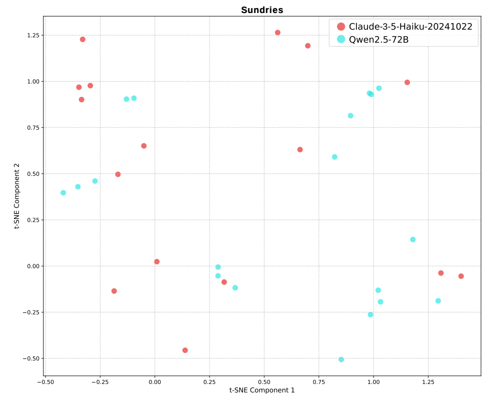

Figure 10: The t-SNE visualization of all embedded statements in the Sundries category for *Claude-3-5-Haiku-20241022* and *Qwen2.5-72B*.

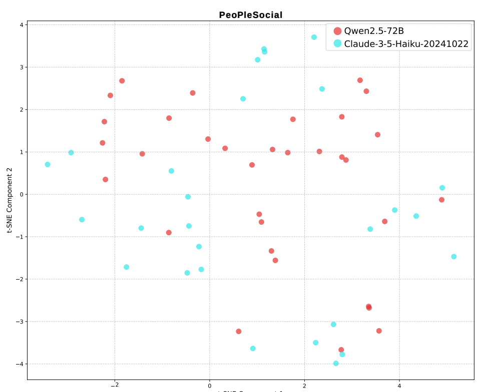

Figure 11: The t-SNE visualization of all embedded statements in the People/social category for *Qwen2.5-72B* and *Claude-3-5-Haiku-20241022*.

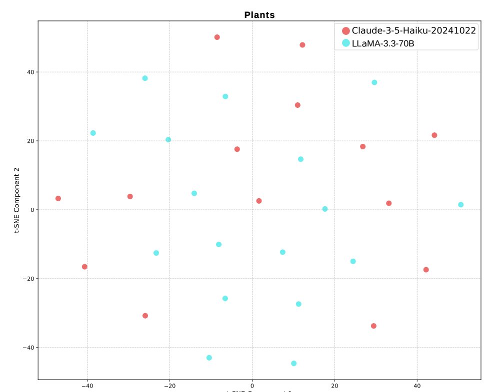

Figure 12: The t-SNE visualization of all embedded statements in the Plants category for *Claude-3-5-Haiku-20241022* and *LLaMA-3.3-70B*.

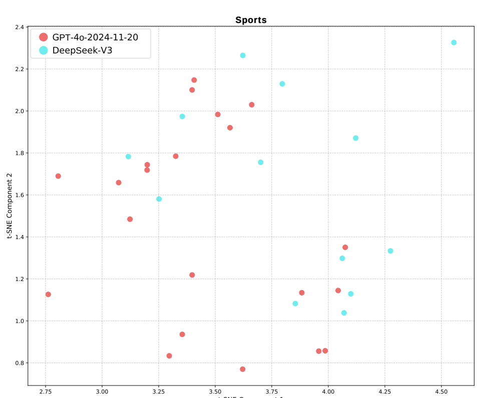

Figure 13: The t-SNE visualization of all embedded statements in the Sports category for *Gemini-2.0-Pro-Exp* and *LLaMA-3.3-70B*.

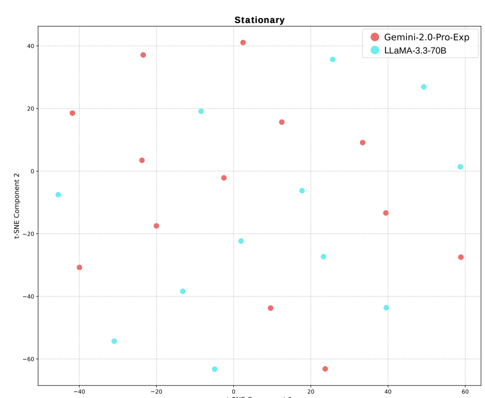

Figure 14: The t-SNE visualization of all embedded statements in the Stationary category for *LLaMA-3.3-70B* and *Qwen2.5-72B*.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In section B of the paper, we discuss the limitations of our work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results in this article are accompanied by complete hypotheses and proofs. We provided a detailed introduction to the benchmark architecture, the underlying principles, and the corresponding experimental data used to support the final results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have fully disclosed all the information needed to reproduce the main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides open access to the data and code in `https://ck-arena.site`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper explains the complete experimental setup, including the selection of baselines, threshold settings, and the process of processing data.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiment in the "Experiment Results" section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For each experiment, the paper provided sufficient information on the computer resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms to the NeurIPS ethical standards in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper discusses the potential impact of this work on society in the context of "Ethical Statement".

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: This paper discusses the needed safeguards of this work on society in the context of "Ethical Statement".

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets (e.g., code, data, models) used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets introduced in the paper are well documented.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Our paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Since our research does not involve any human subjects, there is no need for IRB approval or equivalent.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.