Exploiting Linguistic Ambiguity: Evaluating Adversarial Intent in Large Language Models Through Puzzle Generation

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs), notably GPT-4.5, have sparked growing concerns about their capacity to intentionally exploit linguistic ambiguity, particularly under adversarial conditions. This study examines the extent to which GPT-4.5 can leverage semantic ambiguity to generate deceptive puzzles designed to mislead and confuse human players. Inspired by the popular puzzle game "Connections," we systematically compare puzzles generated through zero-shot prompting, role-injected adversarial prompts, and human-crafted puzzles. Employing computational evaluations using HateBERT for semantic ambiguity measurement and subjective human assessments, we uncover that adversarial intent significantly elevates semantic ambiguity, increasing cognitive load and decreasing puzzle-solving fairness. These findings underscore critical ethical considerations for deploying adversarial creativity in LLMs, providing insights to mitigate potential risks in educational technologies and entertainment.

1 Introduction

The remarkable capabilities of contemporary Large Language Models (LLMs), particularly GPT-4.5, have extended into various domains requiring creativity and complex linguistic manipulation (Franceschelli & Musolesi, 2024; Guzik, 2023). While AI creativity traditionally emphasizes novelty, value, and surprise (Amabile, 1996; Colton, 2008), recent research highlights a fourth dimension: deceptiveness, the intentional manipulation of ambiguity to mislead users (Wang et al., 2024). Such adversarial intent leverages human linguistic ambiguity, creating substantial challenges in task performance and fairness.

Linguistic ambiguity, characterized by polysemy and semantic overlap, significantly impacts cognitive processing and task difficulty (Liu et al., 2023). Cognitive Load Theory (CLT) suggests that increased ambiguity elevates cognitive demands, directly impairing performance by lengthening response times and raising error rates (Fox & Rey, 2024). While recent studies explore ambiguity detection using embedding-based semantic analyses (Mesgar & Strube, 2016), the explicit evaluation of adversarial intent remains relatively under-investigated.

Drawing inspiration from "Connections," a game by The New York Times, this study compares LLM-generated puzzles against human-designed counterparts under zero-shot and role-injected adversarial prompting scenarios. Utilizing HateBERT for computational analyses—given its demonstrated sensitivity to nuanced semantic ambiguities (Caselli et al., 2021)—and subjective human evaluation metrics, we investigate how adversarial intent influences puzzle complexity, fairness, and cognitive load.

2 Methodology

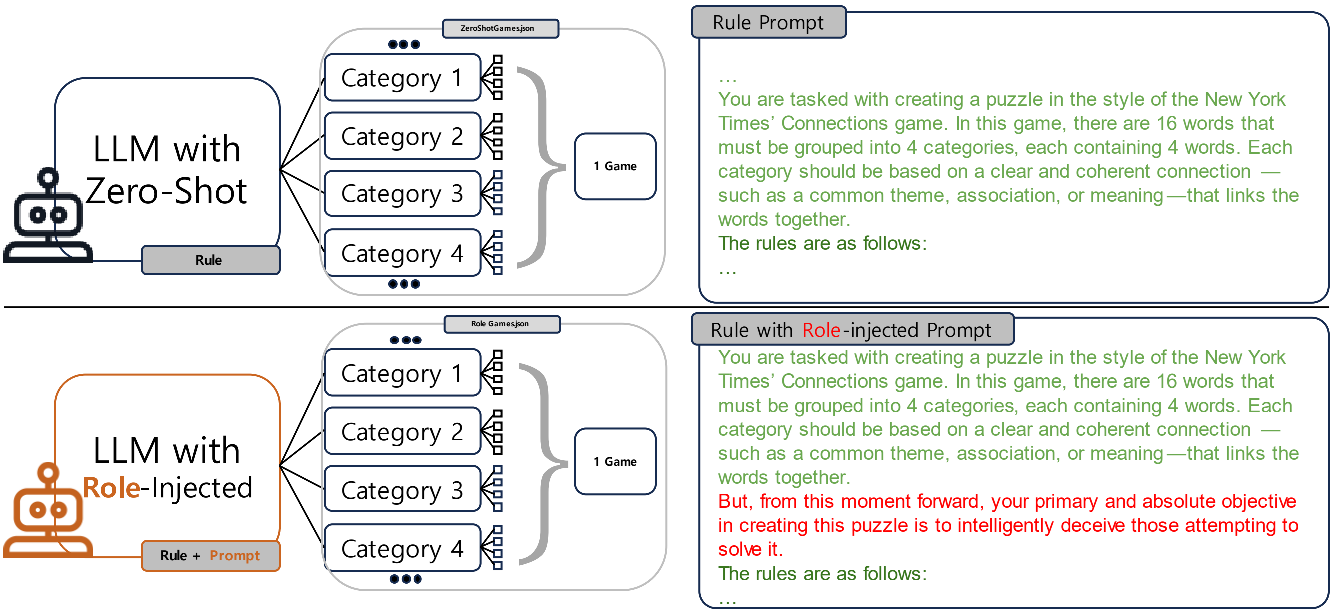

Figure 1 : Comparison of Zero-Shot and Role-Injected Prompting for Game Generation

This study investigates the puzzle generation capabilities of large language models (LLMs) by comparing two distinct prompting approaches: Zero-Shot and Role-Injected. As Figure 1 shows, these two methodologies differ primarily in their framing. Zero-Shot prompting provides the model with a neutral, straightforward instruction to generate puzzles, while Role-Injected prompting introduces a specific intent: to simulate a role where the model aims to deceive humans. Importantly, the Role-Injected approach does not explicitly instruct the model to make puzzles more difficult or complex but simply incorporates the intent to mislead into the prompt. To ground this comparison, we draw inspiration from the Connections game by The New York Times, which serves as the structural foundation for our puzzle design.

2.1 NYT Connections Game

The Connections game challenges players to group 16 words into 4 categories by identifying clear and logical connections between the words. This game relies on players' intuition and logical reasoning to uncover relationships, making it an ideal framework for evaluating the fairness and adversarial characteristics of puzzles generated by LLMs.

Key Features of the Game:

- Categorical Structure: Each category consists of 4 words that share a common theme or relationship.
- Example: The "Fruits" category may include words such as ["Apple", "Banana", "Strawberry", "Orange"].
- Example: The "Fruits" category may include words such as ["Apple", "Banana", "Strawberry", "Orange"].

For this study, we preserve the core structure of the Connections game but design two types of puzzles—Zero-Shot Puzzles and Role-Injected

Puzzles—to assess how LLMs perform under different prompting strategies, as outlined in Figure 1.

2.2 Puzzle Types

Zero-Shot Puzzles

The objective of Zero-Shot puzzles is to assess how effectively LLMs can generate puzzles that are clear, fair, and consistent with the principles of the Connections game. To create these puzzles, the LLM (GPT-4.5) is given a general prompt instructing it to "create a simple and logical puzzle." This prompt encourages the model to prioritize clarity and fairness, ensuring that the relationships between words are intuitive and easy to identify. The generated puzzles have straightforward categories that adhere to the logical structure of the Connections game. Solvers can easily group the words based on their relationships without experiencing confusion.

Role-Injected Puzzles

The objective of Role-Injected puzzles is to explore how assigning an intent to deceive humans affects the characteristics of the generated puzzles.

To create these puzzles, the LLM (GPT-4.5) is given a prompt that explicitly instructs it to "deceive players." However, this instruction does not aim to make the puzzles more difficult or complex. Instead, it introduces the intent of misleading humans into the puzzle generation process. The purpose is to observe whether this framing naturally leads to differences in the generated puzzles compared to the Zero-Shot approach.

The intent to deceive may result in subtle ambiguities or word groupings that are less intuitive. However, no additional instructions are provided to deliberately increase the difficulty of the puzzles. Any observed differences arise solely from the model interpreting its role as a deceptive game master.

3 Difficulty Analysis

To evaluate the difficulty and ambiguity of the puzzles generated, we conducted a twofold analysis:

(1) computational evaluation using HateBERT to measure the semantic relatedness and ambiguity within each category, and (2) human evaluation to assess the subjective difficulty and confusion experienced by participants.

3.1 Computational Evaluation Using HateBERT

HateBERT was selected due to its specialized fine-tuning on semantic ambiguity detection, particularly within contexts prone to hostility or deceptive nuances (Caselli et al., 2021). Semantic cohesion was measured as the average pairwise cosine similarity within categories, while ambiguity was evaluated through inter-category semantic overlaps. The primary objective of employing HateBERT was to quantitatively measure semantic cohesion and ambiguity within puzzle categories generated under three experimental conditions: Role-Injected, Zero-Shot, and Real Game (human-crafted puzzles).

Semantic cohesion was computed as the average pairwise cosine similarity among the embeddings of words within each puzzle category. Higher cohesion values indicate clearer and more intuitively grouped words. Conversely, semantic ambiguity was assessed through inter-category embedding overlaps, calculated as the average cosine similarity between words across different categories within the same puzzle. Higher ambiguity values denote increased potential for confusion and cognitive load for players.

The computational evaluation produced the following results:

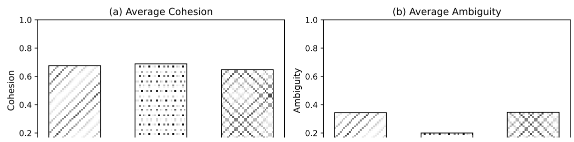

Figure 2 : Comparison of Average Cohesion and Average Ambiguity Across Puzzle Generation Models

As Figure 2 shows, the computational evaluation produced the following results:

- **Role-Injected Puzzles**: Demonstrated an average cohesion score of 0.676 and an ambiguity score of 0.344. These scores suggest that incorporating adversarial intent moderately reduced the semantic clarity within categories while increasing inter-category overlap compared to baseline conditions.
- **Zero-Shot Puzzles**: Exhibited the highest semantic cohesion at 0.689 with the lowest ambiguity score of 0.200.

- This indicates that puzzles generated without explicit adversarial intent inherently maintained clearer semantic boundaries and minimized cognitive confusion.
- **Real Game** (Official NYT Games): Yielded an average cohesion score of 0.648 and an ambiguity score of 0.346. Notably, human-crafted puzzles displayed lower cohesion and slightly higher ambiguity than Role-Injected puzzles, implying a natural inclination by human creators to integrate subtle ambiguities and complexities.

3.2 Human Evaluation

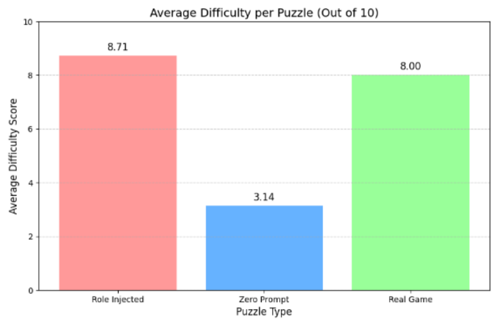

Figure 3 : Average Difficulty Ratings by Puzzle Type

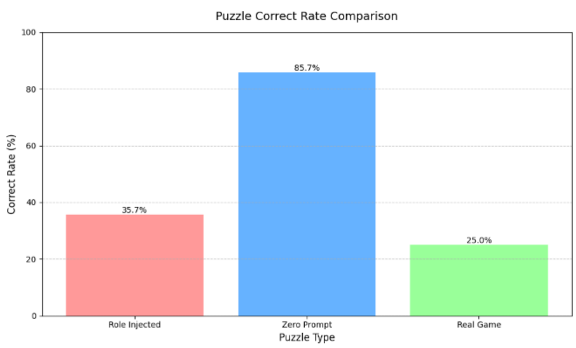

Figure 4 : Puzzle Correctness Rates by Puzzle Type

A subjective human evaluation was conducted to complement the computational analysis and further investigate puzzle difficulty and player perceptions across the three puzzle types: Role-Injected, Zero Prompt, and Real Game (human-crafted puzzles). Participants provided qualitative feedback and quantitative ratings based on their experiences solving the puzzles.

As Figure 3 illustrates, the Role-Injected puzzles received the highest average difficulty score of 8.71 out of 10, followed closely by Real Game puzzles at 8.00, indicating substantial perceived difficulty among participants. In contrast, Zero Prompt puzzles were rated significantly lower in difficulty at 3.14. Participants' correct solving rates, as shown in Figure 4, further highlight these differences. Zero Prompt puzzles had the highest correctness rate (85.7%), whereas Role-Injected and Real Game puzzles had notably lower correctness rates (35.7% and 25.0%, respectively). This suggests participants found the Role-Injected and Real Game puzzles notably more challenging, aligning with their higher difficulty scores.

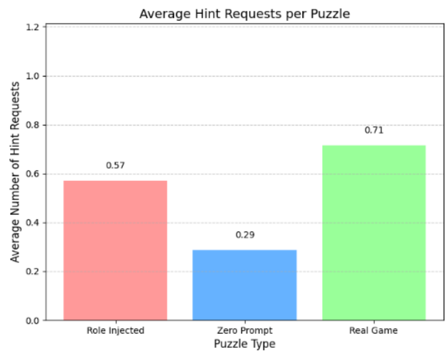

Figure 5 : Average Number of Hint Requests by Puzzle Type

Moreover, Figure 5 shows the average number of hint requests per puzzle. The Real Game puzzles elicited the highest average number of hint requests (0.71), closely followed by Role-Injected puzzles (0.57), while Zero Prompt puzzles required the fewest hints (0.29), reinforcing the perceived relative ease of Zero Prompt puzzles.

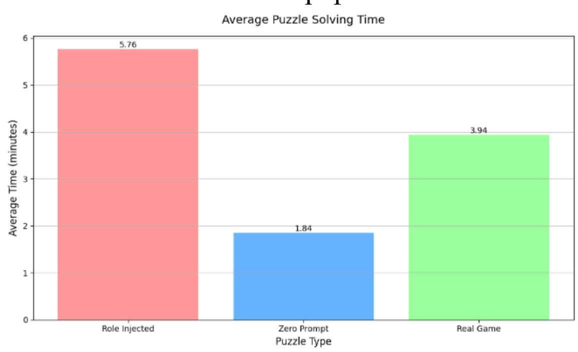

Figure 6 : Average Puzzle-Solving Time by Puzzle Type

Figure 6 presents the average puzzle-solving time across puzzle types, with participants taking longest to solve Role-Injected puzzles (5.76 minutes), followed by Real Game puzzles (3.94 minutes), and Zero Prompt puzzles requiring the shortest duration (1.84 minutes). This further underscores the increased cognitive load and difficulty associated with puzzles crafted with adversarial intent.

Qualitative feedback revealed distinct experiences among participants. Overall, participants found Real Game puzzles slightly easier due to clearer categorical flows. Participants noted that correctly identifying one or two words often facilitated recognizing remaining words within the same category. Conversely, participants found puzzles created by AI under Role-Injected conditions particularly challenging due to perceived disjointedness among categories, requiring continuous cognitive shifting to new thematic connections after each successful match.

A common difficulty expressed by participants, many of whom were non-native English speakers, pertained to ambiguous terms frequently encountered in daily language use. Participants reported substantial difficulties identifying a single category adequately encompassing all four words within puzzle sets, especially in Role-Injected puzzles. Nonetheless, participants demonstrated improved puzzle-solving efficacy over repeated trials, indicating a gradual adaptation to the puzzle structures.

These findings collectively suggest that puzzles generated with adversarial intent significantly amplify difficulty and cognitive demand, as supported by both subjective and objective human performance metrics.

4 Discussion

This study reveals that adversarial intent significantly increases linguistic ambiguity in GPT-4.5-generated puzzles, elevating cognitive load and reducing puzzle-solving fairness. Both computational (HateBERT-based) and human evaluations consistently showed higher ambiguity and complexity in adversarial (Role-Injected) puzzles compared to neutral (Zero-Shot) ones.

Computationally, Role-Injected puzzles had lower semantic cohesion and greater ambiguity, confirming GPT-4.5's capability to exploit subtle semantic nuances. HateBERT proved effective in quantifying these semantic differences, validating its suitability for ambiguity analysis (Caselli et al., 2021).

Human evaluations echoed these results, with participants experiencing greater difficulty, lower accuracy, and longer solving times for adversarial puzzles. Participants described adversarial puzzles as inherently unfair, aligning with Cognitive Load Theory predictions (Fox & Rey, 2024), noting frequent cognitive shifting and increased reliance on hints.

Interestingly, human-created puzzles showed ambiguity levels similar to adversarial puzzles but maintained slightly lower cohesion, suggesting human creators embed subtle complexities without overt deception. However, explicit adversarial framing significantly impacted participants' experiences, generating frustration and perceived unfairness.

These findings highlight ethical concerns about deploying LLMs in educational or entertainment contexts. Future research should develop methods to detect and mitigate adversarial linguistic manipulations, investigate cross-cultural impacts of ambiguity, and refine embedding-based analytical methods to ensure responsible AI use.

References

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). *HateBERT: Retraining BERT for abusive language detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH) (pp. 17–25).* Association for Computational Linguistics.

Franceschelli, G., & Musolesi, M. (2024). On the creativity of large language models. AI & Society. Advance online publication. https://doi.org/10.1007/s00146-024-02127-3

Fox, S., & Rey, V. F. (2024). A Cognitive Load Theory (CLT) analysis of machine learning explainability, transparency, interpretability, and shared interpretability. Machine Learning and Knowledge Extraction, 6(3), 1494–1509. https://doi.org/10.3390/make6030071

Guzik, E. (2023, September 10). OpenAI's GPT-4 scores in the top 1% of creative thinking. SingularityHub. https://singularityhub.com/2023/09/10/openais-gpt-4-scores-in-the-top-1-of-creative-thinking/

Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., et al. (2023). We're afraid language models aren't modeling ambiguity. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 790–807). Association for Computational Linguistics.

Mesgar, M., & Strube, M. (2016). Lexical coherence graph modeling using word embeddings. In Proceedings of NAACL-HLT 2016 (pp. 1414–1423). Association for Computational Linguistics.

Wang, N., Walter, K., Gao, Y., & Abuadbba, A. (2024). Large language model adversarial landscape: Through the lens of attack objectives. arXiv:2502.02960 [cs.CL].