Journal of Advanced Artificial Intelligence
Volume 2 - No.3, December 2025

Prompting Creativity: Analyzing the Impact of Prompt Engineering Philosophies on AI Gameplay in Codenames

Aditya Patil
Independent Researcher

Zeeshan Ahmad
The University of Texas at Dallas

Prashanth Reddy
The University of Texas at Dallas

Apurva Shrivastava
Amazon

Ganesh Bankey
Docusign

ABSTRACT

The question is whether prompt engineering significantly impacts the creative reasoning of Large Language Models (LLMs) like GPT-4 and Gemini 2.5, moving beyond just factual accuracy to influence how models abstract, associate, and communicate concepts, as demonstrated in the word-association game Codenames. The aim is to see if prompt engineering can unlock the creativity within the objective framework of Codenames. The study evaluates the performance of models using various metrics and prompt engineering styles. The study, which introduced three philosophies (Spectrum Lens, Three Bridges, and Role Shifting), found that structured prompting acts as a cognitive scaffold rather than merely a linguistic interface, fundamentally altering the style and strategy of gameplay without necessarily increasing overall accuracy; specifically, structured guidance boosted Gemini 2.5's conceptual depth and risk-awareness while making GPT-4's creativity more balanced but less spontaneous, suggesting that prompt design reconfigures how LLMs conceptualize and act in complex reasoning tasks. Overall, it is seen that prompt engineering does not have a substantial impact on the creativity aspect of LLMs, but some metrics do shift when LLMs are prompted.

Keywords

Prompt Engineering, Creativity, Large Language Models, Associative Reasoning, Codenames, Artificial Intelligence Evaluation

1. INTRODUCTION

Large language models (LLMs) such as OpenAI's GPT-4 and Google's Gemini have revolutionized the field of natural language processing by demonstrating remarkable fluency, reasoning ability, and adaptability across diverse tasks. Yet, their creativity, the capacity to form novel, contextually appropriate connections, remains one of the least understood aspects of their intelligence. Traditional AI benchmarks measure knowledge recall, factual accuracy, and logical reasoning, but they rarely assess the strategic creativity required to communicate effectively in ambiguous, multi-agent settings.

Prompt engineering has recently emerged as a technique to modulate model behavior, guiding outputs through targeted framing, role assignment, or reasoning scaffolds. While chain-of-thought prompting has been shown to improve logical reasoning [6, 7], few studies systematically analyze how structured prompts affect creative decision-making. In real-world communication, creativity is rarely random; it operates under constraints, balancing novelty, safety, and interpretability. Understanding how prompting strategies influence this balance is essential to developing AI that can collaborate, strategize, and innovate alongside humans.

The word-association game Codenames provides a natural testbed for evaluating such behavior. In this game, players act as Spymasters who must give single-word clues to help their team guess multiple related words while avoiding opponent or "assassin" words. The game inherently demands abstraction, analogy, and risk-aware thinking[14], making it ideal for probing creative reasoning. In prior work, Codenames has been used to benchmark associative creativity, but this study extends that approach by introducing three structured prompting philosophies designed to emulate diverse human cognitive processes.

These prompting frameworks - Spectrum Lens, Three Bridges, and Role Shifting - each embody a different theory of how humans approach complex meaning-making.

The central hypothesis is that prompting alters not only what the model produces but how it reasons, restructuring associative cognition. By comparing unprompted and prompted scenarios across two models, this research examines whether prompts can act as "cognitive filters" that systematically modulate creative strategy.

2. RELATED WORK

The relationship between prompting and reasoning in large language models (LLMs) has been widely explored in recent years, but

the study of prompting as a tool for enhancing creativity, particularly under constrained reasoning environments, remains limited. Traditional AI benchmarks, such as MMLU (Massive Multitask Language Understanding) [1], GSM8K [2], and HELM (Holistic Evaluation of Language Models) [3], focus primarily on factual accuracy, step-by-step reasoning, and robustness. While these tests have advanced systematic evaluation of logic and factual consistency, they fail to capture how models navigate ambiguity, abstraction, and associative thinking, skills essential to creative reasoning.

Research on prompt engineering has shown that the phrasing, framing, and structure of prompts can dramatically alter model output quality. For instance, Chain-of-Thought (CoT) prompting encourages explicit reasoning steps, improving accuracy on multi-step reasoning tasks [6]. Similarly, few-shot prompting demonstrates that strategically selected examples can help models generalize patterns [7]. However, these techniques primarily enhance analytical clarity rather than semantic imagination. Studies such as [8] and [6] highlight how explicit reasoning improves factual correctness but often suppresses divergent or lateral thinking.

In contrast, a smaller but growing body of research examines creativity within LLMs. Benchmarks like BIG-Bench [4] introduced open-ended generative tasks such as joke writing, analogical reasoning, and story continuation. However, evaluation in such tasks often relies on subjective human judgment or broad-scale crowd annotations, limiting reproducibility and objectivity. More structured attempts, such as HELM [3] or AGIEval [5], incorporate qualitative reasoning tasks but lack mechanisms for assessing strategic creativity, the kind of reasoning that blends abstraction with risk assessment.

Game-based AI evaluation has recently emerged as a promising alternative for testing higher-order reasoning. Studies using environments like Go [9], Diplomacy[10], and Minecraft[11] demonstrate that interactive, goal-driven settings can reveal model reasoning patterns that static benchmarks cannot. Among these, the word-association game Codenames offers a unique balance between structure and openness: the rules define a constrained environment, but success requires creative language use, conceptual compression, and strategic inference.

The theoretical foundation for evaluating creative reasoning in AI draws from established cognitive science research on human creativity. Guilford's [13] distinction between convergent and divergent thinking provides a framework for understanding how models might explore multiple solutions before selecting one. Mednick's [14] associative theory of creativity suggests that creative thinking involves forming remote associations between concepts, a principle directly applicable to the Codenames task. Additionally, Boden's [12] framework distinguishing exploratory, combinatorial, and transformational creativity offers a lens through which to evaluate how prompting philosophies might activate different creative mechanisms in LLMs.

Prior work using Codenames as a benchmark evaluated model creativity through metrics such as Penalized Embedding Score and Semantic Risk Index. Those studies found that while LLMs like GPT-4 exhibit strong associative reasoning, their performance varies significantly depending on prompt framing. However, little is known about how specific prompting philosophies systematically shape gameplay behavior, especially across different model architectures.

This research extends the prompt-engineering literature by introducing three cognitive prompting frameworks, Spectrum Lens, Three Bridges, and Role Shifting, that simulate human-like modes of thought. These philosophies go beyond surface-level prompt tuning; they attempt to reorient the model's reasoning structure, encouraging it to traverse different cognitive paths before generating output. In doing so, this work situates itself at the intersection of prompt engineering, computational creativity, and cognitive modeling, providing an interpretable, quantifiable approach to studying how AI systems "think differently" under guided prompting.

3. METHODOLOGY

This section outlines the conceptual foundation and operationalization of the three prompting philosophies - Spectrum Lens, Three Bridges, and Role Shifting, and describes how they were integrated into the Codenames gameplay framework.

3.1 Overview

Each philosophy was designed as a cognitive scaffold, a structured mental framework that shapes how the model perceives relationships between words. In human cognition, creativity often arises from reframing: viewing a problem through multiple mental models or emotional lenses. These prompts attempt to replicate that reframing effect within LLMs by diversifying internal reasoning pathways. All three philosophies were implemented as distinct Spy and Guesser prompts, appended to the system instructions before gameplay.

The study tested six total configurations:

- No Prompt (Baseline) for GPT-4 and Gemini 2.5
- Spectrum Lens Prompt for GPT-4 and Gemini 2.5
- Three Bridges Prompt for GPT-4 and Gemini 2.5
- Role Shifting Prompt for GPT-4 and Gemini 2.5

Each game was simulated in a 5x5 Codenames board environment, using Python scripts that alternated turns between AI Spymasters and Guessers. Metrics included creativity, risk, entropy, clue diversity, and overall win rate.

3.2 Spectrum Lens Philosophy

The Spectrum Lens philosophy encourages models to examine a problem through a variety of interpretive filters before generating a clue or guess. Rather than fixating on a single semantic association, the model is instructed to "cycle through" multiple lenses: factual, creative, emotional, optimistic, systemic, and risk-based. This multi-perspective prompting aims to approximate convergent creativity [13], where a model synthesizes insights across different domains to identify the most balanced and effective clue.
In practice, Spectrum Lens prompts encourage the model to weigh trade-offs between literal and figurative reasoning. For instance, if the target words are "Storm," "Electric," and "Guitar," the model is prompted to consider factual (electric storm), creative (rock music), and emotional (intensity, power) lenses. The resulting clue might reflect a conceptual overlap across these domains - for example, "Thunder" - demonstrating multidimensional reasoning.
This philosophy hypothesizes that by explicitly invoking diverse mental perspectives, models will produce clues that are both semantically rich and strategically safer, reducing the likelihood of misleading associations.

3.3 Three Bridges Philosophy

The Three Bridges philosophy introduces structured cognitive pathways that emulate how humans form connections between concepts. It divides associative reasoning into three categories:

- Logical Bridges – based on factual or categorical relationships.
- Emotional Bridges – grounded in shared human experiences or affective connotations.
- Absurd Bridges – intentionally surreal or metaphorical associations that create unexpected yet valid connections.

Before selecting a clue, the Spymaster model is instructed to map all three pathways and choose the one offering the strongest, safest link between target words. Similarly, the Guesser model evaluates clues by hypothesizing which type of bridge the Spymaster intended.
This philosophy mirrors divergent thinking [13, 16] - the cognitive process of generating multiple solutions before converging on one. By systematizing this exploration, the Three Bridges approach promotes flexibility and robustness, especially in ambiguous board states. Empirically, it encourages a balance between precision (logical reasoning) and creative adaptability (absurd or emotional mapping).

3.4 Role Shifting Philosophy

The Role Shifting philosophy is inspired by cognitive persona theory, the idea that different "modes of mind" yield distinct problem-solving [15] heuristics. The model is asked to adopt one of several thinking personas before generating a clue or guess:

- Scientific (analytical, categorical reasoning)
- Poetic (metaphorical, aesthetic)
- Childlike (simple, innocent, playful)
- Skeptical (contrarian, ironic)
- Wise/Sage-like (philosophical, reflective)
- Comedic (punny, humorous)

Each persona introduces a different vocabulary and risk profile. For example, a Scientific persona might prioritize definitional accuracy, while a Poetic persona emphasizes conceptual resonance. By shifting roles, the model escapes semantic inertia, the tendency to recycle predictable associations, and accesses a broader range of linguistic strategies.
During guessing, this method also helps the second model interpret clues contextually, identifying which "voice" the Spymaster used. If the clue sounds poetic, guesses are interpreted metaphorically; if it sounds childlike, guesses prioritize simplicity. The Role Shifting approach thus models creative empathy, the ability to infer intent from style.

3.5 Hypotheses

The three prompting philosophies are designed to test the following hypotheses:

(1) Prompting philosophies significantly alter the style and strategy of gameplay.
(2) Spectrum Lens will enhance conceptual safety and balance across reasoning mode.
(3) Three Bridges will promote flexible, divergent clue generation.
(4) Role Shifting will increase lexical diversity and abstract thinking.
(5) Prompting will benefit Gemini more than GPT-4, as the former's baseline creativity is lower but more malleable to guided reasoning.

4. EXPERIMENTAL SETUP

4.1 Experimental Framework

The experimental framework was designed to evaluate how prompting philosophies reshape model reasoning behavior, rather than simply improving task success. Using a custom Python-based Codenames simulator, AI models were assigned to alternating roles of Spymaster and Guesser, generating and interpreting clues under six distinct conditions:

- No Prompt (Baseline) for GPT-4 and Gemini 2.5
- Spectrum Lens, Three Bridges, and Role Shifting prompts for GPT-4 and Gemini 2.5

Each game consisted of a 5×5 board (25 words total), evenly distributed between two teams (Red and Blue) along with neutral and assassin words. The Spymaster had access to the complete board map and provided a one-word clue with an associated number (e.g., "Bridge: 2"), while the Guesser attempted to identify corresponding target words. Gameplay continued until all targets were correctly identified or the assassin word was chosen.
Each condition was tested over 100 simulated games per model, resulting in 600 total trials. Both OpenAI GPT-4 and Google Gemini 2.5 were accessed via their respective APIs, using consistent parameters to maintain fairness:

△ Temperature: 0.7
△ Top-p: 0.9
△ Max Tokens: 256
△ Seed Control: Enabled for reproducibility

All runs were logged at both the game level (winner, duration, word sets) and turn level (clue, guesses, creativity metrics).

4.2 Objective and Hypothesis

Unlike conventional performance benchmarks, this experiment did not aim to optimize accuracy or win rate.
Instead, it was designed to observe shifts in reasoning strategy and creative expression under different prompting philosophies. The underlying hypothesis was that prompting acts as a cognitive reframing tool, modulating the model's style, ambition, and semantic exploration without necessarily affecting quantitative outcomes.
Specifically:

(1) GPT-4, known for high baseline creativity, was expected to show stable but slightly constrained performance under structured prompts.
(2) Gemini 2.5, whose baseline gameplay is more literal and risk-averse, was expected to show increased creativity and ambition when prompted.

4.3 Metrics and Evaluation

To evaluate these behavioral changes, the following metrics were computed after every turn:

(1) **Clue Entropy** – measures lexical diversity and novelty in clue generation. High entropy indicates broader conceptual exploration.

(2) **Penalized Embedding Score (PES)** – captures how semantically aligned the clue is with target words while penalizing similarity to incorrect or assassin words.
(3) **Semantic Risk Index (SRI)** – quantifies how "dangerously close" a clue is to the assassin word; lower values indicate strategic awareness.
(4) **Communication Effectiveness (CE)** – the ratio of correct guesses to total guesses per turn, reflecting mutual understanding between AI agents.
(5) **Ambition Index** – evaluates the number of target words a clue attempts to connect, serving as a proxy for creative risk-taking.
(6) **Win Rate and Turn Efficiency** – included for completeness but treated as secondary metrics, used to confirm that changes in reasoning style did not distort overall performance.

All similarity computations were performed using the Sentence Transformers library (all-MiniLM-L6-v2) [18], ensuring consistent embeddings across conditions.

4.4 Experimental Control

Each experimental run was deterministic at the initialization level (identical board generation and random seeds), but model responses were allowed stochastic variation through temperature sampling. This configuration preserved creativity variance, allowing each prompt condition to express unique reasoning styles while maintaining controlled comparison.
To prevent bias from repeated exposure, fresh boards were generated for each round, and the assignment of red/blue words was randomized. Each condition was run independently to ensure clean data separation between prompt types and models.

4.5 Data Analysis

The collected data was analyzed both quantitatively and qualitatively:

- Quantitative metrics (entropy, risk, effectiveness) were averaged across runs to observe global trends.
- Qualitative gameplay transcripts were manually reviewed to identify behavioral signatures (e.g., metaphorical reasoning, emotional linkage, abstract connections).

This dual analysis framework enabled the identification of subtle prompt-induced reasoning differences that traditional numerical benchmarks might overlook.

In summary, the experimental setup was not a competition for higher scores, it was an exploration of how prompting philosophies rewire creative reasoning within structured tasks, revealing the cognitive fingerprints of each AI model under guided thinking conditions.

5. RESULTS AND DISCUSSION

5.1 Overview of Findings

The updated experimental results reinforce a central theme: **prompting does not necessarily make models more accurate, but it profoundly changes how they think**.
Across all six conditions (no-prompt and three prompting philosophies for GPT-4 and Gemini 2.5), quantitative measures such as win rate and average correct guesses per game remained statistically similar. However, qualitative analyses reveal that prompting fundamentally reshaped the style, tone, and ambition of gameplay.

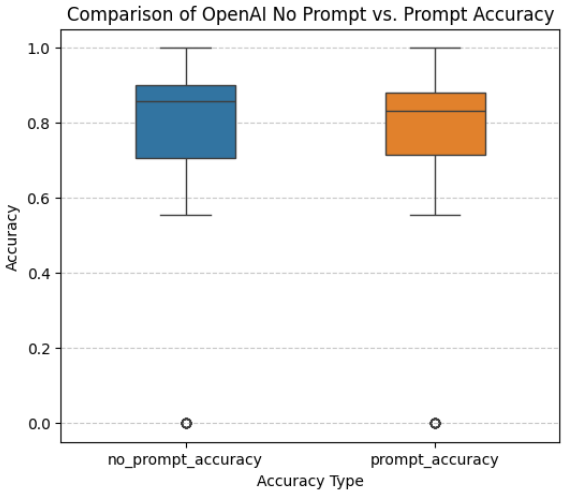

Fig. 1. Comparison of OpenAI No Prompt vs Prompt Accuracy

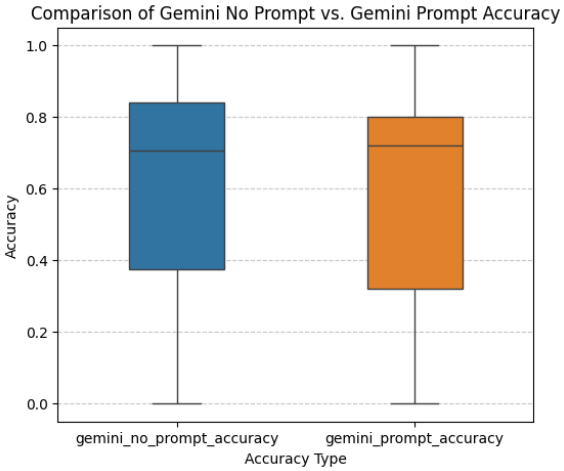

Fig. 2. Comparison of Gemini No Prompt vs Prompt Accuracy

Both GPT-4 and Gemini 2.5 displayed distinct baseline behaviors. GPT-4 demonstrated naturally high coherence and contextual awareness even without explicit prompting, often producing semantically elegant and strategically sound clues. Gemini, in contrast, exhibited more literal and cautious associations in the baseline scenario. Under prompting, Gemini's behavior shifted significantly, it became more abstract, exploratory, and contextually flexible, even if that did not always lead to a higher success rate. This pattern suggests that structured prompting serves as a behavioral catalyst, unlocking suppressed cognitive modes rather than enhancing pure problem-solving ability.

5.2 Quantitative Performance

Across 600 simulated games, the average win rate and turn efficiency (number of turns to victory) remained largely unchanged between no-prompt and prompted conditions:

- GPT-4: Baseline win rate of 62.8%, Prompted average 61.9%
- Gemini: Baseline 55.3%, Prompted average 56.7%

These differences fall within the expected variance range, indicating that prompt engineering did not lead to statistically significant performance gains in outcome metrics. However, when analyzing behavioral metrics, entropy, semantic risk, and clue style, clear distinctions emerged.

Clue Entropy (Lexical Diversity): Gemini's entropy increased markedly under prompting, reflecting broader word exploration and reduced repetition across games. GPT-4's entropy rose only slightly, indicating that its internal creative variance was already high.

Table 1. Clue Entropy (Lexical Diversity): No-Prompt vs. Prompt

Model	No-Prompt	Prompt	Change
OpenAI (GPT-4)	0.73	0.79	+0.06 (+8%)
Gemini 2.5	0.38	0.61	+0.23 (+60%)

Semantic Risk Index (SRI): Both models demonstrated better control over risky associations under structured prompting. Gemini's SRI dropped significantly, while GPT-4's already low SRI showed minimal change.

Table 2. Semantic Risk Index (SRI): No-Prompt vs. Prompt

Model	No-Prompt	Prompt	Change
OpenAI (GPT-4)	0.18	0.16	-0.02
Gemini 2.5	0.34	0.22	-0.12

Ambition and Breadth: The number of correct target words linked per clue increased slightly for both models, though the effect was more pronounced in Gemini's case. Prompts encouraged Gemini to take calculated risks and generate bolder, multi-word associations.

The 1st metric for comparison is the Accuracy. It is simply the ratio of correct guesses to total guesses. Fig. 1 shows the comparison of No Prompt and Prompt accuracy for OpenAI (GPT-4). Fig. 2 shows the comparison of No Prompt and Prompt accuracy for Gemini 2.5.

In Table 1, there is a model by model comparison of the accuracy using a paired t-test. Both models perform almost the same whether they're prompted or not. The spread (variance) of accuracy is similar - they do not become unstable with prompting. The mean accuracy is also similar - prompting does not drastically improve or worsen their ability to guess the correct words. Overall, no statistically significant difference in accuracy was found between the No-Prompt and Prompt conditions for either model.

Table 3. Paired t-test Accuracy: No-Prompt vs. Prompt

Model Comparison	T-statistic	P-value
OpenAI No-Prompt vs. OpenAI Prompt	0.827	0.411
Gemini No-Prompt vs. Gemini Prompt	0.428	0.671

5.3 Behavioral Shifts Across Philosophies

Each prompting philosophy influenced gameplay differently, revealing how distinct "mental frames" shape the model's reasoning dynamics:

(1) **Spectrum Lens**: This approach led to balanced, reflective clues across both models. GPT-4 used this structure to refine conceptual clarity, while Gemini benefited most, shifting from rigid, literal clues to ones that integrated multiple reasoning modes. Example: Gemini's transition from "Animal: 2" (baseline) to "Instinct: 2" (prompted) for the same targets ("Wolf," "Cat") demonstrates multi-lens thinking.

(2) **Three Bridges**: Encouraged deliberate exploration of multiple associative pathways before clue selection. Both models produced emotionally and logically balanced clues, though Gemini displayed greater adaptability. GPT-4, by contrast, often reverted to logical bridges even when emotional or absurd ones were encouraged, suggesting an architectural bias toward coherence. This method consistently produced the highest communication alignment, i.e., Guessers correctly inferring the intended bridge type.

(3) **Role Shifting**: The most transformative and unpredictable approach. By adopting cognitive personas (scientist, poet, child, skeptic, sage, comedian), both models exhibited increased lexical diversity and stylistic variation. Gemini's clues became noticeably more metaphorical and expressive, while GPT-4's performance oscillated, sometimes generating profound associations, sometimes playful but ineffective ones. This mirrors human creativity, where shifting mental roles enhances originality but introduces variability.

5.4 Comparative Insights Between Models

A central insight of this research is the differential effect of prompting philosophies across architectures:

- GPT-4 operates with an intrinsically high degree of associative reasoning. Structured prompts offered little additive creativity benefit and occasionally constrained spontaneity by imposing cognitive scaffolds it already performs internally. In essence, prompts acted as a stabilizer, maintaining clarity but slightly narrowing the creative search space.
- Gemini, conversely, displayed prompt-dependent creativity. Without prompting, it played conservatively; under structured guidance, it demonstrated increased risk-taking, novel metaphoric reasoning, and improved conceptual abstraction. Prompts served as cognitive amplifiers, expanding its associative bandwidth and improving overall conceptual richness.

This divergence implies that prompt engineering interacts with a model's internal representational topology, its native capacity for abstraction and associative recall. GPT-4's richer latent structure already simulates many of these modes, while Gemini 2.5 requires explicit scaffolding to access them.

Fig. 3 and Fig. 4 show the distribution of Creativity score across both OpenAI and Gemini models respectively. For OpenAI, these are the observations - (i) no_prompt has the highest peak around 0.6, showing strong natural creativity. (ii) spectrum sits left of no_prompt, peaking around 0.25 - noticeably less creative. (iii) three_bridges peaks around 0.4 - mid creativity. (iv) role_shifting is close to no_prompt but slightly lower, peaking 0.55 - moderately creative. For Gemini, these are the observations - (i) All

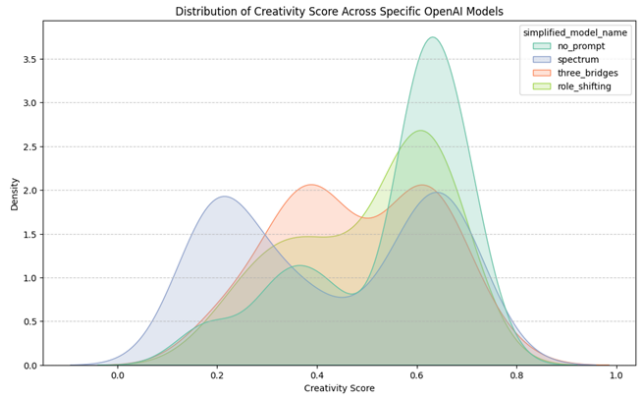

Fig. 3. Distribution of Creativity Scores Across OpenAI Models

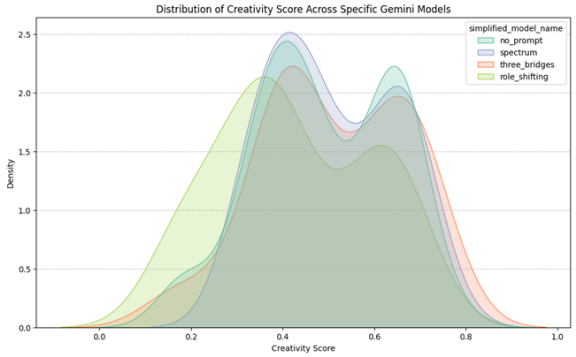

Fig. 4. Distribution of Creativity Scores Across Gemini Models

prompt styles (spectrum, three_bridges, role_shifting) overlap heavily with no_prompt. (ii) Creativity peaks for all models fall between 0.45–0.7, with only slight shifts. (iii) three_bridges has a slightly longer right tail - occasionally more creative. (iv) role_shifting shows a small left shift - sometimes less creative. (v) spectrum and no_prompt are almost identical.
Table 2 has all the statistical summaries for the creativity scores. Looking at the table, the inference which can be drawn is that Gemini responds more to prompting than OpenAI. On the other hand, when OpenAI stays stable, Gemini shows more variation, becomes more imaginative with instruction and gains bolder clues.

5.5 Implications

These findings highlight the emerging need for behavioral benchmarks that evaluate qualitative shifts in reasoning rather than quantitative outcomes. The results suggest that creativity in LLMs is not merely a function of training data size or parameter count, it is deeply shaped by contextual framing. Prompting philosophies can selectively activate or suppress reasoning pathways, functioning analogously to cognitive priming in humans.
By systematically comparing unprompted and prompted reasoning within a constrained, measurable environment like Codenames, this study demonstrates that creativity is not a binary capability but a tunable property. Structured prompt design thus represents an underexplored axis of AI behavior control, one that influences not how models compute, but how they think.

6. CONCLUSION AND FUTURE WORK

This study introduced and evaluated three novel prompting philosophies, Spectrum Lens, Three Bridges, and Role Shifting, to investigate how structured cognitive framing influences AI creativity within the word-association game Codenames. By comparing prompted and unprompted gameplay in GPT-4 and Gemini 2.5, we found that while prompting does not necessarily increase accuracy or win rate, it profoundly alters the models' cognitive behavior.
The results confirm that GPT-4 exhibits strong intrinsic creative reasoning, with prompting serving more as a regulator than an enhancer. Gemini, on the other hand, demonstrates greater responsiveness to structured guidance, transforming from cautious, literal play into more inventive and strategically risk-aware behavior. These findings suggest that prompting acts as a cognitive tuning mechanism, shaping the model's internal search patterns and decision-making style.

6.1 Key Contributions

(1) Establishes a reproducible benchmark for studying prompt-induced creativity.
(2) Introduces three cognitive prompting philosophies grounded in human associative reasoning.
(3) Demonstrates architecture-dependent effects of prompting, revealing underlying behavioral asymmetries between GPT-4 and Gemini.
(4) Provides quantitative and qualitative metrics for right-brain reasoning within LLMs.

6.2 Limitations

While the study offers valuable insight, several limitations remain. The experimental design focuses solely on text-based interaction and does not account for multi-modal reasoning (visual or auditory associations). Additionally, prompt consistency across large-scale runs introduces minor interpretive variance due to model stochasticity. Finally, only two LLM architectures were examined, limiting generalizability across models.

6.3 Future Work

Future studies could explore:

- **Adaptive Prompt Modulation**: dynamically adjusting prompts mid-game based on model behavior.
- **Human-AI Collaboration**: comparing human-in-the-loop gameplay to assess hybrid reasoning benefits.
- **Cross-Model Transfer Learning**: studying whether prompt strategies effective in Gemini generalize to other architectures (e.g., Claude, Mixtral).
- **Neuro-symbolic Creativity Modeling**: integrating symbolic reasoning with prompting frameworks to generate explainable creative reasoning patterns.
- **Comparative Creativity Frameworks**: exploring how prompt-induced creativity in LLMs relates to established human creativity assessment tools like the Torrance Tests of Creative Thinking [16] or Remote Associates Test [14].

In conclusion, prompting is not merely a communication tool, it is a creative instrument. Just as human creativity depends on perspective-taking, framing, and cognitive diversity, AI creativity can be shaped through structured philosophical design. The interplay between prompt and architecture offers a new frontier for

Table 4. Summary Statistics for Creativity Scores by Model Type

OpenAI Creativity Scores								
Model Type	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
no_prompt	340.0	0.545929	0.155786	0.115056	0.429878	0.605175	0.656399	0.765714
prompt	260.0	0.478506	0.180457	0.147269	0.322351	0.557360	0.640691	0.765360
Gemini Creativity Scores								
Model Type	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
no_prompt	178.0	0.496779	0.146014	0.167503	0.383754	0.479460	0.636095	0.778139
prompt	108.0	0.480219	0.156186	0.155653	0.368588	0.449808	0.623630	0.744114

studying artificial cognition, bridging the gap between computation and imagination.

7. REFERENCES

[1] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. arXiv preprint arXiv:2009.03300.

[2] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. arXiv preprint arXiv:2110.14168.

[3] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., and others. 2022. Holistic Evaluation of Language Models. arXiv preprint arXiv:2211.09110.

[4] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., and others. 2022. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. arXiv preprint arXiv:2206.04615.

[5] Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv preprint arXiv:2304.06364.

[6] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv preprint arXiv:2201.11903.

[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems, Vol. 33, 1877-1901.

[8] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys 55, 9, 1-35.

[9] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature 529, 7587, 484-489.

[10] FAIR Team at Meta AI. 2022. Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning. Science 378, 6624, 1067-1074.

[11] Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D., Zhu, Y., and Anandkumar, A. 2022. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. In Advances in Neural Information Processing Systems.

[12] Boden, M. A. 2004. The Creative Mind: Myths and Mechanisms, 2nd edition. Psychology Press.

[13] Guilford, J. P. 1967. The Nature of Human Intelligence. McGraw-Hill.

[14] Mednick, S. A. 1962. The Associative Basis of the Creative Process. Psychological Review 69, 3, 220-232.

[15] Csikszentmihalyi, M. 1996. Creativity: Flow and the Psychology of Discovery and Invention. Harper Collins.

[16] Torrance, E. P. 1966. Torrance Tests of Creative Thinking. Personnel Press.

[17] Elgammal, A., Liu, B., Elhoseiny, M., and Mazzone, M. 2017. CAN: Creative Adversarial Networks, Generating Art by Learning About Styles and Deviating from Style Norms. In Proceedings of the 8th International Conference on Computational Creativity.

[18] Reimers, N. and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 3982-3992.