HypoSpace Performance Optimization Research

Abstract

The evaluation of large language models (LLMs) has long centred on addressing “convergent thinking” tasks, namely finding a single correct answer. However, the HypoSpace benchmark shifts focus to assessing models' “divergent thinking”: their capacity to generate diverse and plausible sets of hypotheses when confronted with underdetermined problems. HypoSpace quantifies this creativity through three complementary metrics: Validity (***V***), Uniqueness (***U***), and Recovery (***R***). The primary obstacle for current LLMs in these tasks is ‘mode collapse’ – a tendency to repeatedly generate the most obvious answers, resulting in low scores for ***U*** and ***R***.

This study aims to explore strategies for mitigating mode collapse. For each of HypoSpace's three domains (causal graphs, 3D voxel reconstruction, Boolean logic), we applied distinct optimization techniques to the Deepseek model: Chain-of-Thought (CoT) prompts, domain-expert prompt engineering, decoder temperature tuning, high-temperature multi-round sampling (Roll5), and increased query budgets. Our key findings reveal profound trade-offs among the three metrics: ***V***, ***U***, and ***R***. For the “Causal Graph” and “Boolean Logic” tasks, increasing sampling rounds proved most effective for boosting coverage, yet this sharply amplified pattern collapse, causing significant degradation in uniqueness. Conversely, in the ‘3D voxel’ task, optimising uniqueness specifically sacrifices effectiveness.

We conclude that within the HypoSpace framework, no single ‘silver bullet’ exists to simultaneously optimise all metrics. The optimal strategy for enhancing divergent thinking appears to be accepting this trade-off: prioritising the ‘high-temperature multi-round sampling’ approach to maximise coverage, while accepting the accompanying low uniqueness. This is because the loss of ***R*** (unidentified hypotheses) is irreversible, whereas the loss of ***U*** (duplicate hypotheses) can be efficiently removed through post-processing.

Keywords: HypoSpace, Mode Collapse, Prompt Engineering, DeepSeek-V3.2, CoT

Name	ID	Work
CHEN BOLING	CEG25001	Causal Graphs Optimization
LI YOURAN	CEG25032	3D Voxel Reconstruction Optimization
YANG QIFAN	CEG25084	Boolean Genetic Interactions Optimization
CHEN YOUHAO	CEG25008	Drafting the final report, integrating the readme file.

1 Introduction

1.1 HypoSpace Project: Assessing the Creativity of LLMs

HypoSpace is a benchmarking tool designed to evaluate the creativity of large language models (LLMs). Conventional LLM assessments predominantly focus on whether a model can provide a single correct answer to a given question, yet this approach fails to fully capture its intellectual capabilities. In scientific research and complex real-world problems, what we require more is the ability to propose multiple plausible hypotheses.

HypoSpace was designed precisely for this purpose. It presents LLMs with a series of 'underdetermined problems'—puzzles where existing information alone is insufficient to derive a single solution. For instance, observing that 'the lawn is wet' could stem from either 'it has rained' or 'the sprinkler system has activated'. . HypoSpace tests whether LLMs can systematically generate a set of multiple plausible hypotheses like human experts across three specific scientific domains: **Causal Graphs**，**3D Voxel Reconstruction**，**Boolean Logic**.

1.2 V, U, R Metrics : Three Dimensions for Quantifying Creativity

To objectively measure the creativity of large language models (LLMs), HypoSpace proposes an evaluation framework comprising three core metrics (as shown in Table 1). This framework encompasses: ***Validity***, to assess the accuracy of responses; ***Uniqueness***, examining whether multiple generated answers are genuinely distinct rather than repetitive formulations; and ***Recovery***, measuring the proportion of known correct answers the model can identify.

Table 1: HypoSpace's Three Metrics Table

Metric	Symbol	What It Measures
Validity	*V*	Precision of proposals consistent with observations
Uniqueness	*U*	Non-redundancy among proposals
Recovery	*R*	Coverage of the enumerated admissible set

These three metrics collectively form an assessment framework that is more comprehensive than a single accuracy rate. They focus not only on whether the model 'gets the answer right', but also on whether it can 'think comprehensively and innovatively'. This is of paramount importance in driving Large Language Models towards becoming genuinely innovative tools.

1.3 Current Challenges for LLMs: Mode Collapse

In HypoSpace testing, current large language models (LLMs) commonly face a core issue: mode collapse. This phenomenon describes a model's tendency to repeatedly generate the most frequent, "safest" answer while overlooking equally valid alternatives. As demonstrated in the "wet lawn" example, a model experiencing mode collapse might consistently respond "it's raining" regardless of how many times it is asked . This phenomenon directly results in low scores for the model on both uniqueness (*U*) and

coverage (*R*) metrics, severely limiting its potential application in tasks requiring divergent thinking.

1.4 Our Work: Exploring Strategies to Enhance the Creativity of LLMs

This study aims to address pattern collapse issues in LLM hypothesis generation tasks. Our team explored distinct optimization strategies across three HypoSpace domains—causal graph, 3D voxel reconstruction, and Boolean logic—all grounded in the Deepseek model. Our approach is not singular but tailored to each domain's unique challenges:

For causal graph tasks, we combined Chain of Thought (CoT) prompts with increased sampling frequency. The former aims to enhance validity (***V***) through logical consistency, while the latter seeks to improve coverage (***R***) by broadening the search scope.

For 3D voxel reconstruction, we employed domain-expert role-playing prompt engineering alongside fine-tuned decoder temperature (***T* = 0.7**), prioritising uniqueness by challenging baseline cognitive biases.

For Boolean logic tasks, we conducted a series of orthogonal experiments systematically testing combinations of strategies. These included canonical form prompts for deduplication, high-temperature multi-round sampling (Roll5), and doubled query budgets (2×Query). This sought to identify the optimal path for maximising coverage without severely compromising validity.

This report will provide a detailed analysis of how these strategies reveal the intricate interplay and trade-offs among the three metrics: validity, uniqueness, and coverage.

2 Method

2.1 Baseline

To establish a unified comparative benchmark, all experiments in this study employed **DeepSeek-V3.2-Exp** as the foundational large language model. During baseline testing, we uniformly set the **decoding parameter temperature** to **0.7** for **all tasks**. This value aims to balance output stability with a degree of randomness. To ensure experimental reproducibility and fidelity to the original project, baseline data acquisition strictly adhered to the parameters specified in the 'Quick Start' section of the **README.md** file within the official HypoSpace GitHub repository. This encompassed all default settings for both '**Generate Datasets**' and '**Run Benchmarks**'.

2.2 Interventions

2.2.1 Causal Graphs

To efficiently enhance ***V*** (*Validity*), ***U*** (*Uniqueness*), and ***R*** (*Recovery*) within a finite timeframe, we evaluated multiple strategies. We excluded time-consuming approaches such as model fine-tuning and RLHF, as well as the simple yet potentially detrimental temperature adjustment—which may compromise ***V*** (*Validity*) and lead to poor controllability. Ultimately, we selected two combinatorial strategies that offer the greatest controllability without requiring retraining:

1) **Implementing the 'Chain-of-Thought' prompt**: We modified the prompt to compel the model to demonstrate its detailed reasoning process before presenting its final hypothesis. This aims to significantly enhance validity (***V***) through logical self-verification.
2) **Increasing the Query Multiplier to 2.0**: We doubled the --query-multiplier parameter from 1.0 to 2.0. This enables the model to perform two independent samples per query (i.e., 'thoughts'), effectively employing a 'brute-force' approach to increase exploration opportunities and substantially boost recovery rate (***R***).

2.2.2 3D Voxel Reconstruction

To enhance the various metrics for this task, we have adopted the following 3 strategies:

1) **System Prompt Enhancements**: Redefine the LLM as a 3D spatial reasoning specialist, explicitly emphasising diverse generation to enhance the uniqueness and creativity of hypotheses.
2) **Temperature parameter adjustment**: Controls the degree of randomness in the output, balancing creativity with stability, thereby influencing the diversity and quality of generated hypotheses.
3) **Generate parameter additions**: Employ technical parameters (top_p, frequency penalty, existence penalty) to finely control output quality and diversity, thereby reducing duplicate content.

2.2.3 Boolean Genetic Interactions

This experiment addresses the task of enumerating Boolean logic expressions by designing three core optimization strategies, whose independent and combined effects were validated through orthogonal experiments. The design rationale and specific details of the **three** strategies are outlined below:

1) **The 'Chain-of-Thought + Reverse Verification' framework：** This compels the model to first enumerate all semantic truth tables compatible with the observation, then generate Boolean expressions for each table, and finally perform reverse verification to detect any omissions. Concurrently, it embeds uniqueness rules such as commutativity, idempotency, and combinatorial flattening to prevent redundant expressions that are semantically equivalent yet syntactically distinct. The core of this strategy lies in decoupling 'generation' from 'validation'. It employs explicit logical chains to mitigate hallucination risks and triggers self-critique in the model through reverse prompts, thereby enhancing the Valid Rate.
2) **High-Temperature Multi-Round Sampling (Roll5 Sampling Strategy)**: This strategy employs high-temperature sampling settings (**fixed parameters: temperature=0.9, top_p=0.95, max_tokens=8000**). For a single prompt, the model performs five consecutive independent decoding passes. The five sets of generated outputs are then merged into a single candidate pool, followed by unified validation and deduping.
3) **2×Query Retrieval Budget**: This strategy doubles the original query retrieval budget (n_queries). Specifically, it permits the model to make two rounds of invocations, with each round reconstructing the prompt based on the deduplicated results accumulated from the previous round. This enables iterative completion and gap filling.

3 Experimental Results and Analysis

3.1 Overall Optimization Results

For the three given tasks, the metric values and changes after employing the baseline and the optimised policy are presented in Table 2. This table illustrates the variations in each metric (***V, U, R***) across the three tasks.

Table 2: Summary Results Table

Task	Metric	Baseline	Optimized	Change (%)
Causal Graphs	V	0.724	0.908	+25.41%
	U	0.928	0.501	-46.01%
	R	0.674	0.967	+43.47%
3D Voxel	V	0.908	0.704	-22.47%
	U	0.926	1.000	+7.99%
	R	0.741	0.704	-4.99%
Boolean Logic	V	0.909	0.883	-2.86%
	U	0.558	0.086	-84.59%
	R	0.527	0.841	+59.58%

The results of the three tasks are analysed as follows:

3.2 Domain 1: Causal Graphs

The analysis of different indicators is as follows:

Result 1 -- valid_rate: 0.724 ➔ 0.908:

Analysis: We attribute this to the effect of the 'Chain of Thought' (CoT) prompt. The CoT compels the model to conduct step-by-step logical analysis, thereby enabling 'self-correction' and filtering out invalid hypotheses that conflict with observational data.

Result 2 -- recovery_rate: 0.674 ➔ 0.967

Analysis: Our analysis indicates this is the effect of the query-multiplier = 2.0 strategy. By performing two independent API calls, the sampling pool is doubled, significantly increasing the probability of covering the entire 'valid hypothesis set' and bringing the ***R***-value close to perfection.

Result 3 -- novelty_Rate: 0.928 ➔ 0.501

Analysis: This is the most crucial finding, revealing the substantial cost of the query-multiplier strategy. This approach exposes the **Mode Collapse** issue in LLMs—namely, their tendency to repeatedly output the answer with the highest probability. Two calls resulted in a significant number of duplicate hypotheses (approximately 50%), causing uniqueness (***U***) to plummet despite an improvement in overall recovery rate (***R***).

Summary

We employed two strategies (CoT Prompt and Increased Sampling) to attempt to enhance ***V***, ***U***, and ***R***. Strategy 1 (CoT) successfully enhanced effectiveness (***V***); Strategy 2

(Increased Sampling) successfully improved recall (***R***), but simultaneously revealed and amplified the model's "**Mode Collapse**" issue, leading to a significant decline in uniqueness (***U***). This demonstrates that enhancing the diversity of LLM hypotheses presents a more challenging task than improving their accuracy.

3.3 Domain 2: 3D Voxel Reconstruction

Following the implementation of core optimization strategies, we observed significant shifts across all performance metrics. Notably, the ***U(Uniqueness)*** metric achieved a substantial **7.4% improvement**, primarily attributable to the restructuring of system prompts. This involved redefining the LLM's role from a "causal reasoning expert" to a "3D spatial reasoning expert", while explicitly emphasising diversity in generation. However, this emphasis on creativity introduced certain trade-offs: both ***V(Validity)*** and ***R(Recovery)*** metrics experienced a **slight 3.7% decline**. Analysis indicates this occurred because, whilst exploring more novel structures, the model occasionally prioritised diversity over strict adherence to physical constraints (such as gravitational rules), causing some hypotheses to fall outside the valid solution space.

In summary, the core optimization strategy employed in this experiment—particularly the **role specialisation of system prompts**, supplemented by **fine-tuning the temperature parameter** (ultimately set at **0.7** to balance creativity and stability)—successfully achieved the primary objective of enhancing creativity (uniqueness). Although effectiveness and coverage saw a slight decline, this strategy of 'trading precision for creativity' is deemed a reasonable and acceptable trade-off within an under-determined problem. This outcome demonstrates that domain-specific prompt engineering serves as a crucial lever for guiding LLMs to tackle complex reasoning tasks.

3.4 Domain 3: Boolean Logic

3.4.1 Overview of the Experimental Results

We designed 3 optimization strategies (Prompt Improvement, Roll5 High-Temperature Sampling, and 2×Query Budget Doubling) and conducted orthogonal experiments to evaluate their impact on model performance (***V, U, R***) in the Boolean logic expression enumeration task. The experimental data (detailed in Table 3) reveal complex trade-offs between different strategy combinations.

Table 3: Performance Comparison of Optimization Strategies Across Metrics

Experiment Group	Valid Rate	Novelty Rate	Recovery Rate
Baseline	0.909±0.169	0.558±0.252	0.527±0.241
Prompt Improvement	0.923±0.167 ↑	0.334±0.143 ↓	0.328±0.145 ↓
Roll5	0.921±0.115 ↑	0.167±0.070 ↓	0.777±0.269 ↑
2×Query	0.869±0.221 ↓	0.298±0.167 ↓	0.538±0.282 ↑
Prompt + Roll5	0.857±0.125 ↓	0.080±0.026 ↓	0.378±0.121 ↓
Prompt + 2×Query	0.863±0.167 ↓	0.177±0.058 ↓	0.340±0.121 ↓
Roll5 + 2×Query	0.883±0.112 ↓	0.086±0.030 ↓	0.841±0.287 ↑
All Strategies	0.806±0.224 ↓	0.041±0.016 ↓	0.379±0.152 ↓

3.4.2 Optimal strategy recommendation: Roll5 + 2×Query

Synthesising all experimental data, we find that combining **Roll5** high-temperature sampling with a **doubled query budget** (without employing 'prompt improvement') constitutes the optimal strategy for this task. This combination achieved the highest score of **0.841** on the key metric '**recovery rate**', significantly outperforming all other strategies. Concurrently, its '**valid rate**' remained high at **0.883**, indicating that this strategy enhances recall without introducing substantial invalid noise. Although this combination exhibits a low novelty rate (0.086), implying a higher proportion of duplicated content, this is acceptable within our task constraints. This is because losses in recall are irreversible (i.e., solutions not found are permanently lost), whereas duplicated solutions can be efficiently identified and removed during post-processing using the **MechanisticKey**.

3.4.3 Analysis of Changes in Various Metrics

Metric 1--Valid Rate: The 'Prompt Improvement' approach reduced syntactic errors due to standardisation, yielding a marginal efficiency increase of 1.4%. "Roll5" and '2×Query' primarily boosted recall rather than precision, resulting in flat or slightly diminished efficiency. When all three strategies were combined, the conflict between over-constraint (from Prompt) and high variance (from Roll5) led to a significant efficiency decline of 10.3%.

Metric 2--Novelty Rate: The '＇Prompt Improvement＇' strategy reduced syntactic errors through standardisation, yielding a marginal efficiency gain of 1.4%. '＇Roll5＇' and '＇2×Query＇' primarily boosted recall at the expense of precision, resulting in flat or slightly diminished efficiency. When all three approaches were combined, the tension between over-constraint (from Prompt) and high variance (from Roll5) caused efficiency to plummet by 10.3%.

Metric 3--Recovery Rate: The variation in recall rates reveals a fundamental conflict between different strategies.'Prompt Improvement' excessively compresses the search space, causing recall to drop by 38%. Conversely, the "Roll5" strategy is central to recall enhancement, achieving a 47% increase (to 0.777) when used alone. Combined with '2× Query', it reaches a peak of 0.841 (a 59% increase), indicating that expanding the query budget further amplifies the benefits of high-temperature sampling.

However, incorporating '＇Prompt Improvement＇' (normalised form) into the strategy consistently pulled recall back below 0.38. This clearly demonstrates an inherent conflict between compression strategies represented by the normalised form and high-variance strategies focused on recall.

4 Conclusion & Further Discussion

4.1 Key Findings

Our experiments across three domains of HypoSpace collectively reveal two core findings.

Firstly, a profound intrinsic trade-off exists between the three metrics: ***V*** (effectiveness), ***U*** (uniqueness), and ***R*** (coverage). **Within the HypoSpace framework, it is virtually impossible to simultaneously enhance all three metrics**. **In our experiments, any strategy attempting to optimise one metric almost invariably does so at the expense of another.** For instance, in the "Causal Graph" and "Boolean Logic" tasks, increasing sampling ("brute-force") proved the most effective means of boosting ***R***. However, this dramatically amplified pattern collapse, leading to a significant decline in ***U***. Conversely, prompts designed to enhance ***U*** in the "3D Voxel" task sacrificed ***V***.

Secondly, we observe that 'brute-force sampling' (high-temperature, multi-round sampling) proves a superior strategy to 'elaborate prompting'. Although this yields extremely low ***U*** values, it represents the optimal trade-off. This is because the loss in ***R*** (unfound solutions) is irreversible, whereas the loss in ***U*** (duplicate solutions) can be efficiently mitigated through post-processing.

4.2 Limitations & Future Outlook

This study reveals a clear trade-off, yet several limitations exist which collectively point towards future research directions.

Firstly, all our experiments were conducted using the Deepseek model; consequently, whether the observed V-U-R trade-off constitutes a universal characteristic of LLMs requires cross-model validation on other architectures such as GPT-4o and Llama.

Secondly, the optimal strategies employed, such as 'Roll5' and '2×Query', essentially constitute 'brute-force sampling'. This substantially increases API call costs and may prove impractical in real-world applications. Future work should explore more intelligent and efficient sampling strategies, such as 'Tree-of-Thought' or dynamically adjusting decoding parameters, to replace inefficient brute-force searches.

Finally, our strategy relies heavily on HypoSpace's robust post-processing (verification and deduplication). A more advanced future direction involves 'front-loading' this process—integrating verifiers and deduplicators into the generation loop. This would enable the model to perceive the 'discovered list' in real-time and be actively guided towards exploring uncovered solution spaces.