Optimization of Qwen-2.5-72b-instruction Model for 3D Reconstruction Task in HypoSpace

Tianshi He

Supervisor(s): Dianbo Liu

2025-11-01

Abstract

HypoSpace has proposed a scheme for evaluating the creativity of LLMS, which utilizes three interesting tasks to evaluate their Validity (precision of proposals consistent with observations), Uniqueness (non-redundancy among proposals), and Recovery (coverage of the enumerated admissible set). In this task, we selected Qwen-2.5-72b-instruction as the experimental model and used the following three methods to improve the model's score in completing 3D reconstruction task: Temperature Optimization, exploring the optimal model temperature by trying different temperature parameters and dynamic temperature mechanisms; COT Optimization, improving the performance of large language models by providing a thought chain tailored to specific problems; Feedback Optimization, by pointing out the shortcomings of the previous responses to enhance the quality of their answers in the next response.

In addition, we conducted multiple rounds of tests for different optimization methods and their combinations. Finally, we obtained the following results: The model performed best when the temperature was 0.5 while the dynamic temperature mechanism did not significantly improve the score; COT Optimization significantly improves model performance; The feedback mechanism has little impact on the improvement of model performance; The simultaneous combination of multiple models does not have a cumulative effect on performance, but when COT Optimization is included in the combination, the performance of the model is remarkably enhanced.

1 Temperature Optimization

The temperature parameter is a hyperparameter in the generative model used to adjust the sharpness of the probability distribution when predicting the next word. Low temperature means that the probability distribution is sharper, resulting in more conservative and coherent outputs. High temperature generates smoother probability distribution and amplifies low probability vocabulary, resulting in more random and diverse outputs at the expense of logicality.

Therefore, we infer that the temperature of the LLMs will have a certain impact on the creativity of the output results, and we set this parameter as the optimization object. We conducted Dynamic Optimization and Static Optimization on this parameter separately.

Static Optimization sets the initial temperature to diverse values, and observes the optimization effect. Dynamic Optimization is a more intuitive optimization idea. Considering that the model may not be able to generate a more diverse range of answers in the later stage, we gradually increase the temperature of the model in stages to pursue better results.

In the verification, we have drawn the following conclusions regarding **Temperature Optimization**: Dynamic Optimization cannot significantly improve the creativity; When the fixed temperature is 0.5, the model performs best for 3D reconstruction task. The complete test results of **Temperature Optimization** is shown in Figure 1.

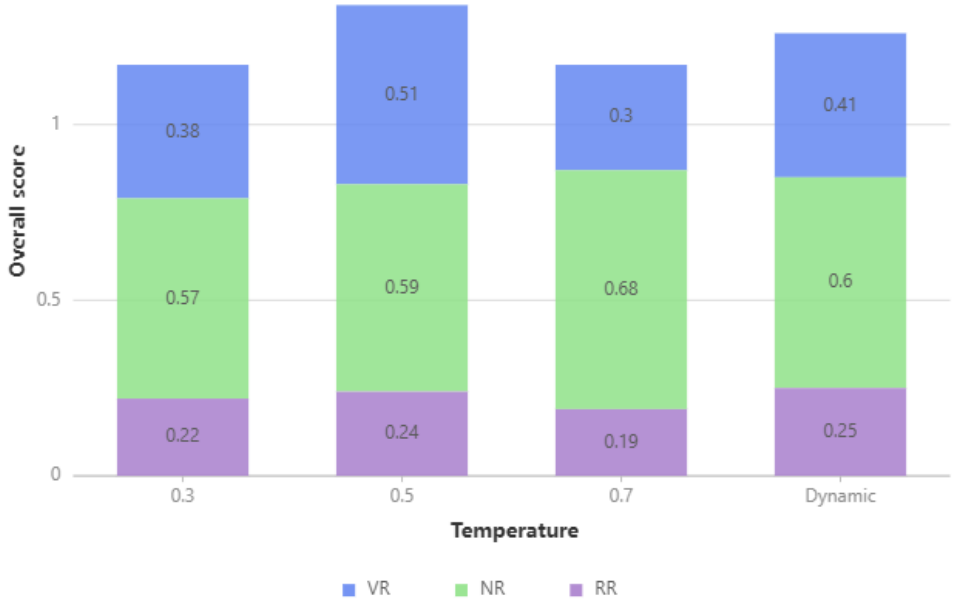

Figure 1: Results for Temperature Optimization

2 COT Optimization

The **Chain-of-Thought** (COT) is a prompt method that enhances the arithmetic, common sense, and reasoning abilities of large models by requiring the model to explicitly output intermediate reasoning steps before outputting the final answer.

In order to enhance the creativity of the model, we have broken down the 3D reconstruction problem into 5 steps and encouraged big language models to analyze step by step.

To evaluate the optimization effect of COT, we used the control variable method. When the default temperature is set as the fixed temperature, we compare the optimized results of COT with the results without optimization. Experimental results have shown that COT has a significant impact on model creativity. Without optimization, the VR, NR, RR index for LLM are **0.3, 0.68, and 0.19**, respectively. After optimization, they are **0.74, 0.68, and 0.48**, respectively. We can observe that under the condition of unchanged novelty, the Validity and Recovery have increased by over 100%, which also indicates that our prompt has played a good guiding role.

3 Feedback Optimization

The **feedback Optimization** strategy informs the LLMs of the evaluation results of previously generated answers, in order to stimulate it to pay attention to avoiding previous problems when generating results, thus improving its creativity evaluation score in HypoSpace.

To implement this feedback mechanism, we recorded the evaluation history of each generated answer (including indicators such as legality and novelty) as well as all previously generated unique structures. We brought the previous historical information in the next question to urge the model to make adjustments.

For the feedback mechanism, we also adopt the strategy of controlling variables. However, it has been proven that the use of feedback mechanisms has little impact on the results, perhaps because the current mechanism is too naive.

4 Testing and Conclusion

In order to facilitate the testing of the optimization effects of various optimization strategies and their combinations, we have added a command-line parameter: optimze, which enables three optimization strategies through a 3 bit binary. Enable temperature optimization at the highest level, COT optimization at the second level, and dynamic temperature optimization at the lowest level. For example, if we want to only activate COT optimization, we should set optimize parameter as 4.

Our testing process is as follows:

1. Generate test data. All top views are filled with two blocks selected from a 3 * 3 matrix, so there are nine different solutions for each test case.

2. Change parameters and optimize strategies to evaluate test results.

3. Record and summarize the result data.

For different parameters and optimization strategies, the complete test results are shown in Table 1.

Optimize Parameter	Init Temperature	VR	NR	RR
0 (0b000)	0.3	0.38	0.57	0.22
0 (0b000)	0.5	0.51	0.59	0.24
0 (0b000)	0.7	0.24	0.68	0.19
1 (0b001)	0.7	0.32	0.64	0.23
2 (0b010)	0.7	0.41	0.6	0.26
3 (0b011)	0.7	0.33	0.59	0.2
4 (0b100)	0.7	0.74	0.68	0.48
5 (0b101)	0.7	0.61	0.69	0.37
6 (0b110)	0.7	0.81	0.66	0.48
7 (0b111)	0.7	0.59	0.68	0.4

Table 1: Results table

By observing the complete result data of the 3D reconstruction task, we can conclude that COT Optimization has the most significant effect. Dynamic temperature changes cannot increase the score while leveraging multiple optimization strategies simultaneously cannot bring palpable creativity improvement.

References

[1] Wei, Jason, et al. "Chain of Thought Prompting Elicits Reasoning in Large Language Models." *ArXiv* abs/2201.11903 (2022): n. pag.

[2] Chen, T., Lin, B., Yuan, Z., Zou, Q., He, H., Ong, Y.-S., Goyal, A., & Liu, D. (2025). *HypoSpace: Evaluating LLM Creativity as Set-Valued Hypothesis Generators under Underdetermination. arXiv preprint* arXiv:2510.15614.

[3] Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). *Is Temperature the Creativity Parameter of Large Language Models? arXiv preprint* arXiv:2405.00492.