Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38166

A Lightweight LLM-Based Method for Generating Game NPC Dialogues, Focusing on Role Consistency and Task Execution

Haocheng Li

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai,China
2235060317@st.usst.edu.cn

Abstract. The big language model can improve the degree of freedom of the game NPC dialogue, but the lightweight model may still have problems such as role tone distortion, knowledge crossover, and task information omission in a specific worldview. This paper is based on Qwen3-1.7 B-Instruct for 6 GB local devices, using QLoRA to fine-tune the domain, and using dynamic constraint prompts to coordinate the six dimensions of role, world state, scene, task, historical dialogue, and player input. The text of 875 empty knights was cleaned and structured, and 4375 samples were obtained after enhancement, including 440 test sets. The evaluation system includes BLEU, ROUGE, Distinct, generative behavior, and LLM-as-a-Judge. The BLEU and ROUGE of the candidate model lr-5e5-epoch2 are higher than the benchmark model current-reference, but the average output is longer, and the EOS termination rate and Distinct are significantly reduced. DeepSeek Judge completed the analysis of 440 pairs of responses and preferred the benchmark model in the non-parlement sample with a winning rate of 55.71 % to 44.29 %. The results show that the improvement of reference similarity does not necessarily bring about the simultaneous improvement of role consistency and generation stability, and multi-dimensional evaluation helps to make a choice between task fit and role robustness.

Keywords: Game NPC Conversation, QLoRA, Dynamic constraint prompt, Role consistency, LLM-as-a-Judge

1. Introduction

Large language models provide the foundation for open text generation and interactive dialogue. Large language models provide the foundation for open text generation and interactive dialogue [1]. LoRA freezes pre-training weights and trains low-rank increments, significantly reducing trainable parameters and memory requirements in GPT-3 175B experiments [2]. QLoRA further combines 4-bit quantization with LoRA, allowing 65B models to be fine-tuned on a single 48 GB GPU [3]. Qwen3 covers scales from 0.6B to 235B and supports 119 languages and dialects [4]. These methods improve the feasibility of adapting to low-resource domains. However, it mainly addresses

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38166

computational costs without directly ensuring NPC role consistency, worldview boundaries, and task execution issues.

In recent years, research on generative NPCs has gradually focused on text controllability. Gao and Emami studied NPC script generation and self-diagnosis, and their model scores showed high correlation with manual scoring [5]. Gao and Emami studied NPC script generation and self-diagnosis, and their model scores showed high correlation with manual scoring [6]. RoleFact alleviates character illusions through fact-checking [7]. CharacterEvalar evaluates character consistency and interaction quality based on 1,785 multi-turn dialogues [8]. These works incorporate roles, world knowledge, and illusions into the generation and evaluation system. However, in lightweight models, identity retention and task information transmission still need to be re-coordinated.

BLEU and ROUGE measured the issue of n-gram overlap and content coverage [9, 10]. The coincidence rate between GPT-4 review and manual preference in MT-Bench is more than 80 % [11]. G-Eval obtained a Spearman correlation coefficient of 0.514 in the summary task as part of the evaluation system, indicating that the LLM review can supplement the open attributes for judgment [12]. Indicates that word indicators struggle to cover role boundaries, naturalness, and task execution. LLM-as-a-Judge is also influenced by the review model and Rubric. In summary, existing studies have promoted low-cost fine-tuning, controllable NPC generation, and multi-dimensional evaluation, respectively. However, in terms of resource feasibility, role constraints and task execution are often handled separately. The lightweight model still needs to coordinate multi-source constraints and use complementary evidence to judge role robustness and task fit.

In view of the above problems, this paper studies the role consistency and task execution coordination ability of a lightweight LLM under the condition of limited memory. The experiment first cleans and structures the corpus of Hollow Knight, and constructs Basic, Character, and Dynamic as samples. Secondly, QLoRA is used to complete domain adaptation. In reasoning, dynamic prompts are used to integrate roles, worlds, scenes, tasks, and historical information and give priority to roles and worldviews. The task is to handle conflicts within the allowed range of principles and, finally combined with automatic indicators. Generative behavior and LLM-as-a-Judge were evaluated to examine the resource feasibility of the scheme and the differences between different training configurations in reference similarity, role consistency, task execution, and generative stability.

2. Data set construction and analysis

2.1. Data sources, cleaning and partitioning

This thesis selects the HollowKnightText Database as the experimental corpus [13]. The data set contains NPC dialogues, roles, and context information, and different NPCs have clear language styles and worldview characteristics. Its scale is suitable for fine-tuning under limited resources. After deduplication, 875 records were obtained from 876 records, which were divided into 699 / 88 / 88 according to 8:1:1. contains fields such as Game, Source, Type, Text, Condition, and Topics, covering 262 entities and 11 categories of scenarios.

2.2. Structured samples and dynamic enhancement

Basic, Character, and Dynamic are added to the basic instructions step by step, including the character and the world, scene, task, history, and other dynamic fields. The Dynamic sample further

DOI: 10.54254/2755-2721/2026.38166

unifies the system, character, personality, world, scene, condition, topic, task, history, player, and output fields. The data is divided before enhancement. Each original record forms five context variants, and the response remains unchanged. The final training, validation, and test set is 3495 / 440 / 440.

2.3. Data risk and control

No cross-set duplication such as prompt, output, character-output, or original-proxy was found upon review. Among these, 80.00% of repeated outputs came from five variants of the same reply, rather than leaking from the training set to the test set. This structure may make reference indicators more sensitive to fixed expressions, so diversity, termination behavior, and Judge preferences must be considered simultaneously.

3. Been designed

3.1. General route of technology

As shown in Figure 1, the text is cleaned, annotated, and layered before entering QLoRA for training. During the inference phase, the context is organized by the dynamic constraint coordinator, and after generating a response, the constraint check is executed. Using automated indicators, decoding ablation, and DeepSeek evaluation to analyze similarity, generating behaviors, and preferences. Finally, the complete process from data to model selection is completed.

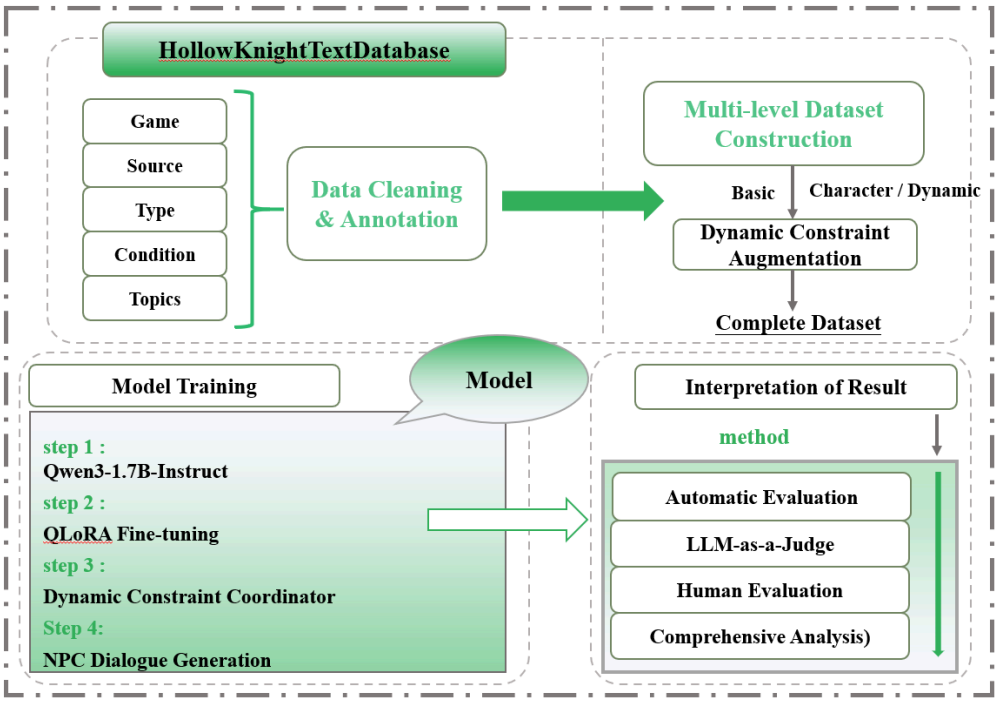


Figure 1. Overall research process (photo/picture credit: original)

3.2. Dynamic constraint prompt and behavior coordination

As shown in Figure 2, the system integrates player input, roles, world states, and scenarios, and then combines task and history for conflict detection and priority judgment. Dynamic Prompt generates

DOI: 10.54254/2755-2721/2026.38166

constraints according to identity, character, fact boundary, and current target organization, and then checks role boundary, task information, repetition, and termination status. When a conflict occurs, priority is given to maintaining the role identity and worldview, and then advancing the task in a way that conforms to the role position to prevent the OOC response from only pursuing the completion of the task.

For example, as a mechanism description rather than an experimental sample. When the player asks the reserved and attentive and concise Hunter to explain the exaggerated humor, the system retains its restrained and observational tone while conveying the necessary facts. That is, the task is only completed within the allowable range of roles and worldviews.

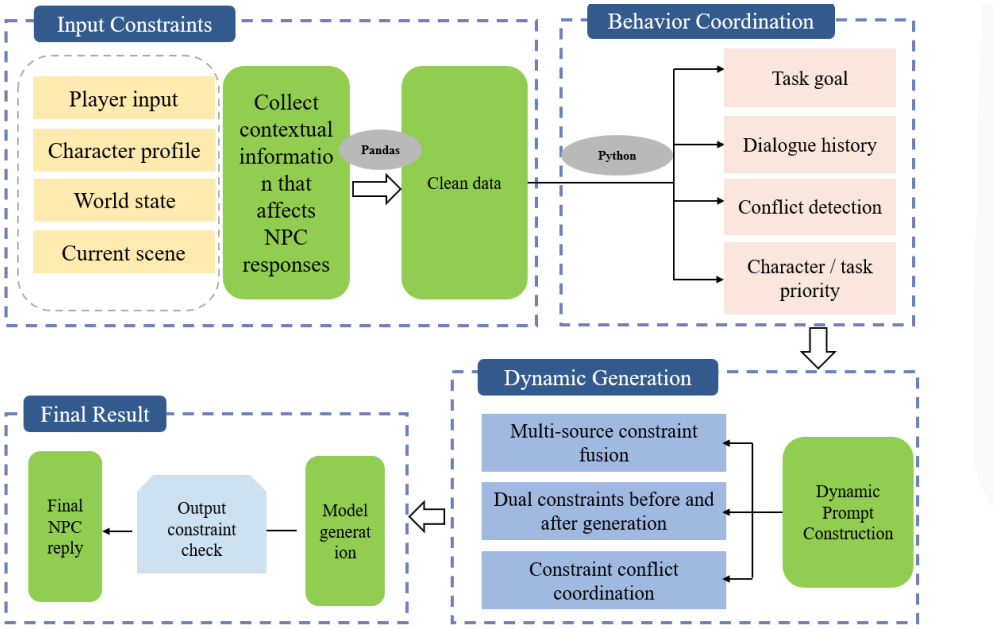


Figure 2. Dynamic constraint behavior coordination process (photo/picture credit: original)

3.3. QLoRA fine-tuning

The base model is Qwen3-1.7B-Instruct. QLoRA is computed using 4-bit NF4, double quantization, and float16LoRA set r = 16, alpha = 32, dropout = 0.05, the target modules are q-proj, k-proj, v-proj, o-proj, gate-proj, up-proj, and down-proj. Batch size is 1, gradient accumulation is 8, maximum length is 1024, FP16 and gradient checkpoint are enabled. current-reference The learning rate of 2 × 10 Ω was used to train 5 epochs, and 5 × 10 Ω was used to train 2 epochs in lr-5e5-epoch2. The remaining key configurations are consistent with seed = 42.

The logical parameters of the model are 1,720,574,976. The ideal FP16 and 4-bit weights are about 3.44 GB and 0.86 GB, and Double Quantization is calculated at about 0.37 bit / parameter, which can save about 0.08 GB of quantization constant overhead. These are theoretical values without LoRA parameters, activation, KV cache, optimizer and CUDA allocator. The final verification CUDA peak in the log is 2.71 / 4.14 GB, which cannot be mixed with the theoretical weight or training memory.

QLoRA freezes quantitative weights and forms effective weights with low-rank increments:

$$W' = W + \frac{\alpha}{r} BA \tag{1}$$

DOI: 10.54254/2755-2721/2026.38166

Among them, W is the basic weight, A and B are low-rank matrices, r is the low-rank dimension, α is the scaling coefficient, W' is the sum of the basic weight and the LoRA increment.

The training only computes the mask negative log-likelihood for the Assistant reply token:

$$L = -\frac{1}{\sum_{i=1}^{T} m_i} \sum_{i=1}^{T} m_i log p_\theta(y_i \mid x, y_{<i}) \tag{2}$$

Here, L the loss function, T is the total length of the sequence token, m_i is the supervised mask of the i th token, y_i is the target token, x is the input context, $y_{<i}$ is the previously generated target token, p_θ is the conditional probability distribution given by the model with the parameter θ.

4. Experimental design

4.1. Contrast experiment and ablation experiment

The main experiment compares current-reference with lr-5e5-epoch2 on the same 440 test samples. The two share data partitioning and generation parameters, and the comparison is used to observe the influence of learning rate and training round adjustment on convergence and generation behavior. Decoding ablation is used to set the complete strategy, remove the repeated penalty and n-gram constraints, and unify the seed and generation upper limit of the test sample.

4.2. Indicators and determination

Combined with automatic indicators, BLEU, ROUGE, BERTScore, Distinct, perplexity and generation length are used to detect the quality of generated text. The EOS rate, maximum length termination rate, repetition rate, delay, throughput and peak memory detection model are stable. Finally, the task completion rate is generated to represent the response ratio that meets the task requirements and the worldview conflict rate statistics and the response ratio of the role character, the story fact or the world state contradiction. The role consistency is judged according to style, character, identity and tone.

4.3. LLM-as-a-judge and artificial evaluation

Roles and task attributes that are difficult to judge in LLM-as-a-Judge supplementary indicators. The 440 pairs of replies are compared by deepseek-v4-flash at temperature = 0 from the eight dimensions of role, world, context, task, naturalness, style, length and repetitive control, and output JSON.

In Figure 3, Candidate A / B is randomly exchanged according to seed = 42, Prompt / Rubric version is 1.1-deepseek-json / 1.0, JSON is verified, Resume skips the existing samples, and 440 are successfully resolved.

Manual evaluation is intended to be independently and blindly evaluated by 3 to 5 evaluators. At present, manual comprehensive evaluation has always judged that the generated text has strong character style and task goal consistency, but the sentences generated under similar conditions are the same.

DOI: 10.54254/2755-2721/2026.38166

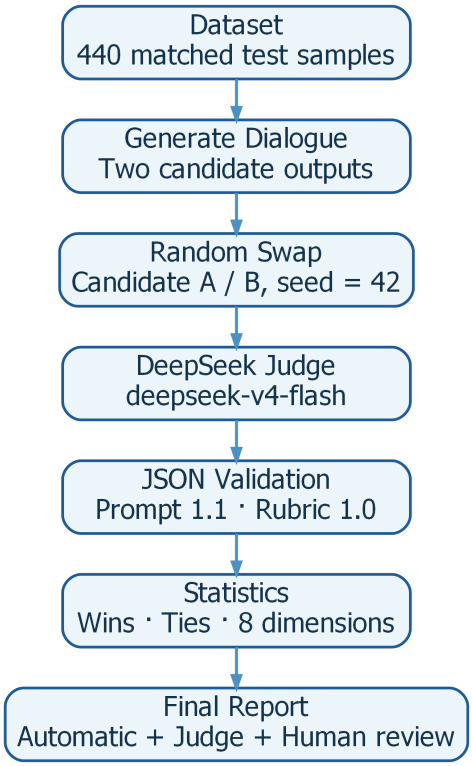


Figure 3. LLM-as-a-Judge pairwise evaluation process (photo/picture credit: original)

In order to ensure repeatability, the experiment unified seed = 42, and saved parameters such as configuration, log, checkpoint, prediction, index and Judge JSONL. Both text and JSONL use UTF-8, and Judge supports Resume to skip completed samples.

5. Experimental results and analysis

5.1. The source of the results and comparability description

The results complete the complete prediction of 440 texts, generating automatic indicators and DeepSeek Judge. The automatic index description refers to similarity, generation behavior and efficiency, and Judge provides eight-dimensional score and pairwise preference. Although the multi-random seed method is used, the significance test and conclusion are limited by the current fixed test set and configuration.

5.2. Training convergence results

It can be seen from Table 1 that the verification loss of the candidate model is reduced from 3.1384 to 3.0810, the perplexity is reduced from 23.0672 to 21.7801, and the Token Accuracy is slightly reduced from 0.4026 to 0.4014. It shows that the candidate model has improved the conditional probability fitting of the verification corpus, but the token-level accuracy rate has not been improved synchronously, so it still needs to be explained in combination with the generated evaluation.

DOI: 10.54254/2755-2721/2026.38166

Table 1. Training and verification summary indicators

Index	Reference Flowsheet	Candidate Model
Best Validation Loss	3.1384	3.0810
Perplexity	23.0672	21.7801
Token Accuracy	0.4026	0.4014

5.3. Automatic evaluation results

The ROUGE-1 / 2 / L and BLEU-1 / 2 / 4 of the candidate models in Table 2 are higher, indicating that their fixed reference responses have more lexical overlaps. However, the enhanced sample keeps the original answer unchanged, and its reference similarity cannot directly represent the role, world or task correctness.

Table 2. Automatic reference similarity index

Index	Reference Flowsheet	Candidate Model
ROUGE-1	0.1437	0.1762
ROUGE-2	0.0160	0.0211
ROUGE-L	0.1032	0.1230
BLEU-1	8.6143	15.9808
BLEU-2	2.5845	6.1598
BLEU-4	0.3640	1.4486

5.4. Generating behavior and reasoning efficiency

The average length of candidate models in Table 3 increases from 28.39 to 40.55 token, the generation time increases from 4.62 s to 6.64 s, and the throughput and peak memory are basically unchanged. The EOS rate decreased from 98.41 % to 86.59 %, the maximum length termination rate increased from 1.59 % to 13.86 %, and the Distinct decreased, in which the ternary and quaternary repetition rates increased. Distinct-n is the ratio of different n-gram numbers to total n-gram numbers, where the output length increases the denominator. If the new content follows the existing expression and the only n-gram does not increase year-on-year, Distinct will decrease. Therefore, it can be seen that the change is affected by both length and repetition rate, but it is not equivalent to the decline of semantic quality.

Table 3. Generation behavior and stability index

Index	Reference Flowsheet	Candidate Model
Distinct-1	0.0889	0.0642
Distinct-2	0.3530	0.2759
Distinct-3	0.5378	0.4509
Average output length	28.39	40.55

DOI: 10.54254/2755-2721/2026.38166

Table 3. (continued)

EOS termination rate	98.41%	86.59%
Maximum length termination rate	1.59%	13.86%
Ternary repetition rate	0.53%	2.75%
Quaternion repetition rate	0.00%	0.55%
Average generation time (s)	4.62	6.64
Tokens/s	6.02	6.01
Peak GPU memory (GB)	1.4406	1.4406

5.5. Reliability check and result discussion

The Prompt / Think Leak, Reference Copy and Empty Output Rate of the two models are all 0. In the ablation experiment, the EOS rate of the complete strategy was 87.27 %, which decreased to 70.68 % after removing the n-gram constraint, and the maximum length termination rate increased to 30.00 %, while the ROUGE-L increased slightly to 0.1300 when the repeated punishment was removed. The results show that the decoding constraint mainly affects the repetition and termination behavior, but the reference similarity, diversity and termination stability are not the same optimization goal.

5.6. LLM-as-a-judge results and comprehensive discussion

As shown in Figure 4, all 440 Judge evaluations were completed. current-reference won 200 times, lr-5e5-epoch2 won 159 times, 81 times in a draw, accounting for 45.45 %, 36.14 % and 18.41 % of the total. The conditional winning percentage of 359 pairs after eliminating the draw is 55.71 % and 44.29 %, but it cannot be regarded as the overall winning percentage of all samples. Nearly one-fifth of the draw shows that when a considerable number of samples do not have a clear overall preference, the model difference is not overwhelming. No winner mapping error was found after random exchange of Candidate A / B.

Table 4 shows that current-reference is higher in role, world, naturalness, style, length and repetitive control, with a comprehensive score of 2.1940. However, lr-5e5-epoch2 was slightly higher in context and task, with a comprehensive score of 2.0777. However, the candidate model is closer to the reference text and task objectives, while the short output, high EOS rate and Judge score of the benchmark model reflect more robust role performance and repetitive control. This shows that loss and evaluation indicators are more sensitive to reference matching, while Judge considers open attributes such as identity, naturalness and length, and the quality of the two types of evidence is different. Therefore, the final model is selected as current-reference, and the task-fitting advantage of lr-5e5-epoch2 is retained.

Table 4. The mean value of LLM-as-a-Judge eight-dimensional score

Dimensional	Reference Flowsheet	Candidate Model
Role consistency	2.2841	2.1205
Consistency of worldview	2.3477	2.2773

DOI: 10.54254/2755-2721/2026.38166

Table 4. (continued)

context dependence	2.0659	2.0750
mission completion	1.6341	1.7273
naturality	2.7477	2.2386
Style fidelity	2.2364	2.0659
Length appropriateness	3.3386	3.1023
repetitive control	4.6500	4.2795
Overall Score	2.1940	2.0777

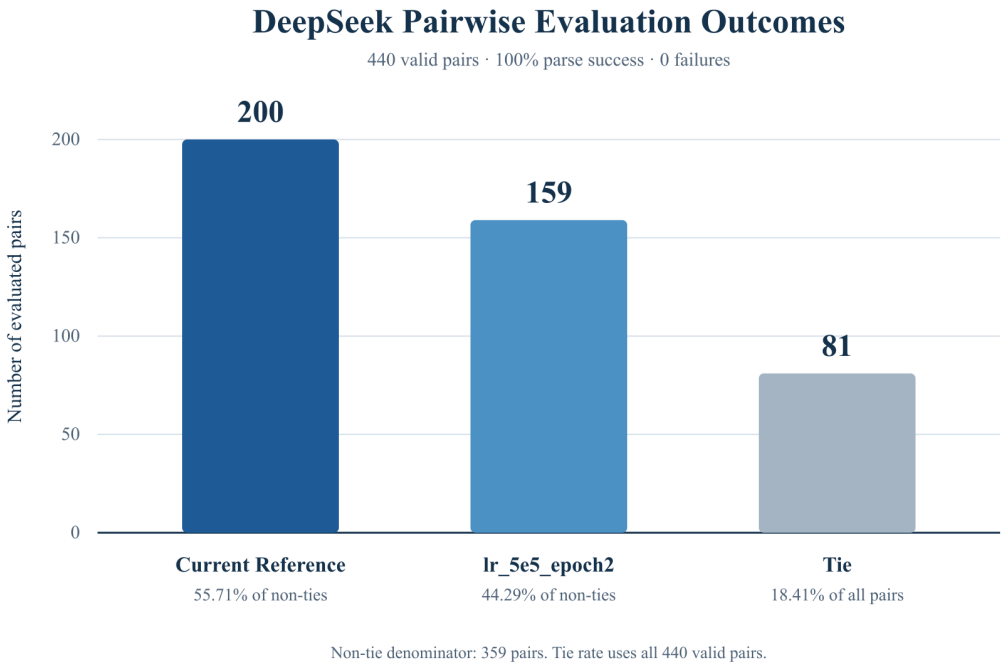


Figure 4. LLM-as-a-Judge pair preference results (photo/picture credit: original)

5.7. Dynamic constraint classification and generation case comparison

Table 5 compares the same test sample by role, world, scene, task and history. The world and task information of candidate replies are more specific, and the role function is more clear, but the connection between history and player questioning is still indirect, indicating that the advantage of a single case cannot replace the overall evaluation.

Table 5. Dynamic constraint classification and comparison of two model generation

Dimensional	Role and style	World and scene	World and scene	Genuinely generate text
Input constraints or judgment basis	Journal-Hunter, reserved, observable, concise	Radiance Infection, lore_revelation	Records Entombed Husk Clues, Undertakes 'Broken Kingdom-Truth'	The test sample was 16f52ac45f3efbcf, and both responses ended with EOS.

DOI: 10.54254/2755-2721/2026.38166

Table 5. (continued)

reference flowsheet	The first person judgment, the hunter 's picture style tone is weak.	When mentioning infection and kingdom, there are few details of the scene.	Echoing the truth, but not describing Entombed Husk.	Yes, I can see the truth now. The infection has taken hold of the kingdom. It will not be undone.
candidate model	Biological entry expression, the role function is more clear.	Clear Radiance, infected body and sealed state.	The task is more fitting, and the historical connection is still more indirect.	The body of a bug infected by the Radiance. It has been sealed away by its own mind, but the infection will not be contained forever.
sample comparison	Strong candidate	Strong candidate	Task improvement, history mixing	25→31 tokens

6. Conclusion

In this study, 4375 samples were constructed with 875 empty knight texts. Domain adaptation is completed based on Qwen3-1.7B-Instruct and QLoRA, and roles and tasks are coordinated with dynamic Prompt. The study combines reference similarity, diversity, termination behavior, efficiency, decoding ablation and pairwise preference as evaluation criteria. Form a closed-loop experiment of data, training, generation and evaluation, and avoid relying solely on BLEU and ROUGE to select the model.

Among them, the loss, confusion, BLEU and ROUGE of the candidate models are better, but the output becomes longer, EOS rate and Distinct decrease. Among the 440 pairs of Judge results, the benchmark model had a higher winning rate. Therefore, the deployment adopts the benchmark model and retains the task fit advantage of the candidate model.

It is mainly limited to homologous enhancement. When each original reply corresponds to five different context variants, its target reply remains unchanged. Although the data is divided before enhancement and there is no cross-set data leakage, the fixed reference structure may still amplify BLEU and ROUGE and limit the generalization judgment of unfamiliar roles, tasks and scenarios. In the future, it is necessary to establish a test set of independent strange roles and scenes, adopt group-based split and increase diversified target responses.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https://doi.org/10.48550/arXiv.1706.03762

[2] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2106.09685

DOI: 10.54254/2755-2721/2026.38166

[3] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems, 36*, 10088-10115. https://doi.org/10.48550/arXiv.2305.14314
[4] Qwen Team. (2025, April 29). Qwen3: Think deeper, act faster. https://qwenlm.github.io/blog/qwen3/
[5] Gao, Q. C., & Emami, A. (2023). The Turing quest: Can transformers make good NPCs? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 93-103). https://doi.org/10.18653/v1/2023.acl-srw.17
[6] Weir, N., Thomas, R., D'Amore, R., & Van Durme, B. (2024). Ontologically faithful generation of non-player character dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 9212-9242). https://doi.org/10.18653/v1/2024.emnlp-main.520
[7] Sadeq, N., Xie, Z., Kang, B., et al. (2024). Mitigating hallucination in fictional character role-play. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 14467-14479). https://doi.org/10.18653/v1/2024.findings-emnlp.846
[8] Tu, Q., Fan, S., Tian, Z., et al. (2024). CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11836-11850). https://doi.org/10.18653/v1/2024.acl-long.638
[9] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). https://doi.org/10.3115/1073083.1073135
[10] Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74-81).
[11] Zheng, L., Chiang, W. L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems, 36*, 46595-46623. https://doi.org/10.48550/arXiv.2306.05685
[12] Liu, Y., Iter, D., Xu, Y., Wang, Z., Zhu, S., & Zeng, M. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2511-2522). https://doi.org/10.18653/v1/2023.emnlp-main.153
[13] Van Nest, J. (2026, January 13). HollowKnightTextDatabase [Data set]. *GitHub*. https://github.com/SAM-stack-star/HollowKnightTextDatabase