香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Natural Language Processing

Grounded NPC Dialogue Generation for Genshin Impact with Persona and Memory Consistency

Authors: Yiyang Lu, Ziqi Chen, Yuxin Tian

Student IDs: 225040537, 123090059, 225040478

Email: 225040537@link.cuhk.edu.cn, 123090059@link.cuhk.edu.cn, 225040478@link.cuhk.edu.cn

Project & Demo: https://chenziqiadam.github.io/Genshin-NPC-Dialogue-Generation/

Introduction

- **Goal:** build a dialogue agent that can role-play Zhongli, an NPC from Genshin Impact, while staying in character during open-ended conversations.
- **Challenge:** generic LLMs often fail on three aspects: role consistency, lore knowledge, and behavior consistency.
- **Key observation:** SFT can improve style but may overfit to short NPC-like lines; RAG can improve grounding but may distract the model if retrieval is noisy.
- **Research question:** how do backbone strength, SFT, and RAG interact when building a game NPC dialogue agent?

Motivation

1. Persona Fidelity

Zhongli-like voice, restraint, contracts, and Mora quirks.

2. Lore Correctness

Accurate Teyvat facts: Archons, regions, characters, history.

3. Task Robustness

Answer the actual request despite style shifts, conflicts,or RAG noise.

A strong NPC agent must both stay in character and answer the user's actual question.

Experimental Setup

qwen3: 8 Qwen3 model variants: 4B Base, 4B Base + RAG, 4B SFT, 4B SFT+RAG, 8B Base, 8B Base + RAG, 8B SFT, 8B SFT+RAG.

62 Chinese evaluation prompts.

496 total model responses.

2 local LLM judges: Llama-70B and Qwen-30B.

7 scoring dimensions: role consistency, factual correctness, RAG grounding, hallucination control, instruction following, naturalness, overall quality.

Evaluation Pipeline

62 Prompts → 8 Model Variants (Responses) → LLM Judges (Llama-70B & Qwen-30B) Score 7 Dimensions → Aggregate & Report Figures

Methodology

A. Data & Knowledge

- **SFT data**: character stories, voice lines, birthday notes, behaviors, traits, lists, relations.
- **RAG knowledge**: base: Zhongli lore beyond world knowledge, character relations, behavior rules.

B. Model Training

LoRA - based SFT on Qwen3-4B-Instruct and Qwen3-8B.

LoRA

C. Retrieval

BAAI / bge-small-zh-v1.5 embeddings + FAISS index retrieve top-k chunks.

D. Inference

System prompt + retrieved knowledge + model output

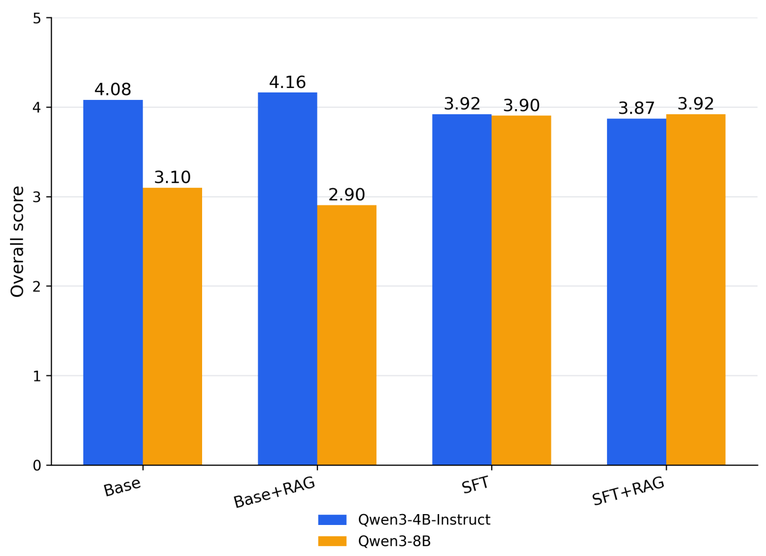
E. Comparison

Simultaneous comparison of Base / SFT / Base + RAG / SFT + RAG.

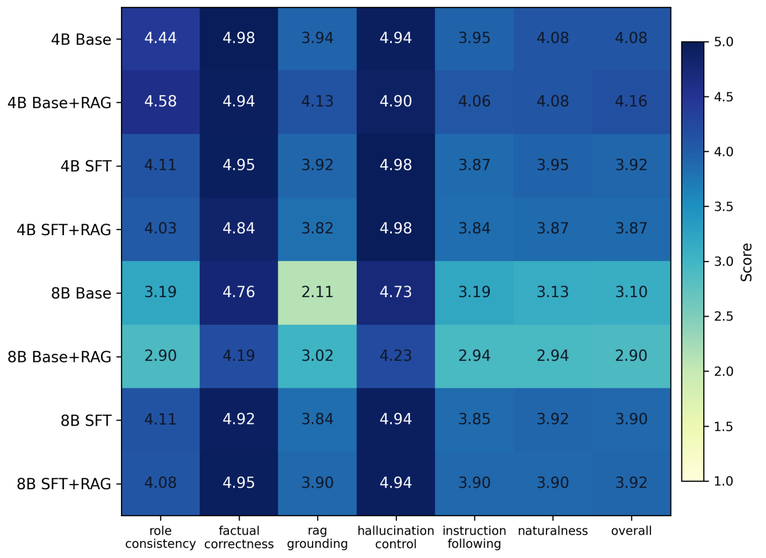
- SFT teaches how Zhongli should speak.
- RAG provides what Zhongli should know.
- Prompt constraints preserve behavior consistency.

Results

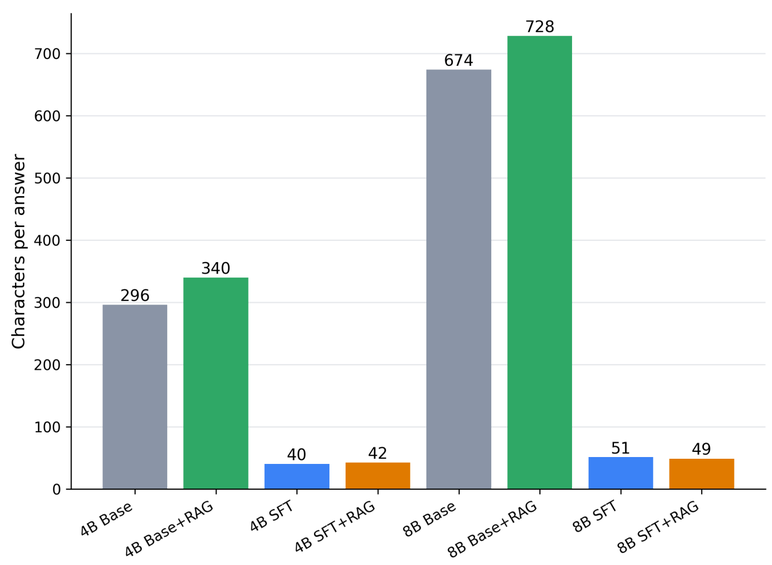
1) Overall Scores by Training / Retrieval Setting (Llama-70B Judge)



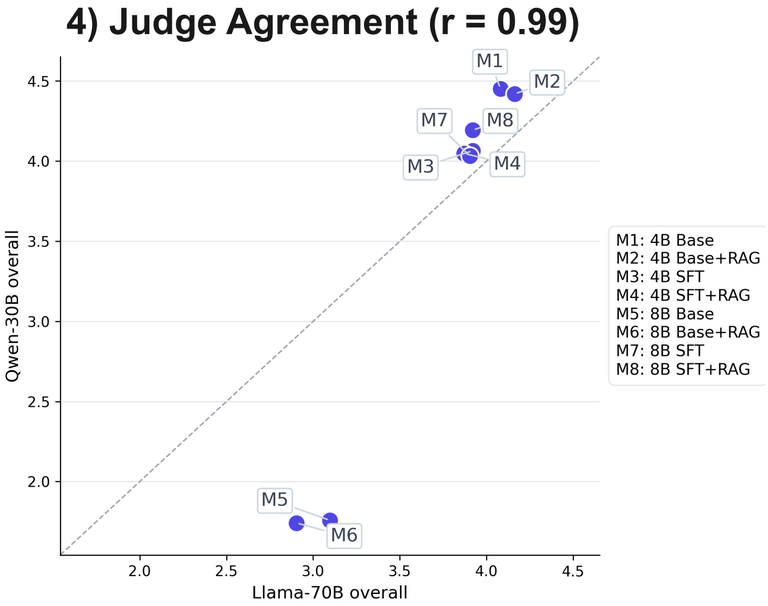
2) Dimension Scores (Llama-70B Judge)



3) Average Answer Length



4) Judge Agreement (r = 0.99)



Key Findings

★ Best primary result: 4B Base + RAG reaches 4.16/5 overall.
★ SFT strongly improves Qwen3-8B: 3.10 → 3.90 (+0.81).
★ SFT hurts Qwen3-4B-Instruct: 4.08 → 3.92, largely because answers become too short.
★ RAG slightly helps 4B Base, but may hurt when retrieved context is irrelevant or too broad.
★ Judge agreement is high, supporting the reliability of the trend.

Conclusion

✔ The effect of SFT and RAG is conditional on the backbone.
✔ Current SFT improves role flavor but may weaken complete instruction following.
✔ Future work: longer instruction-style SFT data, retrieval filtering/reranking, answer-length control, and explicit training against reasoning leakage.