

SENTIENT YET CONSISTENT LLM-POWERED NPC

15 July 2025, Alisa Dammer

ABOUT ME

Freelance MLE by day, hobbyist learner and programmer by night. Love reading about new health-related researches. Love playing around with gadgets. Have time.

WANNA CONNECT?

 @FatDataUnicorn@fosstodon.org

 @FatDataUnicorn

 Alisa Dammer#5584

 alisadammer

 alisadammer.com



STRUCTURE

1. Intro
 - a. Use cases
 - b. Advantages and disadvantages
2. Metrics
 - a. Data collection
 - b. Example: Response novelty
 - c. Example: Quest Consistency Ratio
3. Fine-tuning
 - a. Simple
 - b. Architectural
 - c. Expensive
4. Conclusion

AI IN GAMING

What is it actually useful for?



STATIC CONTENT AUTOMATION

- Based on classic approach
- Scaling factor
- Highest control
- Lowest cost

QUEST GENERATION

- Content scaling
- Mid-level control
- Mid-level cost

DYNAMIC DIALOGS

- Content scaling
- Mid-level control
- Mid-level cost



MULTI-PLATFORM BOTS

- Community building
- Low-level control
- Mid- to low-level cost
- Moderation needed

GRAPHICS AND ANIMATION

- Content enhancement
- Mid- to low-level control
- Highest cost



EXAMPLES

DREAMIO: AI-Powered Adventures

Pause Save Load Options

Status

1. location: convenience_store
2. time: day
3. items: phone, wallet, keys, first-aid kit, water (2 bottles), snacks, baseball bat, kitchen knife
4. companions: none
5. attributes: strength: 5, agility: 6, intelligence: 7

You
Yes, I'm going to try to reach the armory.

Dice roll
The player has rolled the dice. Result: 12 out of 14 (x D20) - Failure.

Warning
The player wants to go to the armory. The store owner warned it's a long shot. The player rolled a 12 on a D20, which is a failure. The roads are jammed, and the aliens are targeting major routes. The player is on foot. The player is likely to encounter danger on the way.

Narrative
You thank the store owner for the information and prepare to leave. He advises his head tech has a brave one, he says, "but think you're making a mistake. It's too dangerous out there." You ignore his warning and step out of the store, into the chaotic streets of Amherst. Suddenly, alien vehicles and the像 quality of his weapons. The roads are a tangle of abandoned vehicles, and the air is thick with smoke and the sound of explosions. As you attempt to navigate the debris-strewn streets, an alien patrol ship descends, its searchlight sweeping across the area. It spots you. An energy blast streaks forward. You don't have time to react.

PLAY FREE

AI Roguelite

Crafting Table

Withering Valley

Area Level 1

You can use space probe to have space eu. Combat will use pots to gain health. A pack of giant bats fly into the valley. The pack of giant bats is the size of a car and has red glowing eyes, large sharp fangs and a large leathery bat wing membrane.

You then cast "Witheress" on the pack of giant bats. All of them fall down onto their backs. You used a spell that causes weakness on the pack of bats. All of the bats lose the strength of their limbs. You now have to fight all the bats against those bats. They have a great laugh. A great spell, but I do not think it will help you with the next wave of monsters.

The AI decides that pack of giant bats was damaged.

You defeated pack of giant bats. You gain 2 XP. You receive as lost: Huge bat wing membrane, Bat wing membrane.

You and the space probe arrive at the forest. A huge, cave-like opening with a strong, constant breeze blowing out of it. You are not sure if this was intentionally designed that way, or if it just happened to be that way.

You and the space probe arrive at the Withering Valley. The Withering Valley is a wasteland. A desolate plain, which looks to be several miles across. A single tree can be seen standing in the centre of the valley.

Places **NPCs** **Enemies**

I'M WATCHING YOU



EVALUATION

Reference based vs Reference free

Data Set vs Single Input

General vs Custom

Specialized vs Free form

Classification, Ranking, etc

BASIC DATA BLOCKS

Limitations:

- Game state
- Mechanics and components
- Characters state
- etc...

Technical Stats:

- Toxicity
- Length
- Relevance
- Sentiment

Additional Stats:

- User engagement
- User survey



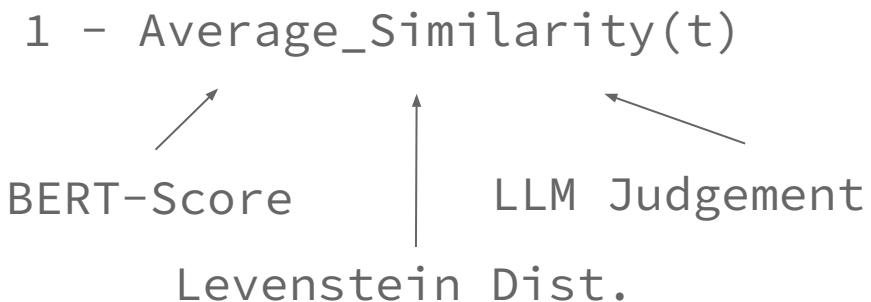
GENERAL LOGGING STRUCTURE

PLAYER INPUT	CLASSIFICATION	RAW RESPONSE	TECHNICAL CHARACTERISTICS (T, S, L, ...)	RELEVANCE SCORE	THRESHOLD
“Hi! How are you”	GET, Free Dialog, Character	“My back is killing me, it might be raining soon”	0.0005, 4, 10, ...	0	pass
“Can you tell me about XXX?”	GET, Free Dialog, Character, World	“Back when the earth was cold... XXX was not yet born, but it was nice”	...	0	fail
“Do you have a task for me?”	POST, World, Quest	“Bring me my left sandal, the red one, the one i bought recently”	...	0.1	pass
“I brought you YYY”	POST, Structured dialog, World, Quest	“Oh! Thx, i don't need it though”	...	0.01	pass

PLAYER SURVEY:

ITEM	COMPLETION	DUPPLICATE POTENTIAL	NOVELTY (1-10)	ESTIMATION (1-10)	FREE INPUT
Quest XXX	ABANDONED	1	2	2	Already did this exact thing for another NPC
Quest XXY	FINISHED	0.1	8	6	Simple, funny
Quest XYY	FAILED	0.2	7	2	Could not finish within provided limitations , NPC kept interfering

RESPONSE NOVELTY SCORE



- Scaling might become a problem
- Can be automated
- Can be enhanced with Player Preference Survey



QUEST-CONSISTENCY-RATIO

(valid quests / total generated) * 100

- Simple enough
- Can be assessed automatically
- Can be enhanced by player feedback and engagement
- Can be more nuanced if “finished/rejected/expired” states are included



I'M MISTER WOLF, I
SOLVE PROBLEMS



SIMPLE

- Inject JSON-structured game state
- Central state engine outside of LLM
- Free-from response on non-modifying events
- Function calling for modifying events
- Use both guidelines and fall back in prompts





ARCHITECTURAL

- RAG for Persist player-NPC-sessions, Lore
- KG for inventory state
- Quest logs
- Functional APIs
- Reinforcement Learning from Human Feedback (RLHF) + Proximal Policy Optimization
- Continuous monitoring

FANCY YET EXPENSIVE

LLM as a judge

Resources:

1. Verga P., et al (2024). Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models,
<https://arxiv.org/abs/2404.18796>
2. Song H. et al (2024). FineSurE: Fine-grained Summarization Evaluation using LLMs,
<https://arxiv.org/abs/2407.00908>
3. Liu Y., et al (2023). G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment,
<https://arxiv.org/abs/2303.16634>
4. Manakul P., et al (2023). SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models, <https://arxiv.org/abs/2303.08896>
5. Wang J., et al (2023). Is ChatGPT a Good NLG Evaluator? A Preliminary Study,
<https://arxiv.org/abs/2303.04048>



CONCLUSION

- Still need tons of data
- Mixture of hard- and custom metrics
- Continuous logging required
- Still in the beginning but promising
- GameDev now needs DS/DE/MLE =)
- Enhanced community building possibility
- Strong cost factor consideration
- Performance limitations

RESOURCES

1. Lin Y., Chen Y. (2023). LLM-EVAL: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models, <https://arxiv.org/abs/2305.13711>
2. Thomas P., etal (2024). Large Language Models can Accurately Predict Searcher Preferences, <https://arxiv.org/abs/2309.10621>
3. Kocmi T., Federman C. (2023). Large Language Models Are State-of-the-Art Evaluators of Translation Quality, <https://arxiv.org/abs/2302.14520>
4. EvidentlyAI (July 2025).
<https://www.evidentlyai.com/llm-guide/llm-evaluation>
5. Microsoft AI Playbook (July 2025). [Link](#)
6. Taubenfeld A., etal (2025). Confidence Improves Self-Consistency in LLMs, <https://arxiv.org/html/2502.06233v1>
7. Park J.S., etal (2023). Generative Agents: Interactive Simulacra of Human Behavior, <https://arxiv.org/abs/2304.03442>
8. Fan J., Gertz M. (2023). Evaluating Factual Consistency of Texts with Semantic Role Labeling, <https://arxiv.org/abs/2305.13309>
9. Song L. (2025). LLM-Driven NPCs: Cross-Platform Dialogue System for Games and Social Platforms, <https://arxiv.org/html/2504.13928v1>