A Symbolic Linguistic Approach to Fine-Tuning Large Language Models for Role-Playing Game Storytelling

Steven Tidwell

Texas A&M – Commerce

AI-510 – Computational Symbolic Linguistics

Dr. Christian Hempleman

5/9/2025

Abstract

This paper presents DM42, an AI-driven Dungeon Master system designed for retro-inspired role-playing game (RPG) environments, integrating symbolic computational linguistics with fine-tuned large language models (LLMs). The system addresses limitations in current LLMs regarding narrative consistency and domain-specific reasoning by incorporating structured metadata, rule-based mechanics, and syntactic analysis into both training and inference pipelines. DM42 employs a hybrid architecture combining LoRA-based fine-tuning for Mistral 7B models with Retrieval-Augmented Generation (RAG), utilizing vectorized campaign lore stored in a PostgreSQL database with pgvector indexing. Key datasets, including structured dialogue logs, annotated gameplay transcripts, and syntactically parsed medieval texts, are preprocessed and enriched with symbolic annotations to enhance contextual relevance and control. Additionally, Penn Treebank syntactic trees are used to support grammatical scaffolding during prompt generation. The system's outputs are evaluated across multiple fine-tuned LLM variants to assess narrative coherence, character consistency, and logical fidelity. This work demonstrates the viability of combining symbolic and statistical methods to achieve robust, domain-aware narrative generation, offering a foundation for future research in AI-driven storytelling and interactive game systems.

Introduction

Large language models (LLMs) have significantly advanced the field of natural language generation, offering unprecedented capabilities in creative writing, dialogue systems, and narrative construction. However, despite these advances, LLMs struggle with structured, domain-specific storytelling tasks, particularly in fantasy and role-playing game (RPG) settings where coherence, consistency, and adherence to world rules are essential. The DM42 project proposes a hybrid approach, integrating symbolic computational linguistics and statistical modeling to address these limitations. By combining fine-tuned LLMs with symbolic rule layers and carefully structured datasets, the system aims to produce an AI Dungeon Master (DM) capable of immersive, logically consistent gameplay narration within a retro-inspired RPG environment.

Symbolic Computational Linguistics in DM42

Symbolic computational linguistics focuses on rule-based approaches to language understanding and generation, contrasting with purely statistical models. In DM42, symbolic methods are incorporated primarily through the careful structuring and preprocessing of datasets. Metadata tagging organizes the narrative space, allowing deterministic retrieval of relevant lore segments. These metadata tags, such as location, scene type, and non-player character (NPC) identity, enable precise RAG (Retrieval-Augmented Generation) queries, improving the AI's contextual awareness. Additionally, syntactic structures derived from Penn Treebank parsing provide scaffolding for prompt construction, enhancing grammatical correctness and narrative coherence.

Dataset Creation and Preprocessing

The DM42 project integrates multiple datasets to train and guide the AI Dungeon Master system, including:

- Critical Role Dungeon (CRD) data, providing annotated examples of in-character dialogue and narrative exchanges.
- Fireball dataset, offering structured gameplay transcripts from Dungeons & Dragons sessions with examples of combat interactions, player choices, and NPC responses.

- Project Gutenberg medieval texts, offering stylistic and thematic grounding, processed with syntactic analysis for metadata enrichment.

These datasets required intensive preprocessing. Raw texts were cleaned of noise, duplicate data, irrelevant dialogue markers, and inconsistencies. They were then transformed into fine-tuning-ready structures using an instruction:input:output pattern, ensuring alignment with the symbolic components needed for effective LLM integration. Metadata headers were embedded directly into the dataset to support symbolic control during both training and inference, allowing the AI to draw upon tagged contextual details, such as location or NPC, when constructing responses.

Metadata Tagging and Symbolic Control

A key aspect of symbolic linguistics in DM42 is the application of structured metadata. Campaign texts were annotated with structured headers (for example, ### location, ### npc, ### scene_type) and chunked using delimiter markers. This ensured that the AI could retrieve, rank, and apply relevant lore during inference. For example, a player prompt about exploring the Blackfen Wilds would automatically trigger lore retrieval marked with matching location and scene type, ensuring the AI's output remained coherent and contextually relevant.

This approach bridges the gap between symbolic control and statistical generation, allowing the system to balance deterministic retrieval with probabilistic text generation. In retrospect, the success of this approach relies heavily on dataset quality, making preprocessing, cleaning, and structured annotation critical components of the overall system.

Penn Treebank Integration

To enhance linguistic precision, the project employed the Penn Treebank (PTB) syntactic parser. Medieval texts were parsed to extract subject-verb-object relationships, enabling more accurate tagging and filtering within the RAG pipeline. For example, a parsed sentence like "The baron was slain by wolves" allowed the system to isolate the key entities and actions, improving the AI's ability to tie responses to specific quest hooks or world events. This approach highlighted how symbolic parsing could reinforce narrative grounding and improve the fidelity of AI-generated dialogue.

Technical Stack

DM42 leverages a layered architecture:

- LLMs: Mistral 7B models fine-tuned using LoRA, producing three tuned versions (v1, v2, v3), alongside Phi-2 as a lightweight, CPU-friendly fallback.
- Dataset Design: CRD dialogue, Fireball combat logs, and Gutenberg texts formatted into JSONL batches, supporting efficient fine-tuning and vectorization.
- Inference Pipeline: Local inference powered by Ollama, wrapped in Python, connected to a RAG layer pulling lore vectors from a PostgreSQL database enhanced with pgvector indexing.
- Symbolic Layer: Metadata-enriched prompts and syntactic scaffolds enhance consistency and rule compliance, ensuring that AI outputs remain logically sound.

Testing and Evaluation

Model evaluation involved comparing outputs across Mistral 7B v1, v2, v3, and Phi-2 using identical prompts. Assessment criteria included coherence (logical narrative flow), memory and consistency (tracking past events and character states), symbolic integration (correct application of metadata and symbolic structures), and immersion (engaging, in-character responses). These evaluations combined qualitative (human judgment) and quantitative (prompt-response) methods, identifying strengths and limitations across model versions. Notably, the importance of well-structured, symbolically enriched datasets became clear during evaluation, as models trained with detailed metadata consistently outperformed those relying solely on raw text.

Future Directions

Future work includes improving symbolic parsing by expanding the Penn Treebank integration, enhancing metadata systems with more granular tags such as emotional tone or narrative perspective, and extending symbolic frameworks to handle multi-user interactions. Additionally, integrating symbolic rules to guide dynamic narrative construction in unexplored regions offers promising opportunities for procedural content generation.

Conclusion

The DM42 project demonstrates the powerful interaction between symbolic computational linguistics and statistical LLM architectures. By emphasizing dataset structuring, metadata annotation, and syntactic preprocessing, the project advances the capacity of AI systems to deliver immersive, domain-specific storytelling experiences. In retrospect, it is clear that symbolic methods provide essential scaffolding for maintaining coherence, logical consistency, and contextual relevance in AI-driven narrative generation. This work offers a foundation for future exploration at the intersection of linguistics, machine learning, and interactive entertainment.

REFERENCES

Basic Fantasy RPG. (n.d.). Retrieved from https://www.basicfantasy.org/

Critical Role Dungeons (CRD) dataset. (n.d.). Hugging Face. Retrieved from https://huggingface.co/datasets/microsoft/crd3

Fireball dataset. (n.d.). Retrieved from https://github.com/zhudotexe/FIREBALL

LoRA: Low-rank adaptation of large language models. (2021). arXiv. https://arxiv.org/abs/2106.09685

Mistral 7B. (n.d.). Hugging Face. Retrieved from https://huggingface.co/mistralai/Mistral-7B-v0.3

PGVector. (n.d.). Retrieved from https://github.com/pgvector

Phi-2. (n.d.). Hugging Face. Retrieved from https://huggingface.co/microsoft/phi-2

Project Gutenberg. (n.d.). Retrieved from https://www.gutenberg.org/ebooks/

RAG implementation. (2020). arXiv. https://arxiv.org/abs/2005.11401