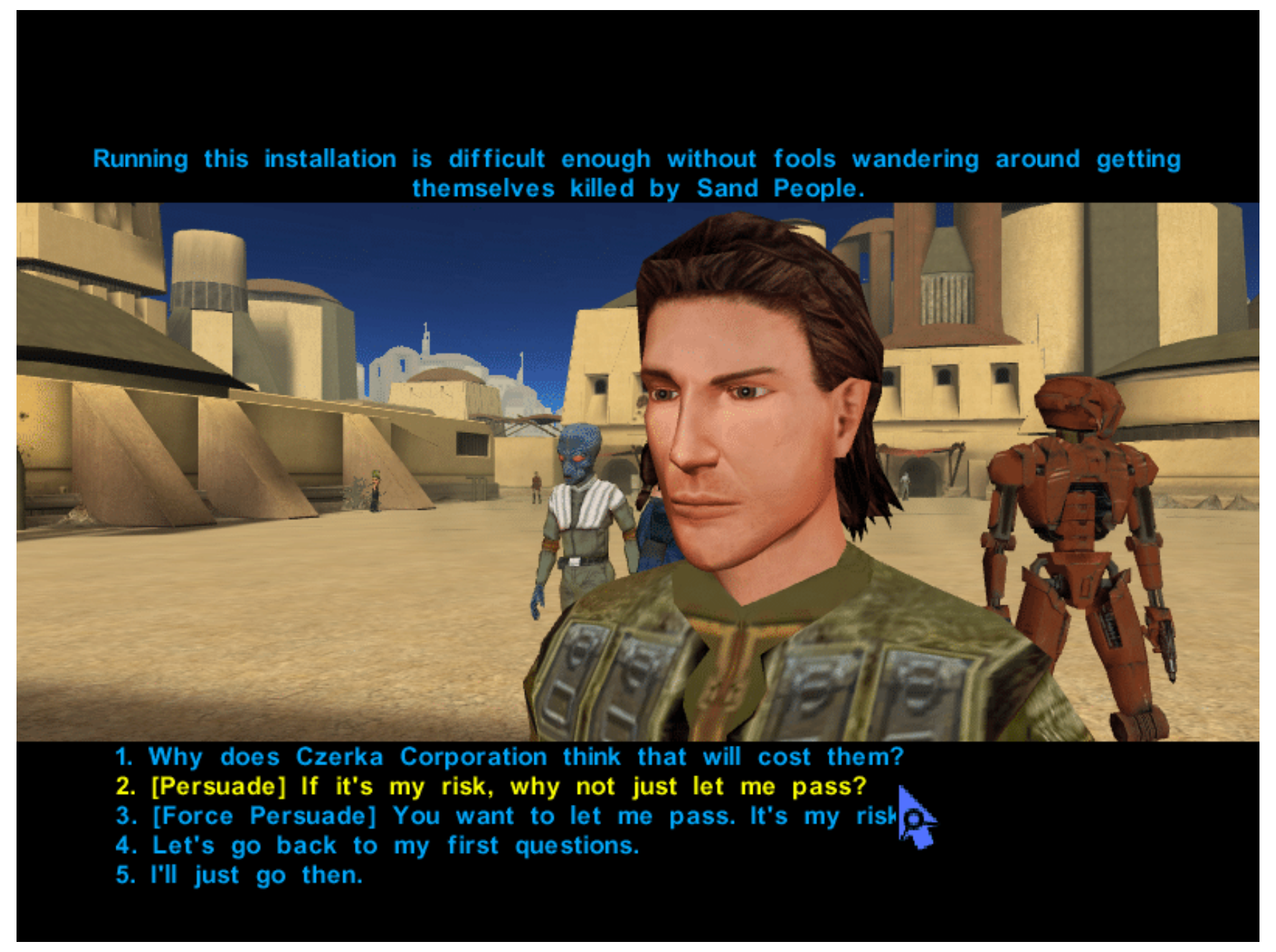
JEDI: Justifiable End-dialogue Driven Interaction for NPC Entities in Role-Playing Games

Willy Chan Omar Abul-Hassan Sokserey Sun

Department of Computer Science, Stanford

Introduction

Story-driven role-playing games (RPGs) are highly popular and financially successful, generating billions annually. They feature complex narratives and allow players to shape game outcomes through their choices.



Running this installation is difficult enough without fools wandering around getting themselves killed by Sand People.

1. Why does Czerka Corporation think that will cost them?
2. [Persuade] If it's my risk, why not just let me pass?
3. [Force Persuade] You want to let me pass. It's my risk
4. Let's go back to my first questions.
5. I'll just go then.

Figure 1. Dialogue Options in Star Wars KOTOR

Creating text and dialogue for RPGs is labor-intensive and costly, requiring human writers to craft responses reflecting various game states and player interactions. While this enhances engagement, it is not scalable for dynamic, real-time interactions. Large Language Models (LLMs) offer potential by generating dynamic, context-sensitive dialogue in real-time, providing a customized experience that adapts to evolving game scenarios and player choices.

Models

We tested and fine-tuned 3 models on their dialogue generation ability.

- **BART**: Employs an encoder-decoder architecture with a denoising objective, versatile for tasks like text generation, translation, and summarization. The encoder processes input bidirectionally, while the autoregressive decoder generates output token by token .
- **GPT-2**: Uses a unidirectional transformer architecture with multi-head self-attention and position-wise feed-forward network, trained on extensive internet data for contextually relevant text.
- **GPT-3.5**: An extension of GPT-3 with significantly more parameters, demonstrating enhanced contextual understanding, few-shot learning, and superior NLP task performance.

Metrics

We used these metrics as they provide a robust evaluation of linguistic accuracy and contextual fidelity, ensuring generated dialogues are coherent and integrated with the game's narrative dynamics.

- **BLEURT**: Assesses semantic similarity between predicted and reference sentences, ensuring alignment with player choices.
- **BLEU**: Measures n-gram precision, crucial for grammatical correctness and contextual appropriateness.
- **ROUGE**:
  - **ROUGE-1 and ROUGE-2**: Capture unique lexicon (e.g., "Wookiee", "lightsaber").
  - **ROUGE-L**: Evaluates structural coherence through the longest common subsequence.
- **DialogueRPT**: Evaluates relevance and human-like quality.
  - **Human-vs-Rand**: Determines relevance based on prior interactions.
  - **Human-vs-Machine**: Assesses how human-like responses are.
- **BERTScore**: Measures semantic similarity through cosine similarity between embeddings, crucial for complex interactive settings.

Dataset: Star Wars - Knights of the Old Republic

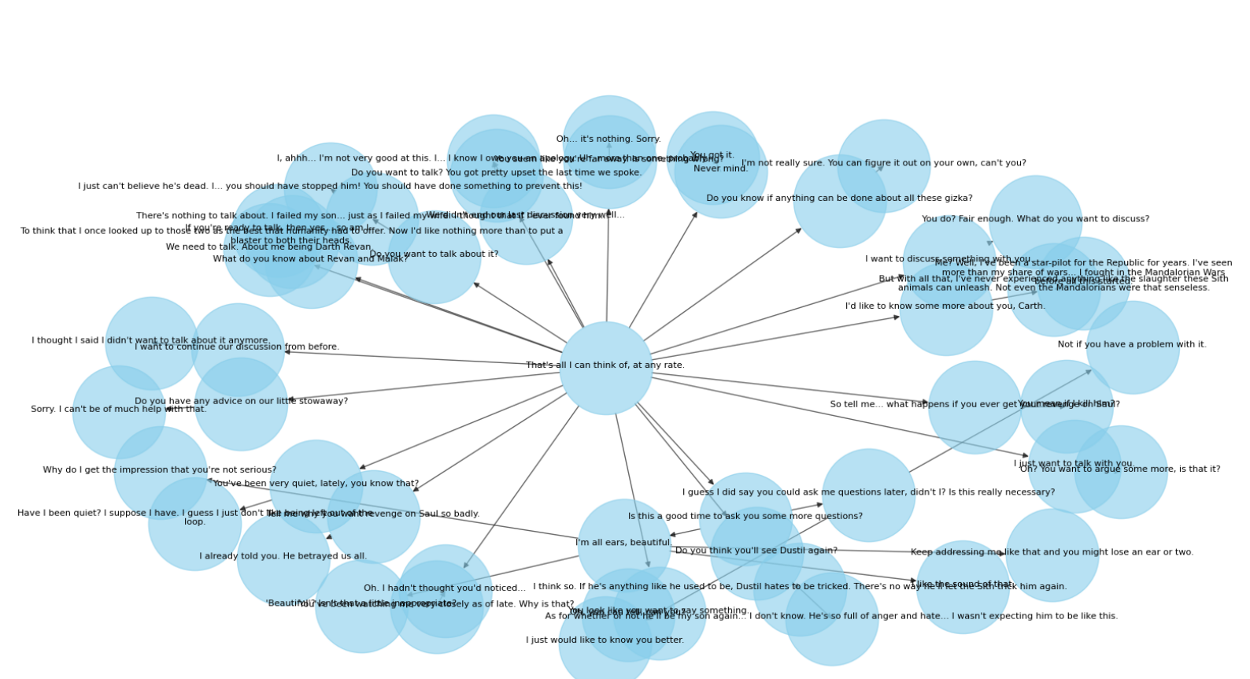
Our project uses a text dataset from "Star Wars: Knights of the Old Republic" (KOTOR). KOTOR, known for its branching narrative and dynamic character interactions, contains over 600,000 words, covering every dialogue interaction in the game. Players navigate a complex array of dialogue choices with non-player characters (NPCs), where these choices and the player's game state influence subsequent NPC dialogues.

Key	Description	Example value
Id	28209	Identifier of this dialogue act in the dataset
Speaker	Judge Shelkar	The character or object that communicates the line
Listener	PLAYER	The character that listens to the line
Text	For your crimes against Manaan and the Selkath you are banned forever from this world, on pain of death!	String literal
Animation	'Judge Shelkar': 'Talk_Forceful'	3D animation that should be played during the delivery of the line
Comment	if the player is exiled	Game development notes
Previous	[28208, 28252, 28314, 28332]	Identifiers of previous dialogue lines
Next	[28210, 28213, 28215, 28218]	Identifiers of next possible dialogue lines, i.e. possible replies
Source DLG	man26_pcexile	The game file in which this dialogue act can be found.

Figure 2. Dialogue (conversation turn) from the KOTOR dataset. Dialogues consist of multiple turns. Dialogues are stored as double linked list and can be reconstructed by walking the linked list, i.e. following the 'previous' and 'next' references.

Dataset: Dialogue Sequence Linearization

We constructed a graph encapsulating all dialogue interactions from the dataset. Random walks, limited to 500 interactions, generate linear dialogue sequences. Additionally, we integrate information related to animations and developer comments to enrich the contextual understanding. Our models are then tasked with generating a masked conversation turn given the preceding and succeeding text.



Input	Target
Player: Maybe I can help your husband. (Animation: [], Comment: nan) **Elora**: <MASK> **Elora**: Please, I beg you to bring Jolee to speak to me about my husband. Sunry's life depends on it! (Animation: [{'Elora': 'Talk_Pleading'}], Comment: if jolee is not in party, subsequent conversations)	Someone is out to destroy my husband. I... I don't know who I can trust. I don't know you. But Jolee - he was always a true friend to Sunry.
Player: You can't escape, Bastila! Wherever you run I'll hunt you down! (Animation: [{'Player': 'Talk_Forceful'}], Comment: nan) **Bastila**: <MASK> **Bastila**: Now you must pay the price. Here on the Star Forge the power of the dark side is at its strongest. This time you will not defeat me! (Animation: [{'Bastila': 'Talk_Forceful'}, {'Bastila': 'Pause'}], Comment: nan)	I have no doubt of that. I was no match for you in the Temple, but when I fled I knew you would follow me here.

Figure 3. Graph of possible conversation turns: performing a random walk through the graph yields a linear dialogue sequence.

Experiments: Modifications

Label Smoothing: Label smoothing was used to prevent the GPT-3.5 model from being overly confident. This technique softens the one-hot encoded ground truth labels, defined as:

$$y_{\text{smooth}} = y_{\text{true}} \cdot (1 - \alpha) + \frac{\alpha}{K}$$

where y_{true} is the original label, α is the smoothing parameter, and K is the number of classes. This improves generalization.

Cross-Attention Mechanism in GPT-2: We enhanced GPT-2 with a cross-attention mechanism (GPT2CA) to improve contextually relevant responses. This mechanism attends to both input dialogue and supplementary context:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

with $Q = W_q X_{\text{dialogue}}$, $K = W_k X_{\text{context}}$, and $V = W_v X_{\text{context}}$. Weight matrices W_q, W_k, and W_v are learned during training. This setup integrates dialogue and game state information for better responses.

Results & Analysis

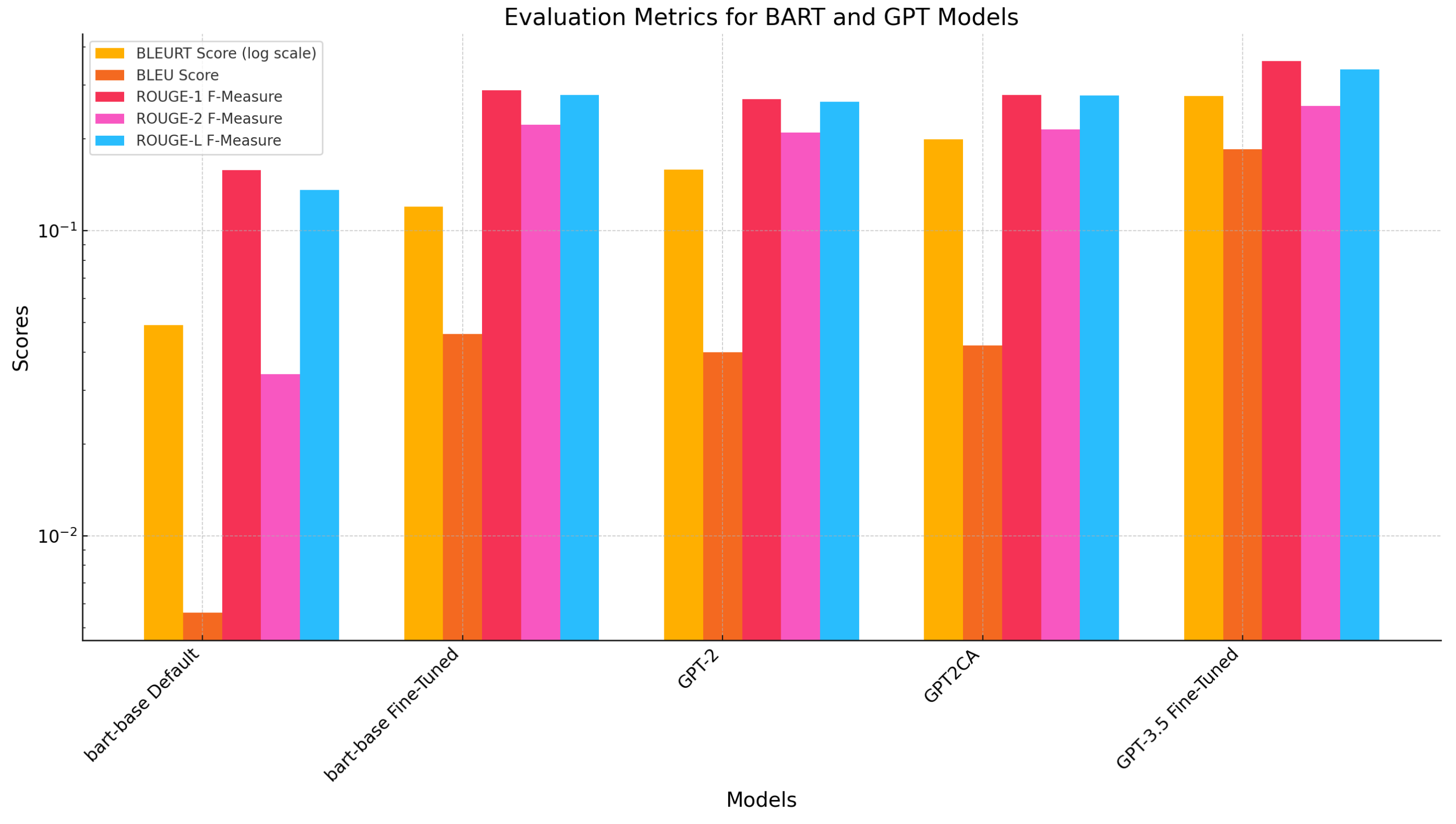


Figure 4. Graph visualizing the difference between models

All models were fine-tuned on our custom dataset using an 80 - 20 training and validation split.

- **Fine-Tuned BART**:
  - Improved BLEURT and BERTScore, indicating better contextual relevance and use of specific terms.
  - Modest improvements in DialogRPT scores suggest limited enhancement in natural, human-like dialogue flow.
  - **Strengths**: Correct word choice and contextual accuracy.
  - **Weaknesses**: Lacks natural flow and subtlety typical of human dialogue.
- **Fine-Tuned GPT-3.5**:
  - Achieved high BERTScore and BLEURT metrics, excelling in semantic similarity and understanding.
  - Strong DialogRPT scores, especially in human-vs-machine evaluations, showing capability in generating fluent and natural text.
  - **Strengths**: Coherent, fluent text with deep semantic understanding.
  - **Weaknesses**: Moderate ROUGE and BLEU scores, indicating challenges in syntactic structure and n-gram matching.
- **Fine-Tuned GPT-2 with Cross-Attention (GPT2CA)**:
  - Significant improvements over standard GPT-2, especially in BLEURT and BERTScore metrics.
  - Better DialogRPT scores, reflecting improved human-like and contextually relevant responses.
  - **Strengths**: Enhanced semantic understanding and context integration due to cross-attention mechanism.
  - **Weaknesses**: Still lags behind GPT-3.5 in overall performance.

Expected Response	Generated Response
He wouldn't dare!	He wouldn't dare!
I figured as much. Get out of here, then. Go on! We have enough to worry about with the shyrack!	Fine. Get out of here! Don't think I'll ever be grateful, because I won't.
[Failure] Critical System Error! Energy Governor Offline! Slimeball Must Die subscript Executing! Isolating unit to minimize damage...	Critical System Error! Energy Governor Offline! Slimeball Must Die subscript Executing! Isolating unit to minimize damage...
K-X12 series hardware offers firepower and stability with unparalleled access to system functions. A battle droid group modified with custom event coding allows for excellent problem-solving and tactics. Caution: malicious code may cause malfunction.	System Status: K-X12a: Critical system failure. No signal...

Table 3: Comparison of Expected and Generated Responses for Finetuned GPT-3.5 Turbo (Best-Performing Model)

Conclusion

With JEDI, we enhanced RPG narrative interactivity by integrating LLMs for dynamic dialogue generation. Fine-tuning BART and GPT models on the KOTOR dataset improved BART's BLEURT scores from -1.3090 to -0.9215 and achieved a BERTScore of 0.8940 with GPT-3.5 Turbo. Despite these successes, challenges like tone inconsistencies and response accuracy remain. Future work will focus on refining attention mechanisms to manage long-range dependencies and maintain narrative coherence in extensive dialogues.