A Conflict-Embedded Narrative Generation Using Commonsense Reasoning

Youngrok Song[1], **Gunhee Cho**[1], **HyunJu Kim**[2], **Youngjune Kim**[3],
Byung-Chull Bae[4] and **Yun-Gyung Cheong**[1]
[1]Department of AI, SungkyunKwan University
[2]Naver Cloud
[3]NCSoft
[4]School of Games, Hongik University
id2thomas@gmail.com, skate4333@g.skku.edu, julia981028@gmail.com, youngjune@ncsoft.com,
byuc@hongik.ac.kr, aimecca@skku.edu

Abstract

Conflict is a critical element in the narrative, inciting dramatic tension. This paper introduces CNGCI (Conflict-driven Narrative Generation through Commonsense Inference), a neuro-symbolic framework designed to generate coherent stories embedded with conflict using commonsense inference. Our framework defines narrative conflict by leveraging the concept of a soft causal threat, where conflict serves as an obstacle that reduces the likelihood of achieving the protagonist's goal by weakening the causal link between context and goal through defeasible inference. Comparative studies against multiple story generation baselines utilizing commonsense reasoning show that our framework outperforms the baselines in creating narratives that distinctly embody conflict while maintaining coherency.

1 Introduction

Conflict plays an essential role in building a compelling narrative, as "a minimal condition for narrative is the thwarting of intended actions by unplanned events, which may or may not be the effect of other characters' intended actions." [Holtzmann, 2016] The plot diagram known as Freytag's triangle (or Freytag's pyramid) also delineates conflict as a core element of fictional narratives by explaining the plot structure with the introduction, climax, and catastrophe (or resolution), where the narrative tension keeps building up to the climax through the protagonist's effort to achieve her goal despite innumerable obstacles. Crafting conflict, however, in a computational manner is a demanding task, requiring the knowledge of the protagonist's goals and obstacles with the understanding of the whole plot structure.

Recent advances in large language models (LLMs) have significantly influenced diverse NLP tasks, including creative writing and story generation [Franceschelli and Musolesi, 2023; Cho *et al.*, 2022]. Pretrained language models can help create or co-author coherent and high-quality fictional narratives with proper prompting or few-shot learning. While existing research has leveraged the commonsense reasoning capabilities of language models to enhance story coherence [Yang *et al.*, 2022], few studies have explored intentionally introducing elements that disrupt coherence, potentially adding depth and interest, as these aspects do not naturally emerge from straightforward commonsense inference. This paper proposes a novel approach to create a coherent, conflict-embedded story in a controllable manner.

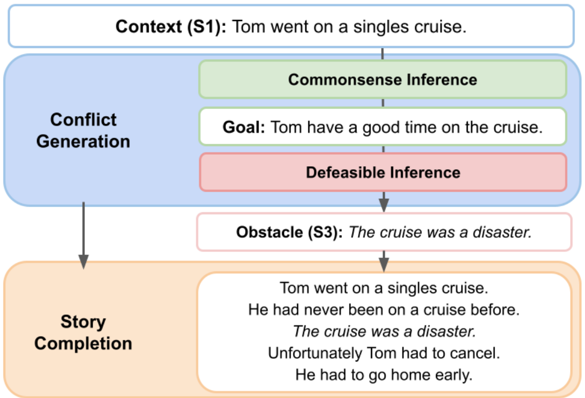
Context (S1): Tom went on a singles cruise.
Conflict Generation
Commonsense Inference
Goal: Tom have a good time on the cruise.
Defeasible Inference
Obstacle (S3): *The cruise was a disaster.*
Story Completion
Tom went on a singles cruise.
He had never been on a cruise before.
The cruise was a disaster.
Unfortunately Tom had to cancel.
He had to go home early.

Figure 1: An overview of CNGCI framework. Sn denotes the sentence in the nth position.

We present a framework with a two-stage (conflict generation and story completion) approach to generating coherent stories embedded with conflict, named the Conflict-driven Narrative Generation through Commonsense Inference (CNGCI). First, our framework employs generative defeasible inference to create conflict within a given context. Next, a fine-tuned GPT2 model and commonsense inference constraints are utilized to complete the story, incorporating the generated conflict and the original context. The human evaluation results with the ROCStories dataset show that the proposed framework can successfully generate stories that explicitly contain conflict while maintaining a coherency comparable to the baselines.

The key contributions of our work are as follows:

1. Introduction of a two-stage framework for generating conflict-embedded stories while preserving coherence, utilizing commonsense inference

2. Utilization of defeasible inference to generate conflict elements within narratives by applying the newly proposed concept of soft causal threat
3. Development of coherence scoring metrics and rules, derived from statistical analysis of story datasets
4. Evaluation of the effectiveness of the proposed framework through comparative human evaluation

2 Related Work

The research on computational models of narrative, a field that has seen extensive exploration of conflict and dramatic tension [Gervás, 2009], has accommodated various innovative approaches. These approaches, including domain-specific symbolic representations, have enhanced the understanding and generation of narrative. From rule-based systems that construct plots based on characters' actions, goals, and potential interventions [Sgouros, 1999; Pérez and Sharples, 2001], to inference-based systems [Szilas, 2003] and planning algorithms that integrate user decisions and character intentions, engaging interactive stories have been generated with inherent conflicts [Mateas and Stern, 2003; Porteous and Cavazza, 2009; Ware *et al.*, 2014].

To generate an interactive narrative with conflict, Sgouros proposed a rule-based approach, forming a plot using potential action sequences driven by each character's goals and roles while considering potential character interventions. Szilas built an inference-based storytelling system that creates interactive stories by incorporating key constraints including consistency, conflict, surprise, and impressiveness. Barber and Kudenko developed GADIN to generate engaging interactive narratives, considering internal conflicts with five distinct categories - betrayal, sacrifice, the greater good, takedown, and favor - collectively termed as 'dilemmas.'

In a planning-based approach, Ware *et al.* developed the CPOCL (Conflict Partial Order Causal Link) planning algorithm, which integrates events that obstruct story characters from achieving their goals while constructing a solution to a narrative planning problem. Song *et al.* adapted the POCL (Partial Order Causal Link) planning algorithm to induce conflicts by imposing ordering constraints, where one character's action threatens the causal link established by another. Gervás *et al.* proposed employing genetic representations to link plot units, termed Axes of Interest (AOI), into a simple story, while certain AOIs implicitly embody elements of narrative conflict.

The approaches mentioned above to generating conflict present a significant challenge to creating such scenarios without using domain-specific symbolic representations. Furthermore, integrating conflicts while maintaining narrative coherence requires complicated computation.

3 The CNGCI Framework

In this section, we present the ***C****onflict-driven* ***N****arrative* ***G****eneration through* ***C****ommonsense* ***I****nference(CNGCI)* framework, designed to create coherent stories featuring conflicts by leveraging commonsense inference. The system operates in two primary stages: **(1) Conflict Generation** and **(2) Story Completion** as depicted in Figure 1. Initially, the framework generates the story's conflict using commonsense and abductive reasoning with an initial sentence given as input. Subsequently, the framework employs a fine-tuned GPT-2 model to expand the narrative, crafting the remaining sentences based on the initial sentence and the conflict established in the first phase.

C	Lana was trying to figure out how to play a song.
G	Lana learn how to play the song.
O	The song is very difficult.
C	Tom went on a singles cruise.
G	Tom have a good time on the cruise.
O	The cruise was a disaster.

Table 1: Examples of conflict tuples. **C, G, O** denotes Context, Goal and Obstacle, respectively.

3.1 Conflict Generation

This step establishes the conflict that will propel the narrative forward in the story being created. We view conflict in a story as an obstacle that hinders protagonists from achieving their goals. Employing the notion of a *soft causal threat*, detailed subsequently, we represent a *conflict* as a triple of (`Context`, `Goal`, `Obstacle`), where `Context` denotes a sentence describing a situation that sets the story, `Goal` is a sentence describing protagonist's desired outcome, and `Obstacle` is a threat that lowers the likelihood of achieving the `Goal`.

Goal Inference

Field emphasizes that what the protagonist wants and desires are pivotal in shaping a story's dramatic structure. Following his view, we posit that the protagonist's goal emerges from their 'want' within a specific context. Consequently, we define `Goal` as the state where the character fulfills their 'want'.

We utilize the COMET-ATOMIC-2020 [Hwang *et al.*, 2021][1] in 'beam-5' setting as our commonsense reasoning model to infer what the character 'wants' in a given `Context`. For instance, given the `Context` 'Lana decided she was finally ready to get a pet', the model predicts Lana's want as 'to go to the pet store and buy a pet', using the *xWant* relation from ATOMIC-2020. Consequently, we define `Goal` as 'Lana goes to the pet store and buys a pet.', thus establishing a soft causal link $Context \rightarrow Goal$ based on the inferred 'want' relation by COMET.

Obstacle Generation

In classical planning, a *causal link* $A \xrightarrow{p} B$ denotes a hard causal relation from A to B, formed when event A fulfills precondition p required for B. This causal link can be *threatened* by an event T that has the effect $\neg p$, which would invalidate the condition p if T take place between A and B.

In contrast, Ammanabrolu *et al.* introduced the notion of *soft causal relation* to describe causal relations derived from commonsense reasoning, which does not enforce strict logical causality between actions. Following this framework, we

[1] We use the BART variant provided in https://github.com/allenai/comet-atomic-2020

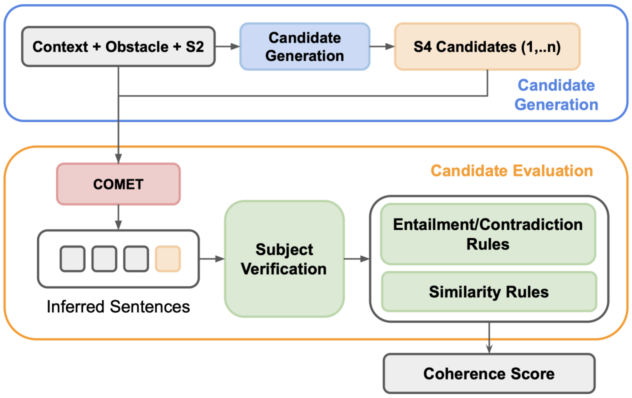

Figure 2: Overview of the Story Completion process for generating the sentence in the fourth position ($S4$), consisting of two stages: candidate generation and evaluation. In the candidate generation stage, n sentences are generated as next-sentence candidates. Then, each candidate is evaluated based on its coherence to identify the subsequent story sentence.

define a *soft causal threat* to denote a potential threat to the *soft causal link* $A \rightarrow B$, reducing the likelihood of B being realized. A *soft causal threat* is represented as a triple of $\langle A \rightarrow B, T \rangle$, where T threatens the *soft causal link* $A \rightarrow B$. The execution of the threatening action T in the sequence does not definitively invalidate the causal link, reflecting the nuanced nature of soft causal relations.

In this paper, we conceptualize conflict as a soft causal link threat, denoted by $\langle Context \rightarrow Goal, Obstacle \rangle$, where the `Obstacle` represents the primary source of conflict, impeding the protagonist from attaining their goal. We define obstacle generation as a task of generating weakeners in defeasible inference, as discussed in [Rudinger *et al.*, 2020]. This reasoning method takes a premise and a hypothesis as inputs and infers evidence that influences the hypothesis's plausibility. Such evidence is called a *weakener* when it makes the hypothesis appear less likely true, and a *strengthener* when it does the opposite. In the context of a *soft causal link* $A \rightarrow B$, defeasible inference identifies a weakener T as a threat that diminishes the probability of the hypothesis B being true by introducing T as an additional evidence to the premise A. Table 1 provides some examples.

Using the `Context` as a premise and the inferred `Goal` as a hypothesis, the defeasible inference model produces a weakener as `Obstacle` that potentially negates the causality between $Context \rightarrow Goal$. For instance, as depicted in Table 1, 'The song is very difficult.' is recognized as an obstacle, which hinders Lana from achieving the inferred goal of 'Lana learn how to play the song.'

We fine-tune the XL variant of GPT2[2] [Radford *et al.*, 2019] with the (Premise,Hypothesis,Weakener) triples in δ-ATOMIC segment of the defeasible-NLI dataset[3] [Rudinger *et al.*, 2020]. This adjustment aimed to generate weakener tokens w in 'beam-5' setting, taking a tuple (p, h) as input, where p, h, w denotes the premise, hypothesis, and weakener tokens, respectively. Then, the trained model is employed as the generative defeasible inference model.

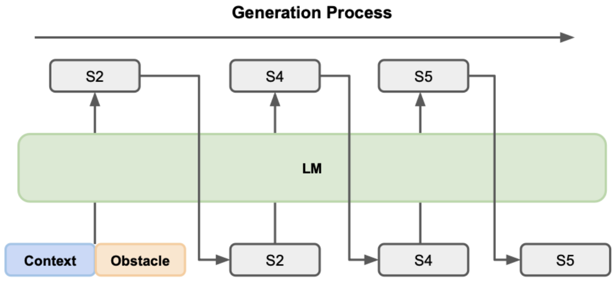

Figure 3: Candidate Generation Process with `Obstacle` introduced as the third sentence. $S2$ denotes the sentence in the second position. Initially, Context and Obstacle generate $S2$ candidates. Then, $S2$ is added to the inputs to produce candidates of $S4$. Finally, $S5$ is generated with `Context,Obstacle,S2,S4` as input.

3.2 Story Completion

The story completion phase extends the narrative by incorporating additional sentences based on the the provided `Context` and `Obstacle` sentences. This approach is derived from the Commonsense-inference Augmented neural Storytelling (CAST) method [Peng *et al.*, 2021], strategically ensuring narrative coherence. It operates by initially generating candidates for the subsequent sentence and then assessing these candidates against commonsense inference criteria to maintain consistency throughout the story.

As depicted in Figure 2, this phase is divided into generation and evaluation phases. In the generation process, a model takes the partially completed story to produce a set of n candidate sentences. Subsequently, in the evaluation phase, the process assesses these sentences to identify and select the one that best aligns with the context of the existing story.

Candidate Generation Model

To develop the candidate generation model, we fine-tune the 124M variant of GPT2[4] using two distinct datasets. Initially, the model is trained using the ConceptNet [Speer *et al.*, 2016] and ATOMIC [Sap *et al.*, 2019] datasets, integrating commonsense knowledge into the system. Then, the model is trained using the ROCStories dataset [Mostafazadeh *et al.*, 2016], which comprises five-sentence stories depicting everyday human experiences and behaviors, to capture narrative structures.

To enable the model to construct sentences in a non-linear order, as outlined in the **Candidate Generation** section, we employ weighted sampling during the training phase. This approach statistically ascertains the placement of the `Conflict` sentence based on observed probabilities[5]. For instance, if the third sentence is identified as the `Conflict` position via sampling, the model is trained to generate a se-

[2] https://huggingface.co/gpt2-xl

[3] https://github.com/rudinger/defeasible-nli

[4] https://huggingface.co/gpt2

[5] Probability weights of each position being the `Conflict` was determined with statistics observed in Appendix C

quence such as `S1,S3,S2,S4,S5`, learning to reorder narrative elements around the conflict.

Candidate Generation

In this phase, the candidate generation model produces five potential sentences as candidates for the next addition to the narrative, given a partially completed story. This step evaluates multiple options for advancing the narrative in a way that preserves the story's coherence.

For instance, to generate candidates for $S5$, as shown in Figure 3, the candidate generation model takes `Context,Obstacle,S2,`$S4$ as input. To guarantee a diverse range of possibilities among these candidates, we employed Diverse Beam Search decoding [Vijayakumar *et al.*, 2016], a technique designed to enhance variety in the generated groups of candidates.

Candidate Evaluation

In the Candidate Evaluation phase, the selection process examines the generated sentence candidates to determine the most appropriate addition to the story in terms of coherency. We employ the COMET model [Bosselut *et al.*, 2019] as a proxy for readers, leveraging nine relation types[6] to collect commonsense insights anticipated by readers during their interaction with the story. Then, relational sentences separately inferred for `Context`, `Obstacle`, and the sentence preceding the candidate are evaluated against those deduced from the candidate. We infer the relational sentences using 'beam-m' setting to enhance their diversity. This means comparing a total of 3 × (9 × m) sentences are compared against the (9 × m) sentences derived from the candidate sentences. From this comparison, a coherence score is calculated.

If this score exceeds a predefined threshold, the corresponding candidate sentence is selected, advancing the narrative to the subsequent sentence generation phase. Conversely, if no candidate achieves the threshold, the one with the highest coherence score among n candidates is chosen. The specific method employed to compute the coherence score is detailed in the next subsection.

3.3 Calculating Coherence Score

This section outlines the process to calculate the coherence score between pairs of sentences.

Assessing Subject Transition across Sentences

The first step analyzes subject transitions, contributing to the overall understanding of sentence coherence assessment in the story. Initially, the NeuralCoref library[7] is employed to extract word clusters referring to the same entity across all sentences. Then, spaCy's[8] dependency parsing feature identifies the word acting as the subject in each sentence, to validates whether it belongs to a different entity cluster compared to the previous sentence.

	Proximity with Candidate Sentence				
Compared Sentence	**± 0**	**± 1**	**± 2**	**± 3**	**± 4**
`Context` `Obstacle`	1	0.775	0.55	0.325	0.1
Preceding Sentence	0.6				

Table 2: Positional Weight. $\pm x$ denotes the proximity between the candidate and compared sentence.

Sentence Proximity in Coherence Scoring

When determining the coherence score, we account for the proximity between the candidate and compared sentences, as closer proximity suggests a stronger causal or inferred relationship. Hence, the second step assigns greater importance to closer sentences by introducing positional weight, as illustrated in Table 2.

Similarity-Based Rules

This process begins with assessing the semantic similarity between a preceding sentence P and a candidate sentence C using the CAST method [Peng *et al.*, 2021], to maintain narrative coherence and thematic continuity.

We evaluate P and C as semantically similar if the sum of the similarities between specific relation sentences of P and C exceeds the threshold value. If the subject changes between P and C, we leverage beam search to find m sentences that link P's oWant and C's xIntent (oWant $\rightarrow$ xIntent), P's oEffect and C's xNeed (oEffect $\rightarrow$ xNeed), and P's oReact and C's xAttr (oReact $\rightarrow$ xAttr). In case the subject remains the same, we adjust the relationship types accordingly; P's oWant, oEffect, and oReact are replaced with xWant, xEffect, and xReact, respectively (xWant $\rightarrow$ xIntent, xEffect $\rightarrow$ xNeed, and xReact $\rightarrow$ xAttr).

Subsequently, we utilize Sentence-BERT [Reimers and Gurevych, 2019] to obtain embedding vectors for each pair of relationship sentences, calculating cosine similarity for all combinations. If the cosine similarity exceeds a pre-set threshold (i.e., 0.5 in our study), the sentences are considered similar and assigned a value of 1; otherwise, a value of 0 is assigned. The average similarity value is then computed for calculating the coherence score.

Implication/Contradiction-Based Rules

We utilize rules based on implication and contradiction to analyze the relationship between two sentences, using Natural Language Inference (NLI) classification through a distill-RoBERTa-based model[9]. This approach is crucial for determining whether sentences support or conflict with each other, thereby maintaining or harming the story's coherence.

We derived the rules using ROCStories as follows. First in each story, sentence pairs are identified while preserving their chronological order. Initially, pairs with sentences that are identified as neutral were entirely excluded, as their context was considered incompatible. Then, utilizing the COMET model, we extract relational sentences from each sentence within these pairs.

[6] oEffect, oReact, oWant, xAttr, xEffect, xIntent, xNeed, xReact, xWant

[7] https://github.com/huggingface/neuralcoref

[8] https://github.com/explosion/spaCy

[9] https://huggingface.co/sentence-transformers/nli-distilroberta-base-v2

		Relation Type			
		Changed Subject Character		Same Subject Character	
Type of Comparing Sentence	**Rule**	**Comparing Sentence**	**Candidate Sentence**	**Comparing Sentence**	**Candidate Sentence**
Context	**Implication**	oReact	oReact	xReact	oReact
		xAttr	xAttr	-	-
		xReact	xReact	oReact	xReact
Preceding Sentence	**Similar**	xEffect	xNeed	oEffect	xNeed
		xReact	xAttr	oReact	xAttr
		xWant	xIntent	oWant	xIntent
	Implication	oReact	oReact	xReact	oReact
		xAttr	xAttr	-	-
		xReact	xReact	oReact	xReact
Obstacle Sentence that Appears Earlier	**Strong Implication**	oReact	oReact	xReact	oReact
		xAttr	xAttr	-	-
		xReact	xReact	oReact	xReact
	Implication	oEffect	oEffect	xEffect	oEffect
		xEffect	xEffect	oEffect	xEffect
		xIntent	xIntent	-	-
		xNeed	xNeed	-	-
Obstacle Sentence that Appears Later	**Implication**	xEffect	xIntent	oEffect	xIntent
		xReact	xIntent	oReact	xIntent
	Contradiction	oWant	xWant	xWant	xWant
		xWant	oWant	oWant	oWant

Table 3: Overview of Similarity-Based and Implication/Contradiction Rules

These relational sentences are paired in all possible combinations and analyzed using the NLI classification model to determine if they imply or contradict each other. Sentences were classified as implying or contradicting if such classifications constituted over 30% of their evaluations. A classification was deemed strong if the proportion of implication or contradiction exceeded 50%. However, if both implication and contradiction rates surpassed 30%, indicating that the context could potentially accommodate any sentence, this relationship was disregarded.

Finally, to guarantee that generated candidate sentences achieve a standard of coherence, the rule set we defined exclusively incorporates relation types of implication or contradiction. Specific rules have been formulated to best suit the type of comparing sentence. For comparing sentence types `Context` and Preceding Sentence, only the pairs that meet the conditions for these sentence types were used. As for when comparing with `Obstacle`, the position of the `Obstacle` sentence was determined statistically according to the statistics in Appendix C, and its appearance before or after the candidate position was separately analyzed.

Computing Coherency Score

The coherency score is calculated through the application of both similarity-based and implication/contradiction-based rules as shown in Table 3. To compute the coherence score, we first determine the fulfillment rates of each rule by comparing the candidate sentence against three specific sentences: `Context`, `Obstacle`, and the one immediately preceding the candidate. For instance, if there are three implication/contradiction-based rules identified between the relational sentences derived from the candidate and those derived from `Context`, and two of these rules are found to satisfy the criteria for implication, the resulting fulfillment rate would be approximately 66.7%

For each candidate sentence, the fulfillment rate is adjusted by a weighting factor k_* to adjust its influence on the overall coherence score[10]. These rates are then further scaled by a normalized positional weight, ensuring that comparisons to the threshold are consistent, irrespective of the sentence's position within the narrative. The coherence score for a candidate sentence is the sum of these adjusted rates. If this score exceeds a predetermined threshold, we consider that the candidate sentence meets the criteria for maintaining narrative coherence.

$$C(S_c,S_o,S_p,S_x)=\sum_{i\in\{c,o,p\}} k_i \cdot w_i \times ICsat(R_i,R_x)+ k_p \cdot w_p \times SIMsat(R_p,R_x) \quad (1)$$

The coherency score is computed using Equation 1, where S_c, S_o, S_p, S_x denote the `Context` (c), `Obstacle` (o), Preceding sentence (p), and Candidate sentence (x) respectively. The weights k_* each correspond to each sentence type's significance in the narrative structure, while w_* are the normalized positional weights. R_* represent the sets of embedding vectors for the relationship sentences extracted by the COMET model for the context, conflict, last, and candidate sentences, respectively.

The function $ICsat(R_a, R_b)$ quantifies the degree to which the implication/contradiction rules are satisfied between two sets of sentences R_a and R_b, indicating relational coherence or discord. Similarly, $SIMsat(R_a, R_b)$ measures the fulfillment of the similarity rule between the sets of sentences R_a and R_b, assessing their semantic closeness. This formulation allows for a comprehensive evaluation of narrative coherence, incorporating both logical consistency and thematic similarity.

4 Evaluation

This section outlines two experiments we conducted to evaluate the CNGCI framework. The initial experiment evaluates the effectiveness of our conflict generation process, while

[10] k_c=1, k_o=0.5, k_p=0.5 were used for this paper

	Agree	Disagree	Other	κ
Obstruction	**64.7%**	26.5%	8.8%	0.46
Topic	**86.8%**	8.8%	4.4%	0.17
Logicality	**72.1%**	22.1%	5.9%	0.25

Table 4: Results of Human Evaluation for `Conflict` Generation.

the subsequent experiment investigates the story completion mechanism.

4.1 Conflict Generation Evaluation

The evaluation of the conflict generation stage involves a human study that assesses the quality of generated `Goal` and `Obstacle` sentences. We recruited 33 participants for our study via the Amazon Mechanical Turk (AMT) platform platform[11]. Participants were presented with a series of questions to evaluate the following aspects of `Obstacle`:

- **Obstruction:** Does the generated `Obstacle` introduce any element or obstacles that STOPS the subject of `Context` from being in a more satisfactory state in `Goal`?
- **Topic:** Is the TOPIC of `Obstacle` coherent with `Context` and `Goal`?
- **Logicality:** Does `Obstacle` make LOGICAL SENSE to occur following `Context`?

The participants were asked to evaluate each aspect on a 5-point Likert scale. The ratings are then classified into three categories based on majority voting: 'Agree', 'Disagree', and 'Other'. We aggregate 'Strongly Agree' and 'Agree' into a single 'Agree' category, and similarly combine 'Strongly Disagree' and 'Disagree' into 'Disagree'. The 'Other' category refers to instances of equal votes between between these two categories or a majority of 'Not Sure' responses. The survey utilized 68 pairs of (`Context`, `Goal`) validated as correctly inferred through a separate human survey, detailed in Appendix A.

Table 4 shows the percentages of samples voted for each category with the inter-annotator agreement measured by Fleiss' Kappa [Fleiss, 1971]. For the obstruction, topic, and logicality aspects of the obstacle, 26 out of 68 samples received a majority vote of Agree. This suggests that the majority of participants perceived the conflict generated by our framework as obstructive, relevant to the context, and logically consistent within the narrative.

4.2 Story Completion Evaluation

The second experiment examines the story completion stage using the Conflict triples validated in Section 4.1. We compare CNGCI against two baselines: Knowledge-enhanced GPT2 and C2PO, examining their performance across the dimensions of conflict, interestingness, and coherency.

Knowledge-enhanced GPT2 [Guan *et al.*, 2019] is a GPT2-small model fine-tuned with the ATOMIC and ROCStories datasets. We selected this model as a baseline since it is trained on both commonsense knowledge and story continuations, which aligns with our candidate generation model.

Context
Lana was trying to figure out how to play a song.

Knowledge-enhanded GPT2:
she knew there was a song in the middle of the song.
she wasn't sure she was going to be able to sing it.
she was able to sing the song.

C2PO:
Lana tries to practice.
Lana starts to show off.
Lana begins to practice.
Lana learn how to play a song.

CNGCI ($n = 5, m = 5$):
Lana was trying to figure out how to play a song.
For some reason, she couldn't figure out how to play the song.
The song is very difficult.
Finally, she decided to ask her friend for help.
She ended up learning how to play the song.

Table 5: Examples generated by Knowledge-enhanced GPT2, C2PO, and CNGCI. Given Context, each method generates the remaining four sentences.

C2PO [Ammanabrolu *et al.*, 2021] is a system that leverages commonsense inference to fill in the gap between two given events.Whereas our system specifically focuses on infilling the Soft Causal Link ($Context \rightarrow Goal$) with an explicit Obstacle excluding the Goal itself, stories generated using C2PO fill the Soft Causal Link without an explicit Obstacle, while including the Goal statement.

Table 5 showcases sample stories generated by each model using identical Context sentences. Notably, we observe that the story produced by Knowledge-enhanced GPT2 contains some logical gaps between sentences. Interestingly, we also notice the establishment of conflict in the form of 'Lana initially expressing doubts but then proceeding to learn how to play the song', even though the model is not explicitly trained to incorporate conflict. This suggests that the model may have been implicitly trained to generate conflict, a characteristic observed in the ROCStory data, but it does not guarantee the consistent generation of conflict-containing narratives in every instance.

In contrast, C2PO generated coherent stories without conflicts, but its narrative sentences tend to be simplistic due to its reliance on COMET for sentence generation. Conversely, the CNGCI framework seamlessly integrated the provided obstacle into the narrative without compromising coherence.

Human Evaluation

To conduct the story completion evaluation, we recruited 41 participants through the AMT platform to assess 38 story pairs. These pairs were split into two groups: 19 pairs from CNGCI versus Guan and 19 pairs from CNGCI versus C2PO, with each model generating 19 stories respectively. To facilitate a more concise reference, the Knowledge-enhanced GPT2 is denoted as 'Guan'. The participants were asked the following questions.

- **Conflict:** Which story contains sentences with conflict?

[11] The details of AMT survey process is provided in Appendix B

- **Interestingness:** Which story was more interesting?
- **Logic Coherency:** Which story had a coherent flow between sentences?
- **Topic Consistency:** Which story exhibited overall consistency in theme?

	CNGCI	Tie	Guan
Conflict	**63.2%**	0%	36.8%
Interestingness	**57.9%**	0%	42.1%
Logic Coherency	42.1%	10.5%	**47.4%**
Topic Consistency	**42.1%**	21.1%	36.8%

Table 6: Win Rate for CNGCI vs. Guan

	CNGCI	Tie	C2PO
Conflict	**84.2%**	1.05%	5.3%
Interestingness	**73.7%**	0%	26.3%
Logic Coherency	**63.2%**	0%	36.8%
Topic Consistency	**78.9%**	0%	21.1%

Table 7: Win Rate for CNGCI vs. C2PO

Table 6 and Table 7 present the results of human evaluation. CNGCI consistently received favorable assessments compared to C2PO across all metrics. While CNGCI outperformed Guan in overall performance, it demonstrated relative weakness in terms of logical coherence. This gap can be attributed to their contrasting objectives. Guan aims to produce narratives that faithfully follow general commonsense-based event flow commonly seen in ROCStories. In contrast, CNGCI primarily focuses on creating conflicts that disrupt this flow which may potentially lead to logical inconsistencies.

5 Conclusion

This paper introduces CNGCI, a neuro-symbolic framework designed for generating coherent stories that embody conflict. Additionally, we propose defining conflict through the concept of a soft causal threat, wherein conflict emerges from an obstacle that reduces the likelihood of the protagonist achieving their goal. A comparative human evaluation against two baseline systems indicates that our framework can generate conflict-embedded stories while maintaining coherency and interestingness.

In the proposed framework, we used the GPT2 model to compare our results against baselines using language models of similar capacity. Although this is preliminary work to test the concept's efficacy, we aim to enhance future models by incorporating the zero-shot reasoning capabilities of LLMs, avoiding the need for manually defined implication and contradiction rules.

A Goal Evaluation

Following questions was designed to evaluate the following three aspects of generated `Goal`.

	Agree	Disagree	Other	κ
Satisfactory	**77.0%**	6.0%	17.0%	0.12
Topic	**87.0%**	5.0%	8.0%	0.35
Logicality	**81.0%**	10.0%	9.0%	0.39

Table 8: Percentage of samples for goal evaluation

- **Satisfactory:** Is the subject character of *Context* in a more SATISFACTORY state in *Goal*?
- **Topic:** Is the TOPIC of *Goal* coherent with the topic of *Context*?
- **Logicality:** Does *Goal* make LOGICAL SENSE to occur following *Context*?

Agree, Disagree, Other categorization follows the process described in Section 4.1 Table 8 reports the percentage of majority votings for the 100 goals and the Fleiss' Kappa.

B Crowdsourcing Evaluation

Our human studies were conducting an Amazon Mechanical Turk (AMT). The data collection process was approved by the authors' Institutional Review Board (IRB), and we obtained consent from the workers to use the annotated results for research purposes. We ensured privacy protection, as no personal information was collected. For goal evaluation, 29 annotators received $0.4 per sample, averaging $12 per hour over 2 minutes per task. In obstacle evaluation, 33 annotators earned $0.7 per sample, with an average of $12 per hour for 3.5 minutes of work per task.

C Identifying the Placement of Conflict Sentences

We assume that conflict elements which hinder the protagonist's goal will result in readers empathizing with the character. Based on this assumption, we analyzed readers' emotional changes for ROCStories, which was validated by [Hyun, 2022]. By identifying the position where emotions are first reversed as the conflict position, the probability of conflict elements appearing in the second position was approximately 61.42%, with third being 25.94%, and fourth 12.65%.

Acknowledgments

This work was partly supported by NCSOFT, National Research Foundation of Korea (NRF) grant funded by the Korea government (MIST) (No.RS-2024-00357849), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00421, Artificial Intelligence Graduate School Program (Sungkyunkwan University)).

Contribution Statement

Youngrok Song and Gunhee Cho contributed equally to this work. Contributions of Youngrok Song and Hyunju Kim were made while at Sungkyunkwan University. Yun-Gyung Cheong is the corresponding author.

References

[Ammanabrolu *et al.*, 2021] Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. Automated storytelling via causal, commonsense plot ordering. In *AAAI*, 2021.

[Barber and Kudenko, 2008] Heather Barber and Daniel Kudenko. Generation of dilemma-based interactive narratives with a changeable story goal. In *Proceedings of the 2nd International Conference on INtelligent TEchnologies for Interactive EnterTAINment*, INTETAIN '08, Brussels, BEL, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[Bosselut *et al.*, 2019] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. *ArXiv*, abs/1906.05317, 2019.

[Cho *et al.*, 2022] JinUk Cho, MinSu Jeong, JinYeong Bak, and Yun-Gyung Cheong. Genre-controllable story generation via supervised contrastive learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2839–2849, New York, NY, USA, 2022. Association for Computing Machinery.

[Field, 2005] Syd Field. *Screenplay: The Foundations of Screenwriting*. A Delta book. Delta Trade Paperbacks, 2005.

[Fleiss, 1971] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

[Franceschelli and Musolesi, 2023] Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. *ArXiv*, abs/2304.00008, 2023.

[Gervás *et al.*, 2023] Pablo Gervás, Gonzalo Méndez, and Eugenio Concepción. Evolutionary combination of connected event schemas into meaningful plots. *Genetic Programming and Evolvable Machines*, 24:1–38, 2023.

[Gervás, 2009] Pablo Gervás. Computational approaches to storytelling and creativity. *AI Mag.*, 30:49–62, 2009.

[Guan *et al.*, 2019] Jian Guan, Yansen Wang, and Minlie Huang. Story ending generation with incremental encoding and commonsense knowledge. In *AAAI*, 2019.

[Holtzmann, 2016] Petra Holtzmann. Routledge encyclopedia of narrative theory. 2016.

[Hwang *et al.*, 2021] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, 2021.

[Hyun, 2022] Jiwung Hyun. Extracting protagonist emotional arcs using deeplearning based commonsense reasoning model. *Sungkyunkwan Univ. Master's thesis*, 2022.

[Mateas and Stern, 2003] Michael Mateas and Andrew Stern. Façade: An experiment in building a fully-realized interactive drama. In *Game Developers Conference (GDC'03)*, 2003.

[Mostafazadeh *et al.*, 2016] N. Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*, 2016.

[Peng *et al.*, 2021] Xiangyu Peng, Siyan Li, Sarah Wiegreffe, and Mark O. Riedl. Inferring the reader: Guiding automated story generation with commonsense reasoning. *ArXiv*, abs/2105.01311, 2021.

[Porteous and Cavazza, 2009] Julie Porteous and Marc Cavazza. Controlling narrative generation with planning trajectories: The role of constraints. In Ido A. Iurgel, Nelson Zagalo, and Paolo Petta, editors, *Interactive Storytelling*, pages 234–245, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[Pérez and Sharples, 2001] Rafael Pérez and Mike Sharples. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(2):119–139, 2001.

[Radford *et al.*, 2019] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[Rudinger *et al.*, 2020] Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

[Sap *et al.*, 2019] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. *ArXiv*, abs/1811.00146, 2019.

[Sgouros, 1999] Nikitas M. Sgouros. Dynamic generation, management and resolution of interactive plots. *Artificial Intelligence*, 107(1):29–62, 1999.

[Song *et al.*, 2020] Youngrok Song, Hyunju Kim, Taewoo Yoo, Byung-Chull Bae, and Yun-Gyung Cheong. An intelligent storytelling system for narrative conflict generation and resolution. *2020 IEEE Conference on Games (CoG)*, pages 192–197, 2020.

[Speer *et al.*, 2016] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *ArXiv*, abs/1612.03975, 2016.

[Szilas, 2003] Nicolas Szilas. Idtension: a narrative engine for interactive drama. In Gobel et al., editor, *Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference*, pages 187–203, 2003.

[Vijayakumar *et al.*, 2016] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan

Lee, David J. Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *ArXiv*, abs/1610.02424, 2016.

[Ware *et al.*, 2014] Stephen G. Ware, R. Michael Young, Brent Harrison, and David L. Roberts. A computational model of narrative conflict at the fabula level. *IEEE Transactions on Computational Intelligence and Artificial Intelligence in Games*, 6(3):271–288, 2014.

[Yang *et al.*, 2022] Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In *EMNLP*, 2022.