LITERATURE REVIEW

Dungeons & AI - An AI Tribute to Dungeons & Dragons

Anuvind M P[1], **R S Harish Kumar**[2] **and Harishankar Binu Nair**[3]

[1] AM.EN.U4AIE22010
[2] AM.EN.U4AIE22042
[3] AM.EN.U4AIE22023

1 Detailed Review of Survey Papers

1.1 Wang et al.(2023) : "Open-world story generation with structured knowledge enhancement: A comprehensive survey"

Wang et al.[11] provide a broad survey of the field of structured knowledge-enhanced story generation (SKESG), focusing on how integrating external, structured knowledge sources (like knowledge graphs) can improve the coherence, relevance, and diversity of automatically generated stories.

1.1.1 Methodologies Discussed

The paper organizes existing research into a taxonomy based on how structured knowledge is integrated:

1. Structured Knowledge as Text: This approach transforms structured knowledge (like triples from ConceptNet or ATOMIC) into natural language and feeds it into language models. Some methods, like KEG and COEP, finetune PLMs on this transformed knowledge. Others, such as COMET and C2PO, build separate reasoning modules that predict logical or causal events to guide story progression.

2. Structured Knowledge as Encoding: Instead of converting knowledge into text, this strategy encodes knowledge graphs using models like Graph Neural Networks. The resulting embeddings are then used to influence generation—either by injecting them directly into models or by modifying word selection probabilities to keep generation contextually grounded and logically coherent.

1.1.2 Key Findings

Models focus on tasks like sentence-prompted generation, ending prediction, and plot infilling. Injecting commonsense knowledge (from ConceptNet, ATOMIC) improves coherence and character logic. Transformers like GPT-2 and BART are commonly used, often with fine-tuning. Models like KEG, COEP, CAST, and MKR show notable gains in logical flow and persona alignment.

1.1.3 Datasets Mentioned

ConceptNet, ATOMIC (commonsense); WordNet, VerbNet (semantic); ROCStories, WritingPrompts (story corpora).

1.1.4 Relevance

Provides a clear framework for using structured knowledge to enhance narrative quality and coherence in story generation.

1.2 Xie et al.(2023) "The Next Chapter: A Study of Large Language Models in Storytelling"

Xie et al.[12] investigate the storytelling capabilities of large language models (LLMs), with a primary focus on GPT-3. They perform a comparative analysis of GPT-3 against several state-of-the-art story generation models using both automatic metrics and human evaluations.

1.2.1 Methodologies Discussed

The core methodology is prompt-based learning with LLMs. Instead of finetuning, the model is guided using a few-shot learning setup where a few example stories are provided in the prompt. The performance of GPT-3 is compared against.KGGPT2, HINT, PROGEN, MTCL, Fine-tuned BART – as a baseline.

1.2.2 Key Findings

GPT-3 outperforms all other models in human evaluations on fluency, coherence, logicality, and engagement. Its stories often rival or exceed human-written ones on short and medium-length tasks, although it tends to generate shorter outputs and sometimes leans on memorized patterns, especially in news-like scenarios.

Vol. , **No.** , 2025.

1.2.3 *Datasets Mentioned*

ROCStories, WritingPrompts (story corpora).

1.2.4 *Relevance*

This paper highlights the strengths of prompt-based LLMs for zero-shot or few-shot storytelling, particularly in comparison to complex fine-tuned models. It emphasizes LLMs' capability to generate human-like stories, raising both opportunities for simplified story generation pipelines and concerns about originality and content control.

1.2.5 *Challenges and Disadvantages Mentioned*

1. Plagiarism and Memorization: GPT-3 often mirrors key phrases or plot structures from training data, raising concerns about originality.

2. Lack of Content Control: Minimal control over output style, structure, or themes makes it hard to tailor stories to specific needs.

3. Inconsistent Output Length: GPT-3 tends to generate shorter outputs, especially for long-form tasks, falling short of desired word counts.

4. High Computational Cost: The model's large size (175B parameters) makes it resource-intensive and expensive to deploy.

5. Difficulty with Long-Term Dependencies: GPT-3 struggles to maintain coherence in long narratives or complex subplots.

1.3 AlHussain and Azmi (2021) "Automatic Story Generation: A Survey of Approaches"

AlHussain and Azmi [2] provide a comprehensive survey of automatic story generation approaches, covering classical symbolic models and modern neural methods. This work situates automatic storytelling at the intersection of AI and psychology and organizes the landscape around structural, planning-based, and machine learning (ML) paradigms. The survey also identifies key challenges, resources, and evaluation metrics, offering a foundational guide for researchers in the field.

1.3.1 *Methodologies Discussed*

The paper categorizes story generation approaches into three major paradigms:

- *Structural Models:* Rely on fixed schemas or grammars (e.g., Propp's morphology, plot graphs, or story grammars) to generate story structure using manually crafted rules and templates.
- *Planning-Based Models:* Frame storytelling as a planning problem. Agents pursue character or author-defined goals. Notable models include TALE-SPIN, AUTHOR, FABULIST, and CPOCL.
- *Machine Learning Models:* Leverage statistical and neural techniques to learn event transitions and complete stories. These include RNNs, LSTMs, Seq2Seq, and reinforcement learning. Models use abstract representations like tuples and dependency chains.

The survey further explores analogy-based and heuristic search methods, including systems like MINSTREL and MEXICA.

1.3.2 *Key Findings*

The paper highlights a shift from rule-based systems to neural models, emphasizing trade-offs between structure, coherence, and creativity. Structural and planning-based models excel at story control but struggle with diversity and believability. In contrast, ML-based models generate more varied and human-like stories but often lack global coherence and narrative control.

1.3.3 *Datasets Mentioned*

Key datasets include ROCStories, WritingPrompts, Children's Book Test, VIST, and STORIUM. The paper also reviews corpora annotated with causal and temporal relationships (e.g., CaTeRS, CATENA, and Causal-TimeBank).

1.3.4 *Relevance*

This survey provides an essential roadmap for developing new storytelling systems, emphasizing the need for hybrid models that combine symbolic reasoning with neural generation. It also calls for reusable knowledge bases and richer evaluations that align with human narrative understanding.

2 Detailed Review of Conference/Journal Articles

2.1 Alabdulkarim et al. (2021) "Goal-Directed Story Generation: Augmenting Generative Language Models with Reinforcement Learning"

Alabdulkarim et al. [1] propose a reinforcement learning framework to steer GPT-2 toward generating stories that achieve predefined goal events. Their approach combines a knowledge-graph-guided policy network (KG-DQN) with reward shaping to ensure logical plot progression while maintaining fluency.

2.1.1 *Methodologies Discussed*

- KG-DQN: A Deep Q-Network that selects GPT-2-generated continuations based on a dynamically updated knowledge graph (extracted via OpenIE) and VerbNet-based reward signals.
- Reward Shaping: Clusters verbs by their proximity to goal verbs using Jenks Natural Breaks, rewarding incremental progress toward the goal.
- Multi-Task Training: Combines language modeling with a discriminative task to distinguish coherent stories from shuffled/repeated-sentence variants.

2.1.2 *Key Findings*

- Achieves 98.73% goal success rate for sci-fi stories, outperforming fine-tuned GPT-2 (50% success).
- Human evaluations favor KG-DQN over baselines in coherence (58% preference) and logicality (56.5%).
- Knowledge graphs reduce repetition (REP-4: 4.45% vs. 29.41% for GPT-2) and improve diversity.

2.1.3 *Datasets Mentioned*

ROCStories (sci-fi subset), ConceptNet and ATOMIC for reward computation

2.1.4 *Relevance*

This work bridges the gap between neural generation and symbolic planning, enabling controllable narrative generation without extensive domain engineering. The framework is applicable to interactive storytelling and scenarios requiring strict adherence to plot constraints.

2.2 Simon and Muise (2022) “TattleTale: Storytelling with Planning and Large Language Models”

Simon and Muise [8] propose *TattleTale*, a storytelling system that integrates symbolic planning with Large Language Models (LLMs) to enhance the coherence and believability of generated narratives. Addressing the issues of logical inconsistency and repetition in LLM-generated stories, the authors use a symbolic planner to provide structured guidance to an LLM, resulting in more coherent text generation.

2.2.1 *Methodologies Discussed*

- **Classical Planning Integration**: Symbolic plans are generated using STRIPS and PDDL, modeling narrative structure with predicates and actions.
- **Story-Agnostic Prompting**: Initial context is established using general prompts independent of the specific story content.
- **Stepwise LLM Input**: Actions from the plan are sequentially fed into GPT-J-6B, producing story content line by line.
- **Manual Entity Curation**: Story elements like characters, locations, and objects are manually extracted for precise domain encoding.

2.2.2 *Key Findings*

- Nearly all planned nouns and verbs are reflected in the generated output, ensuring alignment between plan and story.
- Story coherence and consistency improve significantly compared to prompt-only generations.
- Symbolic planning enables controlled variability in stories by allowing multiple valid plans for the same narrative goal.

2.2.3 *Datasets Mentioned*

- Children’s stories such as *The Way Home for Wolf*, *Robin Hood and the Golden Arrow*, and *The Paper Bag Princess*.
- Manually converted into PDDL format by extracting key narrative elements.

2.2.4 *Relevance*

This study presents a hybrid neuro-symbolic approach to story generation, offering improved coherence and narrative control. The method is applicable to domains like explainable AI, educational content, and interactive fiction, where logical structure and believability are essential.

2.3 Li et al. (2024) "From Words to Worlds: Transforming One-line Prompts into Multi-modal Digital Stories with LLM Agents"

Li et al. [5] introduce the StoryAgent framework for transforming a single-line prompt into rich, multi-modal digital stories. Their approach integrates LLM-based communicative agents with generative models across modalities to achieve coherent, interactive, and editable storytelling. The system emphasizes intervention flexibility, long-duration consistency, and scene interactivity—addressing key pain points in automated narrative generation.

2.3.1 *Methodologies Discussed*

- **Story Cluster with LLM Agents:** A team of LLM-based agents handles narrative decomposition, forming a top-down structure (story arc → characters, settings, beats, scenes, and screenplay), each passed through expert-critic pairs for refinement.
- **Hierarchical Story Representation:** Uses intermediate text formats (e.g., JSON) to connect story planning with generative tools for asset creation, allowing plug-and-play extensibility across visual, audio, and textual components.
- **Scene Understanding and Interaction:** Combines semantic segmentation and depth estimation to understand background scenes and enable spatially coherent placement of characters and objects.
- **Flexible Human Intervention:** Supports fine-grained edits to story elements at any stage, preserving upstream consistency while updating downstream components.

2.3.2 *Key Findings*

- Achieves high consistency across modalities (e.g., visuals match audio descriptions, characters maintain visual identity across scenes).
- Allows diverse modifications (e.g., style change, plot shift, character addition) with minimal degradation to narrative coherence.
- Outperforms naive prompting in story length and modular control—doubling average story length and enabling cinematographic control.
- Enables scalable digital storytelling pipelines that democratize content creation without requiring professional-level technical expertise.

2.3.3 *Datasets Mentioned*

No explicit mention of benchmark datasets; rather, the focus is on the framework and its ability to generate high-quality narratives from scratch using text-to-asset pipelines (e.g., CC2D, AudioGen, ElevenLabs, FreeSound, etc.).

2.3.4 *Relevance*

This work is highly relevant for interactive and multi-modal story generation systems. It advances the state-of-the-art by extending narrative generation beyond text into visual, audio, and animation domains while preserving logical and stylistic coherence. For applications like game design, this framework offers a blueprint for building richly orchestrated, context-aware narratives with user-editable elements.

2.4 Zhao et al. (2023) “More human than human: LLM-generated narratives outperform human-LLM interleaved narratives”

Zhao et al. [15] investigate the comparative effectiveness of narratives generated entirely by Large Language Models (LLMs) versus those created through an interleaved approach—alternating between human-written and LLM-generated segments. The study aims to understand user preferences regarding the coherence, plausibility, understandability, and novelty of these two storytelling paradigms.

2.4.1 *Methodologies Discussed*

- **Interleaved vs. Non-Interleaved Narratives**: Stories were created either fully by LLMs (non-interleaved) or in alternating fashion with human sentences (interleaved).
- **Survey-Based Evaluation**: Two large-scale surveys (500 participants each) were conducted:
  - An absolute Likert-scale survey to rate stories on logical flaws, plausibility, understandability, and novelty.
  - A pairwise comparison survey to determine story preference between the two methods.
- **Statistical Analysis**: Used logistic mixed-effects models (lme4 in R) and exact binomial tests to analyze participant responses.

2.4.2 *Key Findings*

- **Non-interleaved stories** received significantly higher ratings in terms of preference, logical coherence, and understandability.
- Interleaved stories were slightly more **novel** in perception, but at the cost of perceived repetition and weaker sentence flow.
- The study disproved the original hypothesis that interleaving improves quality by showing that it often leads to unnatural transitions and redundancy.

2.4.3 *Datasets Mentioned*

- Topics used for story generation: electric vehicles, beaches, and cooking.

- 20 stories were generated—10 interleaved and 10 non-interleaved—for evaluation.

2.4.4 Relevance

This study provides actionable insights into narrative generation strategies using LLMs, suggesting that fully LLM-generated content may outperform hybrid methods in terms of readability and coherence. These findings are relevant for applications in entertainment, educational content, and automated creative writing systems.

2.5 Xu et al. (2020) "MEGATRON-CNTRL: Controllable Story Generation with External Knowledge Using Large-Scale Language Models"

Xu et al. [13] introduce MEGATRON-CNTRL, a framework that enables controllable story generation by combining large-scale pretrained language models with dynamically retrieved external knowledge. Their approach addresses the limitations of typical LLMs—namely, lack of controllability and insufficient incorporation of factual knowledge—by integrating keyword-based planning and knowledge-based guidance into the story generation process.

2.5.1 Methodologies Discussed

The proposed framework consists of four core components: (1) a keyword predictor that forecasts important keywords for the next sentence based on current story context; (2) a knowledge retriever that uses these keywords to pull related knowledge triples from ConceptNet, transforming them into natural language sentences; (3) a contextual knowledge ranker trained with weak supervision using sentence embeddings to filter the most relevant knowledge; and (4) a conditional generator (GPT-2) that uses the filtered knowledge and context to produce the next story sentence. This process is repeated in a dynamic, step-by-step manner, enabling fine-grained control over generation. The model also allows manual intervention by replacing predicted keywords with custom ones.

2.5.2 Key Findings

Experiments conducted on the ROC story dataset demonstrate that MEGATRON-CNTRL significantly outperforms previous state-of-the-art methods in terms of coherence, consistency, and diversity. Human evaluations show an increase in logical consistency from 74.5% (for the 124M parameter version) to 93.0% (for the 8B version). The model also exhibits strong controllability: up to 91.5% of stories change meaning when keywords are replaced with their antonyms.

2.5.3 Datasets Mentioned

ROC Story Dataset, ConceptNet.

2.5.4 Relevance

This work is highly relevant for enhancing the controllability and factual grounding of LLM-based story generation. By leveraging dynamically ranked external knowledge and keyword-level control, MEGATRON-CNTRL offers a scalable solution to generate more coherent, diverse, and user-guided narratives—complementing symbolic methods like ASP with a fully neural, knowledge-augmented pipeline.

2.6 Venkatraman et al. (2025) "CollabStory: Multi-LLM Collaborative Story Generation and Authorship Analysis"

Venkatraman et al. [9] present CollabStory, the first dataset of stories collaboratively generated by multiple LLMs (up to 5 authors), addressing challenges in authorship attribution and verification in machine-machine co-writing scenarios. The work highlights the risks of LLM-LLM collaboration in evading plagiarism detection and proposes extensions of PAN authorship tasks to multi-LLM settings.

2.6.1 Methodologies Discussed

The authors generate 32,503 stories using five open-source LLMs (Llama, Mistral, Gemma, Orca, Olmo) in single- to multi-author configurations. Stories are created by iteratively prompting LLMs with summaries of prior segments and templates to continue or conclude narratives. The dataset supports four authorship-related tasks: (1) detecting multi-author involvement, (2) predicting author count, (3) verifying authorship at sentence boundaries, and (4) attributing text spans to specific LLMs. Baselines include BERT, RoBERTa, and traditional classifiers.

2.6.2 Key Findings

- GPT-4-based continuity evaluations show 75% accuracy in distinguishing coherent multi-LLM story continuations.
- Increasing the number of collaborating LLMs reduces attribution accuracy due to stylistic blending.

2.6.3 Datasets Mentioned

- CollabStory (32,503 stories, 5 LLMs, 725 avg. words/story)

- Comparison with human-human datasets: STORIUM, CoAuthor, StoryWars

2.6.4 Relevance

CollabStory offers useful insights especially in handling multi-LLM collaboration, maintaining narrative coherence, and evaluating authorship influence. Its methodologies and dataset structure can guide the design of controlled, multi-agent storytelling systems.

2.7 Guan et al. (2020) "A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation"

Guan et al. [4] enhance GPT-2 with commonsense knowledge from ConceptNet and ATOMIC, addressing issues of repetition and logical inconsistency in open-ended story generation. Their model employs multi-task learning to improve causal and temporal coherence.

2.7.1 Methodologies Discussed

- Knowledge Post-Training: GPT-2 is fine-tuned on 1.17M knowledge triples converted to natural language via templates (e.g., "X causes Y").
- Multi-Task Objective: Combines language modeling with a classification task to discriminate between real stories and auto-constructed fake ones (shuffled/replaced sentences).
- Dynamic Knowledge Coverage: Measures alignment between generated stories and ConceptNet/ATOMIC triples during evaluation.

2.7.2 Key Findings

- Reduces confusion by 12% compared to vanilla GPT-2 while doubling knowledge coverage (18.48 vs. 8.04 triples/story).
- Human evaluators prefer the model in 57% of cases for logicality, citing fewer contradictions (e.g., "stopped driving then drove to party").
- Classification auxiliary task reduces repetition.

2.7.3 Datasets Mentioned

ROCStories (98k stories), ConceptNet (600k triples) and ATOMIC (574k triples)

2.7.4 Relevance

This work demonstrates how external knowledge infusion and discriminative training can mitigate the limitations of pure autoregressive models. It provides a blueprint for integrating commonsense reasoning into LLMs for applications requiring narrative logic, such interactive fiction.

2.8 Feng et al. (2024) "SS-GEN: A Social Story Generation Framework with Large Language Models"

Feng et al. [3] present SS-GEN, a novel framework for generating narratives designed to assist autistic children in understanding and navigating social situations. They demonstrate that fine-tuned smaller models can generate structured, safe, and descriptive social narratives comparable to GPT-4o at a significantly reduced cost.

2.8.1 Methodologies Discussed

- **STARSOW Strategy:** A breadth-first, hierarchical prompting scheme inspired by tree structures to guide LLMs (e.g., GPT-4o) in generating chapters, titles, and story content from a curated seed set. This involves multiple layers:
  - *Taking Root:* Generation of diverse story themes.
  - *Branching Out:* Title generation per theme to cover broader goals.
  - *Bearing Star Fruits:* Story generation under strict constraints.
  - *Gardening Work:* Human filtering using ROUGE-L thresholds and expert review.
- **Fine-Tuning of Smaller Models:** Efficient adaptation of Mistral, LLaMA, and Gemma (2B–8B) via LoRA-based PEFT using the curated dataset. Both zero-shot and supervised fine-tuning settings are explored.

2.8.2 Key Findings

- SS-GEN generated over 5,000 Social Stories across 57 chapters using GPT-4o, filtered to meet strict narrative constraints.
- Fine-tuned 2B–8B models (e.g., Gemma 7B) achieved significant performance boosts: e.g., BLEU-4 improved by 42.32% and ROUGE-L by 20.07% for Gemma 2B.
- GPT-4 and expert evaluations confirmed higher coherence, descriptiveness, empathy, and adherence to safety in fine-tuned models vs. zero-shot.
- STARSOW significantly reduced reliance on expensive API calls while maintaining quality and

diversity.

2.8.3 Datasets Mentioned

- **Seed Set:** 179 expert-written Social Story pairs across 14 chapters.
- **Generated Set:** 5085 Social Stories created via STARSOW using GPT-4o.

2.8.4 Relevance

SS-GEN provides both a strategy for scalable story creation and empirical evidence that smaller, fine-tuned models can be cost-effective alternatives to proprietary LLMs. It serves as a blueprint for using LLMs in domains requiring precision, empathy, and structure in generated narratives.

2.9 Saraswat et al.(2024) "Story-Yarn: An Interactive story building application"

Saraswat et al.[7] developed an interactive story building application using Knowledge graphs and LLM. They first customize a knowledge graph using children's stories to capture relationships between story elements. This KG is then combined with a LLM to generate coherent stories collaboratively. They also develop a simple app to enable user interaction, making the tool suitable for both home and school use.

2.9.1 Methodologies Discussed

The methodology involves creating a customized knowledge graph (KG) from a children's stories dataset. The process includes key steps such as node creation, relation deduction, co-referencing, and weight assignment to build meaningful KG triplets. The resulting graph is stored and scaled using the Neo4j graph database. The Story Yarn framework then leverages this KG in combination with a large language model, specifically Gemini 1.0 Pro, to enable interactive story building. The process begins with user-provided keywords, which are used to generate an initial sentence. Keywords are then extracted from each sentence to query the KG, guiding story progression. At each stage, the system presents multiple possible future story paths based on KG triplets and LLM outputs, allowing the user to choose and contribute creatively to the evolving narrative.

2.9.2 Key Findings

The performance of Story-Yarn is evaluated by comparing story paths generated by KG triplets, LLM-only, and the combined Story Yarn approach using similarity, diversity, and entropy metrics. The evaluation showed that KG paths had higher similarity to the story so far but lacked diversity, while LLM-only paths were more varied but less coherent. Entropy analysis found LLM outputs to be more predictable. Story Yarn balanced both, combining KG's structure with LLM's creativity for richer story generation.

2.9.3 Datasets Mentioned

StoryWeaver Dataset, MCTest Keywords.

2.9.4 Relevance

This paper highlights the combination of knowledge graphs and LLMs to generate coherent and diverse story paths, enhancing creativity and narrative exploration. It also emphasizes the effectiveness of the customized KG in supporting the LLM to produce more structured and engaging narratives.

2.10 Wang et al.(2024) "Guiding and Diversifying LLM-Based Story Generation via Answer Set Programming"

Wang et al.[10] combine instruction-tuned LLMs with symbolic story generation to enhance diversity in narratives. By using answer set programming (ASP) to guide LLM-based story generation, their approach produces more diverse stories than unguided LLMs, with improved compactness and flexibility compared to traditional narrative planning.

2.10.1 Methodologies Discussed

The authors propose a two-step process for ASP-guided story generation: outlining with answer set programming (ASP) followed by writing with a large language model (LLM). The ASP step involves pre-generating story outlines by defining narrative functions and using constraints to ensure coherence. In the second step, the generated outline guides the LLM—specifically, GPT-3.5-turbo—in story writing, where narrative functions are translated into writing instructions. Prompts and conversation history are used for sequential paragraph generation. For comparison, the authors also present a baseline unguided LLM-based generator.

2.10.2 Key Findings

The evaluation involved generating ten parallel stories each using the ASP-guided approach and an unguided baseline, then comparing their semantic homogeneity using a sentence embedding model. Results showed that ASP-guided generation consistently produced more diverse stories, with earlier paragraphs being more homogenous than later ones.

2.10.3 Relevance

This paper highlights the combination of symbolic planning through ASP and LLMs to generate diverse and coherent stories. It emphasizes the effectiveness of ASP-guided outlines in enhancing the creativity, structure, and flexibility of LLM-generated narratives.

2.11 Pei et al. (2024) "SWAG: Storytelling With Action Guidance"

Pei et al. [6] propose Storytelling With Action Guidance (SWAG), a two-model LLM framework where one model generates content and another guides the story's direction. Framing storytelling as a search problem, SWAG shows improved coherence and performance over traditional end-to-end methods. This approach allows for greater narrative control and adaptability during story generation.

2.11.1 Methodologies Discussed

The methodology involves a two-model framework consisting of a story generation LLM and an Action Discriminator LLM (AD LLM). The AD LLM is trained using preference data through supervised fine-tuning and Direct Preference Optimization. A feedback loop is established where the story model generates content, and the AD model selects the next best action from a predefined list, updating the story state iteratively. This setup ensures controlled, coherent storytelling and supports modular use of open-source, closed-source, or hybrid LLMs.

- *Action Discriminator LLM:* Fine tuned Llama-2-7B, Mistral-7B, and GPT-4-Turbo.
- *Story generation model:* Base Llama-2-7B, Mistral-7B, and GPT-4-Turbo.

2.11.2 Key Findings

SWAG was evaluated by humans and GPT-4 Turbo against baseline models in terms of interesting-ness, surprise, and coherence. SWAG outperformed its baselines in regardless of its judges and prove to be a better model for interesting and amazing stories.

2.11.3 Datasets Mentioned

Long stories dataset, WritingPrompts dataset.

2.11.4 Relevance

SWAG enhances narrative coherence and creativity by combining story generation and action guidance through a feedback loop. This approach improves modularity and control in storytelling, offering a scalable solution for structured narratives and supporting the effectiveness of hybrid LLM configurations in dynamic content creation.

2.12 Yao et al. (2019) "Plan-and-Write: Towards Better Automatic Storytelling"

Yao et al. [14] propose the *Plan-and-Write* framework, a hierarchical story generation approach that improves coherence and diversity by explicitly separating plot planning from surface realization. Unlike traditional end-to-end generation, this method mirrors human writing practices, enabling more structured and creative storytelling through intermediate storyline representations.

2.12.1 Methodologies Discussed

The Plan-and-Write framework decomposes storytelling into two stages: (1) storyline planning and (2) story generation. It explores two schemas:

- *Dynamic Schema:* Interleaves storyline and story generation. At each step, it plans the next plot keyword and then writes the next sentence, allowing flexibility and contextual adaptation.
- *Static Schema:* Plans the full storyline before writing, akin to a writer outlining a story. It improves coherence by offering a holistic view of the narrative in advance.

Both schemas employ RNN-based Seq2Seq models with attention, utilizing extracted keyword-based storylines as plot representations. The storylines are generated or learned using unsupervised methods (e.g., RAKE keyword extraction), minimizing annotation costs.

2.12.2 Key Findings

Experiments on the ROCStories dataset reveal that both schemas significantly reduce repetition and increase diversity compared to baselines.

2.12.3 Datasets Mentioned

ROCStories dataset — a large corpus of five-sentence commonsense stories annotated with titles.

2.12.4 Relevance

Plan-and-Write introduces a scalable, modular framework for automatic storytelling by aligning machine generation with human creative processes. It highlights the value of explicit planning in neural text generation and opens avenues for more interpretable, controllable, and diverse narrative synthesis.

- **Dataset chosen :** ROC Stories (*link*)

Paper	Models Used	No of Parameters	Dataset Used	Method
Wang et al.(2023)[11]	GPT 2	1.5B	ConceptNet, ATOMIC; WordNet, VerbNet; ROCStories, WritingPrompts.	Two approaches: (1) *As Text* — convert knowledge to natural language (2) *As Encoding* — use GNNs to inject knowledge embeddings.
Xie et al.(2023)[12]	GPT 3	175B	ROCStories, WritingPrompts.	Prompt-based few-shot learning with LLMs using example stories; no fine-tuning.
Saraswat et al.(2024)[7]	Gemini 1.0 Pro	75B	StoryWeaver Dataset, MCTest Keywords	LLM + Knowledge Graph (KG)
Wang et al.(2024)[10]	GPT 3.5 Turbo	200B	-	LLM + Answer set programming (ASP)
Pei et al(2024)[6]	Llama-2-7B, Mistral-7B, GPT-4-Turbo	7B	Long stories dataset, WritingPrompts dataset.	LLM + Action Discriminator LLM (AD LLM)
Alabdulkarim et al.(2021)[1]	GPT 3	1.5B	ROCstories, ConceptNet	Uses KG-DQN to select GPT-2 outputs KGs and VerbNet rewards.
Yao et al. (2019)[14]	RNN with attention	30M	ROCStories dataset	Plot Planner + Story Generator (two-stage Seq2Seq pipeline)
Simon and Muise (2022)[8]	GPT-J-6B	6B	Children's Stories	Symbolic Planning + LLM action
Li et al. (2024)[5]	-	-	-	Hierrarchichal story representation.
Zhao et al. (2023)[15]	-	-	-	Compared interleaved (human+LLM) vs. non-interleaved (LLM-only) storytelling
Xu et al. (2020)[13]	GPT 2	1.5B	ROCStories, ConceptNet	keyword prediction + knowledge retrieval + ranking + conditional generation
Venkatraman et al. (2025)[9]	Llama, Mistral, Gemma, Orca,Olmo	<8B	CollabStories	Multimodal Story Generation
Feng et al. (2024)[9]	Mistral, LLaMA, Gemma	<8B	Custom dataset	STARSOW prompting + LoRA PEFT fine-tuning for structured social story generation
Guan et al. (2020)[4]	GPT 2	1.5B	ROCStories, ConceptNet	Finetuning + Discriminative training

Table 1. Comparison of various papers based on models, parameters, datasets, methods, and disadvantages.

3 Gap Analysis

Despite significant progress in automated story generation using large language models (LLMs), current approaches across the reviewed literature share several limitations that restrict their effectiveness in applications requiring long-term consistency.

Autoregressive generation without persistent memory leads to a common issue where stories gradually drift from their intended themes or characters, particularly in longer narratives. This is especially problematic for interactive storytelling, where maintaining context across multiple turns is essential.

Moreover, **Retrieval-Augmented Generation (RAG)**, a method that dynamically retrieves relevant background knowledge or prior content during generation, is not employed in any of the reviewed works. The lack of RAG leads to over-reliance on parametric memory, which can cause hallucinations, factual inconsistencies, or generic storytelling.

Additionally, **Reinforcement Learning (RL)** is not utilized as a training or adaptation technique in any of the discussed methods. RL could be instrumental in refining generation policies based on user feedback or task-specific rewards.

In conclusion, while the current body of work demonstrates effective story generation through fine-tuning and task-specific prompting, future research can benefit significantly by incorporating long-term memory mechanisms, retrieval-based augmentation, and reinforcement-based adaptation to address the challenges of consistency, controllability, and personalization in story generation.

4 Selection of Best Pretrained Model

Based on the results from SS-GEN [3], we found that smaller open-source models like Mistral can perform just as well as, or even better than, large models like GPT-4o—especially when they are fine-tuned for a specific task. In SS-GEN, Mistral and similar models showed strong improvements in metrics like BLEU-4 and ROUGE-L, and were often preferred by human evaluators for their consistency, safety, and clarity in storytelling.

Because of this, we choose **Mistral** as our main pretrained model. It provides a good balance between performance and efficiency. As a backup, we include **DeepSeek R1**, which is also a powerful open-source model known for its strong reasoning abilities. This setup allows us to build a flexible and reliable story generation system without relying on expensive or closed-source models.

References

[1] Amal Alabdulkarim, Winston Li, Lara J. Martin, and Mark O. Riedl. Goal-directed story generation: Augmenting generative language models with reinforcement learning, 2021.

[2] Arwa I. Alhussain and Aqil M. Azmi. Automatic story generation: A survey of approaches. *ACM Comput. Surv.*, 54(5), 2021.

[3] Yi Feng, Mingyang Song, Jiaqi Wang, Zhuang Chen, Guanqun Bi, Minlie Huang, Liping Jing, and Jian Yu. Ss-gen: A social story generation framework with large language models, 2024.

[4] Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108, 2020.

[5] Danrui Li, Samuel S. Sohn, Sen Zhang, Che-Jui Chang, and Mubbasir Kapadia. From words to worlds: Transforming one-line prompts into multi-modal digital stories with llm agents. In *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games*, New York, NY, USA, 2024. Association for Computing Machinery.

[6] Jonathan Pei, Zeeshan Patel, Karim El-Refai, and Tianle Li. Swag: Storytelling with action guidance. pages 14086–14106, 01 2024.

[7] Hryadyansh Saraswat, Snehal D. Shete, Vikas Dangi, Kushagra Agrawal, Anuj Aggarwal, and Aditya Nigam. Story-yarn : An interactive story building application. In Sobha Lalitha Devi and Karunesh Arora, editors, *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 248–255, AU-KBC Research Centre, Chennai, India, December 2024. NLP Association of India (NLPAI).

[8] Nisha Simon and Christian Muise. Tattletale: Storytelling with planning and large language models. In *ICAPS Workshop on Scheduling and Planning Applications woRKshop*. 2022.

[9] Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. Collabstory: Multi-llm collaborative story generation and authorship analysis, 2025.

[10] Phoebe J. Wang and Max Kreminski. Guiding and diversifying llm-based story generation via answer set programming, 2024.

[11] Yuxin Wang, Jieru Lin, Zhiwei Yu, Wei Hu, and Börje F. Karlsson. Open-world story generation with structured knowledge enhancement: A comprehensive survey. 2023.

[12] Zhuohan Xie, Trevor Cohn, and Jey Han Lau. The next chapter: A study of large language models in storytelling. In C. Maria Keet, Hung-Yi Lee, and Sina Zarrieß, editors, *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia, September 2023. Association for Computational Linguistics.

[13] Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. Megatron-cntrl: Controllable story generation with external knowledge using large-scale language models, 2020.

[14] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling, 2019.

[15] Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre L. S. Filipowicz. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, New York, NY, USA, 2023. Association for Computing Machinery.