Original Research Paper

Volume - 16 | Issue - 08 | August - 2026 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Computer Science

AI AS A SCREENWRITER: A SURVEY OF LARGE LANGUAGE MODELS IN FILM NARRATIVE GENERATION

Saleheen Akhtar School of Computer Science and Applications,Reva University, Bengaluru, India

ABSTRACT Traditionally, screenwriting was seen as the sole filmmaking domain requiring human creativity. Large language models (LLMs) such as GPT-4, Claude, and Gemini now challenge this view by generating dialogue, plot arcs, and scene descriptions from brief prompts. This survey reviews current advances in LLM-driven narrative generation for film and television, examining transformer architectures and comparing model performance against screenwriting benchmarks. We focus on key limitations, including dramatic tension, character consistency, and subtext. The 2023 Writers Guild of America strike, along with concerns about bias, intellectual property, and AI authorship, are central to this discussion. Our aim is to provide an objective assessment of LLMs as narrative tools, rather than predict their potential to replace screenwriters.

KEYWORDS : Large Language Models, Narrative Generation, Screenwriting, Computational Creativity.

INTRODUCTION

Automating screenwriting presents unique challenges. While LLMs excel at generating dialogue and scene descriptions, true screenwriting requires managing dramatic expectations over time. Each scene must build on the last, characters need conflicting desires, and the audience should be surprised in ways that feel inevitable.

Screenwriting is a valuable test case for AI, positioned between short-form copy and unstructured fiction. It requires both strong writing and a practical blueprint for production. Since GPT-3's release in 2020, writers and studios have experimented with LLMs for scriptwriting, often finding results both impressive and lacking. The 2023 Writers Guild of America strike highlighted concerns about AI's impact on writing jobs. This survey examines the technical, creative, and evaluative aspects of LLMs in screenwriting, focusing on their capabilities, limitations, and the ongoing need for human judgment.

LITERATURE REVIEW

A. From the first stories to the Transformers. Automated story generation is not new: Meehan's TALE-SPIN [1], developed in 1977, established the notion that narrative may be formed from planning by generating folk-tale plots from character goals and impediments. Decades later, in 2017, Vaswani et al. [2] unveiled the Transformer; by 2019, enormous pretrained models generated text that was momentarily identical to human writing. In theory, the attention mechanism helps long-form coherence by allowing a model to refer to any previous token; but, in reality, models lose detail over thousands of tokens. Consistency over a long document is different from consistency inside the local context window, and most models lose track of details across long passages. Fan et al. [3] found hierarchical generation (summary then expansion) improved coherence but flattened drama, because the dramatic compression was lost in expansion.

B. Screenplay-specific models. General models were not trained on scripts. Mirowski et al. [4] built Dramatron using hierarchical few-shot prompting rather than fine-tuning, and found outputs were correctly formatted and fluent but lacked subtext AI characters state exactly what they think. Alabdulkarim et al. [5] found models following a preset outline scored higher on coherence but lower on human-rated surprise, confirming that structure-versus-spontaneity is a real computational problem: the scaffolding that helps a model stay coherent also makes it conventional.

C. Dialogue generation. Dialogue is where LLMs look strongest yet are most deceptive. A line of dialogue is a short piece of text, and short text is where these models genuinely perform well. Roller et al. [6] studied open-domain dialogue generation and showed that large models could maintain persona and topic across a conversation. The problem in a screenplay context is that dialogue must do more than communicate information: it must reveal character under pressure, advance plot, and carry subtext simultaneously. Li et al. [7] found persona embeddings improve consistency but compress distinct voices into a shared space, so similar profiles converge the craft of distinct voices, the thing that separates one writer's characters from each other, is not something current models reliably do.

D. Evaluation. Celikyilmaz et al. [8] survey text-generation evaluation and note automated metrics like BLEU and ROUGE measure reference similarity, not story quality. Guan et al. [9] built OpenMEVA, showing existing metrics correlate poorly with human judgment and fail to catch causal-order violations or incoherence a finding that matters for screenplays, where plot logic is central.

LLMS FOR HANDLING SCREENPLAY

A screenplay must communicate the story clearly for production and maintain momentum to engage its audience. LLMs handle formatting effectively, but their performance on more complex requirements is inconsistent.

A. Setting the scene. An entrance condition, a change, and an exit are necessary for a scene. Often, LLMs only produce the midsection. Mirowski et al. [4] observed that although 15 industry professionals co-wrote the study with Dramatron, participants assessed it as surprise (92%) and entertaining (77%). However, only 46% felt proud of the outcome, and more than half expressed no ownership. The main critique was that scenes explicitly revealed motivation rather than allowing it to surface through subtext, which is the space between a character's words and their intentions.

B. Three-act format. With 90–120 pages (about 15,000–20,000 words), a movie script approaches or surpasses context windows from the 2023 era. Act 1 is cohesive, while Act 3 frequently replicates Act 1 in front-load models. In addition to improvements in outline relevance and interest, Yang et al. [10] demonstrated that, thorough hierarchical outline produced a 22.5point increase in human-judged plot coherence over an unscaffolded baseline, demonstrating that coherence deteriorates before a script-length document is finished in the absence of explicit scaffolding. Because conventional beat sheets result in conventional scripts, hierarchical generation in which the model first writes a beat sheet before expanding each beat improves coherence quantifiably at the expense of spontaneity.

C. Character voice. Readers immediately notice when a character's voice is absent. LLMs tend to produce output that gravitates toward the statistical average, making distinctive voices, such as those in early Tarantino or Waller-Bridge's Fleabag, less likely. Fine-tuning on a specific writer's work can help, but this approach raises concerns about data ownership and results in imitation rather than original creation.

D. Genre. Genre is where LLMs perform most reliably, because pattern completion is what they do. A prompt for a psychological thriller in a remote cabin hits the expected beats but will not subvert them. Get Out works because Peele knew the conventions well enough to break them, an LLM trained on horror will comply. Reliability here is exactly the limitation: the model reproduces the genre instead of interrogating it.

EVALUATION: MEASURING STORY QUALITY

NLP has precision, recall, F1, BLEU, and perplexity metrics needing ground truth. Screenplay quality has none; two readers can disagree and both be right, because goodness depends on what the script is trying to do and whether it succeeds on its own terms.

A. Automated metrics and their limits. BLEU measures n-gram overlap against a reference script, which presumes you already had a human one. Perplexity measures fluency, not drama. Guan et al. [9] proposed coherence metrics for causal plausibility, useful but narrow: a script can be causally coherent yet emotionally inert, with nothing at stake for the audience.

B. Human evaluation. The most meaningful approach remains human decision, with all its cost. Mirowski et al. [4] recruited 15 industry professionals rather than crowd workers, their interviews surfaced recurring concerns around subtext, motivation, and interiority. When readers compare AI and produced scripts unlabeled, AI is identified above chance but not overwhelmingly, and the rate improves with longer samples deficits compound over time.

C. A proposed evaluation framework. We propose four dimensions: structural integrity (clear act structure with plot points at intervals), character consistency (stable decisions and voice across scenes), scene necessity (each scene changes direction uncuttable), and dramatic tension (establishes what is at stake). No automated metric measures all four; human evaluation is still necessary for the last two, which are also the ones models fail most often.

WIDER IMPLICATIONS

The aforementioned surface-versus-judgment difference has ramifications that go beyond scholarly curiosity. Industry pros viewed co-written scenes as surprising and engaging but felt little ownership of the outcome, according to the Dramatron research [4], indicating that the tool alters the writer's relationship to the work even when output quality is good. A script's creative agency is just as valuable as its text, and existing technologies do not offer this kind of agency.

Control and structure are closely linked. Yang et al. [10] showed that explicit scaffolding significantly improves coherence, indicating that human oversight is essential for effective long-form production. However, while structured output is reliable, creating distinctive work requires the writer to override the plan, which current models cannot do. Therefore, the most effective workflow is interactive: the writer sets direction and standards, and the model generates drafts and alternatives.

Evaluation remains the main barrier to accountable progress. As noted in Section IV, current automated metrics do not align well with human assessments of story quality [9]. Without better evaluation methods, it is difficult to rigorously test claims about improved narrative AI, limiting both research and deployment. These limitations suggest that AI-generated output should be treated as drafts requiring human review, not as final products.

DISCUSSION

LLMs consistently fail at the judgment layer scene necessity, subtext, and dramatic tension while succeeding at the surface layer format, dialogue fluency, and genre matching. This is not coincidental; while judgment relies on modeling how audience expectations evolve over time, surface properties are readable to next-token prediction. There may be a hard limit to the tension between structure and spontaneity: a plan that is clear enough to execute is by definition conventional, whereas unique turns depart from expectations. The findings converge LLMs are proficient at the mechanical layer and unreliable at the judgment layer, and the most defensible use is as a collaborative aid under human creative control, not as an author.

FUTURE DIRECTIONS

The evaluation challenge is pressing: without standardized metrics beyond fluency, it is impossible to rigorously assess improvement. Long-form coherence remains problematic; while models with larger context windows (such as 128k tokens) outperform earlier versions, coherence does not improve proportionally with context length. Character modeling is also underdeveloped, as most approaches treat characters as static attributes rather than dynamic systems of desires, fears, and contradictions. Effective screenwriting emphasizes these dynamics over fixed traits.

Finally, the human-AI collaboration workflow requires further research. In practice, writers often use LLMs for specific tasks such as brainstorming, testing dialogue, or overcoming creative blocks, while maintaining creative control. The most valuable tools are those that support writers, not those that attempt to replace them.

REFERENCES

1. J. R. Meehan, "TALE-SPIN: An Interactive Program that Writes Stories," in Proc. 5th Int. Joint Conf. Artif. Intell. (IJCAI), vol. 1, pp. 91–98, 1977.
2. A. Vaswani et al., "Attention Is All You Need," in Adv. Neural Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017.
3. A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical Neural Story Generation," in Proc. 56th Annu. Meeting Assoc. Comput. Linguist. (ACL), pp. 889–898, 2018.
4. P. Mirowski, K. Mathewson, J. Pittman, and R. Evans, "Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals," in Proc. 2023 CHI Conf. Hum. Factors Comput. Syst., Art. 355, 34 pp., 2023.
5. A. Alabdulkarim, S. Li, and X. Peng, "Automatic Story Generation: Challenges and Attempts," in Proc. 3rd Workshop Narrative Understanding, pp. 72–82, 2021.
6. S. Roller et al., "Recipes for Building an Open-Domain Chatbot," in Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguist. (EACL), pp. 300–325, 2021.
7. J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, "A Persona-Based Neural Conversation Model," in Proc. 54th Annu. Meeting ACL (Vol. 1: Long Papers), pp. 994–1003, 2016.
8. A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of Text Generation: A Survey," arXiv preprint arXiv:2006.14799, 2021.
9. J. Guan et al., "OpenMEVA: A Benchmark for Evaluating Open-ended Story Generation Metrics," in Proc. 59th Annu. Meeting ACL/IJCNLP, pp. 6394–6407, 2021.
10. K. Yang, D. Klein, N. Peng, and Y. Tian, "DOC: Improving Long Story Coherence with Detailed Outline Control," in Proc. 61st Annu. Meeting ACL, pp. 3378–3465, 2023.