WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Exploring the capability of Large Language Models in Creative Writing:Assessing Their Potential for Generating Original Stories

Rashmita Uchil[1], Bhushan Gaonkar[2]

[1]Student, Institute of Computer Science, Mumbai Educational Trust- MET ICS Mumbai, India
[2]Professor, Institute of Computer Science, Mumbai Educational Trust- MET ICS Mumbai, India

Abstract

As artificial intelligence continues to evolve, Large Language Models (LLMs) are increasingly being used in creative writing, sparking discussions about their ability to craft original and engaging stories. This study explores how well AI can generate narratives that feel unique, immersive, and emotionally compelling. By comparing AI-generated stories with those written by humans, we analyse key elements such as storytelling structure, creativity, and coherence. While LLMs can produce fluent and stylistically diverse text, they often struggle with deeper storytelling aspects like originality, emotional nuance, and narrative consistency over long-form pieces. Additionally, this research examines ethical concerns, including the risk of unintentional plagiarism and the impact of AI-generated content on human creativity and intellectual property. Rather than replacing human authors, LLMs show promise as tools that can assist in the creative process such as helping writers brainstorm ideas, refine their writing, and push the boundaries of storytelling. This research highlights both the opportunities and limitations of AI in creative writing, emphasizing the need for human oversight to ensure originality and artistic expression.

Keywords : Large Language Models (LLMs), Creative writing, Originality, Emotional nuance, Narrative consistency, Ethical concerns, Unintentional plagiarism, Intellectual property, Human authors, Writing tools, Artistic expression

1. Introduction

I. What Are Large Language Models (LLMs)?

Large Language Models, or LLMs, are a type of artificial intelligence that can generate human-like text based on patterns they've learned from huge volumes of written material. These models are trained on everything from books and articles to websites and online conversations, allowing them to understand how we use language—and even how we tell stories. Some of the most well-known examples today include OpenAI's GPT-4, Claude 3 by Anthropic, and Google's Gemini. They're used to write emails, answer questions, create poems, generate code, and, increasingly, to assist with writing entire stories or articles from scratch.

II. Background and Evolution of LLMs

The rise of LLMs is built on a breakthrough that changed the field of natural language processing: the transformer model, introduced in 2017. This new approach allowed AI to pay attention to every word in a sentence—rather than just the nearby ones—making the generated text more coherent and contextually accurate. Over time, these models have grown in both size and ability. What started as tools for basic tasks like translation and summarization are now being used in far more creative areas, like writing short stories, scripts, or full-length articles. The technology keeps getting better, and with each version, LLMs come closer to mimicking the way we think and write.

III. Current State of the Art in Creative Writing

Today, LLMs are making their way into the creative side of writing. They're being used to write blogs, help with marketing campaigns, generate fictional stories, and even suggest poetry. Many content creators use them for brainstorming or drafting ideas because of how fast and fluent their outputs can be. Tools like GPT-4 and Claude 3 are particularly good at generating text that reads naturally. However, while they excel in structure and fluency, they sometimes fall short when it comes to deeper storytelling skills—like building emotional arcs, creating believable characters, or weaving in subtle themes.

IV. Problem Statement: Can AI Replace Human Creativity?

Even with all their progress, there's still a big question hanging over LLMs: Can they really be creative, or are they just really good at sounding like they are? While they can produce stories that look polished on the surface, those stories often lack the kind of originality and emotional weight that comes from personal experience or human insight. Readers often feel there's something missing—a sense of soul, or purpose. This paper is built around that central problem: whether these models can truly create, or if they're just remixing what they've seen before.

V. Purpose and Proposed Direction

The goal of this paper isn't to decide whether AI is better or worse than human writers, but rather to explore how well these tools perform in creative writing—and how they compare to humans. Using existing studies, benchmark data, and content comparisons, this research takes a closer look at what LLMs are capable of when it comes to storytelling. The proposed solution is not to replace human creativity with AI, but to think of AI as a creative partner—a tool that can help writers work faster, overcome blocks, or organize ideas, while still leaving room for the human touch that makes a story feel real.

2. LiteratureReview

Large Language Models (LLMs) have taken centre stage in the field of artificial intelligence, especially in how we work with language. Ever since the introduction of the transformer architecture by Vaswani et al. in 2017, which changed the way machines process language, we've seen a rapid evolution in the capabilities of LLMs. Today's models—like GPT-3 and GPT-4—can generate text that often sounds surprisingly human. While these tools were initially used for tasks like translation, summarization, and question-answering, more recent studies have started exploring their role in creative writing, especially fiction.

Several researchers have investigated how LLMs can assist in writing both informational content and creative material. For example, a study by Ijibadejo and Altamimi (2024) looked at how LLMs perform in generating articles and creative pieces. They pointed out how effective these models are at producing fluent and readable text, while also noting the challenges in maintaining originality and creativity. However, their work covered a broad spectrum of writing tasks, without a deep dive into storytelling itself. More focused studies have looked at the storytelling abilities of LLMs. Roemmele and Gordon (2018), for instance, tested how AI models handle narrative logic using a story-completion task. Others, like Peng et al. (2021), explored how well these models could write full stories. They found that while LLMs can create grammatically sound and coherent text, they often struggle with deeper elements like character development, emotional impact, and keeping the story consistent from beginning to end. There's also growing interest in how a model's "worldview"—that is, the patterns and biases it inherits from training data—shapes the stories it generates. This is important because these models learn from large, often unfiltered datasets that can include stereotypes or cultural biases. If not addressed, those issues can subtly influence the narratives the AI produces. Another area of research focuses on human-AI collaboration in writing. Clark et al. (2018) studied what happens when writers use LLMs as co-authors or assistants. While many writers found AI suggestions helpful for brainstorming or speeding up the drafting process, some felt that the machine's influence led to more generic or formulaic writing. Across all of these studies, one thing becomes clear: while LLMs are powerful tools for generating language, their ability to produce truly original, emotionally rich, and narratively strong stories is still limited. Much of the research so far has focused on surface-level coherence or technical performance, but hasn't gone deep into the creative heart of storytelling. That's where this study comes in. By narrowing the focus specifically to fiction and original story generation, we aim to better understand what these models can—and can't—do when it comes to the creative process. This research builds on earlier work but takes it a step further by evaluating how well LLMs perform as storytellers, not just text generators.

3. Methodology

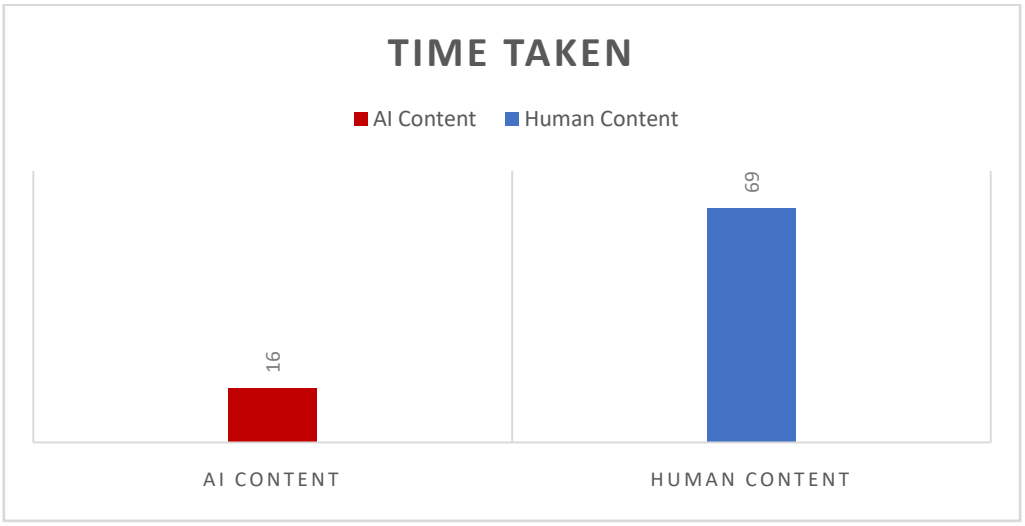
This study takes a qualitative, comparative approach grounded in secondary data analysis. Instead of running original experiments or conducting human evaluations, the research focuses on gathering, reviewing, and synthesizing existing information from a variety of reliable, publicly available sources. The goal was not to build a model or test one in a lab, but to understand—through observation and analysis—how current large language models (LLMs) perform in creative writing tasks, and how that performance stacks up against human writers. To do this, I explored a wide range of industry benchmarks, expert commentaries, blog-based experiments, and comparative reviews. These sources include detailed model evaluations published by *Intensed* (2024) and Harshit (2024), which examine the strengths and limitations of popular LLMs such as GPT-4, Claude 3, and Gemini. These models were assessed across several natural language processing benchmarks, including coherence, stylistic consistency, reasoning, and fluency. The insights gained from these reviews helped form the basis for graphical comparisons that illustrate where each model tends to perform well—and where they don't. To deepen the analysis, this research also incorporates documented comparisons between AI-generated and human-written content. Neil Patel's article, which compares article creation time and long-term traffic performance between AI and human writers, offered an especially useful real-world example. Additional perspectives from platforms like *GravityWrite* (2024), *WSIWorld*, *ClictaDigital*, and *SiteCentre* contributed further context on user engagement, creativity, and emotional tone. These sources provided both quantitative data (such as time to generate content, or traffic over a five-month period) and qualitative feedback (like perceived authenticity or storytelling quality). Wherever applicable, graphs and visualizations were created or adapted to help communicate key findings more clearly. These visuals focus on critical performance factors such as writing speed, narrative coherence, creative depth, and long-term reader engagement. By placing human and AI outputs side-by-side, these charts aim to visualize the subtle and not-so-subtle differences that readers—and algorithms—respond to. It's important to note that all the data used in this research comes from secondary sources, and while care was taken to choose reputable and recent material, the findings reflect the perspectives and testing methods used by the original authors. This methodology, therefore, doesn't claim to be comprehensive or definitive. Instead, it provides a broad and balanced view of how LLMs are currently performing in creative writing, based on the best publicly available data at the time of writing. By leaning on existing studies and expert observations, this approach allows us to examine a wide spectrum of models and writing styles—without the limitations of a single test environment. It also makes room for diverse evaluations, such as storytelling capability, emotional tone, and reader impact, which are harder to capture with automated metrics alone. Through this synthesis, the paper offers a practical understanding of how AI-generated stories compare to human creativity, and where the future of writing might be headed as the two continue to intersect.

4. Comparative Analysis of AI and Human Writers

To gain a clearer understanding of how current large language models perform in creative content generation, we reviewed a range of publicly available benchmarks and case studies. This analysis includes both model-to-model comparisons—such as GPT-4, Claude 3, and Gemini—as well as direct comparisons between AI-generated and human-written content. Factors such as content quality, creativity, coherence, time efficiency, and user engagement were examined. The following graphs synthesize these insights to highlight the strengths and limitations of AI-generated content in relation to both other models and human writers.

4.1 AI vs. Human Content Creation Time

One of the most immediate and impressive advantages of large language models is the speed at which they can produce written content. In an experiment shared by Neil Patel (2023), this strength was put into perspective: it took an AI tool just 16 minutes to generate a complete, full-length article, while a human writer took 69 minutes to write the same type of content. This sharp contrast in time—visualized in the figure below—clearly demonstrates the efficiency of AI, especially for content creation tasks that are time-sensitive or require rapid scaling, such as marketing campaigns, product descriptions, or news summaries. However, while the time savings are compelling, they come with trade-offs that are hard to ignore. The same study observed that although AI could generate content quickly, the human-written article consistently outperformed the AI version in key areas like reader engagement, clarity of tone, and overall uniqueness. Readers were more likely to spend time on the human-written content, suggesting that subtle qualities like narrative flow, emotional nuance, and intentional storytelling still resonate more deeply when created by people. This comparison underscores an important point: while AI tools are incredibly useful for speeding up the writing process—particularly for rough drafts, repetitive tasks, or basic outlines—they still benefit greatly from human input. Refining AI-generated content with a human touch can help ensure it meets higher editorial standards, connects better with the target audience, and reflects brand or personal voice more effectively.



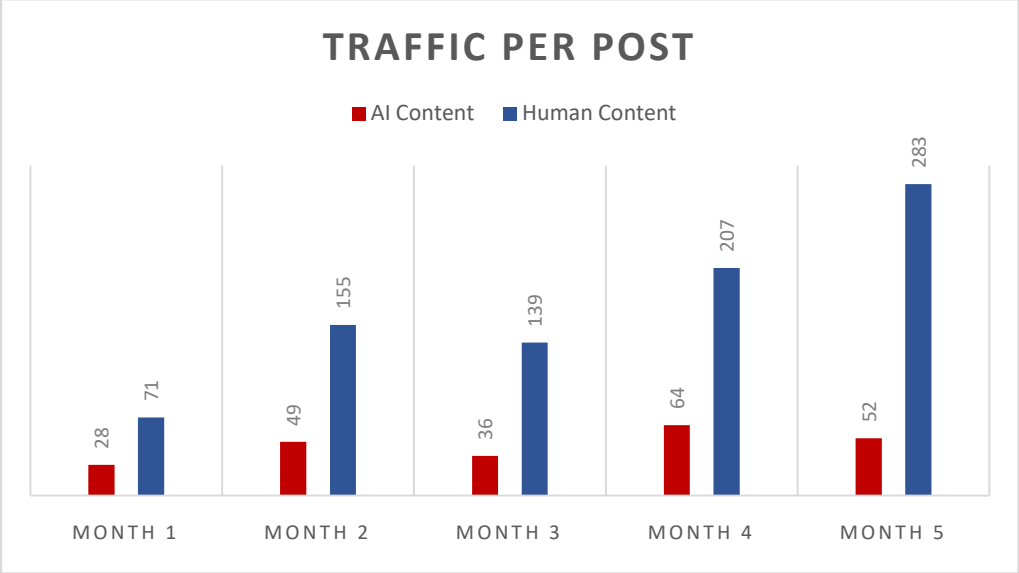
Source: Neil Patel Blog (https://neilpatel.com/blog/ai-vs-human-content/)

4.2 Performance Over Time: Website Traffic Analysis

In addition to efficiency, it's important to evaluate how AI-generated content performs once published. According to Neil Patel's five-month comparison between AI-written and human-written blog posts, human content consistently outperformed AI content in terms of website traffic. Over the observed period, the human-written article generated significantly higher organic traffic, engagement, and time-on-page.

This trend, shown in the below graph, suggests that while AI can produce content rapidly, it may not yet match the depth, nuance, or SEO alignment that human writers bring to the table. Google's ranking systems also seem to favour originality, natural tone, and expertise—all traits more strongly associated with human-authored work. This gap highlights a crucial insight: speed does not necessarily translate to sustained performance.

For content creators and businesses, this means that although AI is a powerful tool for scaling content quickly, a hybrid approach—where humans refine or enhance AI drafts—may be necessary to maintain quality and impact over time.



Source: Neil Patel Blog (https://neilpatel.com/blog/ai-vs-human-content/)

4.3 Content Quality Case Study: AI vs. Human-Authored Health Writing

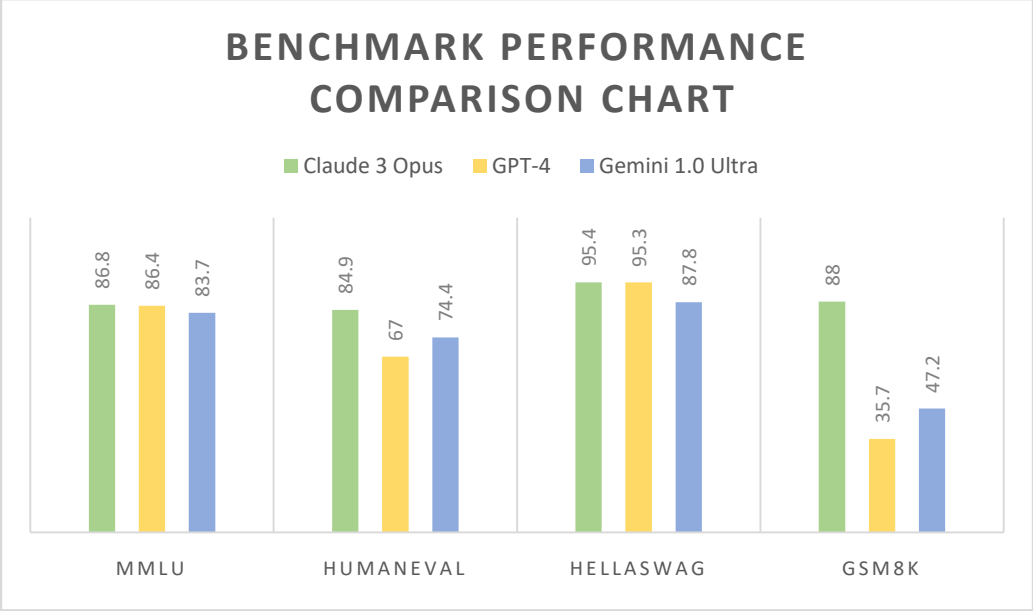
To better understand how AI-generated content compares with human writing in real-world scenarios, it helps to look at a concrete example. One such comparison, shared by Aha Media Group, offers a compelling side-by-side look at content written by an AI tool versus a professional human writer—both tasked with covering the same topic: *How to Prepare for a Women's Health Exam*. The AI version, created using ChatGPT-4, is clear, organized, and efficient. It outlines a few basic steps—gather your medical history, prepare a list of questions for your provider, follow any fasting instructions. On a surface level, it checks the boxes: the content is grammatically correct, the structure is logical, and the information is relevant. For someone looking for a quick, high-level overview, it works just fine. But when you read it closely, the tone feels somewhat mechanical. It lacks warmth, personality, and the subtle cues that help a reader feel understood or reassured. It's informative, but it doesn't make a human connection. Now compare that with the human-written version. It conveys similar advice, but does so with a deeper awareness of context and audience. The writer not only explains what to do but also why each step matters. For example, they highlight the value of understanding your family medical history—not just as a checklist item, but as a meaningful part of personalizing your healthcare. They walk the reader through what specific details to collect, such as ages of diagnosis and lifestyle patterns in the family. The tone is gentle, conversational, and clearly written with the reader's experience in mind. It feels like a person is on the other side of the screen—someone who genuinely cares about guiding you through a potentially sensitive or stressful appointment. This example brings into focus a subtle but crucial difference. AI is excellent at delivering structure and information quickly, which makes it incredibly useful in contexts where speed and clarity are the priority. However, human writing offers something AI still struggles to replicate: empathy, nuance, and emotional resonance. In healthcare—and other fields where trust and relatability matter—this distinction can have a major impact on how the content is received and whether it truly supports the reader's needs. In short, while AI can assist with drafting and organizing content, it may not yet be equipped to handle topics that require emotional intelligence or cultural sensitivity without human guidance. That's why many experts advocate for a hybrid approach, where AI handles the foundational tasks, but humans step in to refine the message and ensure it resonates with its intended audience.

Aspect	AI-Generated Content	Human-Written Content
Tone	Neutral and informative	Personalized and empathetic
Detail Level	Basic guidelines	Specific examples and thoughtful context
Language Style	Formal and polished	Conversational and approachable
Emotional Engagement	Minimal	Stronger, with a human touch
Originality	Generic phrasing	More insight, tailored voice

Source: Aha Media Group. *Human-Written vs. AI-Generated Healthcare Content*. https://ahamediagroup.com/blog/human-written-vs-ai-generated-healthcare-content/

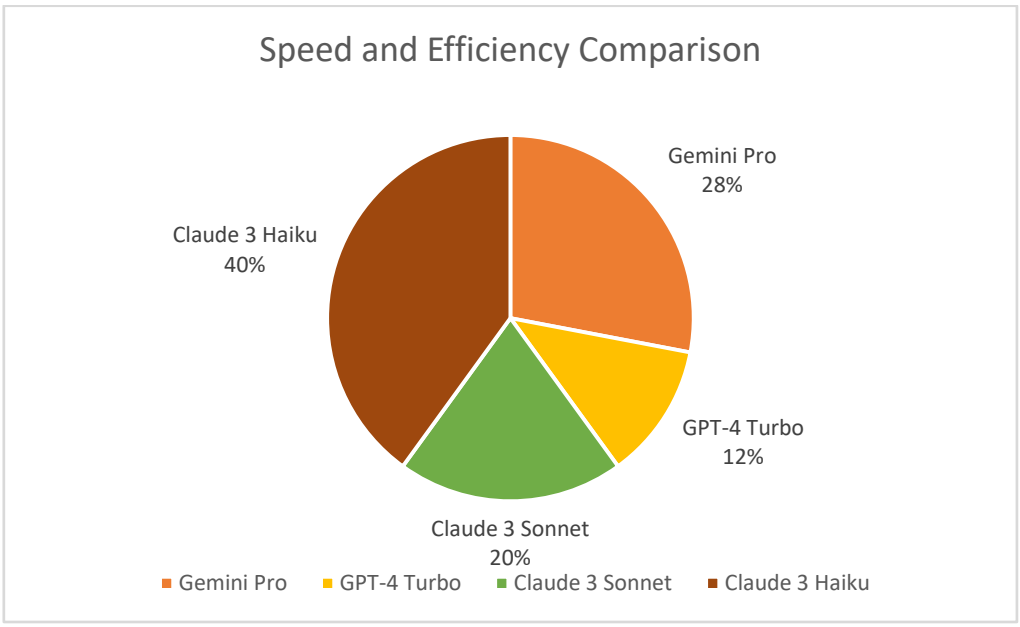
4.4 Benchmark Performance Across NLP Tasks

This bar chart compares Claude 3 Opus, GPT-4, and Gemini 1.0 Ultra across four prominent benchmarks: MMLU (massive multitask language understanding), HumanEval (coding ability), HellaSwag (commonsense reasoning), and GSM8k (grade school math). Claude 3 Opus and GPT-4 perform similarly across most tasks, with Claude slightly outperforming on MMLU and HellaSwag. Gemini 1.0 Ultra shows competitive results on some tasks but underperforms significantly on GSM8k. These results give insight into how general reasoning and language capabilities vary across models—an important factor when assessing creative writing potential.



4.5 Speed and Efficiency Overview

This pie chart offers a qualitative comparison of several leading LLM variants based on public reports of processing speed and efficiency. Models such as Gemini Pro and GPT-4 Turbo are highlighted for faster performance, making them suitable for high-volume or time-sensitive writing tasks. On the other hand, Claude 3 Haiku and Claude 3 Sonnet are noted for balancing speed with creative depth. While not a direct measure of creativity, speed plays a practical role in the usability of LLMs for large-scale content generation.



5. Ethical Considerations and Limitations

As large language models (LLMs) become more integrated into the world of creative and professional writing, they bring with them a number of important ethical concerns that go beyond their technical capabilities. While these models offer undeniable advantages in terms of speed, scalability, and fluency, their use raises deeper questions about authorship, originality, and bias—especially in fields where the integrity of written content truly matters. One of the most pressing ethical issues involves authorship and transparency. When a piece of writing is generated by an AI system, it's not always clear who should receive credit. This becomes especially complicated in contexts like journalism, academia, and literature, where readers often assume the content reflects a human's thoughts, emotions, or creative intent. Because LLMs don't actually possess these human qualities, failing to disclose their involvement can mislead readers and undermine trust.

Another concern lies in the area of originality and plagiarism. Although LLMs are capable of generating new content, they do so by drawing patterns from the massive datasets they've been trained on. This means their outputs can sometimes resemble or unintentionally mimic existing phrases, storylines, or ideas. In academic or literary settings, this raises valid questions about whether the work is genuinely original or simply a remix of what's already out there. Bias is also a critical issue. Since these models are trained on vast collections of internet data, they inevitably absorb the biases and stereotypes present in those sources. In creative writing, this can result in the repetition of harmful tropes or the exclusion of certain voices and perspectives. This highlights the need for responsible use, which includes careful editing, cultural sensitivity, and an awareness of the potential for unintended harm. It's also important to acknowledge the limitations of this research. The study is based entirely on secondary data from credible but publicly available sources such as AI research blogs and benchmark analyses. While these resources provide valuable insights, the findings are interpretive and do not reflect firsthand testing or direct model evaluations. Additionally, because LLM technology is evolving so rapidly, some observations may become outdated in a relatively short time. Despite these limitations, the study offers a meaningful snapshot of where AI currently stands in the realm of creative writing. It sheds light on both the potential and the pitfalls of using LLMs as writing tools. Looking ahead, future research should focus on developing clearer standards for attribution, better methods for identifying and mitigating bias, and collaborative workflows that combine human creativity with the strengths of AI in a responsible and transparent way.

6. Discussion

As large language models continue to gain traction in creative writing, the results from this review suggest a mix of promise and limitation. Tools like GPT-4, Claude 3, and Gemini are no longer just technological novelties—they're becoming common tools for writers, marketers, educators, and content creators. They're fast, accessible, and impressively fluent. But when we look beyond surface-level fluency and start asking whether they can truly "create," the answers become more nuanced. On the one hand, these models are capable of generating content that is structurally sound and grammatically polished. GPT-4, for instance, consistently performs well when it comes to coherence and logical flow. Claude 3 stands out in more emotional or narrative-driven tasks, where tone and subtlety matter. Gemini, while generally competent, tends to fall behind when it comes to narrative richness and depth. However, what these models can achieve with language doesn't always translate to compelling storytelling. While they can imitate the form and rhythm of a story, the content often lacks emotional complexity, originality, or a sense of lived experience. Characters may behave predictably. Plot twists may feel generic. And themes—though sometimes ambitious—often lack the nuance that comes from a human perspective.
Interestingly, the human vs. AI comparison reinforces this divide. AI can generate a full-length article in minutes, but when these outputs are evaluated in terms of engagement, emotional connection, or long-term impact (like website traffic), human-written content still has the edge. This suggests that readers, whether consciously or not, can often sense when a story has a human voice behind it. One takeaway from this analysis is that AI may be most effective not as a replacement, but as a collaborator. Many writers already use LLMs to get started, brainstorm ideas, or structure a draft. These tools can help overcome writer's block, speed up repetitive tasks, or experiment with style. But the final touch—refining characters, shaping tone, and crafting emotional depth—is still best handled by a human writer. This growing dynamic between human creativity and machine assistance isn't something to fear—it's something to work with. Just as word processors didn't replace authors but changed how they work, AI is shaping a new kind of writing process—faster, more iterative, and possibly more collaborative than ever before.

7. Conclusion and Future Work

There's no denying that large language models have changed how we think about writing. What once required hours of effort can now be generated in minutes with the help of tools like GPT-4, Claude 3, or Gemini. These models are fast, efficient, and capable of producing content that, at first glance, can feel remarkably human. But the more we look at what they create—and how people respond to it—the clearer it becomes that speed and surface polish aren't everything. Throughout this paper, we've looked at how AI compares to human writers, not just in terms of how fast it can work, but in how well it can tell a story. And while AI performs well on many fronts—like structure, grammar, and consistency—it still falls short when it comes to originality, emotional depth, and the subtle art of connecting with a reader. That's where human creativity still stands strong. What this means is that AI probably isn't here to replace writers—it's here to work with them. Used thoughtfully, LLMs can help spark ideas, break through writer's block, or speed up the early stages of drafting. But the best results often come when human judgment steps in to guide the narrative, refine the voice, and add the personal touch that machines can't replicate. Looking ahead, future research might explore how AI could improve at long-form storytelling, how it handles cultural nuance, or how writers can better collaborate with these tools. For now, the takeaway is simple: AI can be a powerful writing assistant—but the story still belongs to us.

8.Reference

[1] D. Harshit, "The Best LLM for Content Creation," *Medium*, 2024. [Online]. Available: https://dswharshit.medium.com/the-best-llm-for-content-creation-06dd1ee5d7b9

[2] *Intensed*, "Claude 3 vs GPT-4 vs Gemini: Which LLM Is Better?" Intensed, 2024. [Online]. Available: https://www.intensed.com/claude-3-vs-gpt-4-vs-gemini

[3] "AI vs. Human Content: Which Should You Use to Boost Traffic?" *GravityWrite*, 2024. [Online]. Available: https://gravitywrite.com/blog/ai-vs-human-content-the-role-of-ai-in-content-creation

[4] "The Good, Bad, and the Ugly: AI Writing vs. Human Writing," *WSI World*, 2024. [Online]. Available: https://www.wsiworld.com/blog/the-good-bad-and-the-ugly-ai-writing-vs.-human-writing

[5] "AI vs. Human – Who Is the Better Writer?" *Clicta Digital*, 2024. [Online]. Available: https://clictadigital.com/ai-vs-human-who-is-the-better-writer/

[6] "AI Content vs. Human Written Content," *SiteCentre*, 2024. [Online]. Available: https://www.sitecentre.com.au/blog/ai-content-vs-human-written-content

[7] O. W. Ijibadejo and M. Altamimi, "Large Language Model for Creative Writing and Article Generation," *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/380262942

[8] M. Ismayilzada, C. Stevenson, and L. van der Plas, "Evaluating Creative Short Story Generation in Humans and Large Language Models," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/html/2411.02316v4

[9] M. Roemmele and A. S. Gordon, "Creative Help: A Story Writing Assistant," in *Interactive Storytelling*, H. Schoenau-Fog et al., Eds. Cham: Springer, 2015, pp. 81–92. [Online]. Available: https://doi.org/10.1007/978-3-319-27036-4_8

[10] *Type AI Blog*, "Who Wrote It Better? A Definitive Guide to Claude vs. ChatGPT vs. Gemini," 2024. [Online]. Available: https://blog.type.ai/post/claude-vs-gpt

[11] Aha Media Group. Human-Written vs. AI-Generated Healthcare Content. https://ahamediagroup.com/blog/human-written-vs-ai-generated-healthcare-content/