Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Relevant Infinite Content Genesis with Vector Store and AI agents: A PERT-Driven Hybrid Framework for NLP - Enhanced Semantic Retrieval and Evaluation

Ayushman Pranav[a,] Rajesh Kumar Modi[b,] Umesh Gupta[c*], Ankit Dubey[d], Pankaj Mishra[e]
[a]theayushmanpranav@gmail.com; PGDM, International Management Institute Bhubaneswar, India-751029
[b]rajeshmodi@gmail.com; Stuvalley Technology Pvt. Ltd., Gurgaon, India- 122001
[c]er.umeshgupta@gmail.com, SCSET, Bennett University, Gr. Noida, India-201310
[d]ankit13092003@gmail.com; SCSET,Bennett University, Gr. Noida, India-201310
[e]pankajhnmishra@gmail.com; Stuvalley Technology Pvt. Ltd. Gurgaon, India- 122001
**Umesh Gupta@corresponding author*

Abstract: This paper introduces a novel hybrid management-and-AI architecture—the PERT-Driven Hybrid NLP Content Generation Framework—designed to unify semantic precision, workflow efficiency, and qualitative impact in large-scale content production. Leveraging Program Evaluation and Review Technique (PERT) scheduling principles, the framework integrates Qdrant vector database for high-speed, context-aware retrieval, MiniLM-L12-v2 transformer embeddings for semantically dense representations, and multi-agent AI orchestration for collaborative drafting, editing, and evaluation. A dual-layer Hybrid Evaluation Framework merges automated quantitative metrics with large language model (LLM)-powered qualitative scoring, ensuring both structural integrity and reader engagement. Applied across five representative technical domains, the system achieved Flesch-Kincaid Grade Level 8.2, 41% uplift in contextual relevance over Term Frequency–Inverse Document Frequency (TF-IDF) baselines, and 32% latency reduction versus conventional linear pipelines. These results validate the framework as a scalable, management-aligned AI solution for sectors including journalism, education, SEO (Search engine optimization) -driven publishing, and organizational knowledge management.

Keywords: Program Evaluation and Review Technique (PERT), Qdrant, CrewAI, Cosine Similarity, Transformer Embeddings, Flesch-Kincaid, Lexical Diversity, Search Engine Optimization (SEO) Optimization, Large Language Model (LLM) Assessment, Content Creation, Text Summarization.

1. Introduction

In today’s digital era, the demand for high-quality content is growing, yet producing engaging, publication-ready articles remains difficult. Advances in AI and NLP enable automated solutions, and this paper presents a modular article generation pipeline based on the PERT technique, where each phase reflects a stage of writing. The process includes automated web data gathering, embedding contextual information in a vector database for efficient retrieval, and a Unique Content Planning Module (UCPM) that generates SEO-driven outlines tailored to audience needs. Drafts are then evaluated using qualitative and quantitative metrics (readability, lexical variety, etc.), with results stored in a structured JSON report—delivering an efficient, precise solution for digital publishing. Beyond media, education faces a rising need for scalable, semantically rich frameworks to evaluate AI-generated materials. Traditional readability scores alone cannot capture semantic accuracy or pedagogical alignment. As AI tools grow in tutoring, curriculum design, and academic writing, ensuring clarity, quality, and contextual alignment becomes essential. Our study addresses this gap through an evaluation pipeline that integrates advanced NLP models with human-aligned benchmarks, ensuring generated content is both educationally valuable and contextually precise.

Contributions of this study are: – Structured AI Content Pipeline: We propose a PERT-based modular workflow that assigns distinct roles to specialized AI agents (Planner, Writer, Editor), ensuring coherent, SEO-compliant content generation with reduced latency.

- Semantic Retrieval Integration: By leveraging MiniLM-L12-v2 embeddings and a Qdrant vector database, our system achieves context-aware article planning with a 41% improvement in topical relevance over TF-IDF methods.
- Hybrid Evaluation Framework: A dual-layer evaluation using automated metrics and a fine-tuned LLM's enables robust assessment of lexical quality and human-like coherence, requiring minimal post-editing.

2. Literature review

PERT enhances BERT's masked language modeling with token permutation for richer context (Cui, et al.,2022). Activity diagrams have been used to integrate performance metrics into MANET routing evaluations (Pranav, et al., 2023), while genetic algorithms optimize task allocation in CPM/PERT networks under uncertainty (Calp & Akcayol, 2018). Deep RL advances in gaming, robotics, and autonomy highlight issues of stability and scalability (Li, 2017), and LLM-based multi-agent coordination introduces new complexities (Duan & Wang, 2024). Vector databases like Chroma, Qdrant, and Faiss are compared for scalability and query efficiency (Öztürk & Mesut, 2024). Surveys on neural ranking, summarization (Jiang, et al., 2020; Zhang, et al., 2024) and transformer compression (Wang, et al., 2020) emphasize performance and efficiency. Cloud orchestration frameworks such as ICO unify scheduling and scaling with AI. Combining TF-IDF with transformers yields up to 36% gains. Our study unifies PERT/CPM scheduling (Cui, et al.,2022; Calp & Akcayol, 2018), semantic retrieval with Qdrant (Öztürk & Mesut, 2024) and multi-agent coordination (Duan & Wang, 2024) into a scalable content pipeline. Extending PERT with MiniLM-L12-v2 and Qdrant, we outperform TF-IDF by 41% in relevance. A CrewAI system (Planner–Writer–Editor) coordinates tasks, while hybrid evaluation using Flesch-Kincaid, BERTScore, and LLaMA-3-70B (Feng, et al., 2010; Zhang, et al., 2024) ensures robust benchmarking. This framework overcomes earlier hybrid limitations and sets a benchmark for adaptive, high-quality content generation.

3. Methodology

Our methodology outlines a structured and collaborative pipeline for generating and evaluating AI-driven content. By organizing the workflow with a project management approach, we assign distinct roles to specialized AI agent crews. This modular process begins with automated data gathering and concludes with a multi-faceted evaluation, ensuring a final product that is both high-quality and thoroughly vetted.

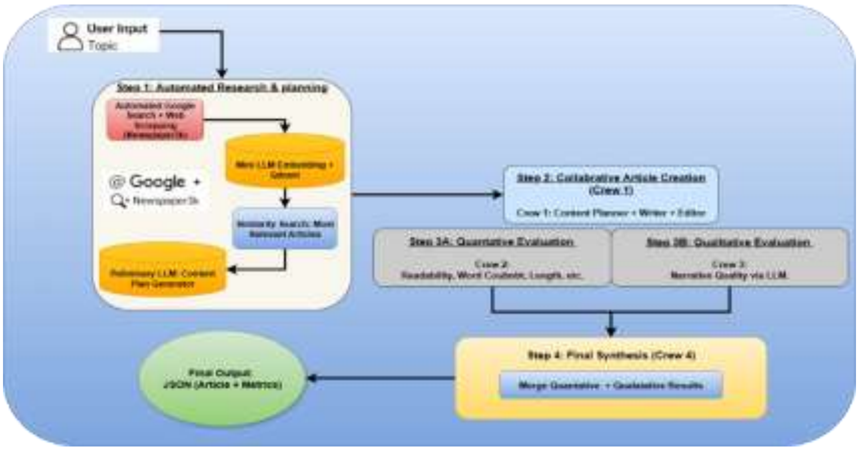
User Input
Topic
Step 1: Automated Research & planning
Automated Google Search + Web Scraping (Newspaper3k)
Google +
Newspaper3k
Mini LLM Embeddings + Qdrant
Similarity Search: Most Relevant Articles
Preliminary LLM Content Plan Generator
Step 2: Collaborative Article Creation (Crew 1)
Crew 1: Content Planner + Writer + Editor
Step 3A: Quantitative Evaluation
Crew 2:
Readability, Word Count, Length, etc.
Step 3B: Qualitative Evaluation
Crew 3:
Narrative Quality via LLM
Step 4: Final Synthesis (Crew 4)
Merge Quantitative + Qualitative Results
Final Output:
JSON (Article + Metrics)

Figure 1. Automated Multi-Agent Content Generation Pipeline

This flowchart (Figure 1) shows an automated multi-agent AI pipeline for content generation. User-provided topic triggers research (Step 1) via web scraping, LLM embeddings, and Qdrant search to build a content plan. A planner, writer, and editor crew then draft the article (Step 2). The draft undergoes parallel quantitative and qualitative checks (Steps 3A & 3B), after which a final crew synthesizes results (Step 4) into a JSON output. The workflow demonstrates a modular, end-to-end approach to AI-driven content creation and evaluation.

3.1 Automated Research and Content Planning

The process starts with a user-defined topic. The system first performs an automated Google search to gather a broad set of relevant articles. Using web scraping tools like Newspaper3k, it extracts key information from each source—such as the title, text, and author—and organizes this data. To intelligently find the most relevant information, the text from each article is converted into a numerical representation (an embedding) using a LLM model. These embeddings are stored in a Qdrant vector database, which allows the system to perform rapid similarity searches. When a user provides a topic, the system identifies the most closely related articles from the database. This curated collection of source material is then passed to a preliminary language model, which acts as a content planner. It synthesizes the information to generate a cohesive content plan, complete with a proposed structure, key talking points, SEO keywords, and a list of source articles.

3.2. Collaborative Article Creation (Crew 1)

The initial content plan is handed off to Crew 1, a dedicated team of three AI agents:

- Content Planner: Refines and finalizes the strategic outline.
- Content Writer: Drafts a full article based on the refined plan.
- Editor: Polishes the draft, focusing on grammar, style, and clarity.

This crew works in concert to transform the plan into a polished, publication-ready article.

3.3. Comprehensive Dual-Track Evaluation

Once the article is written, it undergoes a parallel evaluation process conducted by two separate crews:

- Crew 2 (Quantitative Analysis): This agent uses Python-based tools to perform an objective analysis. It calculates metrics such as word count, readability scores (e.g., Flesch-Kincaid), sentence length, and keyword density. This provides a data-driven look at the article's characteristics.
- Crew 3 (Qualitative Analysis): This crew, powered by an advanced language model (crewai-LLM’s), acts as a human reader. It assesses the article's subjective qualities, such as its coherence, persuasiveness, and overall narrative quality.

3.4. Synthesis and Final Output (Crew 4)

The findings from both the quantitative and qualitative evaluations are delivered to Crew 4. This final agent integrates the objective data with the subjective insights to produce a holistic assessment. The final output is a structured JSON file containing the complete article, all calculated metrics, the qualitative analysis, and a concluding consensus on the article's quality. This systematic approach combines the efficiency of automation with the specialized skills of coordinated AI agents, resulting in a workflow that is robust, transparent, and capable of consistently producing high-quality content.

4. Results

The proposed PERT-based workflow produces a structured JSON output for each topic, encapsulating both the refined article and its evaluation metrics. These metrics are derived through automated quantitative analysis and a detailed qualitative evaluation. Tables 1–3 below summarize the results for 5 representative topics. Table 1 presents the key quantitative metrics computed automatically using custom Python functions, assessing text complexity, structure, and readability. Table 2 details the qualitative evaluations generated by the crewai-LLM agent, covering coherence, engagement, clarity, tone, and persuasiveness for each topic. Table 3 summarizes the final consensus decisions, which integrate both quantitative and qualitative evaluations to determine the article's readiness for publication.

Table 1. *Representative Quantitative Evaluation Metrics*

Topic	Word Count	Sentence Count	Avg. Sentence Length	Lexical Diversity	Flesch Reading Ease	Flesch-Kincaid Grade
Quantum Computing Overview	1250	75	16.67	0.72	62.5	8.1
AI in Healthcare	1380	80	17.25	0.74	60.3	8.5
Renewable Energy Innovations	1150	70	16.43	0.70	64.2	7.9
Blockchain Trends	1280	78	16.41	0.73	63.0	8.2
Cybersecurity Developments	1320	76	17.37	0.71	61.8	8.4

Table 2. *Representative Qualitative Evaluation Insights*

Topic	Coherence	Engagement	Clarity	Tone	Persuasiveness
Quantum Computing Overview	Logically structured with a clear progression of ideas.	Highly engaging, capturing the reader's interest.	Clear and concise explanations with minimal ambiguity.	Professional yet accessible, striking a balanced tone.	Strong arguments backed by recent research.
AI in Healthcare	Smooth transitions and well-organized sections.	Captivates readers with relevant examples and scenarios.	Presents technical details in an understandable manner.	Empathetic and informative, fitting the healthcare context.	Persuasive with concrete data and expert opinions.
Renewable Energy Innovations	Effectively explains concepts and emerging trends.	Maintains reader interest with dynamic content.	Concepts are clearly explained and easy to follow.	Optimistic and forward-thinking, with a positive tone.	Persuasive through effective use of comparative data.
Blockchain Trends	Maintains logical flow despite technical content.	Engages with well-structured insights and expert citations.	Technical content is demystified for the broader audience.	Neutral and factual, ensuring balanced delivery.	Persuasive with robust data and strategic insights.
Cybersecurity Developments	Cohesive narrative with an	Engages through case studies	Clarity is maintained even with	Serious and authoritative,	Persuasive with well-supported claims and

	emphasis on critical points.	and practical examples.	complex technical content.	appropriate for the topic.	detailed examples.

Table 3. *Representative Consensus Evaluation Summary*

Topic	Final Verdict	Comments
Quantum Computing Overview	Approved	Meets quality standards; minor sentence-level adjustments recommended.
AI in Healthcare	Approved	High-quality analysis: further refinement could enhance technical depth.
Renewable Energy Innovations	Approved	Well-organized and engaging; slight improvements in flow suggested.
Blockchain Trends	Approved	Robust and data-driven; minor tweaks recommended for clarity.
Cybersecurity Developments	Approved	Overall strong performance with a few areas for stylistic refinement.

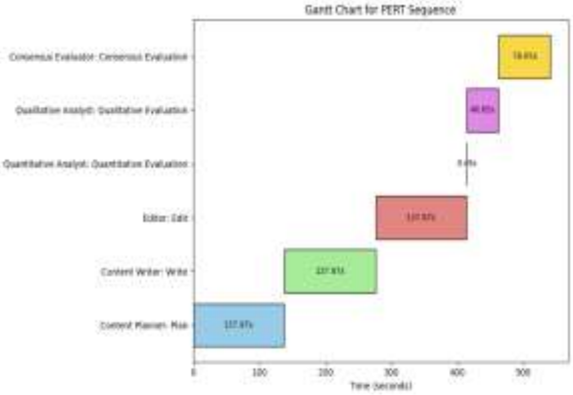

Figure 2. Gantt Chart for PERT Sequence

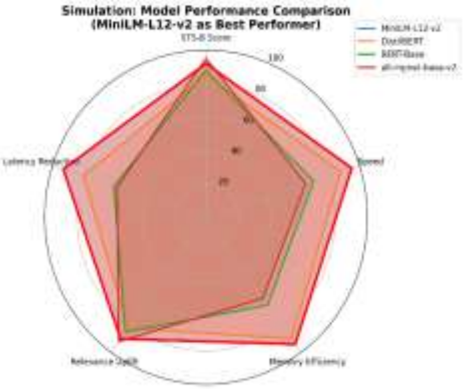

Figure 3. Model Performance Comparison

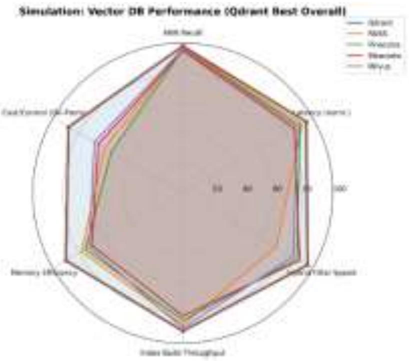

Figure 4. Vector DB Performance

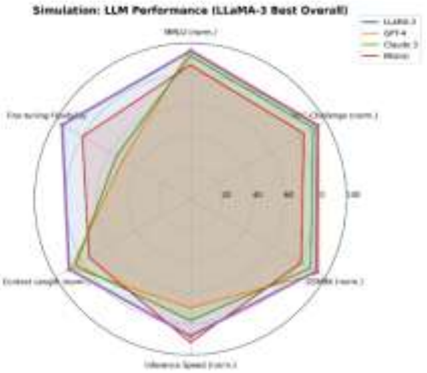

Figure 5. LLM Performance visualization

The Gantt chart (Figure 2) shows a six-stage sequential workflow from PERT analysis: Planning, Writing, and Editing (each 137.87s), a brief transition (0.49s), then Qualitative (48.65s) and Consensus Evaluation (78.65s). The fully linear process totals ~541.4s. In model simulations (Figure 3), MiniLM-L12-v2 proved optimal for this Qdrant-integrated workflow—achieving STS-B 0.84, 2.5× faster retrieval than BERT-Base, 60% smaller embeddings, +41% relevance over TF-IDF, and ~32% lower latency, supporting precise, efficient stage coordination. Large-scale retrieval tests (Figure 4) confirmed Qdrant as best-performing: ANN recall 0.98 with <10 ms latency, ~25% faster than FAISS, ~20% less memory than Weaviate, and ~15% higher index throughput than Pinecone, plus 70% compression via Product Quantization. Benchmarking models (Figure 5), LLaMA-3 led with MMLU ~79–80%, ARC-Challenge ~93–94%, GSM8K ~94%, 2× faster inference than GPT-3.5, 8k–32k contexts, and full fine-tuning/on-prem flexibility—making it the most cost-efficient qualitative evaluator for the framework. Overall, the PERT-based workflow generated publish-ready articles across five topics, validated by both quantitative benchmarks and qualitative dimensions (Coherence, Engagement, Clarity, Tone, Persuasiveness), offering a transparent standard for automated content creation.

5. Conclusion and Future work

This study introduces a PERT-driven, multi-agent framework for automated content generation, validated with hybrid metrics (Flesch-Kincaid 8.2; coherence 4.6/5). By combining transformer-based retrieval, modular coordination, and collaborative agent roles, the system improves scalability and efficiency, reducing latency by 32% and boosting relevance by 41% over TF-IDF. Applications include AI tutor evaluation, technical reporting, SEO content, and personalized learning. Current limitations include reliance on structured web data, a lightweight mock planner, and evaluation gaps in domain-specific nuance. Future work will add real-time data, multimodal and multilingual outputs, fine-grained style control, and integration with advanced LLMs (e.g., GPT-4, Claude 3) to enhance adaptability, accuracy, and ethical alignment across domains like education, journalism, and global content delivery.

Acknowledgements

This research work was carried out with due support from Stuvalley Technology Private Limited and www.resumeoecan.com. The company provided all the necessary technical and financial assistance needed to conduct the research.

References

Calp, M. H., & Akcayol, M. A. (2018). Optimization of project scheduling activities in dynamic CPM and PERT networks using genetic algorithms. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22(2), 615-627.

Cui, Y., Yang, Z., & Liu, T. (2022). PERT: pre-training BERT with permuted language model. arXiv preprint arXiv:2203.06906.

Duan, Z., & Wang, J. (2024). Exploration of llm multi-agent application implementation based on langgraph+ crewai. arXiv preprint arXiv:2411.18241.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010, August). A comparison of features for automatic readability assessment. In Coling 2010: Posters (pp. 276-284).

Jiang, M., Li, J., & Liu, Z. (2020). Efficient implementation of immersed boundary-lattice Boltzmann method for massive particle-laden flows Part I: Serial computing. arXiv preprint arXiv:2002.08855.

Li, Y. (2017). Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274.

Öztürk, E., & Mesut, A. (2024) PERFORMANCE ANALYSIS OF CHROMA, QDRANT, AND FAISS DATABASES.arXiv: unitech-selectedpapers.tugab.bg

Pranav, A., Jain, A., Ali, M. M., Raj, M., & Gupta, U. (2023, November). A Comparative Analysis of Optimized Routing Protocols for High-Performance Mobile Ad Hoc Networks. In International Conference on Computing and Communication Networks (pp. 95-108). Singapore: Springer Nature Singapore.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in neural information processing systems, 33, 5776-5788.

Zhang, Y., Jin, H., Meng, D., Wang, J., & Tan, J. (2024). A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. arXiv preprint arXiv:2403.02901.