

# OPEN-SOURCE AI PIPELINE FOR AUTOMATED SHORT-FORM VIDEO CREATION: A PROOF-OF-CONCEPT STUDY

Author: Marieta Marinova, Miroslav Georgiev, Delyan Dimitrov,  
Linlin Ma

*Abstract: The proliferation of short-form video platforms creates unprecedented demand for content automation. Manual production requires 2-4 hours per video, while commercial platforms impose subscription costs of €20-47 monthly. This study presents ClipClap Factory, an open-source automation pipeline employing n8n workflow orchestration and local LLM inference (Google Gemma 3 12b). Proof-of-concept deployment on Instagram validated the system through 112 workflow executions during summer 2025. Key findings demonstrate: (1) 67% success rate with 75 published videos; (2) €0.35 per video production cost representing 99% cost reduction versus manual production; (3) 4-7 minute production velocity; (4) 5-second square-format videos (1024×1024) optimized for Instagram. Audio truncation occurred in 15% of videos when narration exceeded the 5-second Freepik API constraint. Visual-semantic misalignment appeared in 20% due to unpredictable AI generation behavior. The deployment prioritized technical validation over audience optimization; engagement metrics (peak: 383 views, 21 followers) represent incidental outcomes. The hybrid architecture (local LLM with selective commercial API integration) establishes a cost-optimized pattern for proof-of-concept workflow validation. Future work includes script length optimization and multiplatform expansion.*

*Keywords: workflow automation; open-source AI; video generation; large language models; cost optimization*

*JEL: O33*

## 1. INTRODUCTION

### 1.1 Purpose and motivation for workflow automation in short-form video

The contemporary digital media ecosystem is characterized by an escalating reliance on short-form video as a primary medium for content consumption, marketing, and brand engagement. Instagram Reels, TikTok, and YouTube Shorts now draw billions of users every day. This surge has sparked an intense need for frequent, quality video content. What

began as casual entertainment has become essential infrastructure for journalism, marketing, education, and public relations professionals.

Organizations across all sectors face immense pressure to meet this continuous, algorithmically-driven demand for fresh content. Yet the conventional approach to video production remains time-consuming and expensive. Creating content manually demands scriptwriting, finding visual assets, editing footage, rendering final cuts, and managing distribution across platforms. Professional creators typically spend two to four hours on each short video, handling everything from writing and asset curation to voice work, editing, and format optimization. When calculated at typical freelance rates of €25-50 hourly, production costs reach €50-200 per video. This pricing creates a fundamental scalability problem, particularly for small and medium-sized enterprises (SMEs), independent content creators, and resource-constrained organizations.

## **1.2 The automation gap in short-form video production**

Market analysts project the short-form video sector will hit USD 300 billion by 2027, driven primarily by mobile-first consumption and algorithmic content distribution. Despite this explosive market growth, most content creation still relies on manual, resource-heavy processes. This gap presents a real challenge: while businesses understand they need regular video output, few can afford sustainable production methods at scale.

Several commercial platforms have attempted to bridge this gap, offering subscription services for €20-47 monthly. However, these solutions impose restrictions on output volume, customization options, and data ownership. More concerning is their opaque nature - most function as closed systems offering little visibility into their algorithms, data practices, or content generation methods. For organizations prioritizing data sovereignty, intellectual property protection, or full operational control - such as news organizations, financial institutions, or healthcare providers under regulatory compliance requirements - these proprietary solutions are structurally inadequate.

Open-source software offers a different path forward. Recent breakthroughs in locally-run Large Language Models, especially via frameworks like Ollama, have made powerful text generation accessible without cloud dependencies or sharing data with external services. Workflow tools like n8n let organizations build custom automation pipelines where every component remains visible and modifiable. Yet a critical question persists: can these open-source tools actually work together reliably and economically for real video production? Academic research hasn't answered this yet.

## **1.3 Research value, project hypothesis, and logic of the solution**

The core challenge addressed by this research is the conflict between the requirement for scalable, cost-effective video production and the limitations imposed by both manual workflows and existing proprietary automation tools. This study bridges the critical gap between theoretical possibility and demonstrated feasibility by providing empirical validation of an open-source video automation architecture under real-world deployment conditions.

This study provides crucial empirical validation for the economic viability and technical feasibility of open-source AI-driven video automation. By demonstrating that localized LLM

inference (via Ollama) combined with modular workflow orchestration (via n8n) can achieve per-video production costs under €1 while maintaining production velocity of 4-7 minutes per video, this research directly addresses the accessibility barrier faced by smaller organizations and independent creators. The work contributes empirically-grounded evidence to inform build-versus-buy decisions in content automation technology adoption.

**Positioning and Scope:** As an exploratory proof-of-concept deployment, this study focuses on validating architectural feasibility, economic efficiency, and operational reliability under real-world conditions rather than achieving broadcast-quality production standards. The research deliberately targets high-volume, algorithmically-distributed content scenarios where "good enough" quality at scale provides greater strategic value than perfection at limited volume. This positioning acknowledges the fundamental tension between quality and quantity in algorithmic content distribution, where consistent posting frequency demonstrably outperforms sporadic premium content in platform visibility algorithms.

The central Project Hypothesis posits that an open-source, modular workflow orchestration system integrating local Large Language Model (LLM) inference with commercial multimedia APIs can achieve production velocity and cost efficiency competitive with - or superior to - proprietary SaaS platforms, while maintaining full operational transparency and data sovereignty. This hypothesis challenges the prevailing assumption that high-quality content automation necessarily requires expensive proprietary platforms or cloud-based AI services.

The Logic of the Solution is built upon architectural modularity and strategic cost optimization through a hybrid approach. By decoupling cognitive operations (performed locally via Ollama) from specialized synthesis tasks (delegated to commercial APIs such as ElevenLabs for text-to-speech and Freepik for image-to-video generation), the system minimizes variable costs while preserving customization flexibility. This architectural pattern enables organizations to retain control over semantic content generation - headlines, scripts, and visual prompts - while leveraging commodity services only for lower-level rendering tasks. The economic advantage stems from converting high-cost per-unit cloud LLM inference into zero-marginal-cost local inference, reserving API expenditure exclusively for multimedia synthesis where open-source alternatives remain technically immature.

#### **1.4 Research questions**

This study systematically addresses three interconnected research questions that collectively evaluate the practical viability of open-source video automation:

**RQ1: Production Efficiency and Economic Viability** What production velocity (time per video) and total cost of ownership (TCO per video) can an open-source, hybrid automation pipeline achieve compared to manual production and commercial platforms?

**RQ2: Reliability and Quality Under Real-World Conditions** What reliability challenges (failure rates, error types) and quality issues arise during sustained real-world deployment, and what percentage of generated content meets publishable standards without human editing?

**RQ3: Validated Business Applications and Market Positioning** Under which business conditions and use case scenarios is such a system practically viable - as a primary

production tool, supplementary content filler, or prototyping mechanism - and for which organizational profiles is the approach demonstrably unsuitable?

These research questions are deliberately grounded in pragmatic deployment considerations rather than purely technical performance metrics, reflecting the study's applied research orientation toward informing real-world technology adoption decisions.

### **1.5 Contributions**

This research makes three distinct scholarly contributions:

**Empirical Contribution:** This study provides the first peer-reviewed empirical validation of an open-source AI video automation pipeline deployed under real-world conditions. Through 112 workflow executions over the testing period resulting in 75 published Instagram videos, this research demonstrates achieved performance metrics: 67% success rate, €0.35 average TCO per video, 4-7 minute production velocity, and observable content engagement variance (10-383 views per video). These empirical findings establish a validated baseline for open-source automation performance against which future research can be benchmarked.

**Architectural Contribution:** The study validates a novel hybrid architectural pattern combining local LLM inference (Ollama) with selective commercial API integration (ElevenLabs, Freepik, Instagram Graph API) orchestrated through open-source workflow automation (n8n). This architecture demonstrates that organizations can achieve cost-competitive automation while maintaining data sovereignty over semantic content, addressing a critical gap in existing proprietary solutions.

**Practical Contribution:** By providing reproducible technical documentation, transparent cost analysis, and honest assessment of failure modes and quality limitations, this research enables organizations to make informed build-versus-buy decisions regarding content automation technology adoption. The study explicitly identifies validated use cases (regional news aggregation, content filler strategies) and anti-use cases (high-stakes brand communication, emotionally nuanced content) with empirical support.

### **1.6 Structure of the paper**

The remainder of this paper is structured as follows. Section 2 reviews related work spanning short-form video production dynamics, commercial and open-source automation platforms, and theoretical frameworks for technology adoption in content production. Section 3 describes the system architecture, workflow implementation, and evaluation methodology employed during the proof-of-concept deployment. Section 4 presents empirical results including operational performance metrics, reliability analysis, Instagram deployment outcomes, and quality assessment findings. Section 5 discusses strategic implications, validated business applications, market positioning, and limitations. Section 6 identifies threats to validity and methodological constraints. Section 7 concludes with theoretical and practical implications, future research directions, and recommendations for organizations considering similar automation initiatives.

## **2. LITERATURE REVIEW: THE CONVERGENCE OF MEDIA PRODUCTION AND DECENTRALIZED AUTOMATION**

This critical review synthesizes current trends in digital media production, the necessity of workflow automation, the evolution of open-source AI infrastructure, and theoretical frameworks for technology adoption in content generation. By examining these interconnected domains, we identify a critical research gap: the absence of empirically-validated, open-source alternatives to proprietary video automation platforms with documented performance under real-world deployment conditions.

### **2.1 Market dynamics and strategic importance of short-form video**

The global market for short-form video is experiencing substantial, non-linear growth driven by platform algorithm optimization and shifting user preferences toward mobile-first, attention-optimized content. Industry analyses project the short-form video market to reach USD 300 billion by 2027 (Market Research Future, 2025), with platforms reporting engagement rates 2.5× higher than traditional long-form content (Reddy, 2025). This growth is not merely quantitative but represents a fundamental structural shift in digital communication: short-form video has evolved from entertainment medium to mission-critical business infrastructure across journalism, marketing, education, and corporate communications (IFTTT, 2025).

The strategic imperative for consistent video presence is reinforced by platform algorithm behavior. Research on social media algorithmic distribution demonstrates that posting frequency significantly impacts content reach, with accounts publishing daily achieving 3-4× greater organic visibility compared to sporadic posting patterns (Riley, 2025). This algorithmic reality creates a volume-driven competitive dynamic where production capacity becomes a strategic differentiator.

Empirical studies of AI-assisted short-form video generation show that automated tools can accelerate parts of the workflow but still require careful integration into production practices (Thurman et al., 2025). Manual production workflows remain prohibitively expensive: industry benchmarks indicate 2-4 hours of professional labor per video (scriptwriting, asset sourcing, editing, rendering), translating to €50-200 per video at standard freelance rates (Business Research Insights, 2025). At scale requirements of 20-30 videos monthly for algorithmic competitiveness, this corresponds to monthly production costs of €1,000-6,000 - economically infeasible for SMEs and independent creators who constitute the majority of content producers.

This cost structure creates a critical automation gap: organizations recognize strategic value in consistent video presence but lack economically viable mechanisms to achieve necessary volume. Traditional solutions - hiring dedicated video teams or outsourcing to agencies - preserve the linear cost-per-video economics that make scale unattainable for resource-constrained organizations.

### **2.2 Production cost barriers and scalability challenges**

SaaS platforms offering "AI-generated videos from text or links" have multiplied in recent years, including Pictory, Synthesia, Runway, and InVideo (GetMonetizely, 2025; Sangeetha & Suganya, 2025). These platforms typically abstract away infrastructure

complexity, providing user-friendly interfaces for non-technical users. Standard pricing models range from €20-47 monthly for basic tiers, with volume limitations of 10-50 videos monthly and premium tiers exceeding €200-500 monthly for unlimited generation.

While these platforms demonstrate technical feasibility of automated video production, they impose three structural limitations that constrain organizational adoption:

**Economic Constraint:** Subscription models create fixed monthly costs regardless of actual usage, making them economically inefficient for organizations with variable or seasonal content needs. The per-video economics remain opaque, as providers bundle infrastructure, model access, and platform features into flat-rate pricing that obscures true marginal costs.

**Operational Constraint:** Proprietary platforms operate as "black boxes" with minimal transparency regarding underlying models, prompt engineering techniques, or quality control mechanisms (Anvil, 2025). Organizations cannot inspect, modify, or optimize generation logic, preventing customization for domain-specific requirements or brand-specific stylistic preferences.

**Data Sovereignty Constraint:** Leading commercial platforms process all content - including potentially sensitive headlines, scripts, and brand guidelines - through cloud-based infrastructure operated by third-party vendors (AlphaCorp, 2025). For regulated industries (financial services, healthcare, journalism under GDPR), or organizations prioritizing intellectual property protection, this fundamental architectural characteristic represents an insurmountable compliance barrier regardless of contractual data processing agreements.

Research on automated journalism and AI-generated media consistently emphasizes that fully autonomous content generation without human oversight produces unreliable quality (Sharma & Sharma, 2025; Raghunathan, 2025). Thurman et al. (2025) demonstrate through controlled experiments that automated news video production achieves significantly higher quality when human editors perform post-generation curation, a finding consistent across multiple AI content domains. This empirical evidence suggests that commercial platform claims of "fully automated, broadcast-quality" output should be approached with skepticism, and that realistic deployment architectures must incorporate human-augmented workflows.

### **2.3 Commercial automation platforms: capabilities and limitations**

The open-source ecosystem for AI has matured rapidly over the past two years, with local LLM inference emerging as a viable alternative to cloud-based API services. Ollama, released in 2023, provides a lightweight runtime for executing open-source language models (Llama, Mistral, Gemma families) on commodity hardware without requiring specialized ML engineering expertise (Ollama, 2025). This democratization of LLM access fundamentally alters the economics of AI-driven content generation: cloud API calls costing €0.002-0.02 per generation are replaced by zero-marginal-cost local inference after initial hardware investment.

Recent comparative benchmarking demonstrates that mid-tier open-source models (Llama 3 70B, Gemma 3 12B) achieve 85-92% of GPT-4 performance on content generation tasks while operating entirely offline (Vellum, 2025; Zhang et al., 2025b). For content

automation use cases - where perfect linguistic sophistication is less critical than reliable structure and semantic coherence - these open-source models represent a practical quality-cost optimum.

Open-source workflow orchestration platforms, such as n8n, provide the necessary infrastructure for constructing complex, multi-step automation pipelines integrating diverse APIs and local services (CrossTech Communications, 2025; Project Aeon, 2025). Unlike proprietary iPaaS platforms (Zapier, Make), n8n offers complete architectural transparency, enabling organizations to audit, modify, and extend workflows without vendor dependency. The platform's node-based visual programming paradigm reduces technical barriers while preserving programmatic flexibility through JavaScript execution nodes.

The convergence of local LLM inference (Ollama) and open-source orchestration (n8n) creates architectural conditions for fully transparent, customizable automation pipelines. However, academic literature lacks empirical validation of this architectural pattern applied to video generation at scale.

While open-source alternatives exist for many AI tasks, certain specialized synthesis operations - particularly text-to-speech (TTS) and image-to-video generation - remain dominated by commercial APIs due to superior quality and reliability (Yu et al., 2025). Open-source TTS models (Coqui, Piper) exhibit noticeable synthetic artifacts and limited voice naturalness compared to commercial leaders like ElevenLabs. Similarly, open-source video generation models (Stable Video Diffusion, ModelScope) require substantial computational resources and produce inconsistent quality unsuitable for production deployment.

This quality gap motivates a hybrid architectural strategy: perform cognitive operations (headline selection, script generation, prompt engineering) locally via open-source LLMs to minimize variable costs and preserve data sovereignty, while delegating specialized synthesis tasks to commercial APIs where open-source alternatives remain technically immature. This architecture minimizes API expenditure by restricting commercial service usage to final rendering stages rather than entire generation pipelines.

## **2.4 Platform landscape and economic models**

The build-versus-buy decision framework in information systems research emphasizes Total Cost of Ownership (TCO) as the primary economic criterion, encompassing not only direct licensing or development costs but also operational expenses, maintenance burden, and opportunity costs (First Derivative, 2025; Shaham, 2025). For content automation, TCO must account for infrastructure investment (hardware for local inference), ongoing API costs (per-video variable costs), and human labor for quality curation.

Existing literature on open-source versus proprietary software demonstrates that open-source solutions typically exhibit higher upfront investment (infrastructure, technical expertise) but lower long-run variable costs, making them economically superior at scale (Sangeetha & Suganya, 2025; Getmonetizely, 2025). However, this theoretical advantage has not been empirically validated in the video automation domain with real-world performance data.

Research on automated journalism and AI-generated media converges on the principle of human-augmented automation: systems designed for human-AI collaboration consistently outperform fully autonomous approaches in quality, accuracy, and audience satisfaction (Thurman et al., 2025; Sharma & Sharma, 2025). This finding challenges the technological solutionism narrative that frames full automation as the ideal endpoint. Instead, optimal production architectures position AI as a high-volume content generator subject to human editorial curation, combining machine efficiency with human judgment.

The human-augmented paradigm has significant implications for TCO calculations: if 60-70% of AI-generated outputs meet publishable standards without editing (as informal evidence suggests), the effective cost-per-published-video must account for generation waste. A system producing content at €0.35 per video with 67% success rate achieves an effective cost of €0.51 per published video - still dramatically lower than manual production but higher than naive per-execution costs suggest.

## **2.5 Research gap: Empirical validation under real-world conditions**

Across these literature streams, three critical gaps emerge that this study addresses:

Gap 1: Absence of Open-Source Video Automation Validation While theoretical arguments favor open-source approaches for cost and transparency reasons, academic literature lacks empirical studies validating open-source video generation pipelines under sustained, real-world deployment. Existing research focuses on proprietary platforms (Pictory, Synthesia) or laboratory prototypes without production testing. No prior study documents achieved TCO, reliability metrics, or quality assessment for open-source architectures over multi-week deployments.

Gap 2: Lack of Hybrid Architecture Performance Data The proposed hybrid model - local LLM inference combined with selective commercial API usage - represents a novel architectural pattern not empirically evaluated in prior literature. Research examines either fully proprietary cloud solutions or fully open-source local pipelines, but not strategic combinations optimizing for cost-quality tradeoffs. The economic viability and technical reliability of this hybrid approach remain unvalidated.

Gap 3: Missing Real-World Quality and Reliability Analysis Existing literature on AI content generation predominantly reports controlled experiment results or vendor-provided benchmarks, lacking transparent reporting of failure modes, quality variability, and operational challenges encountered during extended real-world deployment. The absence of honest, empirically-grounded discussion of automation limitations - including failure rates, error types, and quality inconsistencies - prevents organizations from making informed technology adoption decisions.

This study directly addresses these gaps by deploying an open-source, hybrid automation pipeline (combining n8n, Ollama, and commercial APIs) for the testing period, generating 75 published Instagram videos, and systematically documenting achieved performance, encountered failures, and observed quality issues. By providing transparent empirical evidence - including negative results and limitations - this research establishes a validated baseline for open-source video automation against which future work can be benchmarked.

### **3. METHODOLOGY: ARCHITECTURAL DESIGN AND WORKFLOW IMPLEMENTATION**

The methodological approach utilized a modular, containerized architecture designed for cost efficiency, reproducibility, and control. The entire workflow is defined and executed using n8n for orchestration, with Docker providing isolated execution environments.

#### **3.1 System overview: modular architecture and technology stack**

The pipeline is a hybrid solution, integrating robust open-source components for cost-sensitive core functions and select commercial APIs for specialized generative synthesis.

The architecture relies on four core open-source technologies: n8n for orchestration; Docker for containerization; Ollama running Gemma 3 12b for local, zero-cost AI inference; and the custom-made ffmpeg-api wrapper for video compositing. This custom API wrapper was developed specifically to orchestrate the open-source ffmpeg tool, which is used to merge and synchronize the audio and video files efficiently. Furthermore, the system integrates mitmproxy, an internal tool used for monitoring and debugging all HTTP requests, which was essential during the validation phase to identify latency and external API failure points. Commercial APIs, Elevenlabs for high-fidelity Text-to-Speech and Freepik for Image-to-Video generation, are integrated for quality synthesis.

#### **3.2 Detailed, step-based workflow description**

The ClipClap Factory system implements an eight-stage automated pipeline orchestrated through n8n workflow engine, combining local open-source components with selective commercial API integration. Each execution progresses sequentially through the following operational stages:

**Stage 1 - Content Discovery:** The workflow initiates via RSS feed monitoring of dnevnik.bg news portal. The n8n RSS trigger node polls the feed at configurable intervals (default: hourly), extracting article metadata including headline, summary text, publication timestamp, and source URL. Successfully retrieved articles trigger downstream processing.

**Stage 2 - Script Generation:** Article content passes to local Large Language Model inference via Ollama runtime executing Google Gemma 3 12b model. The LLM receives structured prompts instructing generation of concise video narration scripts optimized for short-form social media consumption. Script output undergoes validation for length constraints and content coherence before proceeding.

**Stage 3 - Visual Prompt Engineering:** The generated script returns to the local LLM with specialized prompt templates requesting two distinct outputs: (a) detailed text-to-image generation prompt conforming to Imagen 3 specifications (descriptive, compositionally structured, cinematically styled), and (b) image-to-video motion prompt defining camera movements, scene dynamics, and temporal progression for Kling v2 processing. Prompt quality directly influences downstream media synthesis success rates.

**Stage 4 - Parallel Media Synthesis:** Two concurrent operations execute independently: (a) Text-to-Speech conversion via ElevenLabs Eleven Multilingual v2 API transforms the script into natural-sounding narration audio (MP3 format, 128kbps), while (b) Text-to-Image generation via Freepik API (utilizing Google Imagen 3 backend) synthesizes source imagery (1024×1024 resolution, PNG format). Both operations complete asynchronously; the workflow proceeds only upon successful completion of both branches.

Stage 5 - Image-to-Video Conversion: The synthesized static image (square format) feeds into Freepik's Image-to-Video API (utilizing Kuaishou Kling v2 backend) alongside the previously generated motion prompt. The API generates 5-second video clips in square format. Duration is fixed at 5 seconds due to API pricing structure; 10-second generation incurs significantly higher per-video costs. Motion prompt interpretation by Kling v2 exhibits unpredictable behavior; resulting video motion shows limited correspondence to prompt specifications, representing a current limitation of commercial AI video generation technology.

Stage 6 - Audio-Visual Compositing: Custom-developed ffmpeg-api wrapper receives both the 5-second video clip and the variable-length audio narration track. The compositor truncates audio to exactly 5 seconds to match video duration, potentially cutting mid-sentence if the generated script narration exceeds this limit. No background music is applied in current implementation. The service outputs video optimized for Instagram posts. Approximately 15% of generated videos experience audio truncation when narration scripts exceed the 5-second constraint.

Stage 7 - Storage and Publication: The final composed video persists to local filesystem via custom file-server API, generating structured metadata records (source article, generation timestamp, component costs, quality flags). Successful storage triggers Instagram Graph API publication with automated caption generation (derived from source article headline) and account credential management. Videos publish in 1:1 aspect ratio without Instagram cropping. Publication targets single platform (Instagram) for proof-of-concept validation; multiplatform expansion represents future work.

Stage 8 - Performance Telemetry: Throughout execution, n8n writes granular operational metrics to structured CSV logs capturing: stage-by-stage latency measurements, success/failure status codes, error messages, and quality assessment flags. This telemetry enables post-hoc analysis of failure pattern recognition, and continuous system optimization.

The complete pipeline executes with median duration 5.4 minutes (range: 4-7 minutes) and average variable cost €0.35 per successful publication, assuming standard API pricing and successful completion of all stages. Failure at any stage triggers configurable retry logic or terminates execution with detailed error logging for manual intervention.

The workflow operates as an autonomous, end-to-end factory. The design relies on the modularity of the orchestration system to ensure each step is traceable.

The system incorporates automated error management and detailed logging as integral parts of the workflow. The Log files (written in Step 8 by the "Write Execution Data to CSV Node"), which store metadata and execution outcomes, are a core advantage for process auditability and automated quality control. They enable detailed analysis of failure causes (e.g., API timeouts, incoherent LLM output) and are central to the error recovery mechanisms. The system utilizes n8n's "Retry on Fail" logic within the orchestration for error mitigation. This configuration is applied to external API calls (specifically Steps 4 and 5, visible in the workflow screenshot) to automatically re-attempt failed requests using conditional branching and exponential backoff. This automated mechanism, informed by the

analysis of execution logs, minimizes manual intervention for transient errors and ensures efficient workflow completion.

The total production time for a single video, from initial idea to public posting, averaged between 4 and 7 minutes. This high velocity is critical for competitive deployment. The efficiency is gained by the fast, zero-cost Script Generation phase (~45 seconds), which offsets the latency of the external Media Creation phase (~2 minutes 30 seconds).

The evaluation comprised 112 total workflow executions, validating the architecture's performance, cost efficiency, and reliability.

### **3.3. Evaluation procedure and deployment windows**

The system was evaluated in two separate deployment windows on a live Instagram account. The first deployment consisted of two consecutive days in June 2025; the second comprised two to three days in September 2025, after audio narration had been added. During the final test run in September, the workflow was scheduled to execute every ten minutes. Because the source RSS feed contained relatively few fresh items at that cadence, the pipeline occasionally generated multiple videos about the same news story.

The n8n workflow implements full hands-free operation with explicit error handling. Each execution starts from a manual or scheduled trigger, collects candidate items from the RSS feed, and then proceeds through script generation, image and video prompting, media generation, audio synthesis, compositing, and file export. At each critical stage, the workflow logs intermediate outputs to disk and maintains dedicated log files for errors. If a node fails, a retry mechanism is triggered; only after repeated failures is the execution written to an error log and marked as unsuccessful. All successful executions automatically proceed to publication: every successfully rendered video was posted to Instagram without manual intervention.

Across both deployment windows, the system attempted 112 workflow executions. Each execution corresponds to a single short-form news video that combines a generated visual clip with a brief spoken narration of the news headline. This fully automated, repeated-sampling setup is comparable to emerging industry tutorials and automation templates that convert RSS feeds or structured inputs into short-form videos using multi-step pipelines and third-party API.

## **4. RESULTS: EMPIRICAL VALIDATION THROUGH REAL-WORLD DEPLOYMENT**

This section presents empirical findings from the proof-of-concept deployment of the Clip Clap Factory system on Instagram (@marivmari2025), encompassing 112 workflow executions from June 27 to September 10, 2025. Results are organized into four subsections: operational performance and cost metrics (4.1), reliability analysis and failure characterization (4.2), real-world deployment outcomes including content performance (4.3), and synthesis of key findings (4.4).

### **4.1 Key operational and cost performance indicators (KPIs)**

The system achieved a success rate of 67% across 112 total workflow executions, resulting in 75 successfully published videos and 37 failures. This success rate, while below commercial platform benchmarks (typically 90-95% advertised), demonstrates technical

feasibility for sustained automated operation under real-world conditions with diverse content inputs and external API dependencies.

The system's performance evaluation encompasses three critical dimensions: production velocity, total cost of ownership (TCO) and comparative economic analysis.

Production velocity mean production time from RSS feed ingestion to Instagram publication was 4-7 minutes per video, with variation attributable to external API response latency rather than orchestration overhead. The workflow demonstrated consistent performance across all 75 successful executions:

- Fastest execution: 3.8 minutes (minimal API queue time)
- Slowest execution: 8.2 minutes (elevated API latency during peak hours)
- Median execution time: 5.4 minutes
- Standard deviation:  $\pm 1.2$  minutes

This production velocity represents a 30-50 $\times$  acceleration compared to manual production benchmarks (2-4 hours per video), validating the core hypothesis that automated pipelines can achieve order-of-magnitude efficiency gains.

The average Total Cost of Ownership (TCO) per successfully published video was €0.35, calculated across all 75 published videos with comprehensive accounting of direct variable costs:

Cost Breakdown per Video:

- ElevenLabs Text-to-Speech API: €0.12 (400 characters @ €0.30/1000 chars)
- Freepik Image-to-Video API: €0.20 (single 5-second generation)
- Instagram Graph API: €0.00 (free tier, <200 posts/hour)
- n8n Orchestration: €0.00 (self-hosted open-source)
- Total Variable Cost: €0.35/video

Infrastructure costs (not included in per-video TCO):

- Ollama server (self-hosted): €0.00 marginal cost (existing hardware)
- GPU inference: €0.00 marginal cost (Gemma 3 12B runs on consumer GPU)
- Electricity:  $\sim$ €0.02/video (estimated 0.1 kWh @ €0.20/kWh)

Fully-loaded TCO including infrastructure amortization: €0.48-0.52/video when accounting for GPU hardware depreciation (€1,500 GPU / 5,000 expected videos / 3-year lifespan).

The comparative cost structure for open-source, commercial, and manual production methods is presented in Tab. 1.

Key Finding: The open-source hybrid architecture achieves 86-91% cost reduction compared to commercial automation platforms and 96-99.3% cost reduction compared to manual production at scale (20+ videos monthly).

### 4.2 Detailed reliability analysis and error mitigation strategies

The system exhibited a failure rate of 33% (37 failures / 112 executions) under fully autonomous operation. Systematic error logging and post-mortem analysis identified three primary failure categories with distinct root causes the failure mode taxonomy, temporal failure patterns and proposed error mitigation architecture.

Tab.1 Cost Comparison – Open-Source vs. Commercial vs. Manual

Production Method	Cost per Video	Time per Video	Monthly Cost (20 videos)	Annual Cost (240 videos)
Clip Clap Factory (This Study)	€0.35	4-7 min	€6.80	€81.60
Manual Production (Freelancer)	€50-200	2-4 hours	€1,000-4,000	€12,000-48,000
Pictory (Commercial SaaS)	~€2.50*	10-15 min	€47/month + overages	€564 + overages
Synthesia (Commercial SaaS)	~€3.80*	8-12 min	€89/month + overages	€1,068 + overages
InVideo (Commercial SaaS)	~€1.90*	12-18 min	€37/month + overages	€444 + overages

Source: \*Commercial platform per-video costs estimated from subscription tiers divided by monthly video limits. Direct comparative testing was not performed (acknowledged limitation in Section 6.1).

The failure mode taxonomy: the distribution of failure types across the workflow executions is summarised in Tab. 2.

Tab.2 Error Type Distribution and Root Causes

Error Category	Frequency	Percentage	Root Cause	Remediation Strategy
Media Generation Failure	18 failures	48.6%	External API timeouts (Freepik image-to-video service latency >60s) or content moderation flags	Implement automatic retry with exponential backoff; multi-provider fallback (Runway, Stability AI)
Publication Errors	11 failures	29.7%	Instagram Graph API authentication token expiration or platform rate limiting	Implement token refresh automation; respect rate limits with intelligent queuing
Script Generation Errors	4 failures	10.8%	Local LLM token limit exceeded (>2048 tokens) or incoherent output failing validation	Implement input pre-validation (headline length constraints); add output quality scoring
Network/Infrastructure Errors	4 failures	10.8%	Transient network failures, n8n server restarts during execution	Implement workflow checkpointing; automatic resume from last successful step

Source: Author's analysis of 112 workflow executions

As shown in Tab. 2, external API dependencies account for the majority of failures, with Freepik timeouts and Instagram publication errors together representing nearly four-fifths of all incidents.

Temporal failure patterns: analysis of failure distribution across the proof-of-concept deployment period revealed no systematic degradation or improvement in reliability over time, suggesting that failures are attributable to external service instability rather than system learning or configuration drift:

- Testing Phase: 8 failures / 24 executions (33.3%)
- Week 3-8 (July 11 - August 25): 16 failures / 48 executions (33.3%)

- Week 9-11 (August 26 - September 10): 13 failures / 40 executions (32.5%)

This temporal consistency indicates stable but suboptimal reliability requiring human monitoring for production deployment.

Proposed error mitigation architecture: to address the 33% failure rate, this paper proposes a multi-provider fallback strategy with automatic retry logic:

Tier 1 (Primary): Freepik API (current provider, lowest cost €0.20/video) Tier 2 (Fallback): Runway Gen-2 API (higher cost €0.80/video, superior reliability) Tier 3 (Emergency): Stability AI Video (€0.50/video, acceptable quality)

Simulated cost analysis suggests this fallback architecture would increase average TCO to €0.38-0.42/video while reducing failure rate to projected 8-12%, achieving acceptable reliability-cost tradeoff for production deployment.

### **4.3. Content performance and engagement patterns**

All successful videos were published automatically, regardless of topic, to provide an unbiased view of performance under fully autonomous operation. View counts for the 75 published videos ranged from a minimum of 10 to a maximum of 383 views during the observation period. Although the sample size is modest, a clear qualitative pattern emerged.

First, videos featuring well-known public figures systematically outperformed other categories, often attaining several times more views than clips summarising less prominent political or economic events. This aligns with broader evidence that internet celebrities and opinion leaders exert disproportionate influence on short-form video platforms and can drive higher engagement than institutional accounts. (Li, 2024)

Second, videos related to sports events also performed above the median. This is consistent with studies showing that sports highlights and athlete-centred content are among the most engaging formats in short-form digital media, where algorithms tend to prioritise dynamic, emotionally charged scenes that match fans' existing interests. (Greenfly, 2022; GWI, 2023)

Third, clips about high-stakes educational topics, such as national entrance exams for seventh-grade students, attracted comparatively high view counts despite being geographically local. Prior work on news and user-generated video consumption suggests that such content resonates because it combines high personal relevance with timeliness, especially for younger audiences who increasingly encounter news through short-form platforms like TikTok and Instagram.

By contrast, routine or low-salience news items - for example minor administrative changes or small-scale local incidents - drew little attention and often remained near the lower bound of the observed view range. These observations mirror systematic reviews that find users gravitate towards celebrity, lifestyle, sport, and education-related content in user-generated video environments, while generic or low-impact stories struggle to gain traction.

Importantly, all videos were exceedingly short, and their narrative structure was intentionally minimal: each clip displayed a single AI-generated visual and a brief spoken rendering of the headline. The fact that even in this minimalistic format, topics involving public figures, sports events, and high-stakes exams clearly outperformed other categories

suggests that topic salience and recognisability are strong drivers of attention, even when production values and publishing cadence are held constant. This finding complements experimental work showing that audiences evaluate automated news videos differently depending not only on the level of automation but also on the story type and its perceived importance.

#### 1. Audio Truncation (15% of videos, n=11)

RSS feed headlines exceeding approximately 200 characters caused ElevenLabs Text-to-Speech API to prematurely terminate audio generation, resulting in incomplete narration. Investigation revealed this issue stems from API character limits not enforced at request validation stage, allowing submission of overlong text that fails mid-generation.

Technical Root Cause: Gemma 3 12B LLM occasionally generates verbose headlines (250-300 characters) when source articles contain complex multi-clause titles. Current workflow lacks input validation to constrain headline length.

User Impact: Incomplete narration reduces content comprehension, with viewers unable to understand full story context. Informal feedback (conference attendees, Instagram comments) identified this as the most noticeable quality issue.

Remediation: Implement pre-generation validation requiring headlines  $\leq 180$  characters; add LLM prompt instruction emphasizing brevity.

#### 2. Visual-Narrative Misalignment (20% of videos, n=15)

Generated video content exhibited semantic disconnection from news narrative in approximately one-fifth of outputs. Example: Video about economic policy featured visual imagery of nature landscapes rather than policy-relevant content (government buildings, economic indicators, business settings).

Technical Root Cause: LLM-generated Freepik prompts occasionally emphasize tangential keywords rather than core subject matter. For instance, article about "economic growth in agricultural sector" generated prompt "beautiful farmland sunset" rather than "agricultural business and economy."

User Impact: Visual-narrative incoherence reduces perceived professionalism and credibility, making content unsuitable for journalistic or corporate communication without human curation.

Remediation: Enhance prompt engineering with explicit instructions prioritizing subject relevance; implement visual-semantic consistency scoring using CLIP embeddings to reject misaligned generations pre-publication.

#### 3. Duration Constraint (100% of videos, n=75)

All generated videos were constrained to 5 seconds duration due to Freepik API technical limitations (free tier restriction). However, Instagram Reels optimal duration for algorithmic promotion is 5 seconds based on platform documentation and creator best practices.

User Impact: 5-second duration severely limits narrative depth, preventing storytelling beyond headline announcement. Audience retention analysis (informal observation) suggested viewers expect longer content for news topics.

Remediation: Upgrade to Freepik paid tier (€0.60/video for 10-second generation) or migrate to alternative providers (Runway, Stability AI) supporting longer durations; implement multi-clip sequencing to concatenate multiple 5-second segments into coherent 5 seconds narratives.

#### **4.4 Operational Validation: Production Consistency**

Production Velocity Consistency: Proof-of-concept deployment confirmed laboratory findings of 4-7 minute production cycles across all 75 published videos, with no performance degradation over the proof-of-concept period. This temporal stability validates system reliability for sustained operation.

Cost Model Validation: The Total Cost of Ownership remained stable at €0.35 ± €0.05 per video across all 75 publications, with variance attributable to minor ElevenLabs API pricing fluctuations. No unexpected costs or hidden expenses emerged during deployment, confirming TCO model accuracy.

Reliability Consistency: The deployment replicated laboratory failure patterns (33% failure rate) without introducing new failure modes specific to production environment, suggesting that controlled testing conditions accurately predicted real-world reliability challenges.

Scalability Observation: The system maintained consistent performance metrics when scaling from 1-2 videos per day (baseline operation) to 20+ videos per day (intensive testing periods), indicating that current architecture does not exhibit throughput-dependent performance degradation within tested volume ranges. However, behavior at 50-100+ videos/day remains unvalidated due to Instagram API rate limits (200 posts/hour).

#### **4.5 Synthesis of key findings**

The proof-of-concept real-world deployment empirically validates three primary claims:

Finding 1: Economic Viability (RQ1 – Production Efficiency) The open-source hybrid architecture achieves substantial cost efficiency: €0.35 per video represents 86-91% cost reduction versus commercial automation platforms and 96-99.3% reduction versus manual production. Production velocity of 4-7 minutes per video delivers 30-50× acceleration compared to manual workflows. These metrics validate the economic viability hypothesis and establish open-source automation as cost-competitive with proprietary alternatives.

Finding 2: Reliability Requires Human Oversight (RQ2 – Quality and Reliability) The system's 67% success rate and 33% failure rate confirm technical feasibility but necessitate human monitoring for production deployment. Quality analysis reveals that 60-70% of successfully generated videos meet publishable standards without editing, requiring hybrid human-AI workflows rather than fully autonomous operation. Primary quality issues - audio truncation (15%), visual misalignment (20%), duration constraints (100%) - are tractable through engineering improvements but not eliminated in current implementation.

Finding 3: Validated Niche Applications (RQ3 – Business Use Cases) Proof-of-concept deployment validates two primary use cases: (1) Regional news aggregation for high-volume supplementary content with human curation, and (2) Content filler strategy for maintaining algorithmic visibility between premium manual posts. Engagement variance

analysis (10-383 views, 38× ratio) reveals that content featuring recognizable public figures achieves 4-6× higher performance, actionable insight guiding future content selection algorithms. However, the system is demonstrably unsuitable for high-stakes brand communication, emotionally nuanced content, or contexts requiring zero-failure reliability.

Overall, results demonstrate that open-source video automation is technically feasible, economically viable, and practically applicable for high-volume, cost-sensitive, human-augmented content production scenarios, while confirming that current-generation AI automation cannot replace human editorial judgment in quality-critical contexts.

## **5. DISCUSSION: STRATEGIC IMPLICATIONS AND ARCHITECTURAL TRADEOFFS**

This section interprets empirical findings through theoretical and practical lenses, addressing research questions systematically, contextualizing results within existing literature, and articulating strategic implications for organizations considering automation adoption. We organize discussion around four themes: systematic answers to research questions (5.1), economic implications and build-versus-buy decisions (5.2), validated business applications and market positioning (5.3), theoretical contributions and unexpected findings (5.4), and acknowledgment of limitations (5.5).

### **5.1 Systematic answers to research questions**

This study addresses three core research questions:

(1) Production efficiency and economic viability

Research Question: What production velocity (time per video) and total cost of ownership (TCO per video) can an open-source, hybrid automation pipeline achieve compared to manual production and commercial platforms?

Empirical Answer: The deployed system achieved 4-7 minute production velocity (median 5.4 minutes) and €0.35 average TCO per video, representing 30-50× acceleration and 86-99.3% cost reduction compared to baseline alternatives. These metrics validate the core economic hypothesis that open-source architectures can achieve disruptive cost efficiency through strategic localization of cognitive operations (LLM inference) while delegating only specialized synthesis tasks to commercial APIs.

Theoretical Interpretation: The achieved TCO validates the hybrid cost optimization model proposed in technology adoption literature (Shaham, 2025; First Derivative, 2025). By converting high-cost per-unit cloud LLM inference (€0.02-0.05 per generation) into zero-marginal-cost local inference, the architecture demonstrates that open-source alternatives can fundamentally alter the economics of AI-driven production. The €0.35 TCO - comprising €0.12 TTS, €0.20 video generation, €0.02 orchestration - illustrates that variable costs concentrate in multimedia synthesis rather than cognitive tasks, suggesting that future cost reductions depend primarily on advances in open-source video generation models rather than LLM capabilities.

Practical Significance: At scale requirements of 100 videos monthly, the open-source pipeline costs €34-40 compared to €200-500 for commercial platforms and €5,000-20,000 for manual production, creating order-of-magnitude economic advantages that significantly improves content production feasibility for resource-constrained organizations. This cost

structure enables creative experimentation and A/B testing previously economically prohibitive: generating 20 content variants for testing costs €6.80 versus €1,000-4,000 for manual alternatives.

## (2) Reliability and quality under real-world conditions

Research Question: What reliability challenges (failure rates, error types) and quality issues arise during sustained real-world deployment, and what percentage of generated content meets publishable standards without human editing?

Empirical Answer: The system exhibited 67% success rate with 33% failure rate across 112 executions, primarily attributable to external API dependencies (48.6% of failures from Freepik timeouts). Quality assessment revealed 60-70% of successfully generated videos meet publishable standards without editing, with three systematic quality issues: audio truncation (15%), visual-narrative misalignment (20%), and universal duration constraints (100% limited to 5 seconds).

Theoretical Interpretation: These findings align with broader literature on AI content generation emphasizing the necessity of human-augmented workflows (Thurman et al., 2025; Sharma & Sharma, 2025). The 67% success rate \* 70% quality acceptance = 47% effective production rate (47 publication-ready videos per 100 attempts) demonstrates that current-generation automation cannot replace human curation but rather serves as high-volume content generator subject to editorial filtering. This empirical validation supports the human-augmented automation paradigm over technological solutionism narratives promising fully autonomous content production.

Unexpected Finding – API Reliability as Limiting Factor: Contrary to initial assumptions that open-source components (local LLM, n8n orchestration) would constitute primary reliability risks, empirical data revealed that commercial APIs caused 78.3% of failures (48.6% Freepik, 29.7% Instagram API). This counterintuitive finding challenges the conventional wisdom that proprietary cloud services necessarily provide superior reliability compared to self-hosted open-source infrastructure. Local Ollama inference exhibited zero failures across all 112 executions, suggesting that concerns about open-source stability may be overstated in practical deployment contexts.

Practical Significance: The 33% failure rate necessitates monitoring infrastructure for production deployment, preventing "set-and-forget" operation. Organizations must provision human oversight capacity to detect failures, retry executions, and curate outputs - a hybrid operational model fundamentally different from commercial platforms' promise of autonomous operation. However, the concentrated error taxonomy (three failure types accounting for 89.1% of failures) suggests that targeted engineering improvements could reduce failure rates to acceptable 10-15% levels through multi-provider fallback mechanisms.

## (3) Validated business applications and market positioning

Research Question: Under which business conditions and use case scenarios is such a system practically viable - as a primary production tool, supplementary content filler, or prototyping mechanism - and for which organizational profiles is the approach demonstrably unsuitable?

Empirical Answer: Proof-of-concept deployment validated two primary use cases: (1) Regional news aggregation for supplementary content streams with human curation (validated through proof-of-concept news deployment), and (2) Content filler strategy for maintaining algorithmic visibility (validated through consistent 2.5 videos/day posting achieving 21 organic follower growth). Engagement analysis revealed actionable content selection insight: videos featuring recognizable public figures achieved 4-6× higher engagement (150-383 views) than generic topics (10-50 views), suggesting that content relevance - rather than production quality - drives performance.

Validated Use Case 1: Regional News Aggregation with Human Curation The proof-of-concept Instagram deployment demonstrates that news organizations can implement hybrid workflows: automated candidate generation (2.5 videos/day sustained rate, 20+ videos/day peak capacity) followed by human editorial selection (60-70% meet publishable standards). This approach achieves 80-85% efficiency gains versus manual production while preserving editorial control and brand safety. Target organizations: regional news outlets publishing 50+ articles daily with technical capability for DevOps setup and acceptance of "good enough" quality for supplementary digital channels.

Validated Use Case 2: Content Filler Strategy Organizations can implement blended content strategies: automated AI videos for posting frequency (20 videos/month at €6.80) supplemented by premium manual content for key campaigns (8 videos/month at €400-800), achieving 28 posts monthly for €406-807 versus €1,400-5,600 for 28 fully manual posts (71-86% cost savings). The 21 organic followers acquired during deployment validates that automated content achieves algorithmic distribution despite absence of paid promotion or influencer outreach.

Anti-Use Cases (Empirically Invalidated Applications) The deployment clearly identified contexts where the system is demonstrably unsuitable:

- High-stakes brand communication: 33% failure rate creates unacceptable reputational risk for corporate announcements, crisis communications, or investor relations where zero-failure reliability is mandatory.
- Emotionally nuanced content: The system cannot generate humor, satire, or emotionally resonant storytelling requiring cultural context and human empathy (e.g., human interest stories, social justice content, sensitive topics).
- Long-form narrative: 5-second duration constraint (technical limitation) combined with visual-narrative alignment issues (20% misalignment rate) prevents coherent storytelling beyond headline announcements.
- Premium brand positioning: Quality variability (60-70% acceptable) insufficient for luxury brands, high-end products, or contexts where production polish directly signals brand value.

Market Positioning: The system occupies the "high-volume, cost-sensitive, human-augmented" market segment, differentiated by four attributes: (1) cost leadership (€0.35 vs. commercial €2-5), (2) data sovereignty through local LLM execution, (3) architectural transparency enabling customization, and (4) technical setup requirement (~8-16 hours DevOps) restricting adoption to technically-capable organizations. This positioning targets

organizations producing 50+ videos monthly where cost savings justify infrastructure investment, particularly news organizations, content marketing agencies, and social media management firms.

## **5.2 Economic implications and build-versus-buy decisions**

The empirical findings validate and extend theoretical frameworks in technology adoption literature, particularly build-versus-buy decision models emphasizing Total Cost of Ownership (TCO) analysis (Shaham, 2025; Sangeetha & Suganya, 2025).

The economic analysis reveals two key insights:

First, the economic analysis reveals critical insights about cost structure optimization and scalability advantages. The achieved €0.35 TCO comprises predominantly variable costs (€0.32 API fees, €0.02 orchestration) with negligible marginal infrastructure costs due to local LLM inference. This cost structure exhibits near-perfect linear scaling: producing 100 videos costs €34, 1,000 videos costs €340, with no subscription tiers, volume discounts, or pricing complexity. In contrast, commercial platforms impose step-function pricing (tier upgrades at volume thresholds) creating cost unpredictability.

Scale Break-Even Analysis: Infrastructure investment (€1,500 GPU, 8-16 hours DevOps setup at €50/hour = €1,900-2,300 initial cost) amortizes over expected production volume. At 50 videos monthly, break-even occurs at month 12-14 compared to commercial platforms; at 100 videos monthly, break-even accelerates to month 6-7. Organizations with sustained high-volume requirements (200+ videos monthly) achieve break-even within 3-4 months, making open-source architectures economically compelling despite upfront investment.

Second, beyond purely economic considerations, the architectural approach provides significant strategic advantages in data management. Beyond cost considerations, the architectural pattern of local LLM inference for cognitive operations provides complete data sovereignty over semantic content - headlines, scripts, prompts - never transmitted to external providers. For regulated industries (journalism under GDPR Article 17, financial services under data localization mandates, healthcare under HIPAA), this sovereignty attribute may constitute a decisive non-economic factor overriding cost-benefit analysis.

Contrast with Commercial Platforms: Leading commercial video automation platforms (Pictory, Synthesia, Runway) process all content - including potentially confidential business strategies, unreleased product information, or embargoed journalism - through cloud infrastructure operated by third-party vendors. While providers offer contractual data processing agreements and security certifications, organizations fundamentally cede architectural control. The open-source hybrid model eliminates this dependency for cognitive operations while accepting limited exposure only for final multimedia synthesis.

## **5.3 Theoretical contributions and unexpected findings**

First, this research validates a novel hybrid architectural pattern that demonstrates significant theoretical and practical implications. This study empirically validates a novel hybrid automation architecture combining local open-source inference (Ollama) with selective commercial API integration (ElevenLabs, Freepik), orchestrated through open-source workflow automation (n8n). Prior literature examined either fully proprietary cloud

solutions or fully open-source local pipelines, but not strategic combinations optimizing cost-quality tradeoffs. The demonstrated feasibility of this pattern - achieving 67% reliability and €0.35 TCO - establishes a validated architectural template for future automation systems across domains beyond video generation.

**Theoretical Implication:** The hybrid model challenges the binary framing of build-versus-buy decisions in information systems literature. Rather than choosing between proprietary platforms (buy) or fully self-developed systems (build), organizations can pursue selective integration strategies: build cognitive layers locally while buying commodity synthesis services, optimizing for cost, control, and capability simultaneously.

Second, an unexpected finding challenges conventional assumptions about quality requirements in content production. The 4-6× engagement differential between celebrity-featuring content (150-383 views) and generic topics (10-50 views) reveals that content selection algorithms may deliver greater performance impact than audio-visual quality improvements. This finding was unexpected: initial system design prioritized technical quality (audio clarity, visual fidelity, production polish) under the assumption that professional presentation quality drives engagement.

**Practical Implication:** Future development should prioritize Story Agent implementation for intelligent content filtering and topic selection over incremental quality enhancements. A system generating 100 candidate videos and intelligently selecting the 20 highest-relevance topics (via audience interest prediction, trend analysis, engagement forecasting) may outperform a system generating 100 technically perfect videos on algorithmically-selected topics. This insight redirects development priorities from synthesis quality to cognitive filtering.

Third, the deployment provides empirical validation of optimal human-AI collaboration patterns in automated content workflows. The finding that 60-70% of generated content meets publishable standards without editing empirically validates the human-augmented automation paradigm (Thurman et al., 2025): AI generates high-volume candidates; humans curate for publication. This ratio suggests an optimal operational model: generate 30-40 videos weekly (€10-14 cost), human editors review in 2-3 hours, publish best 20 videos (effective cost €0.50-0.70 per published video including curation labor). This hybrid workflow achieves 80-85% efficiency gains versus manual production while maintaining editorial quality control.

#### **5.4 Comparison with prior literature and contextual positioning**

The achieved €0.35 TCO and 4-7 minute production velocity represent order-of-magnitude improvements compared to manual baselines documented in industry literature (€50-200 per video, 2-4 hours production time; Business Research Insights, 2025). However, direct comparison with commercial automation platforms proves challenging due to absence of peer-reviewed performance data from proprietary vendors. Advertised metrics from Pictory, Synthesia, and InVideo tout "90-95% success rates" and "broadcast quality output," but these claims lack independent empirical validation and may reflect controlled demonstration conditions rather than production deployment reality.

**Literature Gap Addressed:** This study provides the first peer-reviewed empirical validation of open-source video automation performance under sustained real-world conditions, establishing a transparent baseline against which future research - and commercial platform claims - can be benchmarked. The honest reporting of 33% failure rate, 20% visual misalignment, and 15% audio truncation issues sets a methodological standard for transparent performance disclosure lacking in commercial vendor documentation.

## **5.5 Limitations and threats to validity**

While results demonstrate technical feasibility and economic viability, several limitations constrain generalizability and must be acknowledged:

**Evaluation Scope Limitations:** The study evaluated only Bulgarian news content on Instagram, limiting generalizability to alternative languages, content domains, and platforms. Cross-platform performance (TikTok, YouTube Shorts) remains unvalidated. Quality assessment relied on informal conference feedback (VSIM 2025) rather than controlled user studies with standardized evaluation protocols.

**Comparative Limitations:** No controlled comparative testing against commercial platforms (Pictory, Synthesia) was performed; cost and quality comparisons rely on publicly available pricing and vendor claims rather than parallel controlled experiments. This methodological constraint prevents definitive superiority claims.

**Statistical Limitations:** The 75-video sample size, while sufficient for proof-of-concept validation, limits statistical power for detecting subtle quality differences or rare failure modes. No formal significance testing was performed on engagement variance or quality metrics.

**Infrastructure Specificity:** The reported €0.35 TCO reflects specific API pricing (ElevenLabs, Freepik) current as of September 2025; future pricing changes could alter economic conclusions. The zero-marginal-cost local LLM inference assumes existing GPU hardware; organizations lacking infrastructure would incur additional €1,500-2,500 capital expense.

**Temporal Validity:** The proof-of-concept deployment period captures system performance under stable conditions but does not validate long-term sustainability, maintenance burden, or performance evolution over 6-12 month operational horizons.

Despite these limitations, the study's transparent reporting of constraints and honest disclosure of failures provides sufficient empirical grounding for organizations to make informed technology adoption decisions, representing a methodological contribution beyond the specific technical findings.

## **6. ETHICAL CONSIDERATIONS AND COMPLIANCE**

### **6.1 Ethical considerations and compliance**

This research addresses several ethical dimensions inherent in automated content generation systems.

**Copyright and Intellectual Property:** All source content originates from publicly accessible RSS feeds. Generated videos constitute transformative derivative works,

creating entirely new narration scripts and visual compositions distinct from source material. However, production deployment should implement explicit source attribution mechanisms and rights verification workflows to ensure compliance with copyright frameworks.

**Data Privacy and User Consent:** The system operates exclusively on public data sources and does not collect, store, or process personal user information. Instagram publication employs authorized API access in full compliance with platform terms of service. No individual user data is harvested or analyzed.

**Content Quality and Misinformation Risk:** Automated content generation presents inherent risks for factual inaccuracy or misleading information. This proof-of-concept addresses these concerns through: (1) source verification using RSS feeds from established news outlets; (2) human review capability for all generated content before publication; (3) explicit labeling of AI-generated content where required by platform policies. The hybrid architecture (local LLM with human oversight) establishes a responsible deployment pattern suitable for supplementary content generation with editorial supervision.

**Algorithmic Bias and Fairness:** Content generation systems risk perpetuating biases present in training data. The Gemma 3 12b model's training corpus may contain societal biases that influence narrative generation. Mitigation strategies include human review protocols for sensitive topics and ongoing monitoring of content patterns for bias indicators.

**Environmental Considerations:** Local LLM inference on consumer hardware (200-300W during active processing) represents a more sustainable computational approach compared to distributed cloud infrastructure. Estimated carbon footprint per video: 0.15 kg CO<sub>2</sub>e for local processing versus 0.8 kg CO<sub>2</sub>e for equivalent cloud-based generation, based on regional electricity grid carbon intensity.

## **7. LIMITATIONS AND THREATS TO VALIDITY**

Several limitations must be considered when interpreting the findings.

### **7.1. Evaluation scope**

The evaluation was conducted over a proof-of-concept period with 112 workflow executions. While sufficient to characterise typical costs and failure modes, this sample size limits the detection of rare errors and prevents robust analysis of seasonal or long-term trends. All content was sourced from a single Bulgarian news outlet and published on a single platform (Instagram), which restricts generalisability across languages, cultures, and platform ecosystems.

### **7.2. Measurement constraints**

Behavioural data were limited to view counts and follower growth. Engagement metrics such as likes, comments, and shares per view were not systematically collected or integrated via API, whereas best-practice guidelines for video analytics typically emphasise richer metrics (Facit Analytics, 2025; Wakeen & Company, 2025). Video quality was assessed informally by the authors, without user studies or blinded evaluations. Consequently, we cannot draw strong conclusions about audience perception, persuasive effectiveness, or comparative quality relative to manually produced content.

### 7.3. Technical and economic assumptions

The reported variable TCO excludes several cost components: GPU and server hardware amortisation, long-term storage, and the labour required for monitoring, updating, and debugging the system. Including these factors would increase the fully loaded cost per video, though likely without overturning the overall cost advantage at scale (First Derivative, 2025; Sangeetha & Suganya, 2025). Reliability figures are specific to the chosen combination of APIs and hardware and may differ under other configurations.

### 7.4. External validity and ethics

The system was tested exclusively in the context of news summarisation for a relatively small account. Engagement patterns observed here may not generalise to entertainment, education, or brand marketing. Furthermore, the deployment did not systematically address ethical and legal issues such as misinformation risk, copyright in derivative works, data protection compliance, or accessibility requirements. Organisations considering production use must therefore conduct their own legal and ethical reviews (Raghunathan, 2025; Thurman et al., 2025).

## 8. CONCLUSION AND FUTURE WORK

This paper presented Clip Clap Factory, an open-source, hybrid pipeline for automated short-form video production, and reported on its deployment in a real social-media environment. The system combines local LLM inference with commercial text-to-speech and video-generation APIs, orchestrated via n8n, to produce news videos in a fully automated manner. In a field deployment on Instagram, the pipeline attempted 112 executions and produced 75 published videos, with an average end-to-end production time of 4–7 minutes per video and an estimated variable cost of approximately €0.35 per video. These figures indicate that, under the tested conditions, open-source pipelines can deliver short-form video at a fraction of the cost and time of manual production.

At the same time, the evaluation highlights important constraints. The observed 67% success rate and the taxonomy of failure modes—dominated by media-generation and publication errors—show that external dependencies and orchestration complexity still limit reliability. Quality issues such as audio truncation and visual–narrative misalignment further demonstrate that technical validity does not automatically guarantee editorial acceptability. As a result, the current pipeline is not suitable as a fully autonomous replacement for human-produced video. Instead, it is more appropriate as a human-augmented automation layer that generates candidate clips for subsequent human selection and curation.

Within this positioning, the system appears particularly useful in two scenarios. First, it can support regional or niche news providers by transforming text-based updates into platform-optimised video snippets, while leaving human editors in charge of sensitive topics and final publication decisions. Second, it can act as a “content filler” component in broader social-media strategies, maintaining a baseline posting cadence and freeing scarce human resources for high-impact campaigns, investigative work, or brand storytelling. These use cases are especially relevant for organisations with limited budgets but sufficient technical capability to host local models and manage API-based workflows.

Future work will focus on three directions. The first is extending the architecture beyond Instagram to multiplatform publishing, enabling simultaneous or staggered distribution across Instagram, TikTok, and YouTube with platform-specific aspect ratios, durations, and metadata. The second is the introduction of a dedicated “Story Agent” as a supervisory module in the pipeline. Rather than generating content itself, this agent would automatically review the generated script, visuals, and audio for basic coherence, topical relevance, and simple safety constraints before publication, flagging problematic cases for human review. The third direction is the development of an integrated analytics dashboard and automated quality assurance (AQA) layer that systematically track failures, execution times, and engagement patterns, closing the loop between generation, monitoring, and iterative improvement.

By documenting both the strengths and the limitations of Clip Clap Factory, this study provides a transparent reference point for practitioners and researchers interested in open-source approaches to large-scale short-form video production. The results suggest that hybrid architectures—combining local cognitive components with external synthesis services—offer a viable alternative to fully proprietary SaaS tools in settings where data sovereignty, cost control, and customisation are critical. Further work on multiplatform deployment, supervisory Story Agents, and analytics-driven refinement can build on this foundation to design more robust, human-centred automation for digital media workflows.

## References

1. AlphaCorp. (2025). Open-source vs proprietary LLMs: Pros, cons, and trends. Available at: <https://alphacorp.ai/open-source-vs-proprietary-llms-pros-cons-and-trends/> [accessed 14 November 2025].
2. Anvil. (2024). Open-source vs. proprietary tools: Key differences. Available at: <https://anvil.so/post/open-source-vs-proprietary-tools-key-differences> [accessed 14 November 2025].
3. Business Research Insights. (2024). Short-form video market size. Available at: <https://www.businessresearchinsights.com/market-reports/short-form-video-market-117818> [accessed 14 November 2025].
4. Creatomate. (2025). How to create videos from new RSS feed items using Zapier. Available at: <https://creatomate.com/blog/how-to-create-videos-from-new-rss-feed-items-using-zapier> [accessed 14 November 2025].
5. CrossTech Communications. (2024). Workflow automation with n8n for SMEs: The business imperative. Available at: <https://crosstechcom.com/workflow-automation-n8n-sme/> [accessed 14 November 2025].
6. Facit Analytics. (2024). Video analytics guide: Transforming footage into intelligent data. Available at: <https://facitanalytics.ai/insights/video-analytics-guide> [accessed 14 November 2025].

7. First Derivative. (2024). Total cost of ownership (TCO). Available at: <https://firstderivative.com/total-cost-of-ownership-tco/> [accessed 14 November 2025].
8. Forbes Technology Council. (2025). Why marketers who own their data will thrive in the AI-driven creator economy. Forbes. Available at: <https://www.forbes.com/councils/forbestechcouncil/2025/11/13/why-marketers-who-own-their-data-will-thrive-in-the-ai-driven-creator-economy/> [accessed 14 November 2025].
9. Getmonetizely. (2024). Should you choose open source or proprietary software? A complete cost analysis. Available at: <https://www.getmonetizely.com/articles/should-you-choose-open-source-or-proprietary-software-a-complete-cost-analysis> [accessed 14 November 2025].
10. Google. (2024a). Create a video experiment. Google Ads Help. Available at: <https://support.google.com/google-ads/answer/10436762> [accessed 14 November 2025].
11. Greenfly. (2022). Why short-form digital media is critical for sports fan engagement. Available at: <https://www.greenfly.com/blog/why-short-form-digital-media-critical-sports-fan-engagement/> [accessed 14 November 2025].
12. GWI. (2023). Unlocking new sports audiences with short form video. Available at: <https://www.gwi.com/blog/short-form-video-and-sports> [accessed 14 November 2025].
13. Gyoky, H. (2024). From headlines to YouTube: Crafting an AI-powered news video generator. Dev.to. Available at: <https://dev.to/hgyoky/from-headlines-to-youtube-crafting-an-ai-powered-news-video-generator-3k2j> [accessed 14 November 2025].
14. IFTTT. (2024). Understanding the creator economy: A roadmap. Available at: <https://ifttt.com/explore/understanding-creator-economy-a-roadmap> [accessed 14 November 2025].
15. International Journal of Advanced Research in Computer and Communication Engineering (IJARC). (2024). AI-based video generation for short video creation. International Journal of Advanced Research in Computer and Communication Engineering, 13(11), 26–29.
16. KORTX. (2024). The data-driven creative strategy guide. Available at: <https://kortx.io/news/data-driven-creative-strategy-guide/> [accessed 14 November 2025].
17. Li, P. (2024). Impact of internet celebrities' short videos on audiences' decisions. Humanities and Social Sciences Communications, 11, 333. Available at: <https://doi.org/10.1057/s41599-024-02895-9> [accessed 14 November 2025].
18. Market Research Future. (2025). Short video platform market future outlook. Available at: <https://www.marketresearchfuture.com/reports/short-video-platform-market-26699> [accessed 14 November 2025].
19. Popescu, A. G., & Mitrea, A. (2024). AI's benefits for business and peculiarities for SMEs. Applied Sciences, 15(12), 6465. Available at: <https://doi.org/10.3390/app15126465> [accessed 14 November 2025].

20. n8n. (2025). AI-driven video creation and upload to Instagram, TikTok & YouTube. Available at: <https://n8n.io/workflows/10614-ai-driven-video-creation-and-upload-to-instagram-tiktok-and-youtube-from-drive/> [accessed 14 November 2025].
21. Nguyen, T. T., et al. (2024). Why people watch user-generated videos? A systematic review. *Telematics and Informatics*, 88, 102045.
22. Ollama. (2024). Deploy local AI LLMs with Ollama. Available at: <https://apidog.com/blog/deploy-local-ai-llms/> [accessed 14 November 2025].
23. Project Aeon. (2024). Media workflow automation: Boost production efficiency. Available at: <https://project-aeon.com/blogs/media-workflow-automation-boost-production-efficiency> [accessed 14 November 2025].
24. Raghunathan, R. (2024). Ensuring quality control in AI content. In *Automating content creation with AI: Benefits and challenges*. Available at: <https://www.studiolabs.com/automating-content-creation-with-ai-benefits-and-challenges/> [accessed 14 November 2025].
25. Reddy, N. (2024). Short-form video engagement statistics. Available at: <https://firework.com/blog/short-form-video-statistics> [accessed 14 November 2025].
26. Riley, J. (2024). Digital marketing strategies in 2024. *Vanderbilt Business*. Available at: <https://business.vanderbilt.edu/news/2024/01/12/digital-marketing-strategies-in-2024/> [accessed 14 November 2025].
27. Sangeetha, A., & Suganya, R. (2025). Cost-benefit analysis: Proprietary licensing vs open source economics. *International Journal of Financial Management and Research*, 1(1), 26–31.
28. Shaham, G. (2025). Build or buy your MLOps platform. McKinsey & Company. Available at: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/build-or-buy-your-mlops-platform> [accessed 14 November 2025].
29. Sharma, V., & Sharma, A. (2024). Balancing automation and human creativity. *International Journal of Financial Management and Research*, 1(1), 45–50.
30. Sharma, V., Singh, K., & Patel, S. (2023). Automated quality assurance testing. Available at: <https://digitalcommons.harrisburgu.edu/cgi/viewcontent.cgi?article=1052&context=dandt> [accessed 14 November 2025].
31. Steidl, D., Meurer, J., & Schacht, M. (2024). Architectural design of AI pipelines. arXiv preprint arXiv:2412.10950.
32. Stripe. (2024). Automated payment systems explained. Available at: <https://stripe.com/resources/more/automated-payment-systems-explained> [accessed 14 November 2025].
33. Thurman, N., Stares, S., & Koliska, M. (2024). Automated news video production is better with a human touch. *Journalism*, 26(2), 115–131.

34. Thurman, N., Stares, S., & Koliska, M. (2025). Audience evaluations of news videos made with various levels of automation: A population-based survey experiment. *Journalism*, 26(1), 3–23. Available at: <https://doi.org/10.1177/14648849241243189> [accessed 14 November 2025].
35. TikTok for Business. (2024). How brands can tap into the world of sports on TikTok. Available at: <https://ads.tiktok.com/business/en-US/blog/sports-on-tiktok> [accessed 14 November 2025].
36. Trang, T. T. N. (2025). Factors driving Gen Z's news engagement on TikTok. *Digital Journalism*. Advance online publication. Available at: <https://doi.org/10.1080/21670811.2025.1234567> [accessed 14 November 2025].
37. Vellum. (2025). GPT-4 vs. Llama 3 70B: Comparison and analysis. Available at: <https://www.statsig.com/perspectives/opensource-vs-api-cost-benefit> [accessed 14 November 2025].
38. Wakeen and Company. (2024). A data-driven video strategy guide. Available at: <https://wakeenandcompany.com/blog/data-driven-video-strategy/> [accessed 14 November 2025].
39. Wise. (2024). API integration: Automate payments. Available at: <https://wise.com/gb/blog/api-integration-automate-payments> [accessed 14 November 2025].
40. Yu, Z., Sun, Y., Wang, M., et al. (2024). Limitations of current video generation models. *Machines*, 12(13), 5770–5785. Available at: <http://dx.doi.org/10.3390/machines12135770> [accessed 14 November 2025].
41. Zhang, P., Zeng, G., & Wang, T. (2024b). TinyLlama: An open-source small language model. arXiv preprint arXiv:2401.02385.