AMERICAN JOURNALS
Open Access, Peer Reviewed Journals

American Journal of Pedagogical and Educational Research
ISSN (E): 2832-9791| Volume 43, |December - 2025

APPLICATION OF GENERATIVE AI TO AUTOMATE ANALYTICAL REPORTS: OPPORTUNITIES, RISKS, AND QUALITY CONTROL METHODS

Bauyrzhan Beisenbayev,
IT Expert, USA

A B S T R A C T

The article discusses modern approaches to the use of generative models (LLM) for automating the preparation of analytical reports: architectures (including retrieval-augmented Generation), practical benefits and economic effects, key risks (hallucinations , distortions , data leakage, compliance), as well as quality control and verification methods (technical and organizational). Practical recommendations for implementing generative AI in business analytics are offered.

K E Y W O R D S

generative artificial intelligence, automation of analytical reports, RAG, quality control, fact-checking, LLM hallucinations, prompt engineering, credibility assessment, corporate analytics, risk management.

INTRODUCTION

In recent years, against the backdrop of rapid progress in the field of machine learning and neural networks, so-called generative artificial intelligence (Gen) has attracted particular attention. AI [1]is a class of systems capable of creating new data (text, code, images, etc.) based on the analysis of existing information corpora [1].

What distinguishes generative AI from traditional systems is not just the ability to classify, recognize, or predict, but the ability to generate meaningful, human-like content (texts, descriptions, summaries of large documents, etc.) [2].

In the context of business analytics, corporate reporting, and other tasks involving processing large volumes of text and tabular information, the need for automated reporting has long since evolved from a convenience to a virtually insurmountable necessity. Traditional methods (manually collecting, analyzing, summarizing data, and drawing conclusions) require significant time and human resources. Meanwhile, the risk of errors, human bias, and insufficient data freshness remains high.

The use of generative AI in such scenarios promises to significantly accelerate, scale, and standardize the creation of analytical reports. Practical developments have already emerged demonstrating that

[1] Generative Artificial Gen AI (Gen AI) is a field of artificial intelligence that encompasses models and algorithms capable of automatically generating new data (text, images, audio, code, etc.) based on statistical patterns identified in training samples. Gen AI includes large-scale language models, diffusion models, and generative adversarial models; their use is associated with both expanding automation capabilities and the need for quality control and verification of results.

large language models can be used to automatically generate summaries, reports, and analytical overviews based on companies' annual financial statements, saving time and reducing the workload for analysts [3]. With the advent of methods combining generative models with information extraction and aggregation systems (retrieval + generation), it has become possible to ensure that reports not only appear coherent but are also linked to specific data sources, which is critical for corporate and financial reporting.

However, in addition to the advantages, there are also serious limitations and risks, in particular, incomplete reliability, text hallucinations, the inability to “explain” the findings as a person, as well as problems with compliance, audit and corporate control requirements [4].

Therefore, despite growing interest and real-world implementation precedents, the use of generative AI to automate analytical reports requires a systematic approach: careful architectural design, ensuring source transparency, implementing control and verification mechanisms, and adhering to corporate and legal standards.

In this article, we analyze modern technical and organizational approaches to automating reporting using generative AI, assess their potential and risks, and propose quality control methods based on current research and practical examples.

Architecture of the analytical reporting automation system (System Model). Modern analytical reporting automation systems are built using a hybrid approach that combines generative models (LLM) and external knowledge bases/indexes. The central approach is retrieval - augmented. generation (RAG), where generation is performed taking into account the extracted document fragments. This reduces the likelihood of factual errors and increases the transparency of the findings. The RAG concept was first systematically described in the paper "Retrieval - Augmented" Generation for Knowledge - Intensive NLP Tasks» [5]. RAG is the foundation of reliable generation, since it provides reliance on documentary sources, and not only on model parameters.

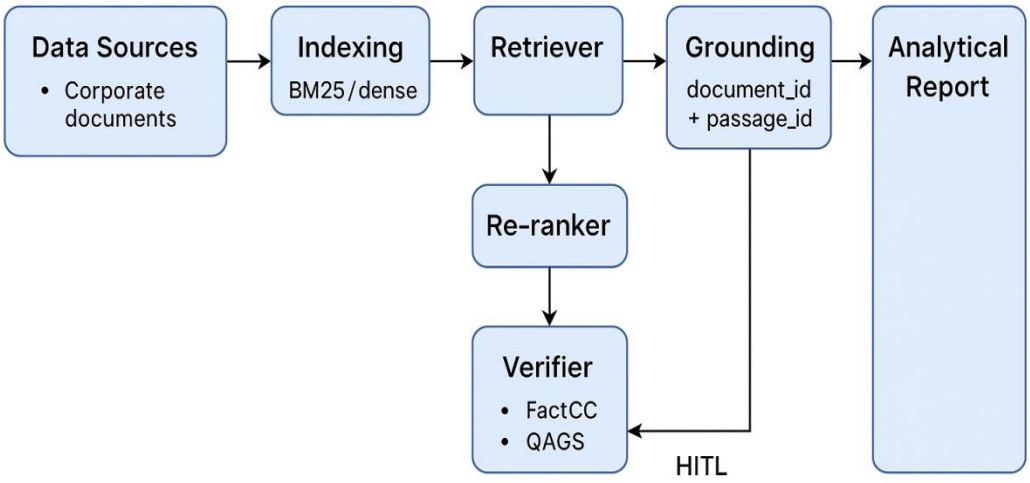

Fig. 1 - Architecture of the analytical report generation system based on Gen AI

Retriever (search module) extracts relevant texts from the index: sparse models (BM 25[2]) are efficient in speed, while dense models based on embeddings provide higher semantic accuracy. Modern RAG systems often use a hybrid approach. An effective retriever is the foundation of quality downstream generation, as it determines which information will be included in the LLM context .

Re - ranking refines the list of extracted fragments, promoting the most relevant ones to the top (often using cross - encoder models). This improves grounding accuracy and reduces the inclusion of irrelevant fragments in the generation. Re - ranking reduces noise in the retrieved data and improves the accuracy of references and interpretation.

Generator (LLM / seq2seq). The generator generates the resulting text based on the retrieved evidence . RAG architectures (RAG - Sequence, RAG - Token) differ in the way they combine knowledge. Generators can operate in zero - shot mode or with domain-specific The generator is responsible for the stylistic quality and coherence of the report, but the correctness of the statements depends on the preceding chain.

Grounding & citation layer. This layer links claims to specific sources (document_id + passage_id) [3]. Recent research emphasizes the difference between "citation correctness" and "faithfulness" - factual correspondence [6]. Without a grounding layer, it is impossible to ensure the verifiability of reports and compliance.

Algorithm 1. Generate – Retrieve – Verify
Input: analytical query q , document corpus D
Output: analytical report T with confirmed factual correctness
1. Retrieve **sources**
Search for relevant documents
R←Retrieve (D,q)R \ leftarrow \ text { Retrieve }(D, q) R←Retrieve (D,q)
2. **Relevance Refinement**
Rearrange the extracted fragments using re-ranker
R′← ReRank (R)R' \ leftarrow \ text { ReRank }(R)R′← ReRank (R)
3. **Text generation**
To generate a draft analytical text based on R′R'R′
T←Generate (q,R ′)T \ leftarrow \ text { Generate }(q, R') T←Generate (q,R ′)
4. **Verification of factual correctness**
O evaluate the consistency of text statements with sources
V←Verify (T,R′)V \ leftarrow \ text { Verify }(T, R') V←Verify (T,R′)
5. **Decision making**
If V< τV < \ tauV <τ, submit the result for expert review (HITL),
otherwise accept the text as the final analytical report.

[2] Sparse models are a class of information retrieval models in which documents and queries are represented as sparse feature vectors (usually based on terms and their frequencies). BM25 (Best Matching 25) is a probabilistic document ranking function based on term frequency weighting, inverse document frequency, and text length normalization; it is widely used as a basic and interpretable search method, as well as a component of hybrid systems, including retrieval-augmented generation .

[3] document_id + passage_id - a text data identification and addressing scheme in which a unique document identifier (document_id) is supplemented by a fragment or semantic segment identifier within the document (passage_id). This decomposition allows for more accurate extraction, citation, and verification of information, including in information retrieval systems and retrieval-augmented systems. generation .

Formal objective of the Generate–Retrieve–Verify procedure.

$$T^* = \arg\max_{T} \ \text{Faithfulness}(T, R') \quad \text{subject to Fluency}(T) \geq \delta,$$

where

T – generated analytical report,

R' – verified set of retrieved evidence,

δ – minimum acceptable fluency threshold.

Formula 1. Equation (X). Objective function of the Generate–Retrieve–Verify framework

The presented algorithm formalizes the process of generating analytical reports in a RAG-oriented architecture and emphasizes the fundamental separation of the generation and verification stages. This scheme improves the reproducibility of results and reduces the risk of factual errors in analytical conclusions.

Verifier / factuality Automatic verifiers compare claims with sources using NLI classifiers, LLM checks , or Q & A approaches (e.g., QAGS [4]). FactCC is a key model for assessing factual consistency [7]. Verification is the critical layer that separates a correct analytical report from a potentially unreliable one .

Human - in - the - loop (HITL). HITL interventions include review, escalation of high - risk approvals, deviation control, and audit. Recommendations for transparency are provided in the work " Model Cards for Model Reporting » [8]. In complex domains, humans remain an important participant in the quality and control loop of interpretations.

Methods for increasing credibility and reducing hallucinations

The reliability of the conclusions generated by generative models is a critical requirement for the application of AI in analytical work, particularly in corporate and regulatory domains. Generation without specifically organized control mechanisms often leads to factual errors, incorrect interpretations, or logical simplifications, necessitating the implementation of structured procedures to enhance reliability. In recent years, researchers and developers have developed a set of techniques aimed at improving factual correctness, the robustness of reasoning processes, and the predictability of final reports. Below, we discuss key approaches that enable the construction of systems that combine a high degree of automation with reliability comparable to expert analysis.

1. Retrieval - augmentation ensures the presence of relevant fragments in context, reduces hallucinations and improves the explainability of results. RAG - the best way to improve factual correctness.

2. Prompt engineering and scaffolding reasoning . Chain - of - Thought and structured generation (draft → revise) improve the model's ability to infer and analyze. Scientists have shown that CoT improves the quality of reasoning tasks [9]. Scaffolded Reasoning makes reports more logical and structured.

[4]QAGS (Question-Answering based Generation Score is a method for automatically assessing the factual consistency of generated text, based on asking questions of the original source and the generated response, followed by comparison of the answers obtained. A high degree of consistency is interpreted as an indicator of fact preservation and semantic correctness of the generated text; this metric is used primarily for summarization analysis and other generative AI tasks.

3. Generate – then – verify. First generation, then automatic and/or LLM -based verification. Such schemes are increasingly used in corporate analytics tools. Separate generation and verification stages significantly reduce errors.
4. Q & A factual evaluation (QAGS) identifies inconsistencies between a summary and the original document through answering questions. QAGS improves the accuracy of complex assertion verification, especially in analytical reports.
5. Semantic and factual -oriented metrics. ROUGE / BLEU are useful, but do not assess factuality; BERTScore , FactCC, and other metrics provide more accurate indicators of source consistency. Metrics should be combined: no single metric covers all types of errors.

Table 1 - Architectural components and quality improvement methods

Component / Method	**Purpose**	**Advantages**	**Restrictions**
Retriever (sparse / dense)	Search for relevant fragments	High relevance (dense), scalability	The index needs updating.
Re-ranker	Rearrangement of candidates	Increases grounding accuracy	Higher computing costs
LLM- Generator	Creating a report text	Stylistics, coherence	Tendency to hallucinate without grounding
Grounding / Citations	Linking statements to sources	Auditability and compliance	"Post-rationalization" of links is possible
Verifier (FactCC , QAGS, LLM-checks)	Fact checking	Reducing errors, QA automation	Does not replace a person completely
HITL	Human expertise	Control of critical findings	High cost
Chain-of-Thought	Structuring reasoning	Improves logic and reasoning	Increases tokenization
Generate – then – verify	Separation of generation and verification	Reliability	Increasing delays
BERTScore, factuality metrics	Automatic assessment	Semantic precision	Does not guarantee factuality

The use of generative artificial intelligence significantly expands the range of tasks amenable to automation in analytical processes. First, models accelerate the preparation of text reports by transforming large amounts of data, documents, and metrics into coherent interpretations and summarized conclusions. Second, AI can automate routine analysis steps: extracting key indicators, comparing trends, forming hypotheses, and identifying deviations. Third, LLM integration with corporate data warehouses, BI platforms, and monitoring systems enables the generation of context-sensitive analytical reports in real time. Finally, thanks to semantic search and multivariate scenario analysis, models support the justification of conclusions, offering alternative interpretations and complementing human expert judgment.
Thus, generative AI provides the foundation for moving beyond simple descriptive analysis to interactive, adaptive, and partially autonomous analytics.

Risks and Limitations of Gen AI Applications. Despite the high effectiveness of generative models in automating analytical processes, their use is accompanied by a number of significant risks. The key limitation remains hallucinations - the generation of factually incorrect or non-existent information,

which is especially critical in analytical reports aimed at making management decisions. An additional threat is the non-determinism of output: the same task can produce different results depending on model settings, context, and temperature, which complicates the reproducibility of analytics.

A significant risk is the bias inherent in training data, which can lead to systematic errors in interpreting events or discriminatory conclusions. Integrating models with enterprise data sources creates privacy concerns, including the potential for sensitive information to be leaked if query context is inadequately managed. Furthermore, high computational costs, dependence on model providers, and the need for expert tuning limit large-scale adoption.

Taken together, these factors require rigorous quality control procedures, fact verification, and model behavior monitoring to ensure the reliable and responsible use of generative AI in analytics.

Quality control and verification of analytical reports. Quality control of generative analytical reports is a multi-level process aimed at ensuring factual correctness, logical consistency, and compliance with corporate standards. The central objective is to separate assessment procedures into automated methods and expert validation , which allows for the scalability of computational checks and the depth of human interpretation. In practice, quality control systems utilize a combination of linguistic metrics, fact-checking, structural analysis, and risk-based monitoring.

Fact-checking methods, which include comparing report assertions with external or corporate data sources, classifying contradictions, and checking document links. Linguistic and semantic metrics such as BERTScore, BLEU, ROUGE, and factuality-specific models provide a quantitative assessment of quality, but their applicability is limited by the lack of a direct link to facts. These are complemented by Q&A-based methods that generate text-based questions and compare the answers with source materials.

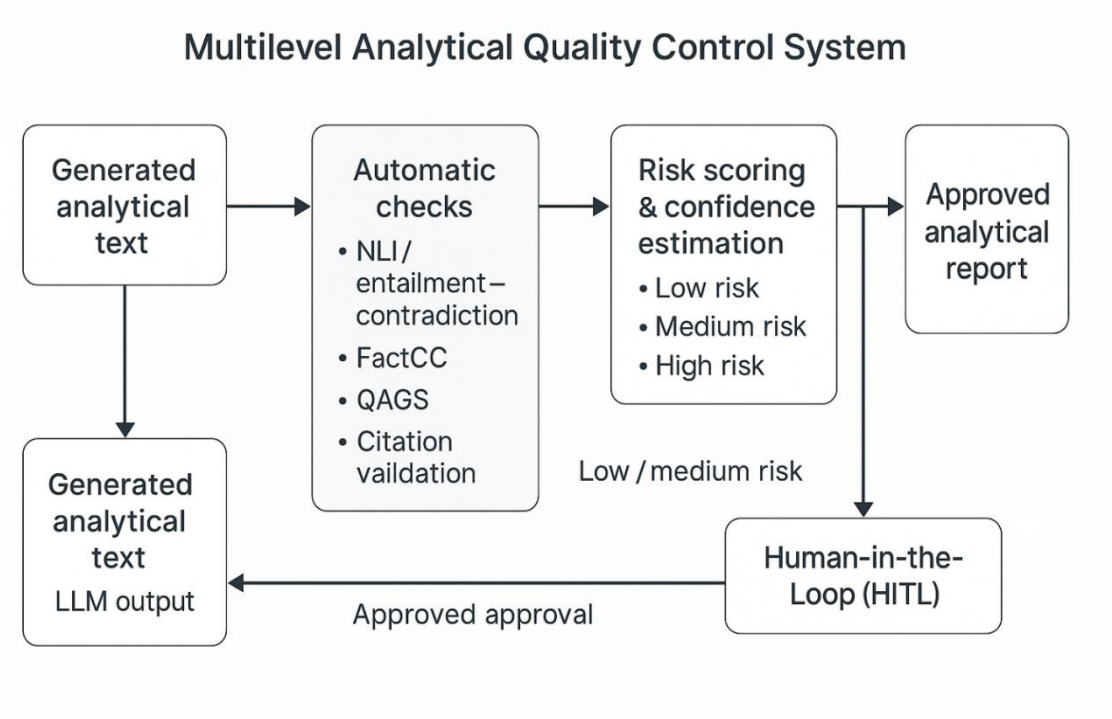

Fig. 2. Multi-level quality control system for analytical reports

Figure 2 illustrates the conceptual multi-level quality control system for analytical reports. This process can be formally represented using the following analytical model.

$$QC = \langle F, S, L, H \rangle,$$

where

F – factual verification,

S – semantic consistency evaluation,

L – logical and structural validation,

H – human-in-the-loop expert review.

The final acceptance decision is defined as:

$$Accept(T) = \begin{cases} 1, & \text{if } F(T) \wedge S(T) \wedge L(T), \\ 0, & \text{otherwise.} \end{cases}$$

Formula 2. Equation (X). Formal representation of the multi-level quality control process

Structural validation also plays a significant role, including the detection of omissions, logical gaps, erroneous interpretations of trends, and incorrect explanations of cause-and-effect relationships. This approach is especially important in analytics, where semantic consistency and explainability are critical. Expert assessment complements automated methods, forming a human-in-the-loop layer : specialists check high-risk conclusions, analyze ambiguous sections of text, and assess the interpretative correctness.

Quality control methods: comparative analysis. Ultimately, an effective QA system is built as an integration of various methods, each covering unique aspects of quality. This ensures the reliability and reproducibility of reports necessary for the application of generative AI in critical analytical scenarios.

Table 2 - Basic quality control methods and their characteristics

Method	Type	Target	Advantages	Restrictions
Auto fact-checking (NLI, verification of claims)	Auto	Identification of factual errors and contradictions	High speed, scalability	Fallible in complex logical conclusions; depends on the quality of sources
Examination references (grounding / citation validation)	Auto	Confirmation of the origin of the statements	Ensures traceability and compliance	Correct but irrelevant links are possible
Q&A approach To validation (question–answer consistency)	Auto	Identifying inconsistencies through questions	Good at revealing hidden contradictions	Requires high-quality question generation
Linguistic and semantic metrics (ROUGE, BLEU, BERTScore)	Auto	Evaluation of completeness, accuracy and semantic closeness	Ease of use, quantitative metrics	Does not guarantee factual correctness
Structural text validation	Automatic/expert	Checking logic, coherence, and explainability	Detects errors in reasoning	Partially subjective , requires domain expertise
Expert verification (HITL)	Human	Assessing risks, interpretations and complex conclusions	High accuracy and contextual depth	High labor costs, limited scalability
validation systems (generate – then – verify)	Hybrid	Improving reliability through separate stages	Significantly reduces errors	Increases latency and computing resources

Practical implementation recommendations. Effective use of generative artificial intelligence in automating analytical reports requires a systematic approach that considers technical, organizational, and regulatory aspects. Based on an analysis of current research, several key recommendations can be identified:

1. Piloting on a limited subject area. It is recommended to begin implementation with a narrow scope or a specific report type. This allows you to assess the quality of generation, identify errors, and adapt the architecture without significant risks to the business.

a grounding mechanism . All model assertions must be linked to specific documents or databases, ensuring verifiability and compliance . Implementing a RAG architecture with versioned document indexes improves the transparency and reproducibility of reports.

3. Hybrid "generation + validation" architecture. The most robust approach is to generate a draft LLM followed by automated and/or manual validation. This combines the speed and scalability of generative models with the precision of expert analysis.

4. Automatic quality checkpoints and metrics. Factual checks (FactCC , QAGS, NLI checks), semantic metrics (BERTScore), and monitoring of discrepancies between text and source data should be implemented. Confidence levels and quality metrics should be transparent to analysts and auditors.

5. Documentation and transparency of processes. It is necessary to maintain records of the model version, data sources, indexes used, and prompt templates. Using the model cards and audit logs provide the ability to verify and reproducibly produce results.

6. Human review and escalation of high-risk assertions. Any assertions with a high probability of error or critical significance must be reviewed by experts. Formalized human-in-the-loop processes minimize the risk of introducing incorrect conclusions.

7. Regular audits and model updates. It's important to systematically evaluate the quality of model generation using new data, test it on challenging cases, and promptly update the model and document index. This helps keep analytical reports current and accurate.

Following these guidelines helps minimize risk, improve factual accuracy, and ensure sustainable integration of generative AI into enterprise analytics by creating a balance between automation, quality control, and expert review.

Conclusion

Thus, generative AI offers powerful tools for automating analytical reports: accelerating the preparation of materials, scaling personalization, and synthesizing complex information. However, implementation requires a robust quality control system: grounding (RAG), automated fact checkers, semantic and factual correctness metrics, transparent documentation, and mandatory human validation. With careful combination of these measures, generative AI can become a reliable assistant to analysts, rather than a source of new risks.

References

1. Generative artificial intelligence // Wikipedia. [Electronic resource]. - Access mode: https://ru.wikipedia.org/wiki/%D0%93%D0%B5%D0%BD%D0%B5%D1%80%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D1%8B%D0%B9_%D0%B8%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B8%D0

%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82 (date of access: 30.11.2025).
2. How does generative AI work // Microsoft AI. [Electronic resource]. - Access mode: https://www.microsoft.com/ru-ru/ai/ai-101/how-does-generative-ai-work (date of access: 11/30/2025).
3. Rizki M., Wibisono Y., Nugroho E. P. Development of an Automatic Summarization System based on Large Language Models for Annual Report Analysis // Brilliance: Research of Artificial Intelligence. - 2025. - Vol. 5, No. 2. - [Electronic resource]. - Mode access: https://jurnal.itscience.org/index.php/brilliance/article/view/6772 (date accesses: 11/30/2025).
4. The evaluation of GenAI capabilities to implement professional tasks // Foresight . Journal of Advanced Research (Forsight). - 2025. - Vol . 19, No. 2. - [Electronic resource]. - Access mode: https://journal-vniispk.ru/1995-459X/article/view/274431 (accessed: 01.12.2025).
5. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Wen-tau Yih , Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // arXiv preprint . - 2020. - [Electronic resource]. - Access mode: https://arxiv.org/pdf/2005.11401 (accessed: 01.12.2025).
6. Wallat J., et al. Disentangling citation correctness and faithfulness in grounded text generation // arXiv preprint. - 2024. - [Electronic resource]. - Mode access : https://arxiv.org/pdf/2412.18004 (date accesses : 01.12.2025).
7. Kryściński W., McCann B., Xiong C., Socher R. Evaluating the Factual Consistency of Abstractive Text Summarization (FactCC) // arXiv preprint. - 2019. - [Electronic resource]. - Mode access: https://arxiv.org/abs/1910.12840 (date accesses: 02.12.2025).
8. Mitchell M., et al. Model Cards for Model Reporting // arXiv preprint. - 2019. - [Electronic resource]. - Mode access: https://arxiv.org/abs/1810.03993 (date accesses: 02.12.2025).
9. Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Chi E., Le Q., Zhou D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models // arXiv preprint. - 2022. - [Electronic resource]. - Mode access: https://arxiv.org/abs/2201.11903 (date accesses : 02.12.2025).