A COGNITIVE FRAMEWORK FOR PRESSURE EQUIPMENT INSPECTION BASED ON MULTI-AGENT AI SYSTEMS AND VLLM MODELS

Aleksandar Cvetić[1]*,
[0009-0001-3562-4660]

Angelina Njeguš[2]
[0000-0001-8682-7014]

[1]Student,
Singidunum University,
Belgrade, Serbia

[2]Singidunum University,
Belgrade, Serbia

Abstract:

This paper investigates the transformation of the traditional pressure equipment (PE) inspection process into an intelligent digital workflow through the application of advanced artificial intelligence technologies. The research focuses on overcoming the issues of inspector cognitive overload and the inefficient management of extensive documentation during the interpretation of the Pressure Equipment Directive (PED) and national regulations. An innovative multimodal architecture is proposed, based on a Multi-Agent System (MAS) and Vision-Large Language Models (VLLMs) utilizing a Retrieval-Augmented Generation (RAG) mechanism for dynamic compliance validation. Through a case study of a Liquefied Petroleum Gas (LPG) tank inspection, it is demonstrated how specialized AI agents—including agents for visual analysis, technical diagnostics, compliance verification, and interactive communication—can autonomously identify defects, calculate the remaining service life, and generate valid reports for registries such as CROPP.

Keywords:

Pressure Equipment, PED Directive, VLLM, AI Agents, RAG Architecture, Digitalization.

INTRODUCTION

The integrity of pressure equipment (PE), such as tanks, pipelines, and boilers, represents one of the most critical pillars of industrial safety and environmental protection. Given that these systems operate under extreme conditions of pressure and temperature, any material degradation can lead to catastrophic accidents with unforeseeable consequences. Therefore, the process of regular inspection is strictly regulated at the European Union level through the PED Directive (2014/68/EU)[1] , as well as through national rulebooks defining strict verification procedures during the service life (e.g., the Rulebook on inspections of pressure equipment during service life)[2].

Correspondence:
Aleksandar Cvetić

e-mail:
aleksandar.cvetic.25@singimail.rs

DOI: 10.15308/Sinteza-2026-638-644

Despite rigorous regulations, contemporary inspection practice faces challenges that slow down the digital transformation of the industry. Previous approaches often come down to the passive storage of documentation in non-searchable formats, which hinders the rapid verification of equipment status in the field. The key problem is the cognitive overload of inspectors who must manually interpret complex regulations and compare measurement results with extensive standard tables, significantly increasing the risk of human error.

Traditional methods based on computer vision (CV) have achieved success in localizing physical damage using algorithms such as YOLO[3], [4] ili CNN[5] arhitektura. However, these models remain "semantically isolated" because they possess the ability to detect defects, but not the ability to make a legal-technical conclusion about their acceptability according to legal norms. Here, a clear technological gap between visual detection and regulatory compliance is identified.

The subject of this paper is the development of an innovative cognitive framework based on multi-agent artificial intelligence systems (Agentic AI)[6] and vision-large language models (VLLM)[7], [8]. Unlike passive systems, the proposed framework utilizes Retrieval-Augmented Generation (RAG) technology[9], [10], [11] to link field measurement results with specific standard requirements in real-time. The application of high-performance vLLM libraries [12] enables efficient processing and system scalability, making it suitable for integration with central registries such as CROPP.

The main contributions of this paper are reflected in:

- The design of a multimodal agent architecture that orchestrates the work of specialized AI agents for data extraction, visual inspection, and engineering calculations.
- The implementation of a contextual RAG mechanism that transforms static legal regulations into a dynamic knowledge base for automatic compliance validation.
- The automation of the integrity assessment process through the synergy of the physical laws of fracture mechanics and the cognitive reasoning of models (Chain-of-Thought)[13], thereby raising the level of industrial safety.

The aim of the paper is to demonstrate, through a practical case study of an LPG tank inspection, how the integration of agentic systems not only accelerates administrative procedures but fundamentally transforms the inspector's role in the digital age.

2. THEORETICAL ASPECTS OF INTEGRITY ASSESSMENT AND FRACTURE MECHANICS

The basis of pressure equipment (PE) inspection rests on maintaining the stress state within permissible limits, where structural integrity is defined by the material's ability to withstand operational loads without the occurrence of plastic deformation or fracture. In the cognitive framework proposed in this paper, these calculations are not merely passive results, but input parameters for cognitive agents performing real-time verification.

2.1. STRUCTURAL INTEGRITY AND STATIC CALCULATIONS

For cylindrical pressure vessels, the most critical is the circumferential stress (tensile stress), known as *hoop stress*. The minimum required wall thickness (*s_min*) is calculated according to standards such as EN 13445-3[14], using the modified Lamé equation:

$$s_{min} = \frac{P \times D}{2 \times f \times z - P} + c \quad (1)$$

Where *P* – design pressure [MPa], *D* – outside diameter [mm], *f* – allowable material stress [MPa], *z* – joint efficiency coefficient (typically 0.7, 0.85, or 1.0) and *c* - corrosion allowance [mm].

The implementation of the agentic system allows the LLM to automatically recognize the value of *f* based on the steel type (e.g., P265GH), thereby eliminating the need for manual searching of engineering tables.

2.2. CORROSION DYNAMICS AND REMAINING SERVICE LIFE

Corrosion is the primary cause of wall thinning and progressive degradation of PE during operation. The corrosion rate (C_{rate}), is defined as the loss of thickness per unit of time:

$$C_{rate} = \frac{s_{initial} - s_{actual}}{t_{inspect}} \quad (2)$$

Based on this value, the prediction of the remaining useful life (RUL) is made, which is a key indicator for predictive maintenance.

$$RUL = \frac{s_{actual} - s_{limit}}{C_{rate}} \quad (3)$$

Where s_{limit} is the lower safety limit defined by the technical documentation.

Modern trends indicate the necessity of using "Physics-Informed" AI approaches, where neural networks are constrained by the physical laws of corrosion to increase prediction reliability[15].

2.3. INFLUENCE OF EXTERNAL FACTORS ON DAMAGE

During the use of pressure equipment, damage to surfaces and welded joints can occur due to mechanical, thermal, and chemical factors, such as impacts from hard objects, fire, or chemical spills. These factors lead to the occurrence of specific defects like cracks, scratches, dents, and damage to anticorrosion protection, which require a precise assessment of their impact on structural integrity.

2.4. VISUAL INSPECTION AND DAMAGE CLASSIFICATION

Visual inspection involves a detailed analysis of external and internal surfaces, as well as welded joints. In the proposed cognitive model, this process is transformed from a subjective human assessment into an objective process where multimodal agents perform automated detection and classification of defects [16]. The system compares captured photographs with historical data and databases of existing defects to detect discrepancies and assess the criticality of the damage for the continued operation of the equipment. The integration of Vision-Language Models (VLMs) enables the generation of natural language explanations alongside each visual detection, thereby increasing the interpretability of the engineering conclusion.

3. AGENTIC AI SYSTEMS AND RAG ARCHITECTURE

The contemporary development of Large Language Models (LLMs) has led to the convergence of two key research domains: autonomous agentic systems and Retrieval-Augmented Generation (RAG) architectures[9], [10], [11]. The theoretical foundation of such systems rests on the synergy of the model's parametric knowledge and external, non-parametric information sources (such as the PED directive and technical standards), ensuring high precision in technically demanding domains.

Unlike static systems, agentic RAG uses the ReAct (Reasoning and Acting) framework [13], which allows the model to generate verbal reasoning traces before taking a specific search action. This iterative process allows the agent to decompose complex engineering queries into subtasks, critically evaluate the relevance of found documents, and, if necessary, redefine the search strategy. In the context of inspection, this means the model first "thinks" about the category of the pressure vessel before extracting specific limit values from the Rulebook.

At the core of the retrieval mechanism lies the theory of vector representations (embeddings). Technical documentation and legal regulations are mapped into a high-dimensional Euclidean space, where semantic similarity is calculated using cosine similarity. Agentic systems extend this concept through dynamic memory management. While standard RAG uses a short-term context window, the agentic approach integrates long-term memory via external vector databases [17], enabling knowledge persistence across multiple interaction sessions and tracking the equipment's damage history.

The theoretical validation of agentic RAG systems requires a rigorous analysis of three key aspects, known as the "RAG triad" [11] :

- Faithfulness: The degree to which the generated answer is firmly grounded in the source context of the regulations, thereby minimizing the risk of hallucinations.
- Answer Relevance: The extent to which the answer directly addresses a specific engineering query.
- Context Precision: The agent's efficiency in selecting exclusively relevant information from extensive standard datasets.

By using the LLM as the "central processor" for logical reasoning, the system ceases to be just a database and becomes an active orchestrator managing communication between specialized agents.

4. ANALYSIS OF EXISTING SCIENTIFIC RESEARCH (*STATE-OF-THE-ART*)

A literature review indicates an accelerated transition from algorithms focused on spatial image processing to cognitive architectures integrating visual perception with logical reasoning. Research in the domain of metal surface inspection began with the application of basic CV (Computer Vision) operations. Classical algorithms like Canny and Sobel operators laid the foundations for edge detection but showed significant sensitivity to noise and lighting variations in the factory environment. A turning point was brought by the introduction of Convolutional Neural Networks (CNNs). In the work of Ren et al., it was demonstrated that CNN architectures can automatically learn hierarchical feature representations, thereby overcoming the need for manual feature engineering..

Regarding real-time detection, the literature makes a clear distinction between speed and precision:

- Single-stage detectors (YOLO series): Studies of YOLOv5[4] and YOLOv8 modifications show that these models achieve a frame rate of over 50 FPS while maintaining high precision (mAP), making them optimal for detecting discrete defects like pitting corrosion.
- Two-stage detectors (Faster R-CNN and Mask R-CNN): Although slower, these models have been evaluated in studies as superior for detecting extremely small damages. Mask R-CNN is particularly significant as it enables precise segmentation, i.e., drawing the defect mask, which is crucial for the engineering measurement of the exact area of damage or scratches.

The latest scientific breakthrough is the integration of visual models with large language models (vLLMs), achieving a semantic understanding of defects:

- Interpretation of welded joints: The research by Olkova i Gavrilova [7] tested Gemma and Qwen2.5-VL models on complex welded joint datasets. The results show that models larger than 20B parameters significantly reduce hallucinations and achieve a recall of 66.36%, outperforming YOLOv12 in domains requiring a natural language explanation of the defect.
- SteelDefectX dataset: In the work by et al. [12] a dataset with 7,778 images enabling zero-shot recognition was presented. The application of the Long-CLIP architecture resulted in an accuracy of 93.63%, with significantly better generalization on imbalanced data classes compared to purely visual models.

The most significant state-of-the-art (SOTA) scientific breakthrough is the transition from simple detection to the semantic explanation of defects via vLLM models (e.g., Qwen2.5-VL, Gemma). These models enable:

- Zero-shot recognition: The ability to identify new, previously unseen types of damage based on textual descriptions of engineering standards.
- Explainable AI (XAI): Generating natural language explanations alongside detection, directly solving the "black box" problem of older AI models. Research on datasets like SteelDefectX confirms that multimodal models achieve better generalization on imbalanced data classes compared to purely visual models.

In the domain of critical infrastructure, SOTA research is moving towards Physics-Informed Neural Networks (PINN). The paper by Zhang et al. [15] explores the application of PINN models for predicting material fatigue (e.g., Inconel 617). The results indicate that integrating physical laws into the AI model training process drastically increases the reliability of the remaining useful life prediction, as the model is not merely statistically driven but constrained by the physical constants of degradation.

5. PROPOSAL FOR A HYBRID INSPECTION MODEL

While traditional software solutions in the industry most often serve as static repositories of technical documentation, the proposed hybrid model transforms the inspection process into a dynamic, cognitive workflow. The foundation of this model is a Multi-Agent System (MAS) architecture that integrates visual perception, engineering reasoning, and strict compliance with legal regulations.

5.1. COGNITIVE ORCHESTRATION AND ROLES OF SPECIALIZED AGENTS

The system ceases to be a passive database and becomes an active "digital engineer-assistant" through the division of responsibilities among five key agents:

- Document Agent: Uses advanced OCR and LLM reasoning for the intelligent extraction of data from scanned PED certificates and manufacturer documentation. It "understands" the context and automatically populates technical parameters such as design pressure and material from non-searchable documents.

Table 1. Comparative overview of key SOTA research and results

Paper (Authors and Year)	Research Focus	Key Methodology	Main Results
Alaa et al. (2024)	Metal detection	Vision Transformers (ViT)	93.5% accuracy; reduced reflection impact.
Olkov i Gavrilov (2025)	Weld inspection	vLLM (Qwen2.5-VL, Gemma)	66.36% recall; generation of textual explanations.
Zhao et al. (2026)	Zero-shot detekcija	Long-CLIP & SteelDefectX	93.63% accuracy on imbalanced classes.
Zhang et al. (2024)	Fatigue prediction	Physics-Informed (PINN)	Increased RUL reliability via physical laws.

- Visual Agent: Implements multimodal vLLM models (like Qwen2.5-VL) for the analysis of surface and welded joint photographs. Unlike classical detection, this agent provides a semantic explanation of the observed defect and compares it with historical images to identify damage progression.
- Technical Diagnostics Agent: Specialized for engineering integrity calculations. Using the principles of Physics-Informed AI, the agent interprets the results of ultrasonic wall thickness measurement, calculates the corrosion rate, and predicts the Remaining Useful Life (RUL) based on fracture mechanics equations.
- Compliance Agent: Acts as the regulatory filter of the system, connected to the RAG base containing the PED directive and national rulebooks. Its task is the autonomous verification of compliance—for example, validating whether the hydrotest pressure or the safety valve setting is within the limits prescribed by the articles of the law.
- Reporting & Communication Agent: Structures the conclusions of all previous agents into professional PDF reports compliant with the requirements of the CROPP registry. It also provides clients with interactive insight into the equipment's status via chat, answering specific queries about revision deadlines.

5.2. IMPLEMENTATION OF THE RAG MECHANISM AND "HUMAN-IN-THE-LOOP" VERIFICATION

The architecture relies on a vector database (e.g., Pinecone) where laws and standards (EN 13445) are indexed as mathematical vectors, enabling semantic search by meaning, not just by keywords. The decision-making process follows the Chain-of-Thought (CoT) method, simulating engineering thinking through four steps:

- Ingestion: Converting raw documentation into a structured JSON format.
- Retrieval: Extracting exclusively relevant articles of the rulebook for a specific equipment category.
- Reasoning: Comparing field findings with the standard and generating explanations (e.g., "Wall thinning exceeds the allowable 15% according to Article X").
- Verification (Human-in-the-Loop): A critical component of the model where the AI agent proposes a conclusion, but the inspector performs the final verification. This step ensures that AI acts as decision support, not an autonomous judge, thereby eliminating the risk of regulatory oversights.

Thus, the basic parts of the RAG (Retrieval-Augmented Generation) architecture of the proposed system are:

- Vector representation of knowledge: Laws (PED directive, Rulebook) are not loaded into the model as text, but are converted into mathematical vectors.
- Semantic search: When a user asks a question or enters data, the agent searches the legal regulations by meaning, not just by keywords.
- Chain-of-Thought: This is a process where the LLM simulates engineering reasoning:
  - Identify vessel type;
  - Find relevant Rulebook article;
  - Compare measured thickness with smin;
  - Make a legal-technical conclusion.

For example, using Chain-of-Thought prompting, agents reason as follows:

- Step 1: The agent receives a UT measurement (e.g., 8.5mm).
- Step 2: The agent retrieves the design thickness (10mm) and the date of the previous inspection from the database.
- Step 3: The agent consults the RAG base for allowable limits according to the standard.
- Step 4: The agent draws a conclusion: "The equipment is safe for operation for the next 12 months, but requires more frequent supervision."

The architectural layers of such a system would include:

1. Infrastructure layer: Database (SQL/NoSQL) containing data on clients, equipment, and measurement history (connected to CROPP).
2. Knowledge Base (RAG): A vector database in which all laws, standards (EN 13445), and internal rulebooks are "indexed".
3. Agent layer (Orchestration):
   - Regulatory Agent: Checks if the equipment complies with the articles of the law.
   - Technical Auditor Agent: Analyzes the results of wall thickness measurements, as well as photographs taken during the inspection of surfaces and welded joints, and calculates the remaining lifespan and whether the PE is safe to use until the next inspection.

- Reporting Agent: Uses the LLM to turn technical data into a fluent, professional report for the client.

The application of such a system would reduce the administrative burden (the inspector spends 40% less time writing reports), eliminate human errors (an AI agent cannot "overlook" a change in the law or misread a standard table), and achieve transparency (the client gets instant insight into the status of their equipment via a mobile app).

5.3. TECHNICAL SCALABILITY AND INTEGRATION

The use of high-performance vLLM libraries and PagedAttention mechanisms allows the system to efficiently process queries from multiple inspectors simultaneously without performance degradation. Data is synchronized via API with the central registry (CROPP), achieving full transparency and traceability of the inspection process. The application of this hybrid model results in a reduction of administrative burden by about 40%, minimization of human error in interpreting tables, and a drastic acceleration of the digitalization process of old pressure equipment archives.

6. CONCLUSION

This paper demonstrates that the synergy of multi-agent AI systems and VLLM models represents the next evolutionary step in the digitalization of pressure equipment inspection. Unlike previous solutions that focused exclusively on isolated damage detection via computer vision, the proposed cognitive framework integrates engineering reasoning, legal compliance, and interactive communication into a unified, intelligent ecosystem. Key research findings confirm the following: (1) Risk reduction of human error: The implementation of RAG architecture and attention mechanisms enables high precision in interpreting complex standards, such as EN 13445-3, with minimal risk of model hallucinations. (2) Operational efficiency: Automating the process of data extraction and report generation reduces the administrative burden on inspectors, thereby accelerating the archive digitalization process. (3) Improved assessment reliability: The use of "Physics-Informed" AI approaches ensures that remaining useful life (RUL) predictions are physically grounded and aligned with material degradation laws. (4) Interactivity and transparency: The introduction of the Communication Agent transforms the system from a passive database into an interactive service that provides timely information to inspectors in the field and allows instant insight into the equipment status via mobile applications. (5) Regulatory security: "Human-in-the-Loop" verification ensures that AI acts as advanced decision support, while the final engineering and legal responsibility is retained by the authorized inspector. Future research will focus on integrating the system with IoT sensors for continuous integrity monitoring in real-time and further optimizing the scalability of agentic orchestration in complex industrial environments.

Table 2. Components and functional division of AI agents

Agent	Key Function	Input Data
Document Agent	OCR and semantic extraction	Scanned certificates and documentions
Visual Agent	Defect detection and classification	Surface and weld photographs
Technical Agent	RUL and integrity calculation	UT measurements and calculations
Compliance Agent	Legal compliance verification	PED directive, Rulebooks (RAG)
Reporting Agent	Report generation for CROPP	Integrated agent conclusions
Communication Agent	Interaction with client and inspector	Chat and e-mail queries

REFERENCES

[1] European Parliament and Council, "'Directive 2014/68/EU on the harmonisation of the laws of the Member States relating to the making available on the market of pressure equipment,' *Official Journal of the European Union, L 189*, pp. 164–259, May 2014." Accessed: Apr. 16, 2026. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2014/68/oj

[2] Ministarstvo rudarstva i energetike Republike Srbije, "Pravilnik o pregledima opreme pod pritiskom tokom veka upotrebe: 114/2021-230, 63/2023-73, 7/2025-3." Accessed: Apr. 16, 2026. [Online]. Available: https://pravno-informacioni-sistem.rs/eli/rep/sgrs/ministarstva/pravilnik/2021/114/3/reg

[3] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, Apr. 2018, doi: https://doi.org/10.48550/arXiv.1804.02767.

[4] D. Wang and X. Liu, "Surface Defect Detection of Strip Steel Based on YOLOv5," *Academic Journal of Science and Technology,* vol. 7, no. 2, pp. 66–71, Sep. 2023, doi: 10.54097/AJST.V7I2.11778.

[5] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data,* vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00444-8.

[6] M. A. Ali and F. Dornaika, "Agentic AI: A Comprehensive Survey of Architectures, Applications, and Future Directions," *Artif. Intell. Rev.*, vol. 59, no. 1, Oct. 2025, Accessed: Apr. 17, 2026. [Online]. Available: http://arxiv.org/abs/2510.25445

[7] I. Olkov, A. Gavrilov, I. Olkov, and A. Gavrilov, "Visual large language models for welding assessment," *Vibroengineering Procedia,* vol. 58, pp. 314–319, May 2025, doi: 10.21595/VP.2025.24983.

[8] X. Chen, "AuroraEdge-V-2B: A Faster And Stronger Edge Visual Large Language Model," Jan. 2026, Accessed: Apr. 18, 2026. [Online]. Available: https://arxiv.org/pdf/2601.16615

[9] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *Proceedings - 2024 Conference on AI, Science, Engineering, and Technology, AIxSET 2024*, pp. 166–169, Dec. 2023, doi: 10.1109/AIxSET62544.2024.00030.

[10] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," A*dv. Neural Inf. Process. Syst.*, vol. 2020-December, May 2020, Accessed: Apr. 17, 2026. [Online]. Available: https://arxiv.org/pdf/2005.11401

[11] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated Evaluation of Retrieval Augmented Generation," EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, *Proceedings of System Demonstrations*, pp. 150–158, Sep. 2023, doi: 10.18653/v1/2024.eacl-demo.16.

[12] S. Zhao, J. Gui, B. Yu, L. Dong, and Z. Gui, "SteelDefectX: A Coarse-to-Fine Vision-Language Dataset and Benchmark for Generalizable Steel Surface Defect Detection," Mar. 2026, Accessed: Apr. 18, 2026. [Online]. Available: https://arxiv.org/pdf/2603.21824

[13] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," 11th International Conference on Learning Representations, *ICLR 2023*, Oct. 2022, Accessed: Apr. 18, 2026. [Online]. Available: https://arxiv.org/pdf/2210.03629

[14] European Committee for Standardization, "EN 13445-3: Unfired pressure vessels - Part 3: Design," *CEN*, Brussels, 2021.

[15] S. Zhang et al., "Physics-informed neural network for creep-fatigue life prediction of Inconel 617 and interpretation of influencing factors," *Mater. Des.*, vol. 245, p. 113267, Sep. 2024, doi: 10.1016/J.MATDES.2024.113267.

[16] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, pp. 1932–1940, Aug. 2023, doi: 10.1609/aaai.v38i3.27963.

[17] Pinecone Systems, "What is a Vector Database & How Does it Work? Use Cases + Examples | Pinecone," *Technical Whitepaper*. Accessed: Apr. 18, 2026. [Online]. Available: https://www.pinecone.io/learn/vector-database/