EUROCOMPLY: Enabling Zero-Touch AI Compliance Auditing via LLM-based Agentic AI

Mazene Ameur, Bouziane Brik, Adlen Ksentini

EUROCOMPLY: Enabling Zero-Touch AI Compliance Auditing via LLM-based Agentic AI

Mazene Ameur*, *Member, IEEE*, Bouziane Brik‡, *Senior Member, IEEE,* and Adlen Ksentini*, *Senior Member, IEEE*
*Communication Systems Department, EURECOM, Sophia Antipolis, France.
‡Computer Science Department, University of Sharjah, Sharjah, UAE.
Emails: firstname.lastname@eurecom.fr, bbrik@sharjah.ac.ae

***Abstract*—In this paper, we present EUROCOMPLY, a novel framework designed to automate regulatory compliance verification in Artificial Intelligence and Machine Learning (AI/ML) systems for the telecommunications sector. With the increasing adoption of AI/ML, ensuring adherence to the European AI Act (EU AI Act) and General Data Protection Regulation (GDPR) has become critical to avoid deployment delays and legal penalties. EUROCOMPLY leverages Agentic AI to inspect datasets and AI/ML pipelines for alignment with the EU AI Act, the GDPR, and the 3rd Generation Partnership Project (3GPP) AI/ML-related standards. The framework employs a dual-mode retrieval architecture combining vector-based and graph-based retrieval for enhanced regulatory interpretation. We validate EUROCOMPLY on 20 telecom use cases across four realistic datasets, demonstrating high faithfulness and strong performance through expert assessments and LLM-as-a-Judge evaluations.**

***Index Terms*—Agentic AI, LLM, MLOps, AI Compliance, EU AI Act, GDPR, 3GPP, Trustworthy AI.**

I. Introduction

THE telecommunications industry is undergoing a major transformation driven by the integration of Artificial Intelligence (AI) and Machine Learning (ML) into its core infrastructure. These technologies are reshaping key areas like network optimization and intelligent service orchestration [1]. In response, standardization bodies have taken proactive steps to align AI adoption with regulatory and technical standards. Notably, the 3rd Generation Partnership Project (3GPP[1]) is leading efforts through technical standard development, stakeholder coordination, and the release of strategic white papers [2]. Nevertheless, this rapid integration also introduces complex legal, ethical, and regulatory challenges, particularly within the European Union. The confluence of the EU Artificial Intelligence Act (EU AI Act[2]) and the General Data Protection Regulation (GDPR[3]) sets forth stringent compliance obligations governing the lifecycle of ML systems. These frameworks mandate transparency, accountability, and data governance standards that, if unmet, can result in substantial financial liabilities and hinder the successful commercialization and scalability of AI-driven solutions within the telecom industry [3]. Recent high-profile cases illustrate the seriousness of this challenge. For example, Meta's use of public data from Facebook and Instagram [4], and regulatory scrutiny over OpenAI's Sora and Vision models due to data handling and consent issues [5]. These developments underscore the need for automated auditing tools integrated into the ML Operations (MLOps) lifecycle. Such tools can provide real-time compliance insights, enabling proactive alignment with legal mandates and mitigating regulatory risks.

Meanwhile, the rapid and ongoing advancements in Large Language Models (LLMs) have laid the foundation for a new computational paradigm known as Agentic AI. The latter refers to systems composed of multiple such agents, each endowed with the ability to reason through tasks, decompose problems, reflect on intermediate outputs, and collaborate with other agents to achieve high-level objectives [6]. LLM-based agents offer distinct advantages over traditional pipeline automation methods by enabling dynamic decision-making and contextual awareness. Their capacity to interpret unstructured regulatory texts, integrate domain knowledge, and execute nuanced reasoning processes positions them as a transformative solution for automating compliance workflows that intersect with evolving AI regulations and telecom standards.

A variety of solutions [3], [7]–[9] have been proposed to assess the compliance of datasets and ML models with regulations. However, these approaches exhibit critical limitations, often focusing on narrowly defined scenarios that limit their generalizability across different applications. Additionally, they predominantly address the GDPR or the EU AI Act in isolation, rather than providing comprehensive regulatory coverage. In this context, we propose `EUROCOMPLY`, which goes beyond state-of-the-art by introducing the first framework specifically tailored to automate end-to-end compliance verification across MLOps pipelines, simultaneously addressing EU AI Act requirements, GDPR mandates, and 3GPP standardization best practices. Notably, `EUROCOMPLY` incorporates a novel dual-mode retrieval architecture that seamlessly alternates between vector-based retrieval for precise extraction of legal clauses and graph-based retrieval for navigating complex regulatory dependencies, thereby enhancing interpretability and compliance accuracy.

To the best of our knowledge, this work pioneers the application of collaborative AI agents for autonomous compliance monitoring, empowering AI developers and stakeholders with real-time validation capabilities for both datasets and codebases against an integrated regulatory framework encompassing GDPR, EU AI Act, and telecom-specific standards such as 3GPP guidelines. It is worth noting that `EUROCOMPLY`

[1]https://www.3gpp.org/
[2]https://artificialintelligenceact.eu/
[3]https://gdpr-info.eu/

can be applied to assess any MLOps pipeline across diverse domains; however, we focus in this paper on demonstrating its capabilities within a telecom use case to showcase its practical applicability in a highly regulated sector.

A. Contributions

The main contributions of this paper can be summarized as follows:

- We propose EUROCOMPLY, a novel framework that automatically audits MLOps pipelines against the EU AI Act and validates datasets for GDPR compliance, while also providing best-practice recommendations from telecom standards such as 3GPP.
- Our approach harnesses the Agentic AI paradigm by orchestrating a system of reasoning agents powered by LLMs. These agents employ self-reflection mechanisms and chain-of-thought prompting to support context-aware, iterative decision-making throughout the compliance validation process.
- We propose a novel dual-mode retrieval architecture that intelligently switches between vector-based retrieval for precise extraction of legal clauses and graph-based retrieval for navigating complex, multi-relational regulatory dependencies. This hybrid retrieval strategy enhances the system's interpretive depth and accuracy in assessing compliance obligations.
- We construct a comprehensive benchmark comprising 20 realistic AI/ML use cases in telecommunications. These scenarios span core, radio, and edge networks, are annotated using four real-world datasets, and are reviewed by subject-matter experts to ensure domain relevance and fidelity.
- Through comprehensive numerical evaluations including assessments by human experts and an advanced Mixture-of-Experts (MoE) model acting as LLM-as-a-Judge, we demonstrate that EUROCOMPLY consistently achieves high levels of faithfulness, low bias, strong answer relevancy, and robust performance in terms of Mean Reciprocal Rank (MRR).

The remainder of this paper is structured as follows. Section II reviews related literature and identifies key research gaps. In Section III, we present the design of the EUROCOMPLY framework. Section IV outlines the experimental setup, including the benchmark datasets and evaluation metrics, and presents a detailed analysis of the results. Finally, Section V concludes the paper.

II. Related Work

In recent years, several initiatives have been introduced to address the complex challenge of ensuring that AI/ML models comply with evolving regulatory frameworks, particularly the European Union's GDPR and AI Act. A recent notable effort, COMPL-AI, developed by ETH Zurich, INSAIT, and LatticeFlow AI, translates the EU AI Act into technical checks for LLM compliance [3]. This initiative focuses specifically on such models, addressing areas such as documentation, bias detection, and transparency, while leaving gaps in cybersecurity and fairness. In contrast, our approach provides broader coverage across the entire MLOps lifecycle, including dataset sourcing, model training, evaluation, and deployment, offering a more comprehensive framework for regulatory compliance in real-world AI/ML systems. In parallel, several GDPR-focused tools have emerged. For instance, LegiLM is a fine-tuned legal language model trained on a comprehensive corpus of global data protection laws and annotated legal texts. It assesses data processing events for GDPR compliance and provides detailed legal reasoning [7]. Similarly, Amaral et al. [8] proposed an NLP-based method to automatically evaluate Data Processing Agreements (DPAs) against GDPR "shall" requirements. While effective, these solutions typically address specific GDPR principles such as Articles 5 and 35, and do not offer end-to-end compliance across the MLOps pipeline. With respect to the EU AI Act, Thelisson and Verma [9] examine the "General Approach" adopted by the European Council in November 2022. They outline governance structures for the conformity assessment of high-risk AI systems. While their framework offers useful regulatory insights, it overlooks the complete MLOps lifecycle, significantly limiting its practical applicability in real-world deployments.

A. Research Gap

Overall, the aforementioned solutions exhibit several significant limitations. (i) Most approaches address either the GDPR or the EU AI Act in isolation, without accounting for how these regulatory frameworks interact. (ii) They tend to focus narrowly on specific artifacts, such as contracts or policy documents. (iii) Many rely on manual, predefined modeling, which limits scalability and adaptability. (iv) They often lack integration with the broader MLOps development lifecycle, resulting in fragmented compliance efforts. (v) Finally, existing solutions are not tailored to the specific demands of the telecommunications sector, which involves complex and overlapping technical specifications.

To address these limitations, EUROCOMPLY introduces a novel and comprehensive compliance framework that automatically assesses both MLOps codebases and datasets against a wide spectrum of regulatory requirements, including the GDPR and the EU AI Act.

III. EUROCOMPLY Framework

In this section, we present the architectural design of EUROCOMPLY and outline its compliance workflow. We then introduce the Dual Agentic Retrieval Augmented Generation (RAG) mechanism, which leverages both vector-based and graph-based retrieval methods.

A. System Architecture

The proposed EUROCOMPLY framework employs an LLM-based self-reasoning swarm of agents to collaboratively address all phases of the MLOps pipeline. As shown in Fig. 1, the architecture is structured into three layers. The Infrastructure Layer provides the foundational data sources, focusing on next-generation 5G and 6G telecom systems, including the Radio Access Network (RAN), Core Network,

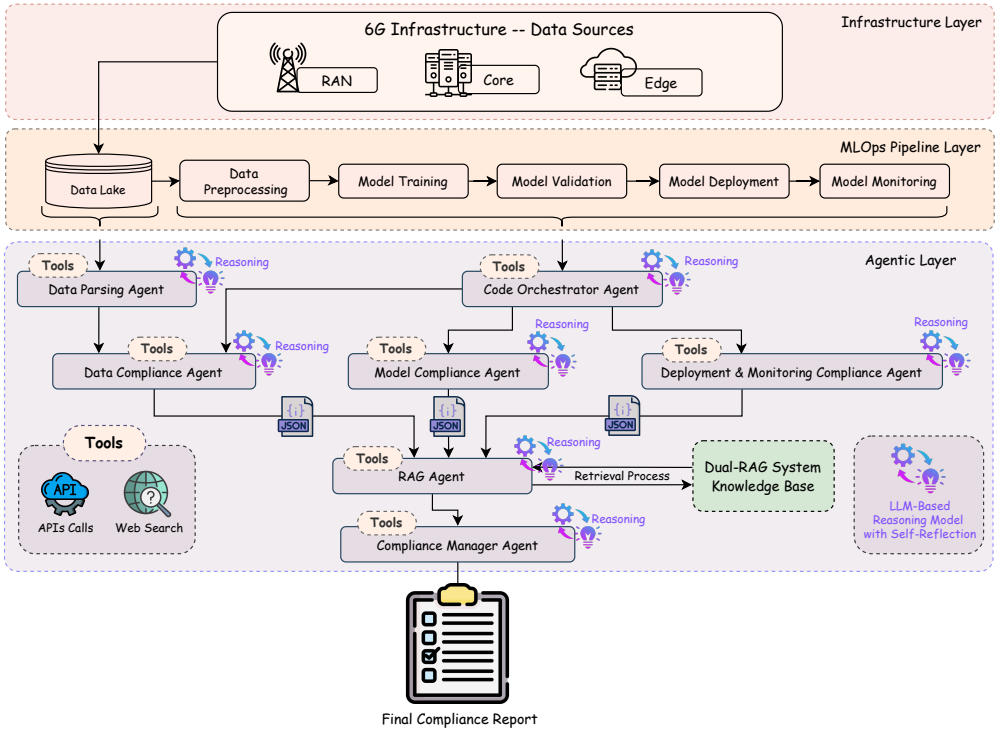


Fig. 1: EUROCOMPLY Agentic System Design for Compliance Verification in 6G-Enabled MLOps Pipelines Using LLM-Based Reasoning.

and Edge Network. The MLOps Pipeline Layer integrates core AI/ML processes, covering data preprocessing, model training, validation, deployment, and monitoring. Building on these, EUROCOMPLY introduces a novel **Agentic Layer**, a compliance-centric layer designed for the rapidly evolving regulatory landscape in AI and data governance. Here, a swarm of autonomous agents continuously performs context-aware compliance assessments across the AI/ML lifecycle. Each agent has a well-defined role, specialized responsibilities, and modular tools, including customized API calls for tasks such as data parsing, code analysis, and web search for real-time retrieval of regulatory and technical knowledge.

B. System Workflow

The system operates through a multi-agent workflow to ensure AI/ML pipeline compliance. The Data Parsing Agent loads datasets from the data lake, parses feature fields, and forwards randomized samples to the Data Compliance Agent, which, alongside preprocessing details from the *Code Orchestrator Agent*, evaluates privacy risks based on GDPR. Meanwhile, the Code Orchestrator Agent autonomously manages the AI/ML codebase, sending preprocessing code for data compliance checks, model code (including hyperparameters and architecture) to the Model Compliance Agent for validation against the EU AI Act and 3GPP standards, and deployment scripts to the Deployment and Monitoring Compliance Agent, which verifies monitoring, robustness, and transparency obligations. These compliance agents are designed around a clear checklist of tasks, ensuring modularity and separation of concerns.

Each compliance agent outputs structured JSON files summarizing their findings. The RAG Agent then enriches these reports by retrieving relevant legal clauses from a hybrid knowledge base combining graph-based and semantic search over the EU AI Act, GDPR, and 3GPP specifications. By linking each result to exact clauses and sections, the RAG Agent enhances traceability and auditability. Its exclusive focus on authoritative retrieval further minimizes LLM hallucinations. The enhanced JSON outputs, with explicit regulatory references, are consolidated by the Compliance Manager Agent into a comprehensive, human-readable Final Compliance Report, offering stakeholders a clear overview of regulatory adherence across the AI/ML pipeline.

C. Dual Agentic Retrieval Augmented Generation Mechanism

One of the defining features of EUROCOMPLY framework is its Dual Agentic Retrieval Mechanism, which enhances contextual compliance reasoning by dynamically switching between vector-based semantic search and graph-based structured queries. Operating over a hybrid knowledge base of vector and graph databases, the retrieval mode is selected by the Retrieval Agent based on query complexity. Simple, direct queries (e.g., retrieving GDPR clauses or AI Act definitions) are handled by the Vector Retrieval Agent for efficient semantic search. In contrast, complex, cross-regulatory queries (e.g., the interplay between GDPR data minimization and AI Act transparency) are routed to the Graph Retrieval Agent, which leverages knowledge graphs to capture interdependencies. This dual approach balances computational efficiency with retrieval depth and accuracy, while enabling seamless adaptation to

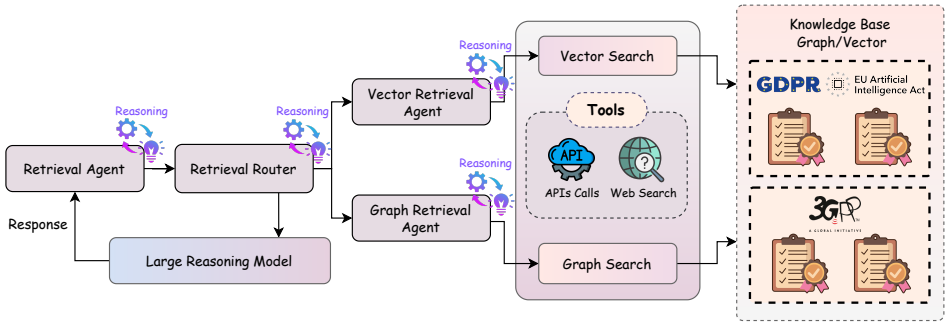


Fig. 2: The proposed Dual Agentic RAG architecture integrating Vector and Graph Retrieval for compliance knowledge reasoning.

regulatory updates by simply refreshing the knowledge base through the ingestion pipeline.

As shown in Fig. 2, the process starts when the *Retrieval Agent* receives a compliance-related query from the *RAG Agent* (see Fig. 1). The Retrieval Router evaluates the query's structure and intent, then routes it accordingly. The Graph Retrieval Agent draws from structured knowledge bases that interlink documents such as the EU AI Act, GDPR, and 3GPP standards. Meanwhile, the Vector Retrieval Agent searches semantically encoded legal and technical documents, retrieving relevant content even when exact terminology is absent. A Large Reasoning Model then synthesizes this information to provide comprehensive, contextually relevant, and traceable compliance responses.

Preprocessing regulatory documents varied in complexity depending on their content structure. While the EU AI Act and GDPR extraction proved straightforward due to their predominantly text-based clauses. In contrast, 3GPP TS 23.482 [2] (our primary 3GPP reference standards for AI/ML integration in telecom) posed greater challenges due to its extensive use of technical diagrams and network figures. To handle this, we used MiniCPM-V-2[4], a state-of-the-art Vision-Text model known for its strong visual comprehension. It converts diagrams and annotated visuals into rich textual descriptions that retain key technical semantics. These outputs are then seamlessly integrated into the hybrid vector-graph knowledge base.

IV. Performance Evaluation

In this section, we introduce the benchmark used to evaluate the proposed EUROCOMPLY framework, followed by an overview of the evaluation setup and metrics. We then present results from human expert assessments and an automated evaluation using an LLM-as-a-Judge approach.

It is important to emphasize that throughout our evaluation, we exclusively utilize **open-source LLMs** to assess the performance of our EUROCOMPLY framework. This decision is motivated by several key considerations. First, the use of proprietary closed-source LLMs can lead to significantly elevated costs, as agent-based systems often require frequent orchestration and generation calls, resulting in substantial billing even for moderately sized experiments. Second, recent advancements in open-source LLMs have produced models that are not only competitive but in some cases surpass their proprietary counterparts in both performance and flexibility, making them a compelling alternative. Finally, our commitment to open-source models aligns with broader goals of promoting transparency, democratizing the LLM ecosystem, and ensuring the reproducibility of research.

A. Telecom MLOps Compliance Benchmark

To rigorously assess the effectiveness of EUROCOMPLY, we constructed a specialized benchmark comprising 20 AI/ML telecom-specific use cases, primarily derived from 3GPP technical specifications and grounded in four realistic, telecom-focused datasets. Each dataset was analyzed to identify suitable AI/ML tasks (such as using LSTMs for traffic forecasting or autoencoders for anomaly detection), ensuring practical relevance. Furthermore, each use case was carefully labeled by human experts with interdisciplinary expertise across compliance and regulation, telecommunications, and AI, and then independently reviewed by external domain experts from EURECOM and OpenAirInterface (OAI), providing an additional layer of impartiality and ensuring high evaluation quality and objectivity. The use cases span a wide range of 5G and 6G infrastructure components, including the core network, RAN, and edge network. The datasets utilized for this benchmark are detailed as follows:

- **AMF Core Network Dataset [12]**: Captures CPU/memory usage, registration times, and session-level resource distribution from EURECOM's 5G testbed using OpenAirInterface.
- **Milano Dataset [13]**: Contains a year-long record of urban internet traffic from Telecom Italia, reflecting realistic usage patterns via normalized indicators.
- **Edge Network Dataset [12]**: Simulates RabbitMQ performance in a 5G edge network, measuring system metrics and message latency under varying workloads.
- **5G RAN Dataset [11]**: Provides multi-day traces from an Irish mobile operator, including detailed cellular KPIs like channel quality, contextual parameters, and throughput.

[4]https://huggingface.co/openbmb/MiniCPM-V-2

B. Evaluation Metrics

To evaluate the proposed EUROCOMPLY framework, a set of state-of-the-art metrics was employed to assess retrieval performance, agent-level performance, and overall system effectiveness. The following metrics are defined [14]:

- **Faithfulness**: Measures the degree to which the generated compliance reports accurately reflect the ground truth information. Scores from 1 (not faithful) to 5 (very faithful).
- **Answer Correctness**: Assesses the factual accuracy of the system's responses. Rated from 1 (factually incorrect) to 5 (factually accurate).
- **Answer Relevancy**: Evaluates how effectively the responses address key compliance topics and regulatory concerns. Rated from 1 (irrelevant) to 5 (highly relevant).
- **Bias Scores**: Captures the level of neutrality and objectivity in the generated compliance analysis and conclusions. Scored on a scale from 1 (highly biased) to 5 (completely unbiased).
- **Mean Reciprocal Rank (MRR)**: Quantifies the effectiveness of the retrieval system in prioritizing critical issues, based on the rank position of the first relevant item. MRR values range from 0 to 1, with 1 indicating perfect ranking.
- **Overall Report Score**: Evaluates the completeness, clarity, and regulatory alignment of the final compliance report and the JSON outputs from the compliance agents. Ratings are from 1 (poor quality) to 5 (excellent quality).
- **Average Response Time**: Measures the total time taken to complete the entire compliance verification process, including reasoning, report generation, enhancement, and final reflection steps.

C. Evaluation Setup

The proposed agentic AI compliance solution was implemented using CrewAI[5], which is an open-source framework designed to coordinate autonomous AI agents by assigning specific roles, objectives, and tools. For the base LLMs, we utilized open-source models hosted locally via Ollama[6]. All experiments were conducted on an NVIDIA Jetson AGX Orin machine, equipped with a 2,048-core Ampere GPU featuring 64 Tensor Cores and 64 GB of memory. For our evaluation, we selected state-of-the-art, open-source LLMs that are among the best-performing at the time of development and aligned with our available computational resources. Specifically, we included Llama 3.1–8B[7], DeepSeek-R1–8B[8], Mistral–7B[9], and Qwen2.5-Coder–7B[10]. For the LLM-as-a-Judge evaluation, we employed the Mixtral 8×7B MoE model [15], which dynamically routes tokens through two expert groups per layer, activating only 12.9 billion parameters from a total of 46.7 billion. For the knowledge database, we implemented the vector database using ChromaDB[11], which is a popular open-source vector database optimized for scalability and efficient embedding storage and retrieval. In parallel, we implemented the graph database using LightGraph[12], which constructs a knowledge graph by extracting entities and relationships from the source documents.

It is worth noting that, throughout all experiments, we used the same LLM for all agents within each setup. Mixing different LLMs across agents was not explored and is left for future work.

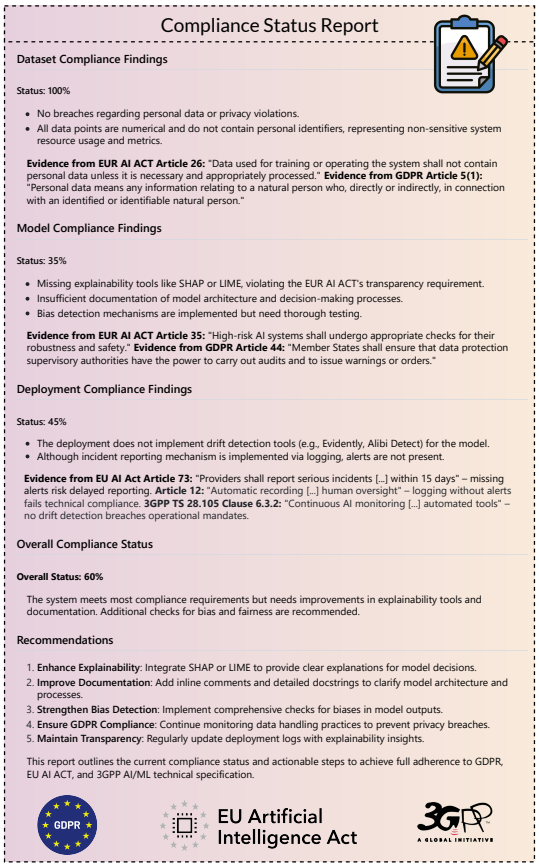
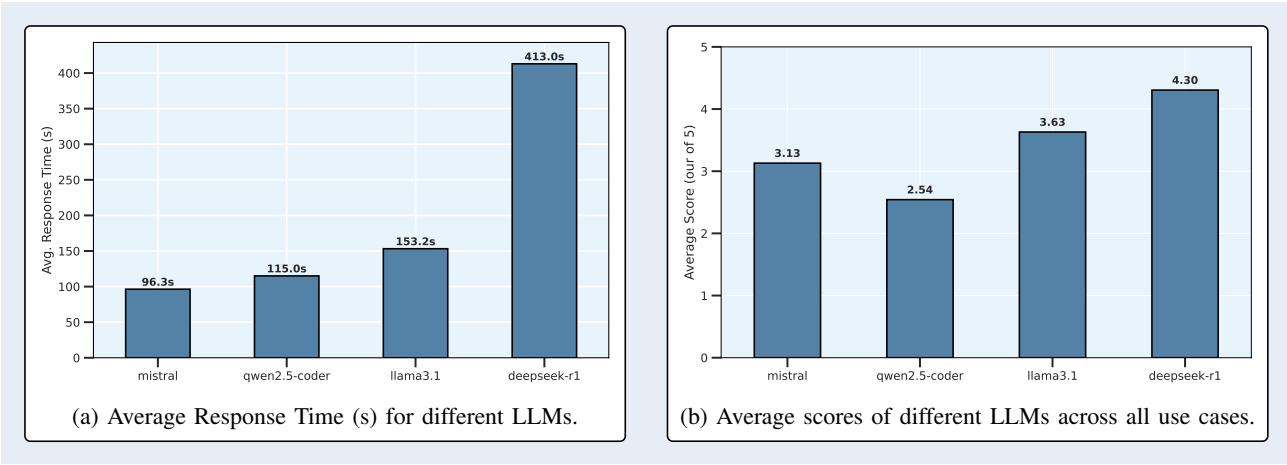


Fig. 3: An example of EUROCOMPLY compliance report using DeepSeek-R1:8B model as base LLM for the agents.

D. Results & Discussion

Fig. 3 presents a representative example of a EUROCOMPLY Compliance Report, demonstrating the system's ability to evaluate and document adherence to regulatory requirements across various stages of the AI lifecycle. The report includes an overall compliance score, complemented by a comprehensive breakdown of detected non-conformities related to datasets, model behavior, deployment practices, and monitoring frameworks. Each issue is explicitly linked to the relevant clauses of applicable AI regulations and 3GPP standards. Furthermore, the report delivers actionable recommendations tailored to address these shortcomings and strengthen the system's alignment with regulatory expectations.

[5]https://www.crewai.com/
[6]https://ollama.com/
[7]https://ollama.com/library/llama3.1
[8]https://ollama.com/library/deepseek-r1
[9]https://ollama.com/library/mistral
[10]https://ollama.com/library/qwen2.5-coder
[11]https://www.trychroma.com/
[12]https://lightrag.github.io/



(a) Average Response Time (s) for different LLMs.

(b) Average scores of different LLMs across all use cases.

Fig. 4: Comparison of LLM performance: (a) response time and (b) average ratings across use cases.

It is worth noting that EUROCOMPLY does not rely solely on non-technical insights from AI regulation. Instead, it enhances its assessments by combining the technical knowledge embedded in LLMs amplified through our chain-of-thought prompting with domain-specific expertise from 3GPP specifications. This approach adds technical depth to the report, making it valuable for both technical and non-technical audiences.

1) Overall Framework Performance: Fig. 4a presents the average response time of EUROCOMPLY during the compliance-checking phase, measured from the initiation of the multi-agent process to the completion of response generation. This evaluation leverages several open-source LLMs as the foundational models for agent reasoning and coordination. The results reveal significant variance in response times across different models, underscoring the impact of model architecture and inference efficiency on overall system latency. Notably, LLMs optimized for deeper reasoning, such as DeepSeek-R1 tend to require more time to generate responses due mainly to their reasoning and context interpretation processes.

2) Compliance Validation via Human-in-the-loop Evaluation: Fig. 4b further consolidates these findings by illustrating the overall average rating per LLM, based on human expert evaluation aggregated across all use cases. We can observe that DeepSeek-R1 stands out with an average score of 4.30 out of 5, significantly surpassing Llama 3.1 (3.63), Mistral (3.13), and Qwen2.5-Coder (2.54). These evaluations underscore the superiority of DeepSeek-R1 in generating accurate, contextually appropriate responses across a diverse set of compliance-checking scenarios. However, as previously discussed, DeepSeek-R1's strength in accuracy and reasoning comes at the cost of increased response latency, averaging 413s compared to sub-160s for other models. This introduces an important trade-off between accuracy and latency. Therefore, selecting an appropriate model depends on the specific requirements of the task. For complex, high-stakes code analysis, DeepSeek-R1's accuracy and response quality justify the increased latency. In contrast, time-sensitive or large-scale applications may benefit from faster models despite their comparatively reduced precision.

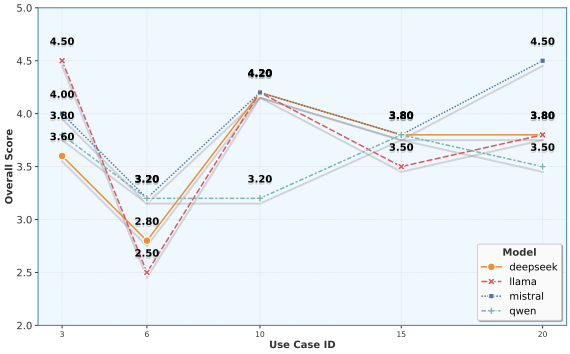


Fig. 5: Overall Framework scores using different LLMs evaluated by LLM-as-a-Judge.

3) Compliance Validation via Automated LLM-as-a-Judge Evaluation: To objectively evaluate the EUROCOMPLY framework, we employed Mixtral 8x7B Mixture of Experts model as a unified evaluator in our LLM-as-a-Judge pipeline. This automated assessment complements expert human review and enables scalable benchmarking across diverse compliance scenarios. Fig. 5 shows the overall framework performance for five use cases leveraging different LLMs: DeepSeek-R1, Llama 3.1, Mistral, and Qwen2.5-Coder. The x-axis represents the use case ID number, while the y-axis denotes the rating generated by the LLM judge. These five representative use cases are selected from four datasets; notably, use cases with IDs 3 and 6 are drawn from the same 5G RAN dataset due to its rich feature set, while use cases 10, 15, and 20 are from the Milano, AMF, and Edge datasets, respectively. As depicted in Fig. 5, the LLM judge consistently evaluates Llama 3.1 and Mistral with higher scores, peaking at 4.50 and 4.20, respectively, while Qwen2.5-Coder and DeepSeek-R1 underperform. These results stem from LLMs' tendency to overweight the beginning and end of reports due to attention and mixture-of-experts effects, often overlooking middle sections. Notably, human experts often disagreed with the LLM-as-a-Judge in such cases, as experts emphasized the overlooked middle content for compliance and contextual accuracy.

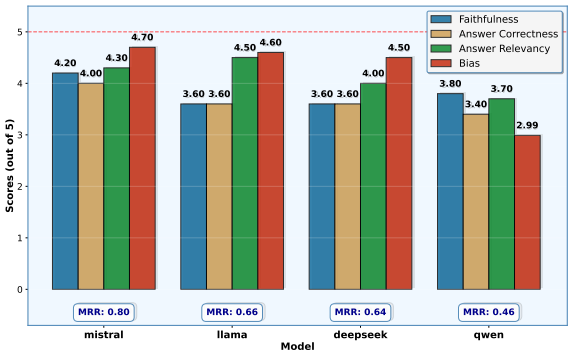


Fig. 6: Performance metrics comparison across models using LLM-as-a-Judge Evaluation.

Additionally, Fig. 6 details performance across four key dimensions: Faithfulness, Answer Correctness, Answer Relevancy, Bias, and MRR aggregated across all the use cases. Similarly, Mistral delivers the most balanced performance, scoring highly across all metrics according to the LLM judge. Llama 3.1 and DeepSeek-R1 show strong relevance and low bias, but somewhat reduced factual precision. Qwen2.5-Coder underperforms across the board, particularly in bias and relevance. From a retrieval standpoint, Mistral demonstrates superior performance in the retrieval process, as reflected by its leading MRR score of 0.80. This is followed by Llama 3.1 and DeepSeek-R1, both with scores around 0.66 and 0.64 respectively, while Qwen2.5-Coder trails with an MRR of 0.45. These results highlight Mistral as a strong candidate for serving as the base LLM in RAG systems. Moreover, the observed variation in responses across different LLMs is due to differences in architecture, reasoning mechanisms, model size, and training data.

It is noteworthy that, despite the substantial results achieved, a key limitation lies in the complexity of maintaining and locally hosting LLMs, which demand significant hardware resources. From a generalizability perspective, EUROCOMPLY can be readily applied to other domains by integrating it into existing MLOps pipelines beyond telecom use cases.

At the end, a demonstration video of EUROCOMPLY is available online[13]. In the video, we showcase EUROCOMPLY in action, highlighting its key features and functional capabilities.

V. Conclusion

In this paper, we introduced EUROCOMPLY, a novel framework grounded in the Agentic AI paradigm, designed to automate end-to-end compliance verification across all phases of the MLOps pipeline. By systematically analyzing data ingestion, training, validation, deployment, and monitoring stages, EUROCOMPLY provides stakeholders with detailed, actionable compliance reports aligned with GDPR, the EU AI Act, and 3GPP standards. Our evaluation, using 20 telecom AI/ML use cases and both domain expert and LLM-based assessments, confirms EUROCOMPLY's effectiveness in detecting regulatory breaches and recommending targeted remediation strategies.

[13]https://youtu.be/nczcZeryCxM

Future work will study heterogeneous LLMs and disagreements between experts and the LLM-as-a-Judge.

Acknowledgments

This work is supported by the European Union's Program under the 6G-DALI project (Grant No. 101192750).

References

[1] K. Trichias et al., "AI/ML as a key enabler of 6G networks: Methodology, approach and AI-mechanisms in SNS JU." Zenodo, 2025.
[2] 3gpp.org. Technical Specifications TS 23.482 [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=4312 [Accessed: 23-May-2025].
[3] P. Guldimann, A. Spiridonov, R. Staab, N. Jovanović, M. Vero, V. Vechev, A.-M. Gueorguieva, M. Balunović, N. Konstantinov, P. Bielik, P. Tsankov, and M. Vechev, "COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act," *arXiv preprint arXiv:2410.07959*, 2024.
[4] A. Weckler, "Meta 'pauses' AI data collection in EU following Irish DPC request," Irish independent, Irish Independent, 14-Jun-2024.
[5] Euronews.com. [Online]. Available: https://www.euronews.com/next/2024/12/10/openai-releases-ai-video-creator-sora-but-it-wont-be-coming-to-europe-yet [Accessed: 12-May-2025].
[6] Acharya, Deepak Bhaskar, Karthigeyan Kuppan, and B. Divya. "Agentic AI: Autonomous Intelligence for Complex Goals–A Comprehensive Survey." IEEE Access (2025).
[7] L. Zhu, L. Yang, C. Li, S. Hu, L. Liu, and B. Yin, "LegiLM: A Fine-Tuned Legal Language Model for Data Compliance," *arXiv preprint arXiv:2409.13721*, 2024.
[8] O. Amaral, M. I. Azeem, S. Abualhaija, and L. C. Briand, "NLP-Based Automated Compliance Checking of Data Processing Agreements Against GDPR," *IEEE Transactions on Software Engineering*, vol. 49, no. 9, pp. 4282–4303, 2023.
[9] E. Thelisson and H. Verma, "Conformity assessment under the EU AI act general approach," AI Ethics, vol. 4, no. 1, pp. 113–121, 2024.
[10] A. Goldsteen, G. Ezov, R. Shmelkin, M. Moffie, and A. Farkash, "Data minimization for GDPR compliance in machine learning models," AI Ethics, vol. 2, no. 3, pp. 477–491, 2022.
[11] Raca, Darijo, et al. "Beyond throughput, the next generation: A 5G dataset with channel and context metrics." Proceedings of the 11th ACM multimedia systems conference. 2020.
[12] M. Mekki, N. Toumi, and A. Ksentini, "Microservices configurations and the impact on the performance in cloud native environments," in 2022 IEEE 47th Conference on Local Computer Networks (LCN), 2022.
[13] Barlacchi, Gianni, et al. "A multi-source dataset of urban life in the city of Milan and the Province of Trentino." Scientific data 2.1 (2015): 1-15.
[14] M. Hindi, L. Mohammed, O. Maaz, and A. Alwarafy, "Enhancing the precision and interpretability of retrieval-augmented generation (RAG) in legal technology: A survey," IEEE Access, vol. 13, pp. 46171–46189, 2025.
[15] A. Q. Jiang, et al., "Mixtral of Experts," arXiv preprint arXiv:2401.04088, Jan. 2024. [Online]. Available: doi.org/10.48550/arXiv.2401.04088

Biographies

Mazene Ameur (M) is a Ph.D. candidate at EURECOM, researching AI integration in 6G networks. He actively participates in European projects such as 6G-DALI and SUNRISE-6G.

Bouziane Brik (SM) is an Assistant Professor at the University of Sharjah. He contributed to numerous European projects such as MonB5G and 5G-Drones.

Adlen Ksentini (SM) is a Full Professor at EURECOM and leads the Network Softwarization Group. He serves as the technical manager for major European projects like 6G-INTENSE and AC3, and is on the board of OpenAirInterface.