UDC 004.8:007:004.056.5
DOI https://doi.org/10.32689/maup.it.2025.3.1

Maryna BAUTINA
Master, Data Scientist, SoftServe
ORCID: *0009-0002-9617-9262*

DEVELOPMENT OF RELIABLE LLM SYSTEMS: DESIGN PRINCIPLES AND APPROACHES TO IMPLEMENTATION

Abstract. Purpose. *The article aims to provide a comprehensive analysis of architectural approaches and system solutions to ensure the reliability of services based on large language models (LLMs), as well as to develop principles and criteria for assessing the level of trust in applied scenarios.*

Methodology. *The study employs an interdisciplinary approach that combines the analysis of modern LLM architectures (zero-shot, fine-tuning, retrieval-augmented generation), a review of their implementation practices in corporate and industrial systems (GitHub Copilot, ChatGPT Enterprise), and a comparative synthesis of regulatory and ethical standards (OECD AI Principles, NIST AI RMF, EU AI Act). Methods of system analysis, comparative modeling, and the trust-by-design concept are applied.*

Scientific novelty. *The paper introduces the concept of building LLM-based services on the principles of trust-by-design, which relies on modular architecture, multi-level validation, and transparent response quality metrics. It is demonstrated that such integration of technical, ethical, and legal solutions enhances the resilience, transparency, and social responsibility of LLM in critical domains.*

Conclusions. *It is proven that establishing trust in LLMs is possible only under conditions of comprehensive integration of technical control mechanisms, ethical approaches, and legal regulation. The obtained results can be used to improve governmental and corporate strategies for artificial intelligence development, aimed at the safe and effective deployment of LLM in sectors with high reliability requirements.*

Key words: *large language models, LLM, trust, transparency, factuality, AI architecture, ethical AI, critical areas.*

Марина БАУТІНА. РОЗРОБКА НАДІЙНИХ СИСТЕМ LLM: ПРИНЦИПИ ПРОЕКТУВАННЯ ТА ПІДХОДИ ДО ВПРОВАДЖЕННЯ

Анотація. Мета. *Стаття спрямована на комплексний аналіз архітектурних підходів та системних рішень для забезпечення надійності сервісів на основі великих мовних моделей (LLM), а також на розроблення принципів і критеріїв оцінювання рівня довіри в прикладних сценаріях.*

Методологія. *У роботі застосовано міждисциплінарний підхід, що поєднує аналіз сучасних архітектур LLM (zero-shot, fine-tuning, retrieval-augmented generation), огляд практик їхнього впровадження у корпоративних і промислових системах (GitHub Copilot, ChatGPT Enterprise), а також порівняльне узагальнення нормативних і етичних стандартів (OECD AI Principles, NIST AI RMF, EU AI Act). Використано методи системного аналізу, порівняльного моделювання та концепцію trust-by-design.*

Наукова новизна. *Запропоновано концепцію побудови LLM-сервісів на засадах довіри за задумом (trust-by-design), що базується на модульній архітектурі, багаторівневій валідації та прозорих метриках якості відповідей. Показано, що така інтеграція технічних, етичних та правових рішень забезпечує підвищення стійкості, прозорості й соціальної відповідальності LLM у критично важливих сферах.*

Висновки. *Доведено, що формування довіри до LLM можливе лише за умов комплексної інтеграції технічних механізмів контролю, етичних підходів і правового регулювання. Отримані результати можуть бути використані для вдосконалення державних і корпоративних стратегій розвитку штучного інтелекту, спрямованих на безпечне та ефективне впровадження LLM у сферах з підвищеними вимогами до надійності.*

Ключові слова: *великі мовні моделі, LLM, довіра, прозорість, фактологічність, архітектура AI, етичний AI, критичні сфери.*

Introduction. Large-scale language models (LLMs) have become the foundational technology behind modern digital services, with widespread implementation across various domains, including education, healthcare, law, public administration, and cybersecurity. However, the extensive deployment of LLMs introduces new challenges: increasing complexity and unpredictability of outcomes, the risk of generating inaccurate or misleading content, susceptibility to adversarial attacks, and significant difficulties in ensuring ethical behavior, transparency, and accountability. These concerns are especially critical in areas where even minor errors may lead to serious consequences for individuals, institutions, or society at large.

In the context of LLMs' rapid integration into public and private information systems, the global community is intensifying efforts to develop approaches that align the technical, ethical, legal, and organizational dimensions of LLM implementation. Developers and regulatory bodies are prioritizing the evaluation and oversight of LLMs, with particular attention to system testing, output validation, interface adaptability, model trustworthiness, and the harmonization of standards with leading international frameworks.

Literature Review. In the scientific literature of recent years, the problems and opportunities related to the development of LLM are systematically studied in an interdisciplinary manner. E. M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell [13] emphasize the risks of generating so-called “hallucinations” and the potential spread of biases in large-scale language models. R. Bommasani et al. [14] analyze in detail the fundamental opportunities and risks of foundation models, noting their impact on various industries and the need for new approaches to ethics and responsibility.

The technical and engineering aspects of creating and testing LLMs are described in the works of OpenAI [15], D. Ganguli, A. Askell, Y. Bai, E. Hubinger, T. Henighan et al. [7] and J. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, H.-F. Song et al. [16], which emphasize the importance of multilevel testing, scaling, red-teaming procedures, and continuous model validation. L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P. Huang and co-authors [19] propose approaches to identifying ethical and social risks, and D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang and others [8] emphasize the development of “superalignment” strategies aimed at aligning models with the common values of humanity.

Regulatory and ethical documents and international standards, such as the OECD AI Principles [12], NIST AI RMF [11], and EU AI Act [6], which set global requirements for transparency, security, accountability, and fairness of artificial intelligence systems, also play an important role.

Thus, the current discourse in the field of LLM outlines the main directions: improving architectures and testing procedures, developing ethical and legal frameworks, assessing social impact, and implementing new technological solutions to increase trust and security.

The purpose of the article is to systematize the principles of designing reliable LLM systems, analyze modern architectural approaches to their implementation, and present the experience of integrating such models into practical services based on domestic and foreign developments.

Results. In recent years, large-scale language models (LLMs) have become not just an engineering achievement, but a key factor determining a new paradigm of digital transformation in the field of artificial intelligence. The evolution of these systems is taking place against the backdrop of rapid changes in the digital economy, education, medicine, public administration, and media space, where LLMs are gradually becoming an infrastructure component of modern information ecosystems. The architectural basis of such models remains transformational approaches, described in detail in the works of leading researchers in the field [2; 16; 20]. It was the development of self-attention and the subsequent scaling of computing power that became a catalyst for the emergence of a wide range of LLM implementations, including GPT, LLaMA, Claude, Gemini, Mistral, Grok, and others. The choice of a model is determined not only by the number of parameters but also by the type of learning environments, the level of openness of the architecture, support for multimodality, and the possibility of flexible customization.

A number of studies have convincingly demonstrated that the performance and quality of language models directly depend on the scale of the architecture, the richness and diversity of training data, and the use of sophisticated engineering solutions to optimize training processes [2; 16; 20]. However, the very growth of scale gives rise to new challenges, including energy efficiency, control of computing costs, transparency of internal processes, and the need to comply with ethical standards. The emergence of the ideas of modular system construction, energy-efficient training, and the use of architectures with dynamic involvement of expert submodels allows flexible adaptation of LLMs to the requirements of specific applications.

Modern researchers pay special attention to the problem of so-called “hallucinations” – a phenomenon when a model produces grammatically correct, but actually incorrect or fictional content [13; 14; 19]. This is due to the over-reliance on statistical patterns in the data and the lack of built-in fact-verification mechanisms, which encourages developers to create additional means of verifying the accuracy of answers. In addition, security issues are becoming increasingly important in the practice of using LLMs, in particular, preventing prompt injection and jailbreak attacks that can push the system beyond the limits of controlled behavior. In such situations, it is extremely important to implement multi-level strategies for protection, query filtering, and result moderation, which is confirmed by research in the field of ethical use of artificial intelligence [7; 8; 11].

The task of ensuring the ethicality and absence of bias in LLM responses is of particular importance today, as these models are increasingly used in sensitive social areas. Studies on the ethical evaluation of language models emphasize that even modern systems can reproduce stereotypes, hidden forms of discrimination, or form a misperception of information [13; 14; 19; 12]. In response to these threats, a powerful area of Fairness-Aware NLP has emerged, which involves the development of specialized approaches to neutralize bias and increase the transparency of model decision-making.

The problem of instability of generation results associated with the probabilistic nature of LLM operation creates additional difficulties in ensuring reproducibility, quality control, and audit. In this context, leading engineers and researchers emphasize the need to implement traceability mechanisms, comprehensive audit, and dynamic monitoring of model performance [15; 20]. The experience of developing and applying LLMs shows that implementation in critical systems requires not only high performance but also proven stability and controllability of behavior.

At the same time, against the background of technological innovations, there is a growing awareness of the socio-technical impact of LLM on the labor market, the structure of professional activity, and educational processes. Analysts warn that automation caused by the introduction of LLM can significantly transform specialized areas, which, in turn, increases the requirements for institutional responsibility, transparency and verifiability of results, as well as for the implementation of ethical and legal regulation standards [14; 12; 11]. It is these aspects that stimulate the development of international regulatory frameworks, such as those of the OECD, NIST, and the European Union, which define the basic principles of transparency, data protection, and accountability in the field of artificial intelligence [12; 11; 6]. In the Ukrainian context, similar challenges are reflected in current applied research and LLM implementation projects in various industries.

To provide a meaningful comparative overview of existing large language models (LLMs), it is insufficient to list only technical parameters such as size or architecture. Instead, it is more informative to analyze how well each model supports the needs of critical applications – where trust, explainability, deployment control, and legal compatibility matter most. Table 1 summarizes the readiness of selected LLMs across five strategic dimensions that are vital for their integration in education, healthcare, legal and administrative services.

Table 1

Comparative characteristics of modern LLMs (as of 2024)

Model	Trust Infrastructure	Customization Options	Explainability Tools	Deployment Flexibility	Regulatory Compliance Potential
GPT-4	Advanced moderation; closed logs	Low (API only)	Limited	Cloud-based only	Medium (black-box limitations)
Claude 3	Ethical alignment focus (Constitutional AI)	Medium (via APIs, fine-tuning under NDA)	Moderate (context tracking, summaries)	Cloud-based only	Medium-High (strong safety by design)
LLaMA 3	Basic filters in open-source weights	High (open weights)	None built-in	Local, hybrid, cloud	High (full audit possible)
Gemini 1.5	Proprietary Google stack with strong sandboxing	Low (API only)	High (for vision-text tasks)	Cloud only (TPU-dependent)	Medium (unknown data provenance)
Grok	Minimal public safeguards known	Low	Not available	Tied to xAI cloud	Low (opaque logic)
Mixtral (MoE)	Sparse MoE routing, open logs	High	Requires external tools	Flexible (self-hosted)	Medium-High (transparent stack)
BLOOM	Full openness, community moderation	High	Optional (via plugins)	Full: local/cloud/hybrid	High (complete reproducibility)

Note: the exact amount of GPT-4 is estimated to be undisclosed.

Source: author's elaboration

The restructured comparison makes it evident that the success of LLM deployment in sensitive domains depends not only on performance, but on the synergy between transparency, customization capacity, and infrastructural control. Open models like BLOOM and LLaMA 3 offer maximal auditability and flexibility, making them suitable for regulated industries with in-house engineering capacity. Conversely, models like GPT-4 or Gemini 1.5, despite their sophistication, are more difficult to align with transparency and explainability requirements due to proprietary constraints.

This shift in evaluation – from parameter-based ranking to implementation-based readiness – reflects the growing consensus in AI policy and research communities that trustworthiness is not a property of the model alone, but of the entire service ecosystem in which it operate.

The issue of system reliability implies that the service must consistently maintain its functionality, fulfill all assigned roles, and respond promptly to changes in load or atypical user behavior. The approach to determining reliability, as proposed by the NIST standard, covers a number of key aspects: ensuring constant access to the service, insensitivity to incorrect or malicious requests, and the ability to respond quickly to incidents by switching to a safe mode [11]. Especially important is the ability of LLM to work with fail-safe logic, a response system that guarantees data safety and prevents the dissemination of incorrect or dangerous information in the event of abnormal situations.

The reliability paradigm of LLM services is increasingly combined with the principles of transparency, manageability, and compliance with ethical standards, as models not only fulfill technical tasks but also act as intermediaries in the interaction between humans and complex digital systems. This requires the creation of architectures that provide for detailed monitoring, flexible management of access parameters, built-in moderation, and automated audit of responses [19; 6]. It is also important to integrate mechanisms for identifying anomalies, storing interaction logs, and dynamically updating system policies in response to changes in the regulatory environment or new threats identified.

Recognition of reliability as a fundamental property of LLM is confirmed both in the regulatory documents of international organizations and in the results of research by leading scientists, who emphasize that only an integrated approach to the design and operation of such systems can guarantee their safety, sustainability, and efficiency in dynamic and potentially risky application scenarios [19; 8; 11; 6].

In current LLM research, trustworthiness is conceptualized as a composite of security, factual accuracy, transparency, response consistency, and personal data protection. As defined by NIST experts, a trustworthy AI system must produce accurate, interpretable, and secure outputs, operate transparently, and preserve user confidentiality [11; 5]. For LLMs, this entails both factual and explainable responses, as well as mechanisms to prevent misinformation and enable traceability.

A promising technical strategy to enhance reliability is model aggregation – combining multiple LLMs within one infrastructure to balance loads, compare outputs, and activate fallback models in case of failure. However, designing such systems involves a trade-off between response speed and explainability. Lightweight models like LLaMA 3 or Mistral offer faster processing but less interpretability, while larger models (e.g., GPT-4, Claude 3) provide deeper reasoning at the cost of latency. More accurate outputs via complex validation pipelines (e.g., RAG) can further slow response time and obscure the logic of generation, reducing user trust [11; 6].

This creates a fundamental compromise: either prioritize transparency with slower, auditable outputs, or optimize speed at the expense of interpretability. Regulatory guidance from NIST and the EU AI Act suggests that in high-risk domains (e.g., healthcare, legal services), verifiability should take precedence [11; 6].

Another foundational design principle is architectural modularity. By organizing services as microcomponents – separating input parsing, generation, validation, and logging – LLMs become more fault-tolerant, scalable, and easier to update without disrupting the whole system [20; 11]. Middleware tools such as API gateways and multi-level caching are essential for load control and efficiency [11].

Finally, personalization is becoming a key vector of reliability. Adaptive mechanisms such as dynamic prompt shaping, context retention, and vector-based user profiling allow models to tailor responses while enhancing relevance [7; 11]. Yet, personalization demands strict adherence to privacy protocols: encrypted storage, time-limited data retention, and full transparency in user data handling [8; 11].

Modern LLM implementations rely on three dominant architectural approaches: zero-shot learning, fine-tuning, and retrieval-augmented generation (RAG). Zero-shot models are applied without task-specific training, allowing rapid scaling but often sacrificing factual accuracy due to lack of contextual adaptation [20]. Fine-tuning enhances precision and relevance in defined domains but raises risks of overfitting and model bias [13]. RAG integrates external knowledge sources or search engines into the generation pipeline, improving transparency and fact-checking, albeit at the cost of integration complexity and potential source inconsistency [10].

Each approach entails trade-offs in trust and performance. As shown by P. Lewis et al. [10], zero-shot offers speed but lacks transparency and reliability. Fine-tuning, according to T. B. Brown et al. [20], delivers domain-specific precision but demands rigorous data governance. RAG is particularly promising for trustworthy applications, combining generative flexibility with verifiable content, though it requires strict safeguards for source quality and data protection [10].

Ultimately, LLM reliability emerges not solely from model architecture, but from the integration of technical, infrastructural, algorithmic, and ethical components that support stable and controllable system behavior [14]. As R. Bommasani et al. emphasize, effective deployment also depends on understanding user needs, contextual risks, and regulatory constraints [14].

Recent trends point to hybrid architectures, where combining different strategies-modularity, traceability, and auditability – enables improved transparency, resilience, and adaptability [5]. Modularization of LLM workflows facilitates granular control over response generation, verification, and logging. Explainability tools and prompt validation systems are increasingly embedded in LLM pipelines to support trustworthy operations [10; 15; 17].

Comprehensive trust metrics now assess not only factual accuracy and process transparency but also resistance to prompt manipulation, user satisfaction, and compliance with privacy standards. These align with global frameworks such as NIST AI RMF and the EU AI Act [11; 6; 5]. As LLMs become integral to sectors like education, healthcare, law, and finance, their deployment demands cohesive architectural and operational strategies that meet high standards of security, ethics, and usability [14; 20; 8].

Modern LLM services are often deployed via bots, web portals, or SaaS platforms using standardized APIs, offering rapid integration, scalability, and support for diverse tasks [15; 11]. Reliable integration with providers like OpenAI or Anthropic requires RESTful APIs with layered authorization, middleware for filtering and routing, and safeguards against abuse [15; 7; 6; 20].

Scalability is achieved through containerization and asynchronous backends using tools like FastAPI, Redis, and Celery, optimizing performance under high load [20; 11]. Prompt engineering and context management – such as using templates, context windows, and intent detection – directly affect response quality [13; 10].

Output validation now combines manual review, automated metrics, and semantic checks to detect hallucinations. Content moderation and logging are integral to ensuring trust and traceability [19; 11; 5].

Security risks include hallucinated content, jailbreaks, and prompt injections, which demand multi-level moderation – both pre- and post-generation [7; 1]. Techniques like red-teaming expose hidden vulnerabilities and improve model robustness [7].

Ethical use in education, healthcare, and law requires strict oversight, transparency, and human control. Systems like GPT-4 and Claude 3 implement layered audits and safety protocols [15; 8; 17].

Ultimately, trust in LLMs depends on robust infrastructure, ethical safeguards, and continuous adaptation to new risks – combining engineering, regulation, and responsible design [19; 8].

The practical adoption of large language models has been enabled by a flexible technological stack that integrates modern AI libraries, distributed architectures, and tools for seamless deployment. As noted by T. B. Brown et al. [20], the Python ecosystem remains the dominant platform for LLM integration, combining FastAPI (REST API development), LangChain (context management), Gradio (UI integration), and Hugging Face Transformers (open-source models).

LLM-based automation is transforming front-end development. N. A. Ikumapayi [9] shows that OpenAI-assisted code generation reduces development time and boosts efficiency. Similarly, S. Shen et al. [18] emphasize the role of domain knowledge in automating JavaScript code.

Efficient orchestration between multiple LLMs is increasingly used to match model capabilities with query types. Middleware-based intent classification enables dynamic switching – e.g., GPT-4 for deep analysis, Claude for extended context, and LLaMA or Mistral for fast, lightweight queries – improving both cost-efficiency and relevance [14].

Preprocessing safeguards are also advancing. Semantic prompt validation and automated filters, as discussed by D. Ganguli et al. [7], mitigate the risk of inappropriate outputs-particularly in high-stakes domains like education, law, and healthcare.

Impact assessments confirm substantial gains: L. Weidinger et al. [19] and A. Almalki & M. Aziz [1] report that basic LLM integration reduces response times by up to 55%, improves satisfaction by 15–20%, and cuts redundant queries in half. These benefits extend to financial, educational, and corporate sectors.

Successful deployments hinge on modular backend tools, adaptive model routing, multi-level validation, and user-centered design. Together, they enhance efficiency, trust, and adaptability across applications. Enterprise ChatGPT offers encryption, moderation, and audit logging, though final validation of outputs often remains user-dependent [1].

These real-world cases demonstrate that robust infrastructure, adaptive logic, and ethical oversight are critical for scaling LLM-based services. Table 2 summarizes key architectural and technological components used in such implementations.

In the context of increasing reliance on large language models in high-stakes domains, the development of a clear and structured framework for evaluating the effectiveness and trustworthiness of such systems is a necessary precondition for their safe and meaningful integration. Given the multifaceted nature of trust in LLMs – combining technical, ethical, and user-centered dimensions – it is advisable to employ a layered approach that links operational performance with design principles such as explainability, transparency, and resilience to manipulation.

Table 2

Comparison of technical solutions for implementing LLM services

Implementation component	Technologies / approaches	Advantages	Limitations / Risks
Server platform	Python (FastAPI), Node.js (Express)	Flexibility, a large base of ready-made libraries, quick launch	Python – lower performance at high loads
UI framework	Gradio, Streamlit	Low-code, integration with ML models, easy customization	Limited scalability, poor style control
LLM integration	OpenAI API, Anthropic API, HuggingFace	Easy connection, support for state-of-the-art models	Cost of tokens, dependence on external servers
Model management	Multiplexer, fallback mechanisms	Flexibility, load balancing, improved relevance	Complication of request routing logic
Prompt processing	Prompt templates, context chaining	Standardization of input, repeatability of results	Vulnerability to prompt injection and jailbreak attacks
Service scaling	Docker, Redis, Celery	Performance control, horizontal scaling	Requires a high level of DevOps competencies
Assessment of response quality	BLEU, cosine similarity, RL feedback	Possibility of automatic monitoring	Does not always correlate with subjective quality of perception
Content filtering	Regex filters, NLP moderation	Reducing toxicity and inappropriate responses	The need to constantly update the rules

Source: author's development

The proposed framework is organized as a stepwise evaluation system that enables both developers and institutional stakeholders to monitor key indicators of system reliability, while adapting the configuration of LLM services to meet evolving demands. This methodology supports continuous assessment across five interrelated dimensions, each of which corresponds to critical functional and governance aspects of LLM use in real-world applications (Table 3).

The systematization of metrics for evaluating the trustworthiness and effectiveness of LLM-based services is an important step toward the operationalization of ethical and technical standards in the deployment of artificial intelligence in critical domains. The five-step model presented in Table 3 outlines an integrated approach that connects the behavior of large language models at the micro level (generation quality, factuality, explainability) with macro-level governance concerns such as security, resilience, user trust, and regulatory compliance.

Table 3

System for evaluating trust and effectiveness of LLM services

Evaluation Dimension	Key Indicators	Tools / Methods
Factual reliability	% of accurate responses; hallucination rate; cross-checks with trusted sources	Fact-checking APIs, knowledge bases, manual annotation
Transparency and explainability	% of responses with source trace; availability of reasoning; logging completeness	Explainability modules, audit logs, prompt tracing systems
Resistance to manipulation	Blocked prompt injection attempts; jailbreak prevention rate	Red teaming, semantic filters, adversarial testing scripts
User-perceived performance	SUS/NPS scores; task completion rate; response time under load	UX analytics, user surveys, load simulation
Security and ethical compliance	Number of flagged outputs; privacy breaches; audit trail availability	Moderation tools, encryption logs, regulatory compliance dashboards

Source: author's development

Each step reflects not only an isolated dimension of system performance, but also a specific aspect of trust formation in digital environments, thereby enabling a layered interpretation of model behavior in real-life applications.

The first stage, focused on factual reliability, addresses the core epistemic expectation from LLMs – that generated responses correspond to verifiable truths rather than plausible-sounding fabrications. This dimension plays a foundational role in ensuring semantic integrity, especially in knowledge-intensive domains such as medicine, law, and education, where the cost of factual errors may be high. Measuring factuality requires both quantitative benchmarking against curated datasets and qualitative assessment through independent validation pipelines.

The second component, transparency and explainability, refers to the model's ability to provide justifiable and auditable outputs. In contrast to black-box generation, transparent systems enable the user or auditor to trace how an answer was constructed, what data or logic informed it, and where potential limitations lie. This aspect is particularly relevant for institutional accountability, where LLM outputs must be interpretable within legal or organizational frameworks. The presence of logging mechanisms, attention visualization, and reasoning trace tools significantly enhances the traceability of LLM behavior and contributes to the auditability of AI systems.

Manipulation resilience, the third pillar of the model, responds to the growing threat of adversarial use, such as prompt injections and jailbreaks, which may bypass system constraints or provoke the generation of harmful content. This dimension evaluates the robustness of LLM configurations under malicious scenarios and provides an early-warning function for identifying system vulnerabilities. Its role is especially significant in public-facing deployments or in sectors with strict reputational and safety requirements, such as healthcare portals or government service platforms.

The fourth dimension, user-perceived performance, introduces the human-centered perspective into the evaluation framework. Here, the emphasis shifts to how users experience interaction with the system, measured through satisfaction indices (e.g., SUS, NPS), task completion rates, and latency under operational loads. This dimension reflects the usability and perceived reliability of the LLM and is essential for long-term adoption, especially in dynamic service environments such as online education platforms, business assistants, or content generation services. The inclusion of this layer ensures that trust is not only engineered but also experienced and sustained over time.

Finally, the fifth stage – security and ethical compliance – consolidates the governance layer of the model. It encompasses indicators related to privacy protection, content moderation, and conformity with regulatory frameworks such as the GDPR, the EU AI Act, or sector-specific AI guidelines. This stage is vital in domains that involve personal data processing, sensitive information handling, or legally binding documentation. It integrates the results of internal audits, external risk assessments, and ethical reviews, allowing the organization to monitor and update its LLM deployment in line with evolving legal and normative expectations.

Taken together, the five-step system offers not merely a diagnostic instrument, but a dynamic infrastructure for continuous evaluation and adaptation. It facilitates informed decision-making for developers, system architects, compliance officers, and institutional stakeholders by enabling early detection of emerging risks, measuring the long-term reliability of LLM services, and supporting iterative improvement. From a practical standpoint, the application of this framework allows organizations to align technical performance with ethical values and user needs, thus forming a coherent and operational definition of trustworthiness in AI systems. It supports the transition from model-centric to service-centric evaluation, where trust is seen as a function of both algorithmic behavior and sociotechnical context. Ultimately, such a multidimensional model contributes to the institutionalization of responsible AI deployment practices, which is crucial for sustainable integration of LLM technologies into sensitive societal and economic infrastructures.

Conclusions. This study demonstrates that the development of reliable LLM systems requires an integrated approach in which architectural modularity is effectively combined with well-designed control and validation mechanisms at every level. The success of such solutions is reflected in their ability to maintain system stability and security under high loads, efficiently manage request processing, and ensure response quality through the use of intermediate control, semantic validation, resource balancing, and modern API gateways. As evidenced by the analysis of practical implementations, LLM services are becoming drivers of personalization and digital platform optimization, significantly reducing users' cognitive load, particularly in education, consulting, automated support, and related sectors.

Nevertheless, considerable challenges remain, especially regarding the potential generation of toxic or inaccurate responses in sensitive domains such as medicine, law, and psychology. Insufficient oversight and inadequate ethical moderation pose risks to user trust and may undermine the social legitimacy of deployed systems. Experience from recent master's and applied research projects confirms the importance of employing advanced technological tools, including FastAPI, LangChain, Docker, and Gradio, in building functional and scalable LLM infrastructures. The use of hybrid architectures that enable dynamic switching between models based on specific requests supports the optimization of quality, cost, and processing speed.

The future of LLM services is closely tied to the expansion of multimodal capabilities, the alignment of interfaces with ethical standards, and the improvement of model relevance through state-of-the-art reinforcement learning techniques. Particular attention should be directed toward research on how different professional and age groups cognitively perceive LLM outputs, as such insights facilitate the customization of systems to individual user needs.

In conclusion, the development of trustworthy LLM systems extends beyond a technical challenge and evolves into an interdisciplinary endeavor that integrates advances in artificial intelligence, cybersecurity, UX design, cognitive science, and ethics. The complexity and coordination of these approaches are essential for creating systems that are not only high-performing but also socially responsible, safe, and acceptable for use in critical sectors.

Bibliography:

1. Almalki A., Aziz M. Exploring the potential and challenges of ChatGPT in enterprise contexts. *IEEE Access*. 2023. Vol. 11. P. 85339–85349. URL: https://doi.org/10.1109/ACCESS.2023.3328700 (date of access: 12.07.2025)

2. Brown T. B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P. et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*. 2020. Vol. 33. P. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (date of access: 12.07.2025)

3. Brundage M., Avin S., Clark J., Toner H., Eckersley P., Garfinkel B. et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint* arXiv:2004.07213. 2020. URL: https://arxiv.org/abs/2004.07213 (date of access: 12.07.2025)

4. De Angelis S., Cirillo F., Mazzocca N., Palmieri F. A trustworthy AI framework for explainable artificial intelligence in critical domains. *IEEE Access*. 2023. Vol. 11. P. 44792–44806. URL: https://doi.org/10.1109/ACCESS.2023.3275093 (date of access: 12.07.2025)

5. European Union. Regulation (EU) 2024/1687 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *Official Journal of the European Union*. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1687/oj (date of access: 12.07.2025)

6. Ganguli D., Askell A., Bai Y., Hubinger E., Henighan T. Red teaming language models to reduce harms: methods, results, and lessons learned. *arXiv preprint* arXiv:2309.00603. 2023. URL: https://arxiv.org/abs/2309.00603 (date of access: 12.07.2025)

7. Hendrycks D., Burns C., Kadavath S., Arora A., Basart S., Tang E. et al. Overview of the Superalignment Plan. *OpenAI Blog*. 2023. URL: https://openai.com/blog/superalignment (date of access: 12.07.2025)

8. Ikumapayi N. A. Automated front-end code generation using OpenAI: empowering web development efficiency. *Available at SSRN 4590704*. 2023. URL: https://doi.org/10.2139/ssrn.4590704 (date of access: 12.07.2025)

9. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*. 2020. Vol. 33. P. 9459–9474. URL: https://arxiv.org/abs/2005.11401 (date of access: 12.07.2025)

10. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST. 2023. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (date of access: 12.07.2025)

11. OECD. OECD Principles on Artificial Intelligence. Organisation for Economic Co-operation and Development. 2021. URL: https://oecd.ai/en/dashboards/ai-principles (date of access: 12.07.2025)

12. On the Dangers of Stochastic Parrots / E. M. Bender et al. *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada. New York, NY, USA, 2021. URL: https://doi.org/10.1145/3442188.3445922 (dat e of access: 14.07.2025).

13. On the opportunities and risks of foundation models / R. Bommasani et al. URL: https://samuelalbanie.com/files/digest-slides/2022-06-foundation-models-opportunities-and-risks-intro.pdf (date of access: 12.07.2025)

14. OpenAI. GPT-4 Technical Report. 2023. URL: https://cdn.openai.com/papers/gpt-4.pdf (date of access: 12.07.2025)

15. Rae J., Borgeaud S., Cai T., Millican K., Hoffmann J., Song H. F. et al. Scaling language models: methods, analysis & insights from training Gopher. *arXiv preprint* arXiv:2112.11446. 2021. URL: https://arxiv.org/abs/2112.11446 (date of access: 12.07.2025)

16. Sandoval G. GitHub Copilot has a copyright problem. *The Verge*. 2023. URL: https://www.theverge.com/23602854/github-copilot-ai-copyright-microsoft-openai-lawsuit (date of access: 12.07.2025)

17. Shen S., Zhu X., Dong Y., Guo Q., Zhen Y., Li G. Incorporating domain knowledge through task augmentation for front-end JavaScript code generation. *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2022. P. 1533–1543. URL: https://doi.org/10.1145/3540250.3558965 (date of access: 12.07.2025)

18. Weidinger L., Mellor J., Rauh M., Griffin C., Uesato J., Huang P. et al. Ethical and social risks of harm from language models. *arXiv preprint* arXiv:2112.04359. 2021. URL: https://arxiv.org/abs/2112.04359 (date of access: 12.07.2025)

19. Zhuang F., Qi Z., Duan K., Xi D., Zhu Y., Zhu H. et al. A comprehensive survey on transfer learning. *Proceedings of the IEEE*. 2020. Vol. 109, No. 1. P. 43–76. URL: https://ieeexplore.ieee.org/document/9153870 (date of access: 12.07.2025)

Дата надходження статті: 09.09.2025
Дата прийняття статті: 20.10.2025
Опубліковано: 04.12.2025