Received: 13 August 2025 | Revised: 1 December 2025 | Accepted: 15 January 2026 | Published online: 3 February 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–11
DOI: 10.47852/bonviewAIA62027195
BON VIEW PUBLISHING

Risk Assessment of Large Language Model Implementation in the Electric Power Sector of Ukraine

Hryhoriy Kravtsov[1], Oleksandr Kravchuk[1,*], Artem Taranowski[1], Dmytro Sinko[1], and Victor Samoylov[1]

[1] Department of Mathematical and Computer Modeling, G.E. Pukhov Institute for Modelling in Energy Engineering of the NAS of Ukraine, Ukraine

Abstract: This study explored the key aspects and risks associated with the implementation of large language models (LLMs) in the electric power sector of Ukraine. We propose a unique taxonomy of risks, along with a hierarchical structure that enables their assessment using the analytic hierarchy process (AHP) developed by T. Saaty. The LLM lifecycle is described with a focus on both human and technological factors (from knowledge selection and training to operational deployment). The study addresses critical concerns related to confabulations, sensitive information leakage, compliance with personal data protection regulations, and the safeguarding of trade secrets. The paper highlights the importance of employing tools for hallucination detection, sentiment analysis, and legal compliance monitoring. A separate section presents an in-depth analysis of LLMs' readiness to accurately digitize graphical content—such as schematics, diagrams, and technical drawings, which are common for documentation in the energy sector worldwide. A series of experiments using the state-of-the-art generative AI systems revealed significant limitations in interpreting complex diagrams, logical structures, and semantic relationships between elements. The findings demonstrate both the potential and the critical limitations of LLMs in energy-related applications, particularly in processing graphical content, making decisions based on synthetic data, and managing risks associated with model training, operation, and upgrades.

Keywords: electric power sector, large language model (LLM), generative AI, risk assessment, accountability

1. Introduction

Large language models (LLMs) are predominantly generative artificial neural networks [1] built on a decoder-only transformer-based architecture. Because of their ability to process, interpret, and generate domain-specific technical information, LLMs are increasingly adopted across industries. Within the energy sector, LLMs are used for tasks such as forecasting and optimization, anomaly detection, predictive maintenance, cybersecurity assistance, operator decision support, and automation of reporting [2, 3]. This expanding applicability demonstrates the growing role of LLMs in supporting both operational and strategic processes in energy systems.

Recent publications highlight that LLM-based solutions can improve fault detection accuracy, enhance situational awareness, assist in early hazard identification, and reduce cognitive workload for personnel in energy enterprises [4]. However, despite these advancements, existing studies primarily focus on demonstrating the technical potential and performance improvements delivered by LLMs. High-level AI governance frameworks—such as the NIST AI Risk Management Framework [5] and the EU AI Act risk-based classification scheme [6]—provide general guidelines, but they do not offer quantitative, domain-specific methodologies tailored to assess the risks associated with LLM deployment in the energy sector.

A gap remains in the literature regarding structured and quantitative approaches to evaluating LLM-related risks within critical infrastructure environments. While several studies have highlighted the benefits and potential use cases, none provide a method for systematically identifying, categorizing, and prioritizing risks relevant to the unique operational, technological, and regulatory context of the energy sector.

2. Problem Statement

The absence of a tailored methodology introduces uncertainty for organizations considering the integration of LLMs into critical infrastructure. To address this gap, the present study develops a method for assessing the risks of implementing AI in the form of LLM at critical infrastructure facilities. The method is designed to satisfy the following key requirements:

1) adaptability of the methodology to the needs of specific industry organization;
2) transparency, visual clarity, representativeness, and the ability to clearly communicate the results;
3) reliance on established and validated analytical methods.

To guide the development of the method, the study examines the following questions:

1) Which human- and technology-related risks are most relevant to the implementation of LLMs in the energy sector?
2) How can these risks be systematically identified, categorized, and prioritized using established analytical techniques?
3) How can the assessment methodology remain rigorous while still being adaptable to different organizational contexts?
4) What limitations do current LLMs exhibit when processing graphical technical documentation?

***Corresponding author:** Oleksandr Kravchuk, Department of Mathematical and Computer Modeling, G.E. Pukhov Institute for Modelling in Energy Engineering of the NAS of Ukraine, Ukraine. Email: oleksandr.kravchuk@pimee.ua

3. Background

Majumder et al. [2] highlighted in their recent study the significant potential of using LLMs to solve a wide range of problems in the energy sector and identified the following key applications:

1) detecting defects in generation and/or distribution network components based on video or photo analysis;
2) performing correlation analysis and forecasting of time series data using fine-tuned neural networks (e.g., predicting network load, equipment failures due to service life);
3) assessing the risk of external (e.g., wildfires, hurricanes, earthquakes, tsunamis) and internal (e.g. employee strikes, insider fraud) factors that may hinder or make electricity generation or distribution impossible;
4) analyzing documents at all stages—from the construction of an electric energy facility to its operation and eventual decommissioning.

Meanwhile, the key weaknesses were identified as follows:

1) non-guaranteed availability of domain-specific data for real-time decision-making;
2) lack of security guardrails;
3) poor adaptability to surrounding physical processes;
4) potential vulnerability to cyberattacks.

In 2024, the National Institute of Standards and Technology (NIST) introduced the Artificial Intelligence (AI) Risk Management Framework: Generative AI Profile [5]. While this issue is in the early stage of development, LLMs are considered a subclass of generative AI (hereinafter—GenAI), capable only of reproduction and limited extrapolation of information presented in text, without deep understanding of the principles that underlie such text generation [7], and if such a premise is accepted then the risks associated with GenAI are also inherent to LLMs. GenAI-related risks then differ from those associated with traditional software, may exacerbate these risks, and may be unique and multidimensional:

1) Stage dimension: may arise during design, development, deployment, operation, or decommissioning;
2) Scope dimension: can exist at the level of a specific model or system, at the implementation or deployment level, or at the environmental level—beyond a single system or organizational context;
3) Source dimension: may stem from factors such as GenAI architecture, training methodology, operator training, or model usage, whereas the most significant ones are often related to human factors and the human-AI interaction.

Challenges in risk assessment are exacerbated by the lack of access to the training data used for a given AI model and by the immature state of science regarding the quantitative evaluation of AI and its safety [5]. It becomes evident that a comprehensive risk analysis must cover the full lifecycle—from the formulation of the task to the deployment of the AI model in production.

The following categories of risks are identified as inherent to GenAI (including LLMs) or amplified by its development and use [5]:

1) information on chemical, biological, radiological, and nuclear (CBRN) weapons and their capabilities;
2) confabulation, i.e., generation of incorrect or fake content (commonly referred to as "hallucination" or "fabrication"), which can mislead or deceive users;
3) harmful, violent, or hateful content (i.e., easier generation and access to violent, inflammatory, radicalizing, or threatening material, including recommendations for self-harm or illegal activity, and difficulty in controlling the public display of hateful, disparaging, or stereotypical content);
4) data privacy (i.e., leakage, unauthorized use, disclosure, or de-anonymization of biometric data, health information, location data, or other personal data or confidential information).

Some of these risks appear to be more important than others.

When it comes to information on CBRN weapons, one must focus on how AI acquires knowledge about such weapons in the first place. From an "anatomical" standpoint, LLMs are stochastic generators that produce output based largely on the probability of the next word or sequence of words [1]. This means that an AI system that has never encountered words about weapons in its training corpus of texts cannot produce such information. In this context, a corpus refers to a body of texts selected and processed according to certain rules [8].

Although confabulation (generation of factually incorrect content) is a serious challenge for LLM developers and users [9–11], Sui et al. [12] view it as an underexplored opportunity. However, in the context of the energy sector, LLM confabulations are unacceptable due to the risk of erroneous decisions based on "fabricated data," potentially leading to catastrophic consequences. While, following the categorization, extrinsic hallucinations (which cannot be verified from the source content but neither contradict nor are supported by that) may well not be as problematic as intrinsic hallucinations (which contradict the source content) [13], as well as factual contradiction (factuality hallucination where a generated response is grounded in real-world information but is contradictory), factual fabrication (factuality hallucination where a generated response can not be verified with real-world information), instruction inconsistency (faithfulness hallucination where a generated response violates instructions given in the prompt), context inconsistency (faithfulness hallucination where a generated response drifts from the context given in the prompt), or logical inconsistency where a generated response contains internal reasoning contradictions that are logically flawed) [14]. In order to mitigate this risk, it is advisable to use hallucination and confabulation checkers.

As for the generation of harmful or hateful content, this is considered a minor risk in the energy sector because the ability of highly trained professionals in this high-tech field to ignore such content, as well as strong corporate policies and industry standards. The risk can also be effectively mitigated using sentiment analysis tools [15].

Data confidentiality is a critical issue—especially concerning personal data and trade secrets. When classifying data, one must also rely on the legal framework of the country considered. The data confidentiality issue and addressing that are therefore sector and jurisdiction dependent. An important consideration here though is the access of AI to such confidential data during training, piloting, and deployment to production, and proper oversight of LLM outputs is crucial. While it is unlikely that pre-trained models contain personal data as reference information, in the event of data leakage, liability will most likely fall on experts who fine-tuned the model using such real-world data or those who granted AI access thereto. The very idea of training on personal data raises numerous questions, especially given regional peculiarities such as naming conventions (surname, first name, patronymic).

4. Methodology

The methodology proposed in this study is designed to address the identified gap by providing a structured approach to assess the risks of implementing LLMs in the energy sector. It is based on establishing analytical techniques and considers both human and technological factors.

The approach includes three main components:

1) Risk taxonomy development:

   a. Risks need to be grouped into two categories: human and technological factors.
   b. The taxonomy should cover the entire lifecycle of LLM implementation.

2) Risk prioritization and evaluation:
   a. Expert judgement and pairwise comparison are used to establish weightings for different factors.
   b. Structured decision-making technique(s), such as the analytical hierarchy process (AHP), is applied to quantify the relative importance of identified risks.

3) Analysis of graphical images using LLMs:
   a. A series of experiments to process and digitize graphics materials (e.g., drawings, illustrations, diagrams, charts) are conducted using state-of-the-art generative AI tools.
   b. The goal of this stage is to determine whether current LLMs can reliably convert visual information into structured textual representation.
   c. The results of these experiments provide insight into the limitations and risks of relying on LLMs for autonomous representation of technical documents.

The combination of these steps—risk taxonomy, prioritization, and empirical testing of LLMs—forms a comprehensive framework for assessing the readiness and risks of applying LLMs in critical infrastructure tasks. The introduced mathematical formalization is not intended to constitute a predictive or empirically validated statistical model of risk. Instead, it provides a normalized interpretive framework that enables conceptual analysis of how risks evolve during the LLM lifecycle.

4.1. Process of LLM implementation

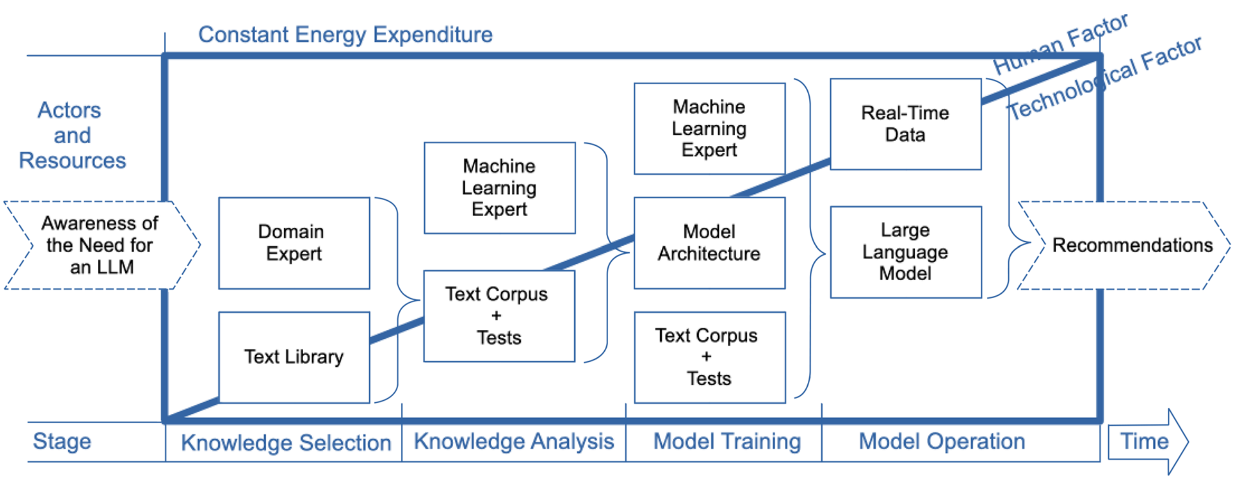
Evidently, the risks associated with the use of LLMs in the energy sector involve multiple factors and stakeholders. Let us visualize the process of selecting, training, and operating an LLM, indicating on the diagram a notional distribution of energy expenditures—both human and computational resources.

Figure 1 illustrates the processes as follows:

1) Knowledge selection is the initial stage that involves choosing a domain expert (or group of experts) and compiling the most comprehensive possible text corpus, referred to here as the text library [8]. The output of this stage is a corpus of texts for training the model and a set of tests (questions and answers) to be used for LLM validation.
2) Knowledge analysis is the second stage of LLM preparation, which involves finding machine learning experts and familiarizing them with the library. During this stage, the machine learning experts must select those texts from the library that (in their opinion) are most suitable for LLM training. Additionally, they choose the neural network architecture, algorithms for transforming concepts into vectors, and define the metric vector space used to calculate semantic distances between concepts. It is critical to ensure that the chosen distance function is a strict mathematical metric, satisfying the axioms of identity, symmetry, and the triangle inequality. The output of this phase is a correctly chosen model architecture. At this stage, particular attention should be paid to the model's ability to interpret graphical content, such as diagrams, which are prevalent in documentation within the energy sector overall and the electric power sector particularly. A dedicated experiment on the performance of modern AI models in handling schematic data is discussed later in this article.
3) Model training is the stage when the machine learning expert defines criteria for evaluating the success of the training and initiates the actual training process. Model validation is performed using the tests prepared by the domain expert during the knowledge selection stage. This stage is iterative and continues until satisfactory results are achieved in passing the tests, or until it is reasonably concluded that an adequate model cannot be created given the development objective and available validation data. A model that successfully passes all validation tests, as defined by the domain expert, is deemed ready for deployment.
4) Model operation stage includes two sub-stages: test deployment and deployment to production. In both cases, the model generates recommendations based on real-time data streams. These recommendations are reviewed by the domain experts, who determine whether the LLM is allowed to be used in production.

Note that Figure 1 intentionally omits the last two stages of the AI model lifecycle: upgrade and decommission. Upgrade is essentially the replacement of one model with another that offers better functional or non-functional characteristics (e.g., faster inference time, lower resource consumption). Decommissioning occurs either when the

Figure 1
Model of energy expenditures in the development and use of LLMs



model is replaced by a more powerful version or when there is no longer a need for the AI to solve the given problem.

Figure 1 shows the boundaries of the LLM development lifecycle (from project initiation to deployment into production) and includes two hypothetical lines: one representing constant normalized energy expenditure (normalized unit or 100%) and the other representing the notional distribution of energy expenditure between human and technological contributions.

The concept of constant energy cost supports the idea that, at the initial stages, most of the energy investment comes from humans who recognize the need for an LLM, set the development goals, process the text library, and prepare the training and validation datasets. The slope of the human–machine energy investment line reflects significant energy input from humans in the early stages.

With each subsequent stage, human energy input decreases, while computational energy expenditure increases. If one closely examines Figure 1, it becomes apparent that the amount of energy invested at each stage correlates with specific risks that directly impact the success of LLM development and deployment.

4.2. Risks classification

We will now analyze the potential risks at each stage as presented in Table 1.

The risks and their classification presented in Table 1 reflect the authors' perspective and may be expanded or completely revised by relevant departments within organizations in the sector.

The risks described in Table 1 logically form a hierarchy, in which each stage of model preparation and operation can be treated as a stage-specific risk, while the corresponding factors can be interpreted as risk categories (i.e., technological risks and human factors risks).

Table 1
Risks and mitigation measures in the implementation of LLMs

Factor	Risk	Risk Mitigation Measures
Knowledge Selection Stage		
Human	Poor representativeness of knowledge in the text library	Invest in the development of the text library by adding texts from the relevant and adjacent fields of knowledge
	Low qualification of the domain expert (or expert group) in the domain area	Collaborative selection of candidates from a broader list
Technological	Inadequate digitization of knowledge or low quality of digitization	Invest in high-quality digitization of texts from the relevant and adjacent domains
Knowledge Analysis Stage		
Human	Insufficient representativeness of the corpus	This issue results, on one hand, from poor representativeness of the library and, on the other, from excessive narrowing of the corpus during expert analysis. Invest in the text library and the preparation or selection of highly qualified domain experts.
	Presence of personal data in the corpus	Implement control mechanisms for the selected text corpus
	Presence of trade secrets in the corpus	Implement control mechanisms for the selected text corpus
	Presence of confidential information in the corpus that may appear in model outputs	Confidential information (excluding personal data and trade secrets) may be highly valuable for training the model. To reduce the risk, it is recommended to implement both additional control over the corpus and over the model's outputs
	Implementation of tests that fail to detect confabulations or the leakage of confidential information which implicitly presents in the corpus (e.g., metaphorically)	Prepare or select highly qualified domain experts and train them in AI prompt engineering
	Insufficient competence of those responsible for selecting the machine learning expert (or expert group)	Engage top specialists in the relevant machine learning field based on economic feasibility
	Insufficient competence of the machine learning expert in model architectures, selection, and training	Select a machine learning expert based on a theoretical interview and a practical test task
Technological	Limited access to the text library	Invest in secure infrastructure for the storage and processing of digitized knowledge
	Limited tools for working with digitized texts	Invest in high-quality tools for working with digitized texts (e.g., tools capable of identifying information that should be excluded from training)
Model Training Stage		
Human	Compromised validation of the trained model (due to limited resources or administrative pressure)	Strictly adhere to the model validation protocol using the test set and ensuring independent oversight of the training process Allocate sufficient resources Prevent administrative pressure

Table 1
(*Continued***)**

Factor	Risk	Risk Mitigation Measures
Technological	Insufficient resources for high-quality model training	Invest in appropriate hardware and sufficient time resources
	Poorly prepared training infrastructure, causing frequent training interruptions and restarts	Invest in appropriate hardware and its configuration
	Poor implementation of the model	Use only those implementations (libraries, software) that have been time-tested and vetted by the ML community
Model Operation Stage		
Human	Use of data that is highly sensitive from an information security perspective	Be aware and accept the risks associated with using the model for analyzing sensitive data
	Unauthorized access to the working model	Access control to the operational model via privilege management
	Unauthorized access to the model's outputs	Access control to model outputs to ensure result preservation for further use
	Poorly prepared infrastructure for model deployment	Invest in appropriate hardware and its configuration
	Delays in deployment due to bureaucratic procedures	Exclude individuals or departments not directly involved in the subject matter from the decision-making process
Technological	Issues with hardware used for model operation	Invest in high-quality hardware and its proper configuration
	Weak cybersecurity, allowing attackers to disrupt operation or reconfigure the model to produce incorrect outputs	Ensure robust cybersecurity for both the AI solution and the data it uses to generate recommendations
	Lack of automated quality assurance of the model's performance	Periodically execute validation tests during industrial operation to monitor the model's functionality and adequacy Continuously monitor confabulations and hallucinations using specialized tools [13] (approaches to hallucination prevention and a taxonomy are discussed by Zhao [16])

Given this hierarchy, it is reasonable—by analogy with Li et al. [17]—to apply AHP by Saaty [18] for building the risk assessment methodology, as it aligns well with the study's objectives.

Before proceeding to the practical part, let us formulate a mathematical model for risk (energy) distribution during the implementation of GenAI. The correlation between normalized risk and normalized energy expenditure is hypothetical and based on the assumption that avoiding energy expenditures results in zero risk.

In Figure 1, we show four stages of LLM implementation: knowledge selection, knowledge analysis, model training, and model operation.

Assuming each stage is indivisible, we define the length of each stage as a unit of 1. This means that for any stage, 0 represents the beginning and 1 represents its full completion. With this, we normalize the duration of each stage to percentages or fractions, ignoring the actual time required for completion. As a result, the domain of the risk distribution function $f(t)$ can be defined as $D(f(t))$, $t \in [0;4]$.

With this definition, it becomes easy to interpret both the stage and its degree of completion:

1) The completed stage index is calculated as $t \div 1$, where $\div$ denotes the floor (integer-division) operation.
2) The completion share of the current stage is calculated as $t \ mod \ 1$, where *mod* denotes the modulus (remainder) operation.

For example, $t = 3{,}25$ means the third stage is completed, and the fourth stage is 25% done.

For interpretive convenience, let us assume the area of the rectangle in Figure 1 (bounded by the start of the knowledge selection stage, the end of GenAI deployment, the time axis, and the constant energy line) is normalized to 1.

This implies the rectangle formed by the time axis, energy axis, the endpoint of the final stage, and the start of the first stage (marked with bold lines) has an area equal to 1.

We formally define the energy or risk distribution function $f(t)$ as $D(f(t))$, $t \in [0;4]$, $0 \leq f(t) \leq 1$ with $0 \leq \int_{D(f(t))} f(t) \leq 1$. This formal definition gives the energy or risk distribution function a probability density function—like nature over the defined domain.

Such a formalization of risk $f(t)$ is a convenient and effective tool for:

1) setting expected risk distributions,
2) approximating observed risks,
3) assessing the level of unexpected risk.

Let $f_0(t)$ represent the function of expected risk distribution, and $f_1(t)$ the function of observed risk distribution.

We introduce a binary function:

$$d(a,b) \equiv \begin{cases} a-b, & if\ (a-b) \geq 0 \\ 0, & if\ (a-b) < 0 \end{cases}$$

A positive value of $d(f_0(t), f_1(t))$ indicates observed technological risks exceeded expectations.

Similarly, a negative $d(f_1(t), f_0(t))$ indicates human-related risks were underestimated.

The case $d(f_0(t), f_1(t)) = d(f_1(t), f_0(t)) = 0$ corresponds to fully aligned observation and expectation $f_0(t) = f_1(t)$.

4.3. GenAI-aided analysis of graphical images

Even before LLMs acquired multimodal capabilities, they demonstrated prerequisites for exploring images effectively. When the content of a highly specialized energy sector related illustration was literally reproduced in text, a GenAI tool was able to extract the very essence of what was depicted in the illustration [19]. Following significant advancements in the image-to-text techniques overall [20] and the emergence of multimodality in particular [21], LLMs have expanded translations from image to text, become affordable, moved closer to end users, and found their application far beyond the computer science domain, including the spheres where health and life are at stake [3, 22, 23]. While LLMs do demonstrate significant progress in accuracy [23, 24] and multilinguality [25], it seems more important to explore their effectiveness in analyzing images at the intersection of an arbitrary domain, a non-English environment, and under real-world conditions. A dedicated experiment in handling schematic images related to the energy sector in the Ukrainian language from the standpoint of an end user appears to address the issue in focus.

The readiness of LLMs to accurately digitize graphical images can be examined by means of GenAI. More specifically, the focus is on the suitability of such models, given the state-of-the-art technology, to process visually represented materials (drawings, illustrations, schemas, diagrams, charts, etc.) and convert them into textual representations, which can be examined by conducting experiments on processing a graphic image using various GenAI tools.

Figure 1 was selected as a hypothetical graphic image to compare human understanding, based on this article and the outputs generated by the mentioned tools. Since this study was contextualized on the electric power sector of Ukraine, the labels in Figure 1 are in Ukrainian as in the Ukrainian electric power sector drawings, schemas, diagrams and other visually represented materials in documentation. However, we expect the study's findings to be applicable to the energy sector beyond Ukraine.

The GenAI tools used in the experiment are publicly accessible and thus suitable for evaluation outside the controlled laboratory conditions:

1) Google Gemini using 2.5 Flash (preview) and 2.5 Pro (preview)
2) OpenAI ChatGPT using GPT-4o and o4-mini
3) Microsoft Edge Copilot using the Think Deeper parameter
4) xAI Grok using the DeepSearch parameter
5) Perplexity AI using the Research parameter

The GPT-4.1-mini model for OpenAI ChatGPT was excluded due to declared availability but practical inaccessibility. The Quick response parameter for Microsoft Edge Copilot was not used, as it simply reproduces the image without following instructions. The Search parameter for Perplexity AI was excluded since it behaves more like a search engine. The DeepSearch parameter for xAI Grok was excluded despite producing quality results because the time to first token exceeded 61 minutes, making it unsuitable for categorization as a widely accessible tool.

Likewise, the supposed availability of Anthropic Claude could not be confirmed, as usage attempts failed due to user resource limits being exceeded.

As prompts, we use four natural language instructions in Ukrainian that can be translated as follows:

1) "Describe the content of the given material"
2) "Interpret the content of the given material"
3) "You are an expert in software-hardware implementation, describe the content of the given material"
4) "You are an expert in software-hardware implementation, interpret the content of the given material"

This variety was chosen to minimize the potential sensitivity of LLMs to specific wording and to simulate real-world conditions, where an average user may not possess prompt engineering expertise.

5. Results

The area containing the elements of the model preparation process in Figure 1 is enclosed within a rectangle, and we impose the condition that the area of this rectangle equals 1. We assume that probability has a geometric interpretation as area.

A line dividing the rectangle diagonally (from the bottom-left corner to the top-right) serves as the boundary between risks associated with technological factors and those related to human factors. This division aligns with the principle of maximum entropy, assuming equal probability for both categories of factors. The probability distribution between these factors was treated as hypothetical.

5.1. Risk analysis

In accordance with Saaty [18], we chose the comparison scale shown in Table 2:

Table 2
Scale for pairwise comparison of risks according to the AHP method

Intensity	Definition	Explanation
1	Equal importance	Both activities contribute equally to the objective
2	Weak or slight	
3	Moderate importance	Experience and judgment slightly favor one activity over another
4	Moderate plus	
5	Strong importance	Experience and judgment clearly favor one activity over another
6	Strong plus	
7	Very strong or demonstrated importance	One activity is strongly favored over another
8	Very, very strong	
9	Extreme importance	The evidence favoring one activity over another is of the highest possible order of affirmation

To assess risks using the AHP method [18], we constructed pairwise comparison matrices $A = \{a_{ij}\}$ for each level of the hierarchy (stage/factor/risk), such that $a_{ij} = a_{ji}$.

We then computed the eigenvector

$$\omega_i = \frac{\left(\prod_{j=1}^{n} a_{ij}\right)^{1/n}}{\sum_{k=1}^{n}\left(\prod_{j=1}^{n} a_{kj}\right)^{1/n}},$$

the maximum eigenvalue $\lambda_{max} = \frac{1}{n}\sum_{i=1}^{n}\frac{(A\omega)_i}{\omega_i}$,
the consistency index $CI = \frac{\lambda_{max}-1}{n-1}$,
and the consistency ratio $CR = \frac{CI}{RI}$ for each comparison matrix.

Here, n is the dimension of the matrix, and RI is the random consistency index for a given matrix size.

The AHP comparison matrices used in this study illustrate the methodology rather than represent an industry-wide consensus. The weightings reflect the assessment of the authors based on synthesized findings from the literature and domain knowledge. The approach was designed so that organizations in the energy sector can substitute their own expert judgements and derive project-specific weightings. To

Table 3
Pairwise comparison matrix of risks for LLM implementation stages

	Knowledge selection	Knowledge analysis	Model training	Model operation	Normalized weight
Knowledge selection	1	1/2	2	3	0.31
Knowledge analysis	2	1	2	4	0.42
Model training	1/2	1/2	1	4	0.19
Model operation	1/3	1/4	1/4	1	0.08

Note: consistency index: CR = 0.049

illustrate, we present several constructed comparison matrices along with their respective consistency indices (Table 3)[1].

As an example, we also present the pairwise comparison tables for human factor risks during the knowledge analysis stage (Table 4) and technological factor risks during the model training stage (Table 5).

Table 4
Pairwise comparison matrix of human factor risks for the knowledge analysis stage

	$\mathbf{R}_{2.1.1}$	$\mathbf{R}_{2.1.2}$	$\mathbf{R}_{2.1.3}$	$\mathbf{R}_{2.1.4}$	$\mathbf{R}_{2.1.5}$	$\mathbf{R}_{2.1.6}$	$\mathbf{R}_{2.1.7}$	Normalized weight
$\mathbf{R}_{2.1.1}$	1	7	3	3	1/5	1/9	1/7	0.08
$\mathbf{R}_{2.1.2}$	1/7	1	1/2	1/5	1/9	1/9	1/7	0.02
$\mathbf{R}_{2.1.3}$	1/3	2	1	1/3	1/7	1/5	1/3	0.04
$\mathbf{R}_{2.1.4}$	1/3	5	3	1	1/5	1/3	1/2	0.08
$\mathbf{R}_{2.1.5}$	5	9	7	5	1	1	2	0.31
$\mathbf{R}_{2.1.6}$	9	9	5	3	1	1	1	0.27
$\mathbf{R}_{2.1.7}$	7	7	3	2	1/2	1	1	0.20

Note: consistency index: CR = 0.097

Here,

1) $\mathbf{R}_{2.1.1}$ is poor representativeness of the corpus
2) $\mathbf{R}_{2.1.2}$ is presence of personal data in the corpus
3) $\mathbf{R}_{2.1.3}$ is presence of trade secrets in the corpus
4) $\mathbf{R}_{2.1.4}$ is presence of confidential information in the corpus that may appear in the model output
5) $\mathbf{R}_{2.1.5}$ is test design that fails to detect confabulations or implicit disclosure of confidential information (e.g., metaphorically)
6) $\mathbf{R}_{2.1.6}$ is insufficient competence of those selecting the machine learning expert (or expert group)
7) $\mathbf{R}_{2.1.7}$ is insufficient competence of the machine learning expert in model architectures, selection, and training

Table 5
Pairwise comparison matrix of technological factor risks for the model training stage

	$\mathbf{R}_{3.2.1}$	$\mathbf{R}_{3.2.2}$	$\mathbf{R}_{3.2.3}$	Normalized weight
$\mathbf{R}_{3.2.1}$	1	1/5	1/7	0,08
$\mathbf{R}_{3.2.2}$	5	1	1	0,44
$\mathbf{R}_{3.2.3}$	7	1	1	0,49

Note: consistency index: CR = 0.012

Here,

1) $\mathbf{R}_{3.2.1}$ is insufficient resources for high-quality model training
2) $\mathbf{R}_{3.2.2}$ is poorly prepared training infrastructure leading to frequent failures and the need to restart training
3) $\mathbf{R}_{3.2.3}$ is poor implementation of the model

The pairwise comparison procedure assumes a panel of 3–5 domain experts from the energy sector and 1–2 machine learning specialists, consistent with typical AHP applications in critical infrastructure studies. Experts should be selected based on (1) domain knowledge, (2) familiarity with AI-assisted systems, and (3) absence of conflicts of interest. In this conceptual study, the authors provide a reference comparison matrix to demonstrate the methodology; however, the framework was designed to be applied using experts internal to sector organizations.

Thus, coefficients in the resulting pairwise comparison tables were selected at the authors' discretion, based on personal experience and logical reasoning. In addition, all matrices were mutually consistent (each matrix had a consistency ratio CR < 0.1). The local normalized risk weights were obtained for each level of the hierarchy. The global weights are computed by multiplying the weight of the stage, factor, and specific risk:

$$\omega_{global} = \omega_{stage} \times \omega_{factor} \times \omega_{risks}.$$

The risk impact was “high” if ω_i > 0,15; “medium” if 0,05 < ω_i ≤ 0,15; and “low” if ω_i ≤ 0,05. Table 6 shows a summary of risks with high and medium levels of impact.

Table 6
Risks with high and medium levels of impact

Stage	Factor	Risk	Impact
Knowledge selection	Human	Poor representativeness of knowledge in the texts' library	Average
		Low qualification of the domain expert (or expert group)	High
Knowledge analysis	Human	Test design that fails to detect confabulations or the implicit disclosure of confidential information (e.g., metaphorically)	Average
		Insufficient competence of those selecting the machine learning expert (or expert group)	Average
		Insufficient competence of the machine learning expert in model architectures, selection, and training	Average
Model training	Technological	Poorly prepared training infrastructure, leading to frequent failures and the need to restart training	Average
		Poor implementation of the model	Average

[1] The full list of matrices and coefficients calculation according to T. Saaty method is available at the following link: https://github.com/oleksandrkravchukatpimee/LLM-risks-evaluation/blob/3443a610d9b2f9a9f8db52b280f2f4fb247525c1/AHP.xlsx

Table 7
Evaluation of processing a graphic image using GenAI tools (understanding and generation dimension)

	Google Gemini 2.5 Flash (preview)	Google Gemini 2.5 Pro (preview)	OpenAI ChatGPT GPT-4o	OpenAI ChatGPT o4-mini	Microsoft Edge Copilot Think Deeper	xAI Grok Grok-3 DeepSearch	Perplexity AI Research
Usability in terms of accessibility	+	+/-	+	+/-	+	+	+/-
Usability in terms of free-tier reliability	+/-	+/-	+/-	+/-	+/-	+/-	+/-
Understanding of textual elements	+	+	+	+	+	+	+
Understanding of graphical elements	+/-	+/-	+/-	+/-	+/-	+/-	+/-
Understanding of combined textual and graphical elements	+/-	+/-	+/-	+/-	+/-	+/-	+/-
Non-English text processing	+/-	+	+	+	+	+	+
Comprehensive text generation	+	+	+	+	+/-	+/-	+

Table 8
Evaluation of processing a graphic image using GenAI tools (extrinsic and intrinsic hallucinations dimension)

	Google Gemini 2.5 Flash (preview)	Google Gemini 2.5 Pro (preview)	OpenAI ChatGPT GPT-4o	OpenAI ChatGPT o4-mini	Microsoft Edge Copilot Think Deeper	xAI Grok Grok-3 DeepSearch	Perplexity AI Research
Extrinsic hallucination	+	+	+	+	+	+	+
Intrinsic hallucination	-	-	-	-	-	-	-

Table 9
Evaluation of processing a graphic image using GenAI tools (factuality and faithfulness hallucination dimension)

	Google Gemini 2.5 Flash (preview)	Google Gemini 2.5 Pro (preview)	OpenAI ChatGPT GPT-4o	OpenAI ChatGPT o4-mini	Microsoft Edge Copilot Think Deeper	xAI Grok Grok-3 DeepSearch	Perplexity AI Research
Factual contradiction	+/-	-	-	-	-	+/-	-
Factual fabrication	-	-	-	-	-	-	-
Instruction inconsistency	+/-	-	-	-	-	-	-
Context inconsistency	+/-	+/-	+/-	+/-	+/-	+/-	+/-
Logical inconsistency	-	-	-	-	-	-	-

The results of the study indicate that the most significant factor influencing the success of LLM implementation in the electric power sector taking into account the risk level is the involvement of a qualified domain expert (or group of experts). For the first two stages (knowledge selection and knowledge analysis), the human factor had the greatest impact, whereas the success of model training was heavily dependent on the quality of the infrastructure and the implemented model, which were components of the technological factor.

The presented analysis is illustrative and requires further expansion to build a more accurate and detailed risk assessment model for the implementation of LLMs in the electric power sector particularly and the energy sector overall.

External expert data collection was not within the scope of the study because it focuses purely on constructing a structured methodology rather than producing organization-specific risk weights.

5.2. Analysis of graphic images using LLMs

In order to consider the readiness of LLMs to accurately digitize graphic images, we conducted a series of experiments on processing a graphical image using GenAI tools.

Each prompt alternative (4 in total) was tested with each GenAI tool (7 in total), using the same file of Figure 1 (originally in Ukrainian). The obtained results[2] allowed us to evaluate such processing (Table 7, Table 8 [13], and Table 9 [14]) and comment on the readiness of LLMs for graphic digitization in the following way.

In general, the outputs across all the GenAI tools were comparable, although results from Microsoft Edge Copilot (Think Deeper) and xAI Grok (DeepSearch) were consistently less comprehensive; the latter also exhibited errors at the language-level errors.

All prompts were conducted in Ukrainian, and results were obtained in Ukrainian as well (i.e., there was no need to use translation tools).

Exceptions were two cases when using Google Gemini 2.5 Flash (preview), where results were returned in English (in both cases the prompt contained the word “інтерпретуй”—which may have triggered English word “interpret”, which has “translate” as an alternative meaning).

[2] The set of experimental data of graphic digitization using the state-of-the-art language models where images are described textually is available at the following link: https://gist.github.com/taranowskiatpimee/174973d140a84da2b5c3b365a34f949c

This highlights an issue of true accessibility of GenAI tools to general users.

Access under free-tier resource-limited conditions—where computational resources are redistributed in favor of paid-tier users—excludes such tools from being widely available, and together with output unpredictability under such conditions makes these tools unreliable for expected outcomes.

Similarly, limited query counts in Gemini 2.5 Pro (preview), OpenAI o4-mini, and Perplexity AI Research mode also further limit usability of these tools.

Regarding the textual rendering of the graphical content, most responses listed key elements of the diagram, but not exhaustively, and omissions were not explained.

While the image was incorrectly interpreted as a chart with labeled "stages" and "actors and resources," which was not the case.

Indeed, textual descriptions mentioned the overall process of the LLM implementation, including knowledge selection, analysis, model training, and model operation, as well as the participants involved and resources utilized. However, even when the omitted stages of the lifecycle were mentioned, there was no explanation for their omission.

It is not uncommon that, across different prompt alternatives and different GenAI tools used, the model training stage was involved in the selection, design, or development of the model architecture, although the diagram only references the model, with architecture as an element without any specification as to when it is selected or developed.

Some critical elements of the diagram received insufficient attention: the lines that formed the rectangle with the diagonal line and the labels "constant energy expenditure," "human factor." and "technological factor".

There is simply no holistic interpretation of these as a reallocation of effort between human and technological resources under constant energy expenditure throughout the whole process. This is exactly the case where what's not mentioned is much more important than what is. Indeed, where "You are an expert… interpret the content…" prompt alternative is used, some output contains fragments like:

1) "Human/Technological factor (diagonally right)—shows the balance between human impact (in early stages) and technological impact (in later stages)"
2) "the upper line labeled 'constant energy expenditure' can be interpreted not literally as electricity, but as steady effort, resources (including financial, human, temporal ones), and attention required throughout the whole lifecycle of the project"

However, this is the case only for prompt alternative 4 and only for 2 of the 7 GenAI tools—and notably, not the same tools, each providing just one of the two relevant insights.

The curly braces on the diagram that group elements within a stage and indicate which element they jointly contribute to in the next stage are ignored completely. The only exception was a correct partial interpretation by AI Grok used with the DeepSearch parameter under the prompt "Describe the content…". Yet, even here, not all the curly braces relationships were properly covered. This is the case just for 1 prompt alternative out of 4, and just for 1 GenAI tools of 7.

Thus, to use a well-known expression: LLMs not only miss the forest for the trees – they do not even see all the trees. The readiness of modern models to accurately digitize graphical images can be considered relative at best. This highlights a risk when applying LLMs in the energy sector, particularly regarding their ability to perform proper analysis of graphic images, with schemas and diagrams being an essential part of regulations and technical documentation in this field.

The conducted analysis has shown the dominant impact of the human factor within the early stages (knowledge selection and preparation) and the dominant impact of the technological factor within the later stages (implementation and operation).

The proposed taxonomy of risks proved effective for representing the main categories of challenges. The mathematical modeling of energy expenditure added an additional dimension to risk evaluation, supporting a more comprehensive assessment.

Empirical experiments with generative AI systems demonstrated that modern LLMs are not yet ready for fully autonomous analysis of graphical technical documentation. These limitations in interpreting complex diagrams, structures, and logical relationships indicate a high risk of using LLMs without human oversight in critical domains where accuracy and contextual understanding are essential.

6. Conclusion

This study proposes a method for assessing the implementation of LLMs for tasks related to the functioning of critical infrastructure in the electric power sector. The findings highlight both the potential of applying LLMs in the energy sector and several limitations that must be considered during the development, validation, and deployment of such solutions in practice.

The presented methodology lays the groundwork for a systematic approach to risk management during the implementation of LLMs, but the current version of the model remains illustrative and requires further refinement to improve its accuracy. Future work should address the identified challenges, especially in the limitations of LLMs in processing graphical documentation.

A practical point of intervention for policymakers is to define the conditions under which LLMs may be used in the energy sector, including risk level categories (from unacceptable to minimal) and the required degree of human oversight. Once such a regulatory framework is established, energy companies and other stakeholders can determine how to implement LLMs by developing internal policies and industry standards; the methodology proposed in this paper can serve as an initial template that organizations may refine to address context-specific risks.

The identified limitations in processing graphical technical documentation indicate a clear direction for further work by energy-sector organizations: focusing on real documentation containing sector-specific graphical content, fine-tuning LLMs on such data, and expanding evaluation procedures through domain-expert review, automated metrics, and consistency checks across repeated queries.

Acknowledgement

The presented text was written in English by Ukrainian speakers who then used ChatGPT for proofreading. The authors take full responsibility for the accuracy of the resulting text.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Github at https://github.com/oleksandrkravchukatpimee/LLM-risks-evaluation/blob/3443a610d9b2f9a9f8db52b280f2f4fb247525c1/AHP.xlsx and https://gist.github.com/taranowskiatpimee/174973d140a84da2b5c3b365a34f949c.

Author Contribution Statement

Hryhoriy Kravtsov: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Oleksandr Kravchuk:** Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing. **Artem Taranowski:** Methodology, Investigation, Resources, Writing – original draft, Writing – review & editing. **Dmytro Sinko:** Methodology, Investigation, Writing – original draft. **Victor Samoylov:** Conceptualization, Supervision.

References

[1] Ozdemir, S. (2023). *Quick start guide to large language models: Strategies and best practices for using ChatGPT and other LLMs*. USA: Addison-Wesley Professional.

[2] Majumder, S., Dong, L., Doudi, F., Cai, Y., Tian, C., Kalathil, D., ..., & Xie, L. (2024). Exploring the capabilities and limitations of large language models in the electric energy sector. *Joule*, *8*(6), 1544–1549. https://doi.org/10.1016/j.joule.2024.05.009

[3] ElSayed, M., Shultz, J., & Kurtz, J. (2025). User-friendly AI-driven automation for rapid building energy model generation. *Energy and Buildings*, *345*, 116092. https://doi.org/10.1016/j.enbuild.2025.116092

[4] Ji, X., Zhang, L., Zhang, W., Peng, F., Mao, Y., Liao, X., & Zhang, K. (2025). LEMAD: LLM-empowered multi-agent system for anomaly detection in power grid services. *Electronics*, *14*(15), 3008. https://doi.org/10.3390/electronics14153008

[5] National Institute of Standards and Technology. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

[6] Regulation 2024/1689. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[7] Ferrag, M. A., Tihanyi, N., & Debbah, M. (2025). *Reasoning beyond limits: Advances and open problems for LLMs*, *11*(6), 1054–1096. https://doi.org/10.1016/j.icte.2025.09.003

[8] Holoshchuk, S. (2021). korpusna lingvistyka: Suchasnyi stan ta perspektyvy doslidzhen [Corpus linguistics: Modern approach and research perspective]. *Current Issues of Linguistics and Translation Studies*, *22*, 33–36. https://doi.org/10.31891/2415-7929-2021-22-7

[9] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

[10] Emsley, R. (2023). ChatGPT: These are not hallucinations – They're fabrications and falsifications. *Schizophrenia*, *9*(1), 1–2. https://doi.org/10.1038/s41537-023-00379-4

[11] Liu, K., Casper, S., Hadfield-Menell, D., & Andreas, J. (2023). Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? *Conference on Empirical Methods in Natural Language Processing*, 4791–4797. https://doi.org/10.18653/v1/2023.emnlp-main.291

[12] Sui, P., Duede, E., Wu, S., & So, R. (2024). Confabulation: The surprising value of large language model hallucinations. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, *1*, 14274–14284. https://doi.org/10.18653/v1/2024.acl-long.770

[13] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ..., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 248. https://doi.org/10.1145/3571730

[14] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ..., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 42. https://doi.org/10.1145/3703155

[15] Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, *36*(4), 102048. https://doi.org/10.1016/j.jksuci.2024.102048

[16] Zhao, H. (2025). How to prevent AI hallucinations. GPT–trainer blog. https://gpt-trainer.com/blog/how+to+prevent+ai+hallucinations

[17] Li, R. Y. M., Chau, K. W., & Zeng, F. F. (2019). Ranking of risks for existing and new building works. *Sustainability*, *11*(10), 2863. https://doi.org/10.3390/su11102863

[18] Saaty, T. L. (1982). *Decision making for leaders: The analytical hierarchy process for decisions in a complex world*. USA: Lifetime Learning Publications.

[19] Taranowski, A. O., & Samoylov, V. D. (2023). ChatGPT as expertless test generator. *Èlektronnoe Modelirovanie*, *45*(2), 44–60. https://doi.org/10.15407/emodel.45.02.044

[20] Sharma, A., & Aggarwal, M. (2025). A holistic review of image-to-text conversion: Techniques, evaluation metrics, multilingual captioning, storytelling and integration. *SN Computer Science*, *6*(3), 225. https://doi.org/10.1007/s42979-025-03719-6

[21] Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., ..., & Cucchiara, R. (2024). The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics*, 13590–13618. https://doi.org/10.18653/v1/2024.findings-acl.807

[22] Adnin, R., & Das, M. (2024). "I look at it as the king of knowledge": How blind people use and understand generative AI tools. *Proceedings of The 26th International ACM SIGACCESS Conference on Computers and Accessibility*, 1–14. https://doi.org/10.1145/3663548.3675631

[23] Zhang, A., Zhao, E., Wang, R., Zhang, X., Wang, J., & Chen, E. (2025). Multimodal large language models for medical image diagnosis: Challenges and opportunities. *Journal of Biomedical Informatics*, *169*, 104895. https://doi.org/10.1016/j.jbi.2025.104895

[24] Fu, B., Hadid, A., & Damer, N. (2025). Generative AI in the context of assistive technologies: Trends, limitations and future directions. *Image and Vision Computing*, *154*, 105347. https://doi.org/10.1016/j.imavis.2024.105347

[25] Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., ..., & Yu, P. S. (2025). A survey of multilingual large language models. *Patterns*, *6*(1), 101118. https://doi.org/10.1016/j.patter.2024.101118

How to Cite: Kravtsov, H., Kravchuk, O., Taranowski, A., Sinko, D., & Samoylov, V. (2026). Risk Assessment of Large Language Model Implementation in the Electric Power Sector of Ukraine. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62027195