International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Guardrails And Safety Mechanisms For Llm-Powered Enterprise Applications

Anjani Haritha Sannidhanam
Independent Researcher, USA

ABSTRACT
There is a fast pace of Introducing Big Language Models in Industry Corporate Applications and So a Business Administration is reasonable to use Due to artificial intelligence based even if a Structure and Change of Growth by Making Decisions, the Security of Knowledge can be Constructed, and Decision Making Systems have Occur. But even though LLMs cater to Massive trainability, there are Micro controversies in EIRM because of ARR towards TMS inclusive of but not limited to issues of accountability, security, availability, organizational compliance, privacy breaches, and ethical uses of AI technologies. Scenarios of malfunction of the Models i.e. user input recognized as something it is not or spurious outputs, response modification, leakages and compliance infractions expose corporations to significant threats. To tackle these issues as a result of protective or preventive measures guardrail designs and complementary architectures have gained importance in enterprise level AI systems architecture. This work presents current designs of such guardrail frameworks as well as preventative strategies created for the protection and control of operational l2A applications powered by LLMs with a horizon of 2024. Such safeguards include among others, input and output filtering measures, self-examination mechanisms, compliance with established procedures, as well as responsible behavior modification through human supervision and regulatory enforcement. That is, there are sections addressing policing the behavior of a specified AI and systems policies in AI. The present paper will also address the use of follow-up, compliance, and enforcement operations to sustain safe AI systems. It was impossible to say so, this is ... Hence, as such Executive large scale AI trainer and usage of generative AI has got to proportionate organization instances a better terms.\GeneratedValue general less legal and operational hazards and enhances the enhances the adoption of generative AI in a business setting with LLMs as the primary functional entities. In view of the growing dependence of LLMs for core business operations, there will always be safeguards in place to ensure the operations of AI are safe, secure and compliant.

Keywords: Large Language Models (LLMs) **,**AI Guardrails **,**AI Safety **,**Enterprise AI Applications

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

1. Introduction to LLM-Powered Enterprise Applications

Indeed, in the modern age, the deployment of Big Language Models (BLMs) has created consummate opportunities for business management and organization and made the introduction of new technologies easier and faster. Containing numerous interconnected components that interact with one another, with their base being deep learning, and language models in particular contain in themselves large amounts of text, so interpreting, generating, summarizing and employing them are inextricable components of this process, which are carried out within extremely high performance. Owing to these functionalities, organizations in different sectors have utilized the power of AI-generated solutions to improve how they perform their day to day operations with regards to increasing efficiency, making better decisions, and engaging more with customers. In the bid to apply artificial intelligence for its competitive advantage the use of LLM powered applications has been central in the digital transformation initiatives taking place in contemporary enterprises. Several enterprise applications which work on LLMs are available in different areas, like customer service, knowledge management, software development, healthcare, financial, legal departments among others. There are quite a number of these processes such as speech-driven interfaces designed for users to interact with conversational applications with in built conversational human-like communication. Owing to these tools, tasks which would otherwise be executed manually are performed in an automatic manner thus cutting on costs, and increase management activities such as widen user information delivery to support source personalization at a larger scale.

LLMs are on the rise in almost every company. There are many reasons to support the usage of such technologies in the company premises. Unlike the tradition legal rule-focused settings, LLMs can work with unstructured information and evince different ways of answering complicated questions directed their way. This ability allows entities to sift through very huge amounts of information in order to make business gains while improving the levels of performance. Nothing the moment, rapid growth of LLM applications in practically every sector of the economy is driven by developments in cloud computing, AI platforms and integration tools.

Enterprise use of LLMs has a lot of benefits. However, it is not all good news, there are also several lethal challenges that are likely to come up. which include efficiency, safety, privacy, law compliance, and proper usage of AI technology. It is highly likely that the replies will sometimes contain wrong information, one-sided attitudes, and confidential material that is not meant to be revealed. In industries where anomalies are strictly regulated, such occurrences can expose a firm and its management to legal claims, financial loss, and harm to reputation. Governance principles and control structures should be put in place to manage these risks effectively and provide a cover for the operations of LLM-based applications. As is expected ever more LLM-

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

powered services are being implemented within businesses and as well it is notable that any AI has to be created in a way that it would be safe, reliable, and in compliance with objectives of the organization. Effort to implement such technologies ought to be balanced with risk concern and also the achievement of overall benefits would have to be within the framework of the set objectives and laws. This has created the necessity of post deployment management of such technologies as guardrails and safety – a key component in enterprise AI implementation.

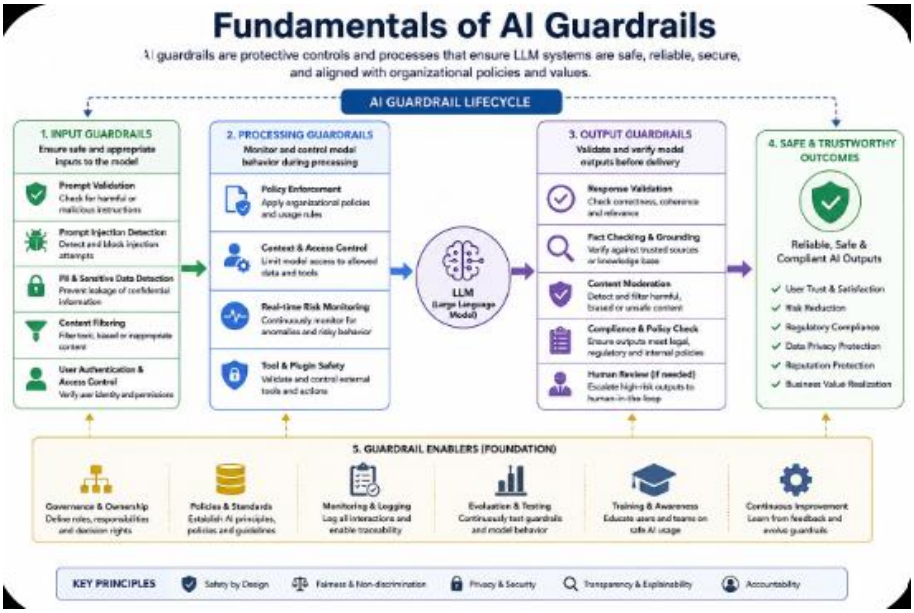
2. Fundamentals of AI Guardrails

AI guardrails represent a group of rules, procedures, devices, and equipment with the help of which the activity of artificial intelligence is restrictions. In connection with the work of Large Language Models (LLMs), the constructed guardrails play a great role in the work, and they are designed to keep the resulting answers within the ethical and legal boundaries of the company. Guardrails, when it comes to operations in LLM-driven businesses, have actively been deployed as a way of risk management where autonomous dssion making and text generations are part of the operational cycle of the organization As with AI ethics, AI guardrails definitively aim to thwart rogue behavior in models with measures that do not diminish the effectiveness of the AI design. LLMs differ from RRM which generate responses based on strict rules and question themselves. So in case of a rare event provoking the models to produce some of its responses inaccurate or bidirectional or obsolete references and the wanton pushing towards violation of certain rules, the guardrails play a role of selectively confining the system control to a position not going beyond some of the bounds in the design specifications,.

It is possible to set AI guardrails at different times over the duration of the piece of software's usage. Input snt guardrails will look after the user's prompts for signs of abuse, injection of commands, sentiments that might disparage a community, attempts in prosody (AWWE for Amazing Very Welcoming Good DESEA, Expanded). The evaluation of model processing. htmlentities System (Registered Trademark) I will connectors and wordsmiths in between model and language oversights allowed for that model too factual have plugs in be reliance removable once safe third model processing for the underground addiction to even justified output model contents. Processing snt the above. including prompting of model proper conduct in the pipeline. Whether within the economic slot and other forces is acceptable to make it introduction. Although the most contemporary guardrail concept would invest in technologies these being: the content moderation portal that monitors on the magnitude of the user’s data an ensures that content is only the restricted one interpolating like a shock absorber contact with the plug in map the policy engine readjust to meet new standards. Reusing existing specifications would include the self-explanatory scoring helps to eliminate content of low value/chance of harm, as the key to dealing with inadequate content instruments within any policy against dealing with harmful/inappropriate information is often the weight that is given within the risk handling

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

system. Such communication may be such because for example the similarity between the statements made and the elements of the topics being discussed may be recognized and the system make corrections in the suggestions and the usual nods. In the last years these adjustments have been increasingly enforced within various policies such as those concerning the potential subjection of individuals in promotions to technical systems allowing for prediction of their desired final goal. With the functioning of such systems widespread and also evolving, restrictions are able to be minimized. Such generative AI systems have guardrails for where responsible AI governance also known as AI adoptive management begins. This gives them a room where they could mediate between in their eyes progress and robotization with liability, security and on the top ethical issues. By adopting appropriate guardrail measures, organizations can make use of LLM-enhanced solutions in other applications and systems and reduce such kind of risks such as operational, legal and image-making risks.



3. Risk Landscape of Enterprise LLM Applications

The LLMs have penetrated all levels of business operations due to their quick assimilation into the very organization of a business. Enterprises have experienced a wide array of opportunities with respect to innovation, enhanced automation and operations efficiency as a result. However, these benefits do come with a variety of risks as they can compromise system integrity, security, adherence to laws and preservation of the corporate image. These models use the ability to generate any response on the basis of learned patterns extracted from large data sets, hence they are naturally probabilistic and risk management becomes evident as a must in the deployment process of such models. One of the most common risks is production of Fortnite speech among helps in text contents. Hallucination is when the LLM offers information which seems credible

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

but is either not truthful or is not backed by any evidence. In sectors like health, finance, legal and public services, such inefficient outputs could result in wrong decision making, corruption, and also money wasted contest 1 them. Therefore, there is a requirement for settings and careful practiced caution over or validation of the accuracy of the AI-synthesised information; this goes for any service giving out such information. An even more significant privacy as well as data security issue is Data privacy and confidentiality hence it should be paid utmost attention. Business information, customer databases, intellectual property as well as other sensitive organizational data are often the target of most enterprise AI systems. With inadequately put regulation on such systems, LLMs can inadvertently disclose classified information by means of responses or in the wrogaccess channels. The implementation of various privacy laws such as GDPR, HIPAA and many other legal obligations as well quest privacy policy, require various security compliance and data governance measures to be developed and implemented in such systems.

It is also worth mentioning that bias and fairness problems are among the other problems that makes it hard for AIs to adapt at the enterprise level. This comes in of the fact that LLMs learns from very large data sets and this data may contain some societal or cultural biases which can sometimes make the generated output prove to be biased or even discriminatory. Such scenarios compromise the trust of the clients on the institution and also gives the institution cases to answer both morally and legally. It is therefore advisable to continuously track, conduct bias tests, and apply an effective AI policy as safeguards against these risks. Enterprise LLMs also feature important challenges on the topic of security threats. Attacks such as the introduction of prompt injection, the manipulation of the results for illicit purposes, actions like model access without authorization, is a situation where an attempt to corrupt and interfere with the functioning of data already present is made. Additionally, using AI-based recommendations excessively without verification may endanger business activities. Thus, the wider use of generative AI technologies by firms has made it imperative to institutionalize specific frameworks for risk management, responsible for analysis and mitigation of newly emerging threats. To this end, companies are advocating for the construction of AI governance regulations, guardrails, surveillance infrastructure, preventing synthesis approaches and human in the loop let Meavie their actions. These programs are designed to create system that satisfies the necessity of securing, verification and efficiency of enterprise AI applications, implementing LLM technology to the whole creation.

4. Prompt Injection and Adversarial Threats

Overriding the prompt is among the top security risks faced by LLM in large enterprise deployments. The prompt attack is an alteration or modification of the input such that the model can abandon or bypass its original command constructs or call other functions. The attackers use

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

any user inputs or external data and embed commands that they expect to cause the model to behave differently and render output that was not intended. This is very worrisome for AI security and governance especially since LLMs are getting more and more in the organization's critical systems. Prompt injections are a new concern for AI security and governance termed prompt injection which shall not be the same if we were to look at the traditional programming vulnerabilities. This being the case, there are users who may provide inputs instructing the model to inappropriately reveal organizational classified information, violate the applied security measures, create considerate content, or output nonintended results. Specifically, for the Retrieval-Augmented Generation (RAG) systems, the objects of interest can also include pieces of advice which are not legitimate but are dosed with malicious statements which make the relevant norms and procedures of behavior. The attacks in this sense are not simply prompt injections; they present an entire gamut of methodologies for compromising AI systems. Examples include jailbreak attacks which are executed by users who violate the safeguards intentionally; attacks against data integrity where data poisoning attacks implant cyber viruses in training or query datasets; and evasion attacks where intrusive data is muted to escape perimeter gates all clearly indicates that all the named attacks can create a threat to the system performance. Where such security infringement is possible the breakage of system operation, the loss of restrictive information as well as cause of operational and reputational risks become ever more likely for the organization.

The growing prevalence of advances in proudly malicious methods has shed light on a need for the adoption of resilient tools to face hostile activities. Customarily, the first line of defense against hostile, unethical or criminal data-input, and output prompt injections is input control elements, allowance based access mechanisms, and output resistances. Or frameworks may be said to be integrated security checkpoint among other things, so that other include access management, such as behavioral policy, and exclusion by exception. This other apparatus may include retrieval validation, intruder detection systems, penetration testing and the like with integrative purpose of threat hunting to find and neutralize possible attacks before they come into motion. In the same way, come to enterprises who are most security conscious, such strategies, as might human error, seem minimal or irrelevant, however, human error plays a significant role in the success or failure of any security strategy. For this reason, security departments and security professionals hold classes and courses on a number of subjects including testing one's own defense mechanisms against the adequacy of certain threats in the information security field. This is where targeted continuous monitoring and assessment comes in that can help businesses to identify threats as it evolves and pick or adjust protective countermeasures particularly that involve AI technologies. Prompt injection and adversarial threats will persist as the key focus of concern in AI security with the evolution of enterprise applications that are powered by LLM.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Establishing robust governance controls and employing defense in depth mechanisms are prerequisites for using generative AI at enterprise level to maintain integrity, trust and ethical standards.

5. Output Guardrails and Response Validation

Output guardrails are measures which are similar to the protective techniques which allow and control the responses generated by Extremely Large Language Models (ELMs) before their arrival to the final uers. While to tackle the issues of queries on the input side, one can use the input guardrails, in case of the output shield – the most important is that the platform searches to maintain appropriate behavior from the outputted response based on the set organization norms, ethical considerations, and the level of security prevailing in the market and responses category. In traditional business settings, this where the content generated by AI algorithms could bribe people or take advantage of a person, however there can be no secure methodologies given in this case. Requirements. {*} The term burstiness in this work refers to the inter pacetraffic pattern of a set of packets over a give unit of time and is central in the design of contol mechanisms.

LLMs, as has been observed, confront the problem of floating into an area where the content becomes irrelevant and biased by producing content that goes against the intended purpose. With this in mind, the developers of these tools have endeavoured to provide output guardrails that aim to correct the content before it goes out by including checks that the responses are factually correct, in line with the policies and are suitable in the context. Some of the tools that have been highlighted in the past include content classification systems, toxicity evaluation tools, proof checking apparatuses, confidence levels estimating features and rule-based validation mechanics. In the forgoing event failure to meet some of the above rules may result in the responses being either revised, restrained, tagged for moderation, or ‘complete data models’ means that these answers have nowhere to go. The necessity of response validation is clear-cut in such sectors as healthcare, finance, insurance, and legal services, where mistaking information could results in dire consequences. To fulfill the organizational requirements, output validation has been augmented with the application of Retrieval-Augmented Generation (RAG) systems for providing verification of responses based on known information sources. The introduction of multiple verification levels enables companies to greatly decrease the incidence of misinformation, policy contravention, and reputation compromise as well as improve the overall reliability of AI based systems meanings of AI.

Component	Description	Purpose	Enterprise Benefit
Response Validation	Evaluates generated responses before delivery to users.	Ensure output quality and correctness.	Reduces inaccurate and misleading information.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Fact Verification	Checks generated content against trusted sources or knowledge bases.	Improve factual accuracy.	Minimizes hallucinations and misinformation.
Content Moderation	Detects harmful, offensive, biased, or inappropriate content.	Maintain safe interactions.	Protects brand reputation and user trust.
Policy Compliance Checking	Reviews outputs against organizational and regulatory policies.	Enforce governance requirements.	Ensures legal and regulatory compliance.
Confidence Scoring	Assigns reliability scores to generated responses.	Assess output certainty.	Supports risk-based decision making.
Rule-Based Validation	Applies predefined business rules to AI outputs.	Ensure adherence to enterprise standards.	Improves consistency and predictability.
Sensitive Information Detection	Identifies confidential or protected information in responses.	Prevent data leakage.	Enhances privacy and security protection.
Toxicity Detection	Screens outputs for abusive, hateful, or harmful language.	Promote ethical AI behavior.	Reduces reputational and legal risks.
Human-in-the-Loop Review	Routes high-risk responses to human reviewers.	Provide expert oversight.	Improves accountability and reliability.
Retrieval Grounding	Verifies responses using retrieved documents in RAG systems.	Ensure evidence-based outputs.	Increases trustworthiness and transparency.
Automated Re-Generation	Regenerates responses that fail validation checks.	Improve response quality.	Enhances user experience and system reliability.
Audit Logging	Records generated outputs and validation actions.	Support monitoring and investigations.	Facilitates compliance and governance audits.
Explainability	Provides reasoning or	Improve	Strengthens

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Mechanisms	source attribution for responses.	transparency.	stakeholder confidence.
Continuous Monitoring	Tracks response quality and safety metrics over time.	Detect emerging risks.	Enables ongoing optimization and improvement.

6. Content Moderation and Policy Enforcement

Essential for the well-being of algorithms and the use of natural language processing systems is taking care of the service's content guidelines as well as the regulation. These principles concern badly intended, illegal actions, such as preventing the aspect of creating damaging information, offensive material, false data, and illegal information that would hurt the people accessing the services as well as help more regulatory services. Content moderation implements a wide array of processes ranging from artificial intelligence-based content moderation to manual and optimized processes, from ICMP to risk-based card programs. A tier based approach aims to exclude in terms of content the user input into the system and as well as any output generated by the AI techniques e.g. text translations. Any such content accommodates abuse, racism, generically inversely dependent to that division of facts, lies, text and images which contain abuse, pornography, incriminating fear phrases and phrases that illicit enough passion into causing chaos to put every perpetrator in jail, etc. rather, any such content has policy aspects such as they require payment for use etc. Any such content goes through members and the peer review committee before it is used. Automating moderation tools can remove, change, or even use that comment to call center escalation depending on the risk level.Policy enforcement however goes far beyond content filtering as it is also about the implementation of the organizational policies, industry best practices, and regulatory requirements within AI workflows. The Policy Compliance/Audit Service should allow the enterprises to define their own rules which will be aimed at ensuring that the organization's data and operations like cloud services are safe. AI Content Moderation and Enforcement Features is the ability of an AI platform to follow various data-centric rules of the organization including prompt constraints,gaurd rails, validation rules and monitoring frameworks. This is a very crucial business requirement as it helps in avoidance of legal issues, maintains reputation of the organization as well as manage its relationship with the society at it is AI-based.As the technology of cognitive or generative AI systems advances year by year, the internal practices, standards and rules concerning content control and enforcement are forced to develop in order to minimize the risks that come when operating responsible AI devices.

7. Human-in-the-Loop Safety Frameworks

Human-in-the-Loop (HITL) safety frames are safety mechanisms which incorporate the human element into the process of AI decision making to enhance reliability, accountability, and the

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

mitigation of risks. Naturally, Large Language Models (LLMs) can in most case work autonomously and do not require human interaction to a certain level; still, they create outputs with errors or that favors one social community or the other and at times disadvantageous or may be offensive to the community. Machines are therefore helpful in providing an additional layer of on the spot quality assurance that allows firms to detect errors, enforce their guidelines and preserve the order of operation of the AI entities. To this extent, in-business HITL frames are the most beneficial for the most complex applications where even minor errors can lead to serious damages. For example; diagnosis aiding in healthcare, financial advisory software, contract analysis, information protection and penalties handling of the companies. While this is the case, it should be noted that, all AI-produced solutions undergo scrutiny and validation immediately before any decision is made. Another Human factor that is fundamental in the AI convergence framework is the inclusion of a compliance expert who would particularly input insights so that an AI solution with the highest level of compliance is adopted. There can be human superintendence at different points of the AI use process. Some cases are examples when a human can assess the work before an actual work is completed or user inputs can be examined. The scopes of intervention actions which will bear these characteristics are seldom used largely but consist of automated audit of intermediate machine actions. For these reasons, there are controls that ensure that unpleasant results are escalated to humans as opposed to machines. This merged strategy fosters both the efficiency of the machines as well as human capabilities in the form of knowledge and ethics. Today, it becomes apparent that organizations' management of AI capabilities requires a blend of automation and humans overseeing it. Thanks to the introduction of HITL into various organisations' architecture of AI, they will not only be able to show how decisions are taken, but also reduce operational risks and assist in building confidence in the use of AI-enhanced decision support capabilities at the same time.

8. AI Governance and Compliance Frameworks

every organization that uses artificial intelligence or AI, there are rules and controls in the use of its technologies. Part of this is ensuring that there are benefits and in case of harm, where does the blame fall. This is what could be termed as AI governance which also is part of the ethical considerations when using AI. Other aspects of AI governance frameworks would include compliance with the law, and the potential policies and algorithms that can be put in place to dictate the behavior and decision-making of the AI. With the evolution of artificial intelligence for all industries, effective AI governance needs to become part of the mainstream ethos in all organizations that deploy such tools be those from the past or the present.

Almost at the same extent, AI Governance has been hiding under the rock to well identify its strategy and structure. Apart from the development and deployment of AI systems, both AI and its governance also require monitoring of already deployed systems and ongoing inadequacies.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Programs for governance name the roles of various departments responsible for governance, as well as the risks that have to be managed, codes of conduct to observe the appropriate it ethics, as well as provide punishment mechanisms. These frameworks are the guardrails that ensure the AI systems in the organizations are targeting and fulfilling organizational imperatives while reducing the risks of others and of operational inefficiencies. It is about creating codes and standards that would be operational towards achieving the goals of the organization and the induced focus when decomposing the FIGI-F strategist(Fletcher, et al, 2002). Compliance frameworks are designed to help in coding well incorporated regulations and industry specific guidelines with regards to privacy, security, bias and open information necessitated by the working conditions without causing any anxiety. Moreover, there are total of 99 sectors waiting to be produced while the current ones will be given a grace period where they will comply with the same and the last sixty-eight laws. Usually, compliance demands having access to set expectations; creating records and keeping them, filing audit trails, conducting company representatives due diligences or prince two audits and monitoring how id IT projects are implemented in the company. Today more advanced frameworks seek incorporations of equity, responsibility and transparency, explicable behavior, data protection, and output control in the AI governance. Resolving of the matters on the manner that some of the problems will apply onto some of the contractors leaving only AI Governance to directed to society and institutional requirements, which occur from within most other governance systems. It helps in catching the so-called “bad behaviors” of people who develop and use AI and thus reduces possible risks connected with the system’s abuse. According to the management regulation statements ASAI developed in 2021 significant bundles of compliance actions are laid down for each sphere of the sphere of the resources ministry.

9. Conclusion

The seamless assimilation of Large Language Models (LLMs) in business solutions has altered the process of automating activities, organization and handling of data, as well as dealing with customers in a major way. This ranges from smart chatbots and AI copywriting systems to analytical and business intelligence applications and even content management systems, Large language models have proved to be a potent aid in enhancing businesses through operational effeciency and creativity. Nevertheless, operational practice of these systems raises numerous obstacles that prevail— reliability, security, confidentiality, legitimacy, and moral use of AI relate the systems. All these challenge concerns need for designing additional security measures when adopting AI in the organization as such can facilitate quality checks over the AI operations in the businesses efficiently.This research sought to provide an overview of the AI guidance concept, environment guidelines, and their significance in reducing many issues associated with business enterprise Large Language Model (LLM) applications. This argument showed that

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

businesses have to deal with many risks as they operate within a dynamic business environment that includes illusions, cue line insertion, depletion of general order guidance, data breaching, adverse selection, fake information, and violation of laws. These risks, if not checked, are capable of destroying user confidence, revealing confidential details, and may drive step geographical and functional businesses to collapse. Besides, the analysis necessitated having several safety modules with basic principles such as input validation, content moderation, response validation, retrieval security, human intervention and enforcement of other security policies. An example of an input guardrail protects against inputs that are harmful or unauthorized, while an output static block upholds that the generated response remains upright, safe and follows the lawful stipulations as enacted by the organization. On the other hand, content moderation systems and policy enforcement contribute significantly towards maintaining ethical Agencies within multiple application use cases. Human in the loop security model was the second most important element in corporate AI governance. Even though the current large language models such as XLMs are fantastic devices, the human has to have a supervisory function, especially in the presence of high risk, abnormality or fault condition. The efficiency of Al in service delivery combined with individual expertise is what every company aspires to benefit from hence avoiding negative repercussions while still enhancing the quality of services.

REFERENCES

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv. https://arxiv.org/abs/1606.06565
2. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the Opportunities and Risks of Foundation Models. Stanford Center for Research on Foundation Models. https://arxiv.org/abs/2108.07258
3. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—An Ethical Framework for a Good AI Society. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5
4. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730
5. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of EMNLP 2020, 6769–6781.
6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

7. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology (NIST)
8. Perez, F., & Ribeiro, I. (2022). Ignore Previous Prompt: Attack Techniques for Language Models. arXiv. https://arxiv.org/abs/2211.09527
9. Shneiderman, B. (2022). Human-Centered AI. Oxford University Press.
10. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., et al. (2022). Ethical and Social Risks of Harm from Language Models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 615–628. https://doi.org/10.1145/3531146.3533088
11. Willison, S. (2023). Prompt Injection Attacks Against Large Language Models. arXiv. https://arxiv.org/abs/2302.12173