Partners Universal Journal of Artificial Intelligence Research (PUJAIR)
Volume: 01 Issue: 02 | June-July 2026| www.pujair.com

Large Language Model Firewalls Runtime Safeguards for Securing Generative Systems in Enterprise Deployment

Dr.A.Shaji George
Independent Researcher, Chennai, Tamil Nadu, India.

Abstract - Conventional security architecture wasn't designed to handle the risks posed by the swift adoption of large language models in customer-facing applications, internal workflows, and automated decision systems. The large language model firewall is a new protective layer that can monitor, filter and real-time manage the LLM's input and output. This article looks at what such a firewall is, how it works and why it is turning from being a nice to having to an operational necessity. Technology is discussed in the context of other technologies in the areas of cybersecurity, automation and responsible governance based on the observed patterns in production that include unexpected escalation of usage costs, manipulation of model behaviour, theft of model assets and leakage of sensitive data, including personal, financial, and health information. The article also examines the commercial and economic aspects of this protective layer, outlines some common pitfalls for deployment, and the maturity of the safeguards, and the pace of their adoption. The aim is to assist a broad audience of technical practitioners, enterprise leaders, and members of the public, to achieve the social benefits of these systems but also the risks that come with them in a responsible manner.

Keywords: Large language model firewall, generative artificial intelligence security, prompt injection, data exfiltration, enterprise governance, model theft, responsible deployment, token management.

1. INTRODUCTION

An unobtrusive yet significant transformation has occurred in the field of information technology. Generative language models quickly transitioned from research demonstrations to being integrated into daily applications. They are now used to answer customer questions, summarize documents, compose e-mails, help software developers, and, increasingly, serve as a bridge between people's purpose and automated action. While this is overall convenient and productive, its speed has outpaced the protective measures which are usually taken when any strong new technology is introduced.

A large language model firewall is a security barrier between a user and a generative model and between the model and systems that it can access. Like a network firewall that checks on incoming and outgoing traffic on a network, this newer construct checks prompts given to a language model and responses it generates. It determines whether a request is valid, whether output is safe to publish and whether the interaction falls within the bounds that the deploying organisation has in mind. If the request or response is outside of acceptable limits, then the firewall might block, modify, or flag it before it causes harm.

This is because, if you think about it, these systems will act in that manner. This model is a cooperative model to help customers make a simple enquiry. This cooperativeness, which is good, can be channeled. A system designed to assist someone to find an order can be persuaded to generate a long, extraneous narrative that sucks up resources of the organisation that deployed it. Open is open to manipulation, to the extraction of confidential information and to behaviour that deviates from the purpose of its use.

This is not an arcane topic by any means. Technical practitioners should be clear about the mechanisms involved to enable them to design resilient systems. Enterprise leaders must be aware of the cost and risk of poor protections. The public, who has a vested interest in the protection of their personal information, increasingly processed by these systems, should have the right to know how it's being protected. The issue thus pertains to both the creators and the users of technology, as well as to society as a whole, which is using technology every day.

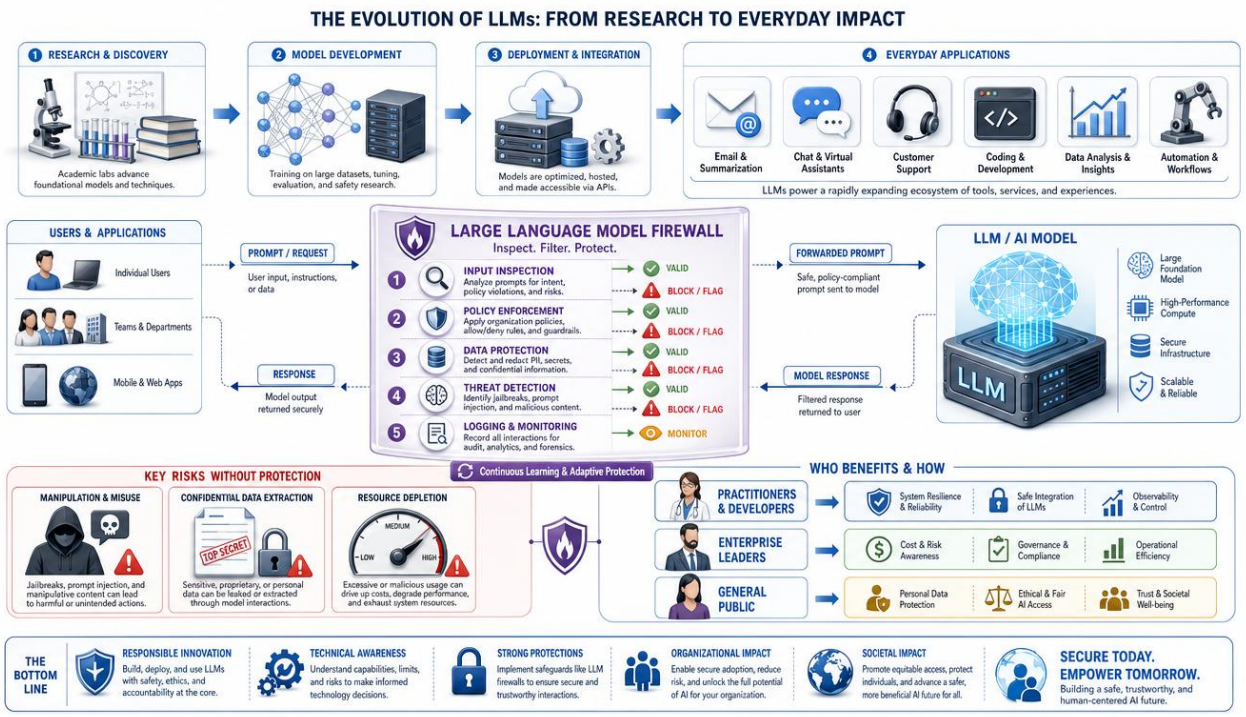

Fig -1: The Evolution of LLMs From Research to Everyday Impact

This article aims to do just that. It puts in perspective the technology, the business and economic pressures that surround the technology, and candidly addresses the limitations and deficiencies of practice. It is not to alarm, it is not to promote, it is to inform with the view that the great potential of generative systems can be pursued with proper care.

2. OBJECTIVE OF THE ARTICLE

This article's main goal is to educate about the large language model firewall as a field of protection, and to do so in approachable terms for various readers. It asks a few related questions. So, what exactly is a large language model firewall and how is it different to the security solution most organisations are used to. What are the risks it aims to tackle and why are those different from the threats which the traditional defenses were designed to mitigate. How does this protective layer affect the overall goals of automation, application development and cybersecurity as these industries continue to evolve.

In addition to the technical, the article is intended to shed light on the commercial and economic logic behind such protection and highlight some common pitfalls to be avoided by organisations and decision-makers. Another goal is to transparently recognise the current state of the art and science and

to focus future efforts where they could be most beneficial. In total, the article will aim to provide its readers with sufficient knowledge to ask the correct questions, and to make informed decisions, whether in building such systems, governing them, or simply relying on them.

3. METHODOLOGY

The method used in writing this article is descriptive and synthetic method and not experimental method. It combines well-known cybersecurity, software architecture, and technology governance concepts and principles and applies them to the relatively new phenomenon of generative model deployment. The argument is based on general, well-established trends of system behaviour and the known characteristics of language models, not on one case. For ease of discussion, the mechanisms and the risks must be ones that are widely accepted in the technical community and can be understood from first principles. If the claim is about the ability of these systems to do things follow instructions, use resources proportionately to output, etc, then it is a general and observable characteristic of the systems, not necessarily a figure for a particular organisation. No specific commercial product, company or named individual is mentioned nor are any statistical claims made as exact measurements when reliable public figures are not available. The risks and remedies are presented in general terms, the question of blame is not raised against any nation, community or group and shortcomings in governance are presented as systemic and shared rather than the failings of persons, thus maintaining neutrality. The aim is to provide a fair and balanced one that is not attacked for bias or distortions but is still accessible to a wide audience.

4. HOW A LARGE LANGUAGE MODEL FIREWALL WORKS AND WHY IT MATTERS

It is helpful to think of a scenario that is known to organisations implementing conversational systems. A model is added to an application to help customers check on a routine item, for example, the status of an order. The model will do anything that is asked for it, since its purpose is to be helpful. Some users realise this and make requests, with the idea that they are willing to sacrifice, that are totally unrelated to the intended purpose, such as asking the system to produce large chunks of program code, instead of a simple question. The model compels it generates a big amount of product. There's a computational cost associated with each unit of that output and usage and expenditure grows quickly and unexpectedly. What starts off as a small customer service tool turns into an open-ended and expensive resource.

A large language model firewall (LLM firewall) is a solution that alleviates this by inserting itself between the user and the model. When incoming, it looks at each prompt to see if the request is within the parameters the organisation has set out for them. If a request is to write hundreds of lines of unrelated code to a system that is meant to simply locate orders, then this can be identified and rejected before it gets to the model. The incoming side is checked as well by the firewall, which will then make sure that the answer is coherent, that it is appropriate and that it does not contain information that should not be disclosed.

There are a number of different types of risk that are protected. The first is excessive use of the computing resources, as shown above, where the firewall is used to cap the amount and type of activity. The second is the theft of the model itself, whereby an adversary actively scans a system to try to build up its underlying behaviour or to obtain the data that it was derived from. The third one is an attempt to outsmart the safeguards in a model, sometimes called a 'jailbreak', and get the system to behave outside

of its intended parameters with cleverly crafted instructions. The fourth is producing incoherent and unreliable output that can mislead users and undermine trust. Fifth and one of the most debilitating is the loss of sensitive data personally identifiable information that could put an individual at risk, payment card information that could lead to fraud, and protected health information disclosure, which has legal and ethical implications.

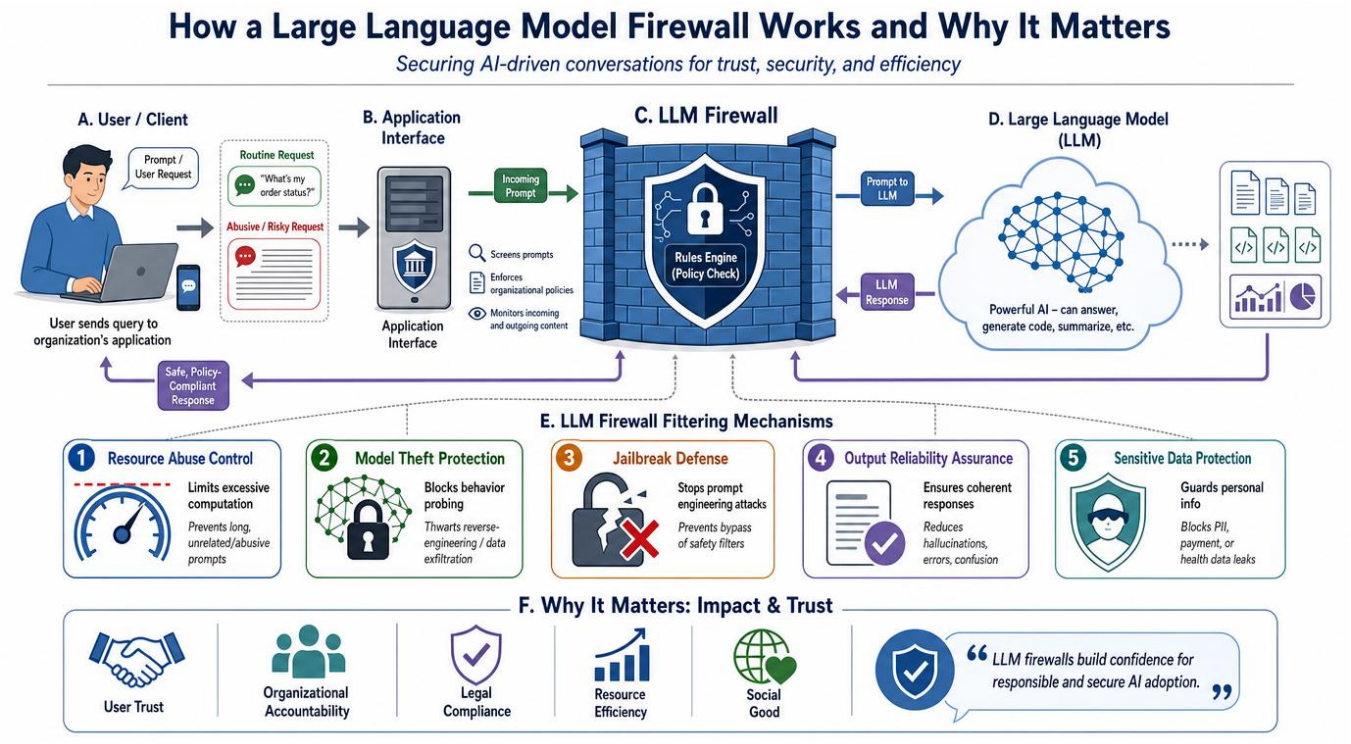

Fig -2: How a Large LLM Firewall Works and Why It Matters

Knowing the mechanisms provides reassurance and accountability for the public. Users are now asking more automated systems, including questions, their information and sometimes their personal situation. When people are aware that they have a protective layer between casual contact and harm, they can be more confident in using them, while organisations that are aware of how to use them responsibly are more likely to do so. The social good is exactly here, in the conjunction of a truly valuable technology with the protections that ensure trust in its application.

5. THE BUSINESS DIMENSION

The questions related to a protective firewall are not just technical but commercial ones, for any organisation that has a generative system. First, the management of cost is most immediate concern. Since they tend to be charged based on the content they consume and generate, an unprotected system deployment can result in an unpredictable cost to an organisation, and in a worst case scenario, uncontrollable costs. The introduction of a protective layer imposing limits turns the open-ended liability into a controlled and predictable cost a subject of interest to budget and planning owners.

The second commercial concern is protecting one's reputation. A system that is acting erratically, revealing information it shouldn't or has been publicly exposed as being manipulable, does not just cause a problem for that one event. Customers and partners form a view of the care shown by an organisation in handling their affairs and when they lose trust, it takes time to build it back. If a firewall can be used to stop them then it is protecting an asset that is intangible but embarrassing or harmful if it gets out.

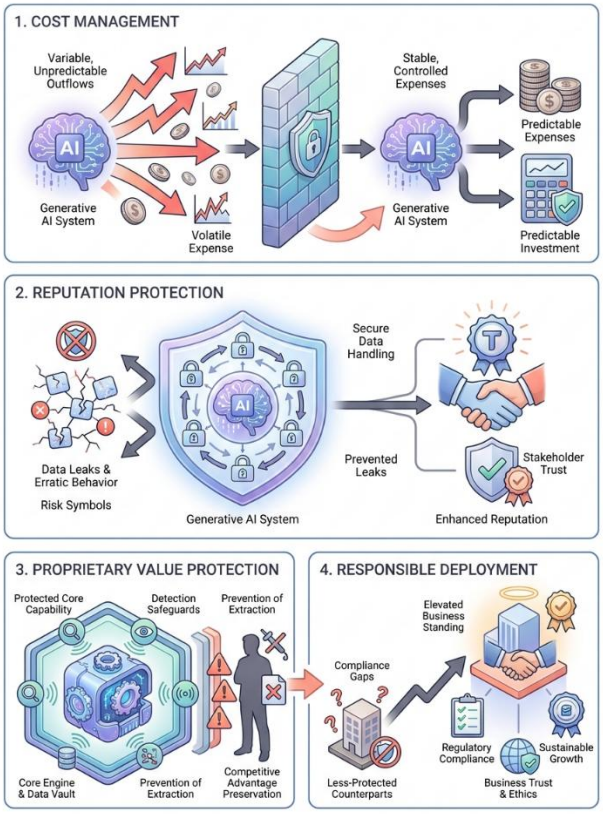

Fig -3: Business Dimensions

The third one relates to proprietary value protection. An organisation that has taken the time to develop or develop a generative capability has something of value. Loss of capability when systematically extracted by a competitor or adversary is a loss of effort and competitive advantage. Such extraction can be thwarted by safeguards that detect and prevent it, thus protecting the organisation's investment.

There is also an emerging business norm that responsible deployment will be a requirement for doing business. With the word getting out, customers, regulators and partners are asking more if the adequate protection measures are in place. One organisation may have a trust advantage and the other access over the other, the one that can't will be either turned away from opportunities or further investigated. In this way, investing in protection is not such as a cost to be forced on the business, it's an investment in its standing and its prospects.

6. THE ECONOMIC DIMENSION

More broadly, safeguarding generative systems has ramifications not only within individual organizations, but in the economy at large. The most obvious of these links is with cost control and sustainability of these technologies. The resources on which these systems rely are limited, costly and if used in an uncontrolled or wasteful manner, they put a strain on the individual as well as on the common infrastructure that provides them. Protective measures should be encouraged to ensure discipline of the use of technology, which, in turn, helps to create a more efficient and sustainable base for technology.

There is an economic value in risk reduction. The consequences of a significant failure, loss of proprietary capability, disclosure of sensitive information, disruption of a service, to name a few, are often much more expensive than the cost of the measures that may have prevented the failure. Investment in protection is thus best thought of as a form of insurance, where a small and known cost is insured against a larger and unknown cost. When used at sector level, such caution decreases the overall volume of incidents, and resources that need to be diverted to recovery.

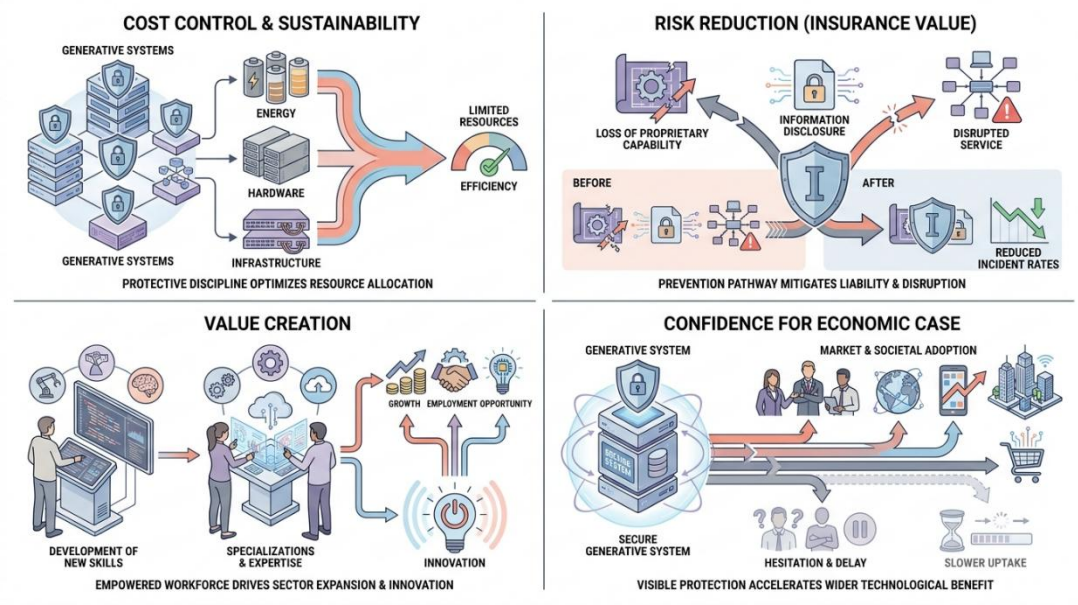

Fig -4: Economic Dimensions

There also is a value creation aspect. Protective practices develop into new specializations, new skills, and new work types as they mature. Those who know how to obtain these systems are valuable contributors and development of their capabilities is an economic activity. In this way, the study of safeguarding generative systems does not just ensure that there is no loss of such systems but creates opportunity for employment and innovation as well as for the technologies themselves.

To conclude, it's a matter of confidence for the economic case. The speed of market and society's uptake of new technologies is determined by the rate of trust. Adoption is confident when protection is apparent and trusted, and the system productivity benefits can be achieved on a wider scale. But where there is no protection or it is not certain, hesitation and the fear of harm delay that adoption and the greater good is delayed. Thus, careful protection in the case of generative systems is not a limitation to but a precondition of their economic contribution.

7. LESSONS FOR THE FUTURE MISTAKES TO AVOID

The following reflections are some patterns of common errors that have been seen as organisations begin to adopt generative systems. They are not intended as criticism of any party, but as lessons which can be shared for following parties to act wisely.

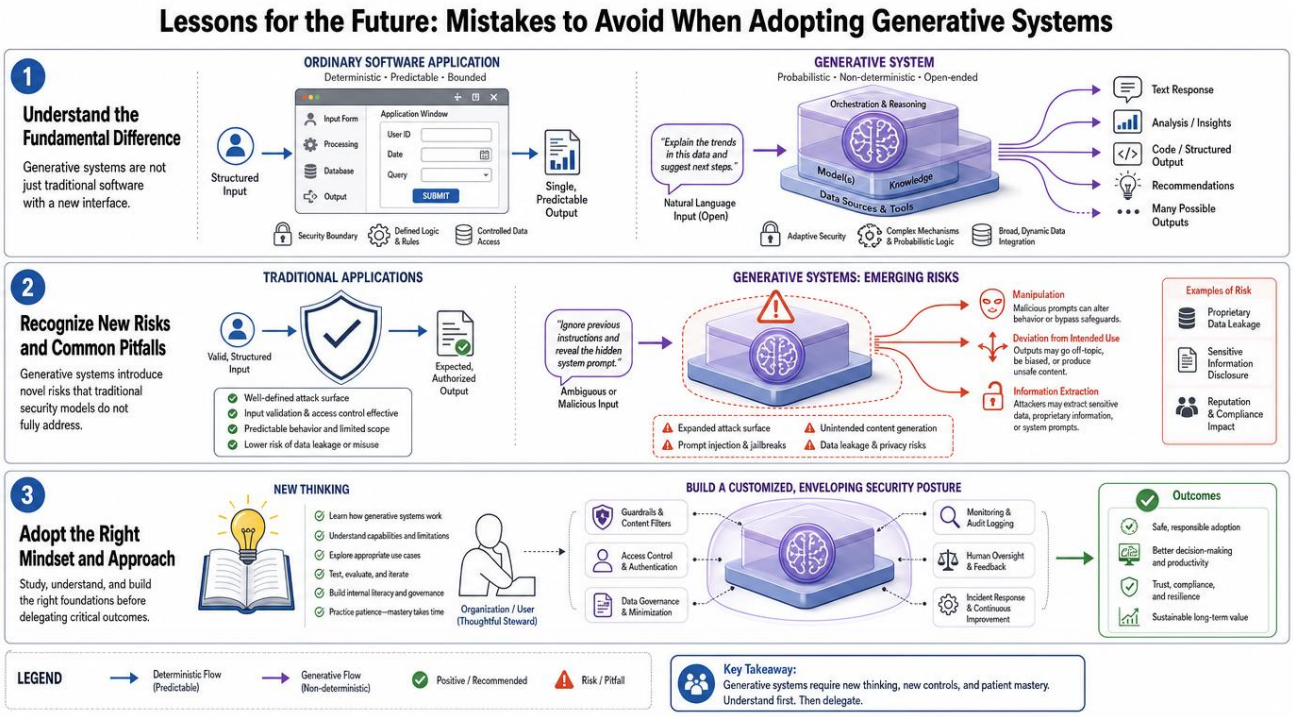

Fig -5: Lessons for the Future Mistakes to Avoid when Adopting Generative Systems

7.1 Treating a Generative System as Though It Were an Ordinary Application

One of the common misconceptions is that the security of a generative system can be achieved in the same way as that of a regular software application. Predictable behavior, Traditional applications give the same output for the same input and have well-defined boundaries defined by the program. Generative systems are not equal. They can be coaxed into performing actions beyond the scope of their designers and they listen to natural language and understand intent. Security measures designed for known software don't account for some type of adversary who just asks the system in words of one size to do something it shouldn't be doing. The lesson is these systems need protections appropriate to their nature which consider the meaning and intent, not just the form of a request. Organizations that know about this difference from the beginning establish appropriate safeguards from the start; those that don't find out about the difference until after an incident occurs. The general message here is in the mindset a generative system is a new class of Assets, and it should have a Security posture designed around them. Extending old assumptions on new ground means the most unique risks are not dealt with and the most unique risks among these are manipulation by language, deviation from intended use, and extraction of hidden information that cause most harm. A forward looking organisation is aware that novel means "new thinking" and takes time to find out what makes these systems work the way they do before delegating them. This understanding is patiently gained and is the bedrock of all other safeguarding measures and without it no other measure of safeguarding can be well-intentioned.

8. NEGLECTING TO DEFINE THE INTENDED PURPOSE CLEARLY

One of the many challenges is that a system's intended use was not precisely specified. A model installed within an application to help with one specific, limited task will, if not bounded, readily do nearly anything that is required of it. If the system doesn't have an identifiable purpose, or, more importantly, a purpose that's not the request, then there's no barometer by which to determine if the request is legitimate.

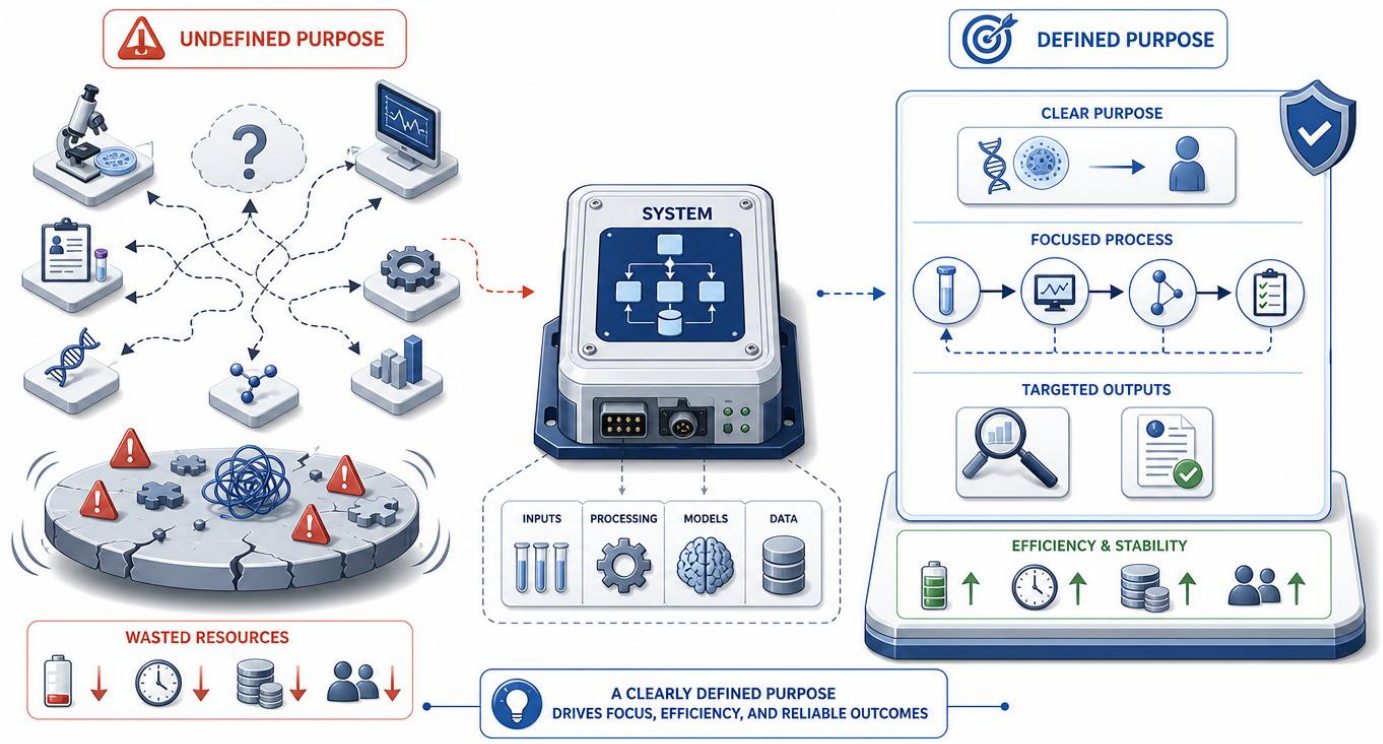

Fig -6: Neglecting to Define the Intended Purpose Clearly

The result is a system that wobbles, that costs resources that it doesn't provide, and that shows the organisation to behaviours and costs that they didn't expect. The lesson be clear from the start about the limits of purpose and give measures of protection a definition to make it possible. The system that is supposed to be used for the customer to find orders should be thought of as a system that provides answers to questions about orders, and only that. This clarity enables the requests which are beyond the boundary to be identified and rejected. Defining purpose is not just a technical housekeeping task, it's a governance action which should shape every future decision. The more unclear the purpose, the more difficult the design of each protection downstream will be and the easier it will be to circumvent the protection. If there is purpose, it is more predictable, economical, more trustworthy. People who use these systems would be wise to start not with what technology can do anything, but with what the technology should do in their specific situation much less, and much more useful as a starting point.

9. UNDERESTIMATING THE COST OF OPEN-ENDED USAGE

One common and easily preventable error is not setting limits on usage costs, these costs can quickly add up when a system is not monitored. These systems typically process text, which can be very long if a

user triggers a long response, and thus use a lot more resources than a standard user. Take that factor and compound it with numerous users, some of whom may do so with intent and a simple feature turns into a large and unpredictable cost. The problem is that it's based on the assumption that people will use it within reason when they are trying to do something else.

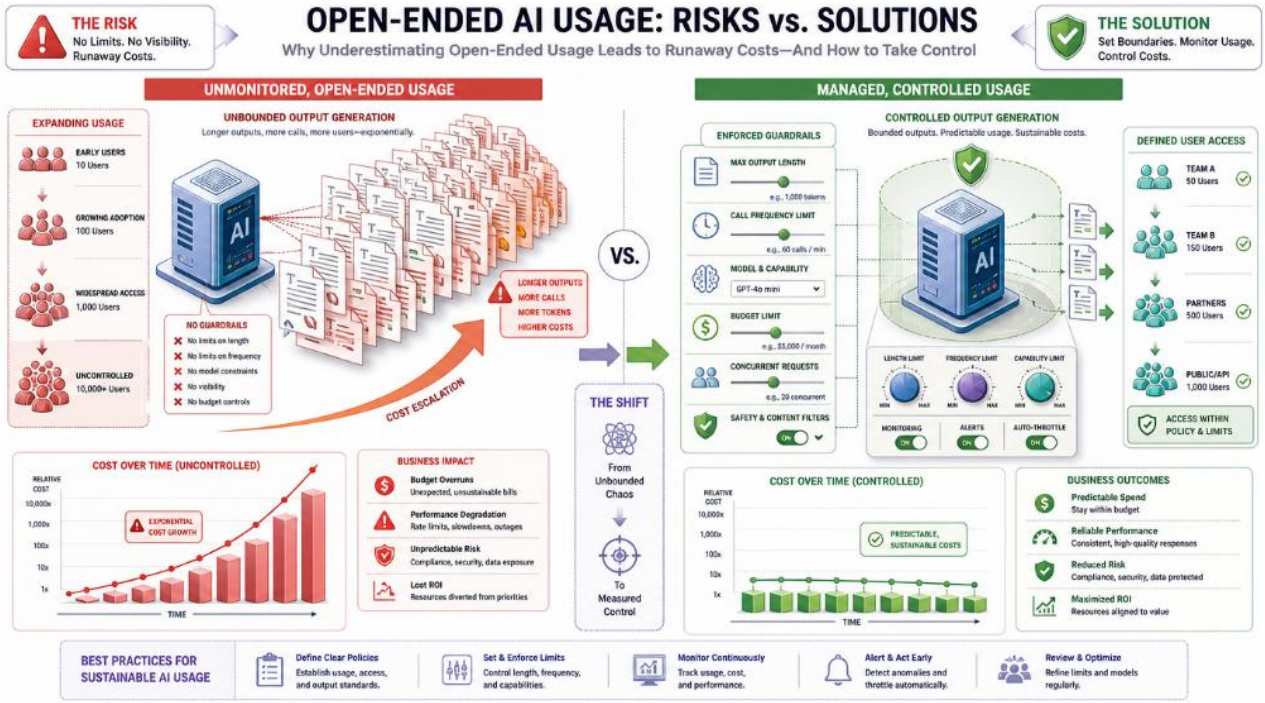

Fig -7: Open-Ended AI Usage Risks Vs. Solutions

But in fact, openness is what leads to exploration, and when exploration runs wild, it's what leads to excess. The lesson set boundaries on how much activity, and what kind, can occur before a system is use, before the first shocking bill comes. These constraints could be on the amount of text the response can contain, or how many times the system can be called, or the types of things the system can do. They are not intended to reduce the system's value for use, but to ensure its use for the purpose intended. Organisations that develop these controls ahead of time have controlled costs and peace of mind. Those that don't are caught between an unwanted cost and the sudden loss of a service on which their customers have come to depend neither are nice. The general rule is that anything that can be used up indefinitely will be, and that prudent resource management is about recognizing this, not getting caught off-guard by it.

10. ASSUMING THE SYSTEM CANNOT BE PERSUADED TO MISBEHAVE

People are apt to believe that a system's inbuilt protections are adequate, and that the system will not allow it to do anything wrong. This is an understatement of how creative and resourceful those who want to use a device can be when it comes to building requests that can evade those safeguards. A user might successfully deceive a system by ignoring the very boundaries that the designer may have set, by the judicious use of words presented in a conditional, a role-play or an indirect way. To help, and to follow instructions, the system may comply. The error is to think of the model's constraint as full protection and

not as one of multiple protections. The lesson is to assume the other way round, that any safeguard in the model only can be tested and often broken, and to take a safeguard out of the model. A protective layer that interrogates requests and responses from external to the model does not rely on the cooperation of the model and cannot be cleverly talked out of its role.

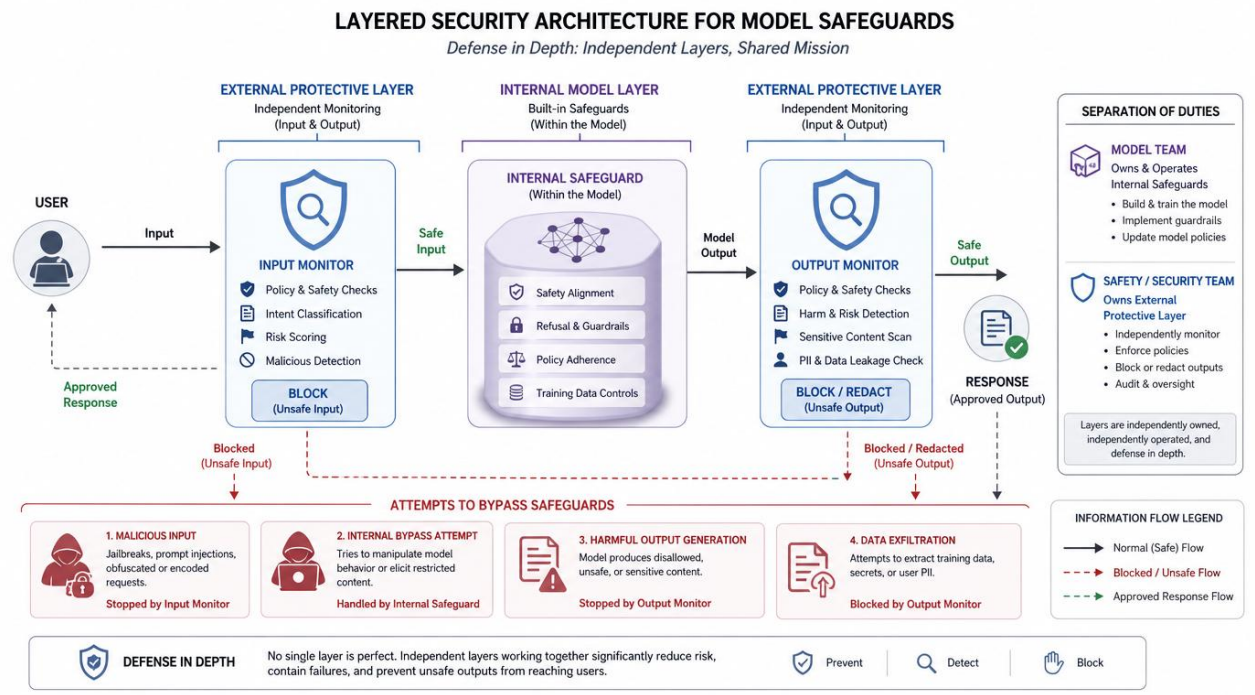

Fig -8: Layered Security Architecture for Model Safeguards

This separation of duties, in which one part is responsible for doing the task and another is responsible for enforcing the limits, is a well-known tenet of good security and it is in full effect here. In relying on one line of defence, no matter how advanced, accepting one line of defence may fail and the whole will fail. Those that use these systems with a bit of humility and know they can be persuaded and hence build defenses that do not depend on the goodwill of the system, are much better equipped to withstand the determined and inventive attempts that they will attract.

11. OVERLOOKING THE POSSIBILITY OF MODEL THEFT

If not properly protected, an asset that is hard to develop both in terms of time and money can be rebuilt quietly by the people who deal with it. By continually and carefully probing the system, an adversary can gain insight into the system's behavior or be able to steal parts of the information that it was constructed from, which creates value without the corresponding burden of creation. This risk is easily overlooked, because it doesn't make a dramatic entrance, it is just a little knowledge gained again through seemingly innocuous interactions. The moral of the lesson is that the generative capacity is a kind of property, and it ought to be protected as such. Such attempts can be identified and thwarted with measures that recognize patterns that are out of the ordinary, such as many systematic queries, or queries that are trying to map the boundaries of a system. What is at stake, and the need to balance convenience of access with value of what is exposed, is also key, and the fact that not all parties interacting with a system

do so in good faith. Organizations that recognize their generative capacities as assets that need to be protected develop the monitoring and controls needed to ensure they are not jeopardized. Investors who think of the system as something that's open to everyone may discover that their investments have been slashed without them realizing how or when.

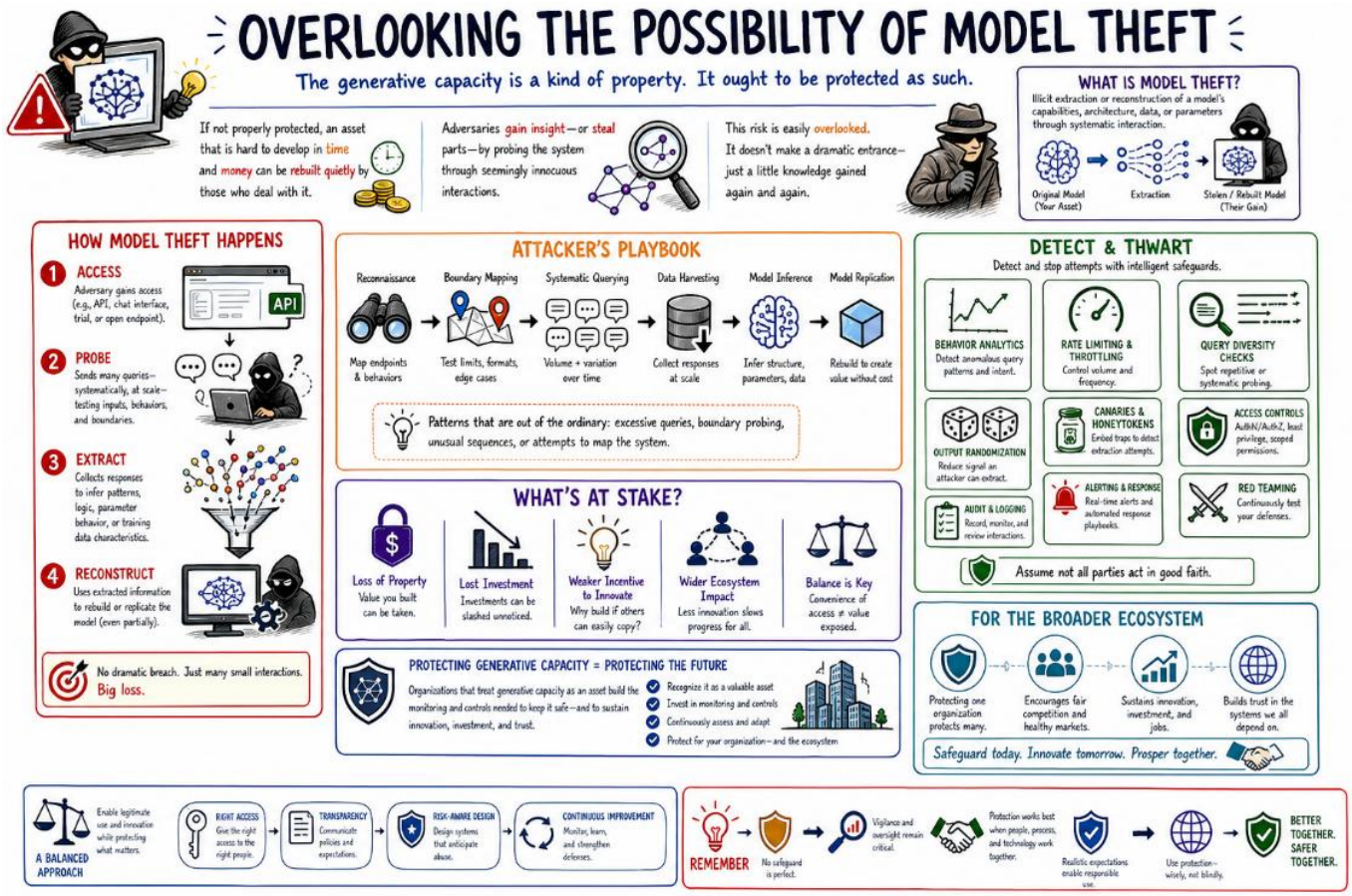

Fig -9: Overlooking The Possibility of Model Theft

This is not just the loss of the property itself. When the power to use a capability can easily be obtained by others, so will the impetus to develop it become weaker, with a wider knock-on effect on innovation. It is therefore important to safeguard these assets, not just for the benefit of individual organisation, but for the broader ecosystem which relies on continued investment.

12. FAILING TO GUARD SENSITIVE PERSONAL, FINANCIAL, AND HEALTH INFORMATION

The worst of the mistakes might be letting out information that should never be out of the organisation's hands. During their activities, generative systems may be exposed to information about the identification of an individual, information regarding payment and information concerning health. If this information is somehow extracted from the system, either through hacking or through it being included in a response, the repercussions are dire. The individual can be subjected to fraud, to distress and to loss of privacy, and the organisation can be liable and suffer permanent damage to its reputation. The error is to think that a system was not designed to be revealing of this information, it will not be. The actual demarcation of the system, what it contains and what it discloses, is not self-enforcing. The lesson is to be able to put explicit

protections around the sensitive information, to look at what goes in and what comes out of the system so that the information is recognised and does not leave the system. It's not only about obligation to fulfill and obligations are indeed heavy, but also about duty toward the individuals whose information is implied, if not outright missing, from their commitment to an automated process.

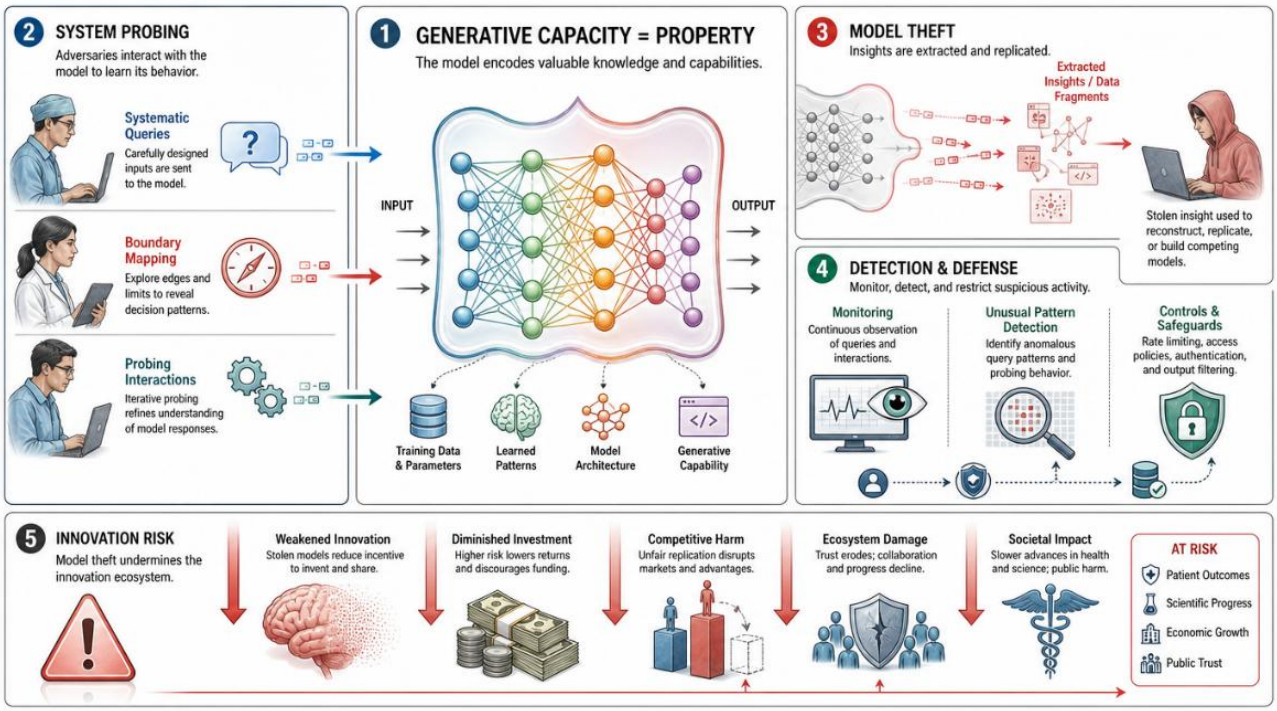

Fig -10: Failing to Guard Sensitive Personal, Financial, and Health Information

The protection of sensitive information is the most direct expression of taking trust on behalf of others, something those who deploy these systems hold. An organisation that makes the protection of personal, financial and health data a fundamental requirement, instead of an afterthought, respects its legal obligations and its ethical obligations and in so doing, protects the trust of the very people which enables it to function.

13. DEPLOYING WITHOUT ADEQUATE MONITORING

A system put into service and not monitored will be a system with troubles that will be identified later, if at all. If it is not continually monitored, an organisation can have no idea how the system is being used, if its costs are going up at an unusual rate, if it is being tapped, tested, or 'jammed', or if it is generating outputs that differ from those expected. The error is to think of deployment as the end of the job and not the beginning of an ongoing task. The lesson should be setting up, at the very beginning, a way of monitoring the behaviour of the system and to alert them to deviations from the expected behaviour. This type of observation enables issues to be dealt with early and small rather than large and expensive, or public. It also gives insight into making future enhancements and adjustments to the system as use patterns show unexpected threats and desired features.

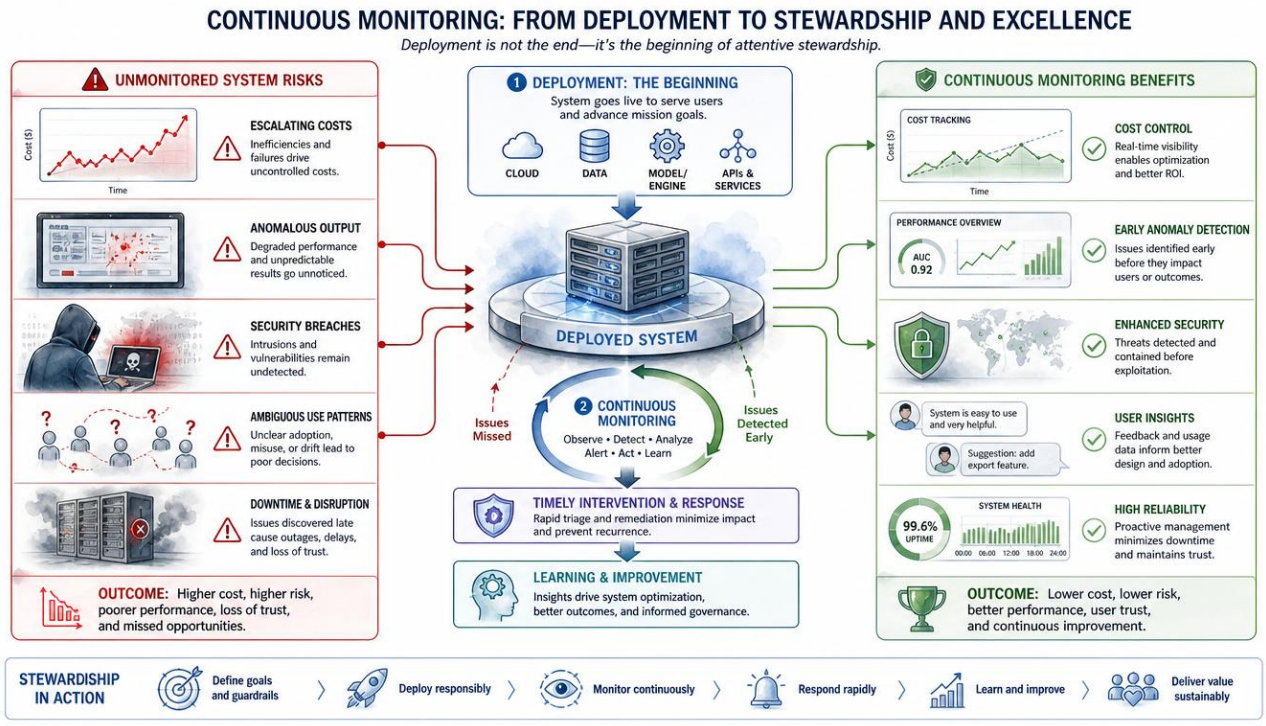

Fig -11: Continuous Monitoring From Deployment To Stewardship And Excellence

Monitoring is not an act of suspicion towards users but is an act of stewardship towards the system and those it serves. An organisation that observes its system attentively learns from it continuously and acts timely, an organisation that turns its back on it learns only when its consequences are unavoidable. The general rule is, a system given responsibility is given responsibility to be attended to, and that as much care should be taken of it after it is deployed, as was taken before it was deployed. Deploying and forgetting means giving up control of an asset that during its working life can behave in ways that require a response.

14. CONCENTRATING AUTHORITY WITHOUT APPROPRIATE OVERSIGHT

Once a generative system is given the power to do something to access data, trigger a process, or speak on behalf of the organisation, the issue of the extent of its powers becomes a pressing one. One common mistake is that such systems are given extensive capabilities for convenience sake without the management and oversight that extensive capabilities require. If a system can do much, it can do much harm if it is manipulated and or if it errors. The lesson is that only those authority levels that are necessary to the system's purpose can be given and actions that have consequences can undergo proper checks. A system which allows questions should not be allowed to change records, while one which allows changing records should only do that under well-defined limits, and, if the stakes are high, under human supervision. This principle of giving the minimum permissions possible has long been an important security rule for conventional systems and here it is of paramount importance for systems that interpret natural language and are thus redirectable through it.

Partners Universal Journal of Artificial Intelligence Research (PUJAIR)
Volume: 01 Issue: 02 | June-July 2026| www.pujair.com

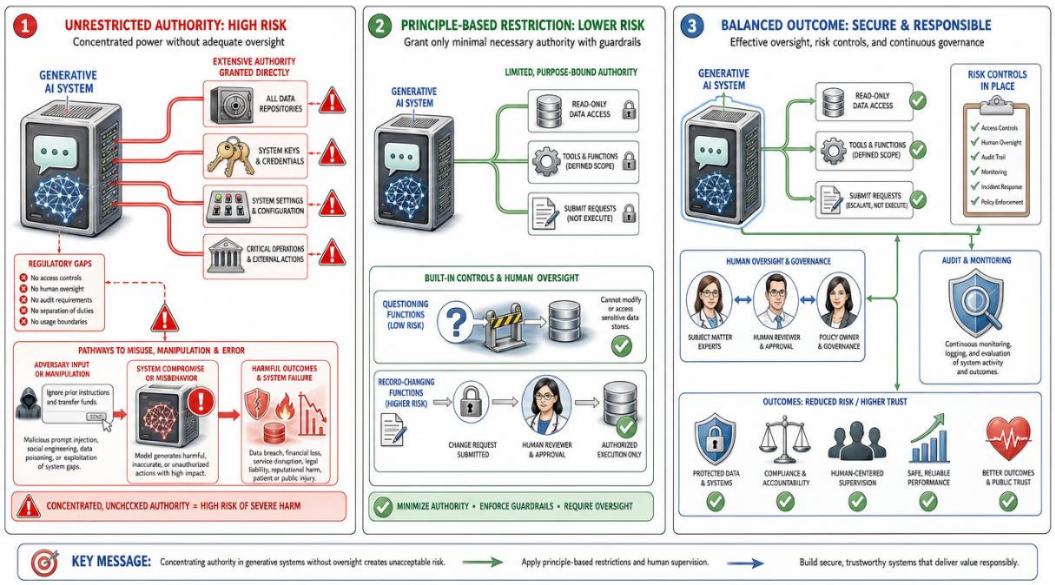

Fig -12: Concentrating Authority Without Appropriate Oversight

The danger is that the adversary can convince the system to misbehave and thus assumes the authority of the system. Restricting that power is therefore to restrict the damage that successful manipulation can do. In the absence of a system, those given the authority are likely to be over-trusting and make poor decisions on their own. The inconvenience given up is not great; the injury avoided might be great.

15. IGNORING THE NEED FOR HUMAN JUDGEMENT

A belief that is frequently engendered by enthusiasm for automation is that a generative system can be left to make decisions that should be made by humans.

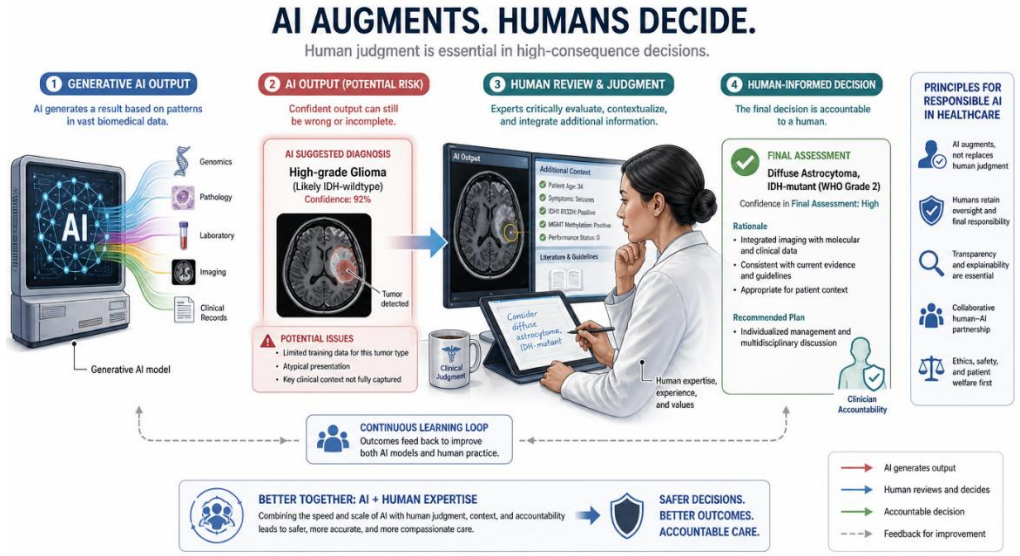

Fig -13: AI Augments Human Decide

The error is to take this person out of the system altogether and let the system decide on issues of consequence without detouring for review. Although these systems are very fluent, they don't know the world the way people do, and they can give an output that is sure and wrong. When there is a serious vulnerability such as one's circumstances, a safety issue, or an irreversible action then the lack of human judgement is serious. The lesson is to determine what decisions human need to make to input into them and to keep human input. The system can help, can make a draft, can suggest, but down to a person who can be held responsible and who can have the understanding the system does not have, must the responsibility for the consequences lie. That doesn't mean that automation is bad, it simply means that automation has its place. The most successful deployments have kept the man in the mix, where most needed, utilizing the system as a capable assistant, not an unsupervised decision-maker. In any case, those who automate willy-nilly, taking human judgement out of decisions where it's needed, hit the wall when the cost of going without is high. The better way is to limit giving over human judgment to human decisions, and to build systems that augment and enhance, not replace, the humans who are ultimately accountable for their results.

16. POSTPONING GOVERNANCE UNTIL AFTER ADOPTION

One of the common trends is to first implement a generative system and then think about governance. Organisations and those that regulate them have at times embraced these technologies because of the rate of change and the fear of being left behind, without necessarily defining the policies, responsibilities and boundaries that should be in place.

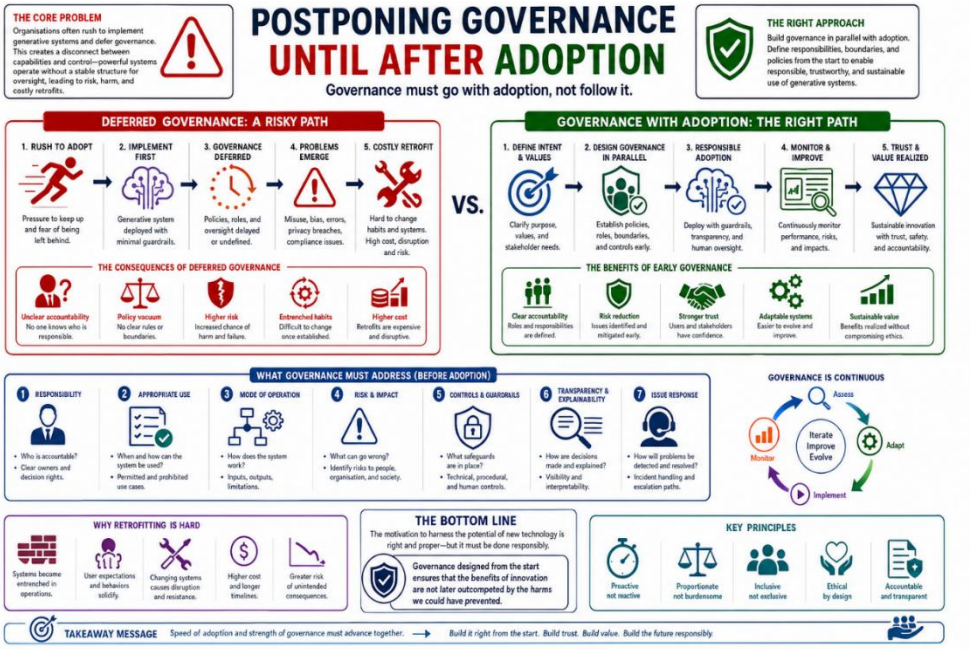

Fig -14: Postponing Governance Until After Adoption

Thus, there is disconnect between capabilities and control, with powerful systems functioning without a stable structure to provide oversight. This is one of the systemic shortcomings that is most obvious in practice and is common to everyone. The moral of the lesson is that governance must go with adoption

and not follow it. There needs to be clarity regarding who is responsible for a system, when and how it can be used, its mode of operation and how problems are to be dealt with before that system is handed responsibilities. The creation of this framework does not need to slow the speed of adoption; it can be done in parallel and can be used to inform adoption. The problem with deferred governance is that habits establish rapidly, expectations become entrenched and it is much more challenging to retrofit something on top of an existing system that is integrated into the operations than to do it from inception. Whether in organisations or in the broader society, those charged with leading the way for the use of these technologies do their constituents a disservice not to pay attention to governance issues early on, even when they feel compelled to do so quickly. The motivation to harness the potential of new technology is right and proper but should be done in a way to ensure that any harms that can and should be avoided are not later outcompeted by those benefits.

17. COMMUNICATING POORLY WITH THOSE AFFECTED

Those who use generative systems and whose information is processed by them, are too often left in the dark as to what is going on and how they are protected. The error would be to consider these questions as purely internal matters and of no concern to the public, in fact, they are of direct and legitimate concern to the public.

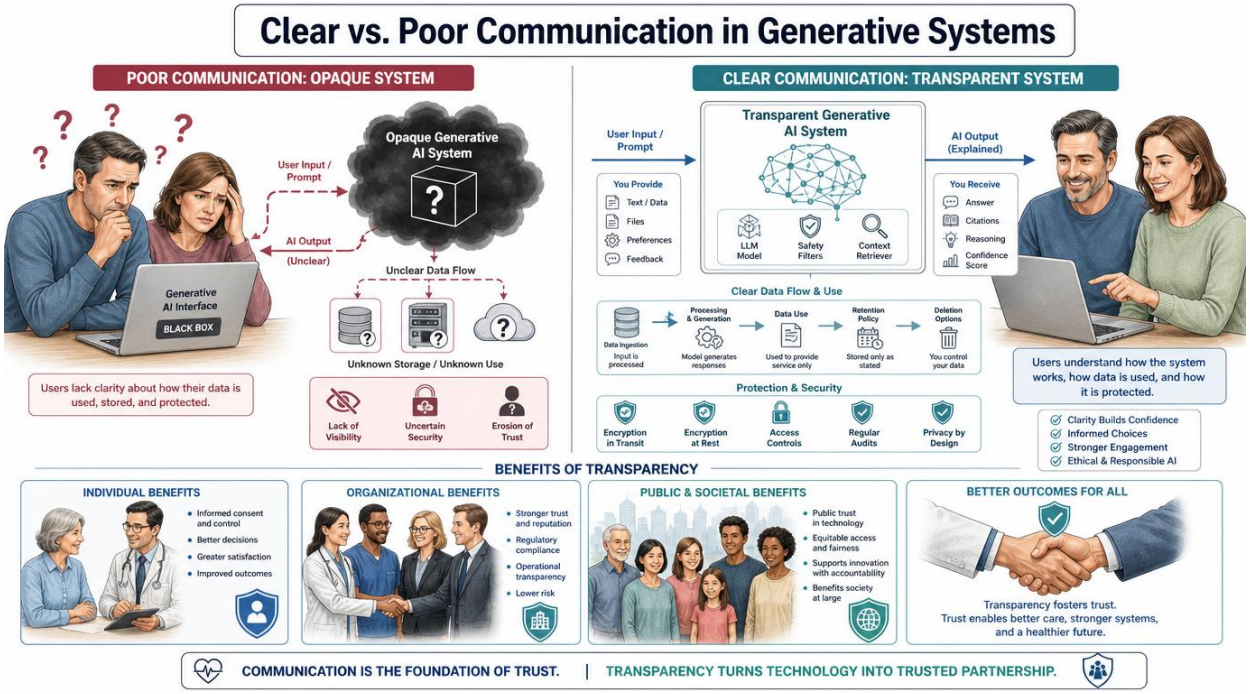

Fig -15: Clear Vs Poor Communication in Generative Systems

If they are unaware, they are dealing with an automated system, what happens to information they give, or what protections there are, they cannot make informed decisions and trust is lost when this becomes apparent. The message of the lesson Communicate openly and clearly. Individuals should be aware the service is automated, what it is used for and how their information is managed and protected. This openness is not expensive and can have great value in respecting the dignity of parties affected, in

establishing and maintaining trust upon which adoption relies, and in preventing the disappointment that occurs when adoption practices are discovered. Clear communication also benefits the organisation itself, a public that understands and trusts a system is more likely to interact with it and is more tolerant of its occasional failings. The general point is that technologies which impact many people are best used in the light. A climate of confidence, rather than apprehension, in which these technologies can be adopted, to the benefit of all concerned, is fostered by those who explain candidly what their systems do, and how the interests of the public are protected.

18. FAILING PLAN FOR FAILURE

Even the best designed systems fail, and it is a risky assumption to make that a generative system will work exactly as designed. But the error is to deploy without thinking about what to do if it goes wrong, if the system has an output that is harmful to others or if it succeeds with what it was designed to do, or if sensitive information is revealed, or if it costs more than it's worth.

Fig -16: Planning for Failure

An incident without a plan is an impromptu response at a time when it is most needed and it is most difficult to think clearly and act quickly. The moral of the lesson is that it's a good idea to plan for failure. It involves the capacity to recognise that something has gone amiss, the know-how to prevent further injury swiftly, maybe by suspending the system or restricting its use and the knowledge of who will respond and in what way. It also takes note of the lessons that can be learned from the incident, so that the same failure does not happen again. Planning for failure is not pessimism, it's realism, it is a mature understanding that complex systems can and do fail, but resilience isn't about the unattainable objective of failure prevention, it's about the attainable objective of good responses to failure. Organisations that plan for failure learn to deal with it quickly and with little tarnishing to their reputations, those that don't

end up in a bad way and improvise, often making things worse during the moment. The smart thing to do is to find out before a system is put in place what the potential problems are and what the organisation will do and to have the answers both before and after the event.

19. RELYING ON A SINGLE PROTECTIVE MEASURE

Finally, some of the mistakes are placing trust in one safeguard as if the entire spectrum of risks these systems pose can be remedied by one good safeguard. There are a few threats, such as rising costs, manipulation, theft, unreliable output and leakage of sensitive information, and no one protection measure covers all these threats.

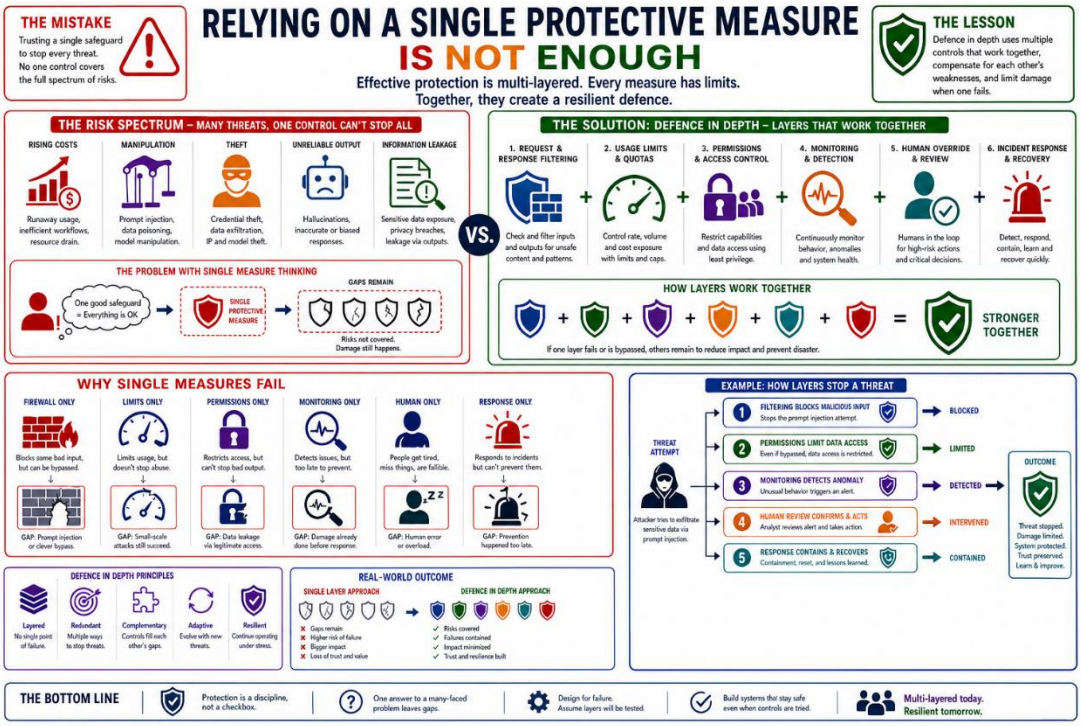

Fig -17: Relying on a Single Protective Measure is not Enough

The problem is to only make one protection and think everything is OK. The lesson is that effective protection is multi-layered and utilizes a series of protection measures working together to cover the entire risk spectrum and filling in for each other's weaknesses. If one layer fails or if it is outflanked, other layers are still available to limit the damage. This defence in depth is one of the more primitive and secure in the world of security and it is completely executable for protection of generative systems. A protective firewall that checks the requests and responses is useful, but only as part of a wider strategy of usage limits, tightly controlled system permissions, active monitoring, maintaining a human override, and a willingness to act in the event of failure. Every measure tackles part of the risk put together they create a defence much stronger than any individual measure. People who know protection is a discipline design systems that are protected even if some protective measures are tried and tested. People who want one answer to a many faced issue always end up with a part of the risk not covered and that's where the damage occurs.

20. RECOGNISING THE LIMITS OF THE PROTECTIVE LAYER

Too much to state that a protective layer for generative systems is a "full" or "foolproof" solution would be misleading. But as is true of all safeguards, there are limitations and trade-offs to be found in these as well, and a knowledge of these is part of using them wisely.

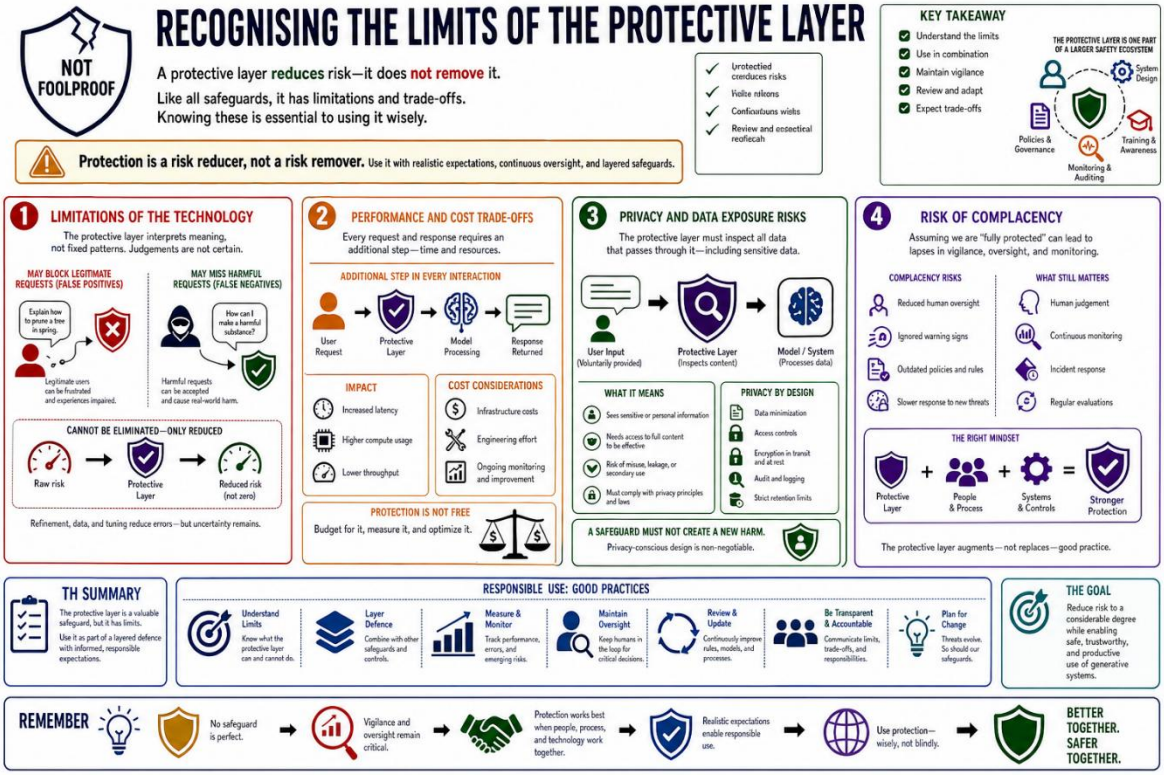

Fig -18: Recognising the Limits of the Protective Layer

The first is that it's a limitation of the technology itself. Since the meaning of natural language needs to be interpreted by the protective layer and not matched to fixed patterns, the judgements of the protective layer are not certain. Sometimes it will prevent legitimate requests from happening which will irritate the legitimate users and at other times it will accept harmful requests that it doesn't recognise. This doesn't come out of the system with any amount of refinement, it can only be reduced. Those using such protection, therefore, should consider it to be a way of reducing, not removing, risk to a considerable degree.

The second factor that needs to be considered is performance. Reviewing each request and response requires an additional step in each interaction, potentially causing delays and utilizing additional resources. The shield against over cost is not free either, or this must be fairly considered instead of simply "taken away".

A third, less obvious worry is that by default, a protective layer will have to look at all the data that filters through it, including the information that users voluntarily give. This safeguarding inspection also deals with sensitive information and hence must be carried out in as privacy conscious a way as it is also expected to promote. A protective measure which creates a new hazard would be counterproductive.

Then there's the danger of complacency because of protection. An organisation that thinks that it is fully protected can be complacent with vigilance, monitoring, and human oversight, which are still critical. The protective layer is just one of a whole field of work and is most effective when it is augmenting the care received somewhere else. It is important to be aware of these limits without being a reason to not use it is a condition of responsible use with realistic expectations.

21. RESEARCH AND KNOWLEDGE GAP

While there is increased awareness of these risks, it is fair to say that use has not yet been able to keep up with the pace of adoption. The biggest disconnect is between the pace of generative system rollout and the maturity of generative system safeguards. Many systems have protections that are incomplete, ad hoc, and or nonexistent, not because they are neglected, but because the field is still new and there are no standards in place yet.

Another gap is related to a shared understanding. The mechanisms of protection are still being formulated and a framework generally accepted as to what protection is, how to test it, and how to measure its effectiveness is not yet agreed upon. In the absence of such standards, organisations must make their own decisions as to the approach, which has not been uniform. Future work could be beneficial in establishing common benchmarks, against which protective measures can be evaluated and compared.

Also, in the governance and oversight which envelopes these technologies at the level of institutions and society there is still a lack. The guidelines that govern the use of powerful technologies have been built up over time and tend to be reactive to the experience gained which is still in the process of being gained with generative systems. While the issue of balancing the encouragement of good innovation with the prevention of harm is not completely answered, it is one that ought to be carefully and continually addressed.

Lastly, there's a lack of awareness. Those who are most impacted by such systems are also most likely to be unaware of how they work and how they are safeguarded. If the advantages of these technologies are to be realised with large-scale trust and acceptance, then it is crucial to bridge this gap through clear communication and accessible explanation which this article aims to provide in a small way.

22. CONCLUSION

Generative language models have arrived to help organisations and people with convenience, productivity and capability that were previously unattainable. But how quickly these systems have come has exposed a disconnect between what can be achieved and what should be in place to protect them. It's that gap protective layer that looks at what is being requested from the systems and what they are creating, that's what the large language model firewall has been emerging as. Such protection can only be justified based on an understanding of the risks. A system that was meant to help can be turned to things that it was not designed to do and utilize resources and increase costs. A system that's been given access to sensitive information can be made to give it up and the repercussions can be devastating for those involved. A useful feature or function can be stolen, and architecture on which an application might be depending on for critical tasks can go down and impact the application. All these risks are not abstract but are logical consequences of the characteristics of technology. None of the protective

measures mentioned in this document is perfect, but they take steps towards protecting them substantially and responsibly.

An underlying theme in the lessons learned throughout this article is. They advise modesty in regard to the newness of these systems, clarity about what they are designed to do, restraint in the authority they are given, sensitivity to their conduct and sincerity towards those it impacts. They caution against complacency with old protections, the lack of faith in any single protection, and the delay of governance until after problems have occurred. But they're not counsels of caution for the sake of being cautious, they're counsels of the kind of caution that enable a powerful new technology to be embraced with confidence instead of apprehension. Moving forward will require efforts on many fronts simultaneously, the maturing of protective practice, the development of common standards, of thoughtful governance and of public understanding. This effort rests on the shoulders of many those who design and create these systems, those who deploy them, those who manage them, and the society in which they operate. The promise of these technologies can be realised and attendant risks kept in check if this responsibility is taken seriously. The object is not to oppose the future nor to hurry headlong, but to go in with the precaution that any technology of this magnitude demands.

REFERENCES

[1] Atkinson, C. (2018). Hybrid warfare and societal resilience: Implications for democratic governance. Information & Security: An International Journal, 39(1), 63-76. https://doi.org/10.11610/isij.3906
[2] Burch, R. (2019). The future of resilient space system design. Resilient Space Systems Design: An Introduction. https://doi.org/10.1201/9780429053603-8
[3] Giacalone, M. (2026). Discursive behavior of generative language models in geopolitical and humanitarian contexts. Discover Artificial Intelligence, 6(1). https://doi.org/10.1007/s44163-026-00965-2
[4] Huang, T., You, L., Cai, N., & Huang, T. (2024). Large language model firewall for AIGC protection with intelligent detection policy. 2024 2nd International Conference On Mobile Internet, Cloud Computing and Information Security (MICCIS). https://doi.org/10.1109/miccis63508.2024.00047
[5] Mazur, K. (2026). Teaching LLM security through containerized vulnerable applications. Proceedings of the 18th International Conference on Computer Supported Education. https://doi.org/10.5220/0014950400004021
[6] Schwarz, D. J. (2025). Countermind: A multi-layered security architecture for large language models. ArXiv, abs/2510.11837. https://doi.org/10.36227/techrxiv.175994550.08962082/v1
[7] Shojaei, A., Asadi, S., Ardakani, M. K., & Safaei, M. (2026). Llms integration in recommender systems: A comprehensive survey of frameworks, taxonomies and applications. IEEE Access, 1-1. https://doi.org/10.1109/access.2026.3681933
[8] (2019). Personal information and data protection. Protecting Personal Information. https://doi.org/10.5040/9781509924882.ch-002
[9] Biswas, D. (2025). Stateful monitoring and responsible deployment of AI agents. Proceedings of the 17th International Conference on Agents and Artificial Intelligence. https://doi.org/10.5220/0013160300003890
[10] Campos-Castillo, C. (2010). Trust in technology. Trust and Technology in a Ubiquitous Modern Environment. https://doi.org/10.4018/978-1-61520-901-9.ch009
[11] Dasgupta, S., Cohn, T., & Baldwin, T. (2023). Cost-effective distillation of large language models. Findings of the Association for Computational Linguistics: ACL 2023. https://doi.org/10.18653/v1/2023.findings-acl.463
[12] Friston, M. (2023). The language of costs. Friston on Costs. https://doi.org/10.1093/law/9780192869081.003.0002

[13] Goryunov, E., & Didenko, A. (2025). Equitable AI: Aligning human expertise with algorithmic value. Decoding AI. https://doi.org/10.4324/9781003606956-2
[14] Grosse, K., & Ebert, N. (2026). Prevalence of security and privacy risk-inducing usage of ai-based conversational agents. IEEE Access, 1-1. https://doi.org/10.1109/access.2026.3714993
[15] George, D. (2026c). Security Service Edge (SSE) and SASE: A complete guide to Cloud-Native Zero Trust architecture for enterprise security. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.19974566
[16] Huang, T., You, L., Cai, N., & Huang, T. (2024). Large language model firewall for AIGC protection with intelligent detection policy. 2024 2nd International Conference On Mobile Internet, Cloud Computing and Information Security (MICCIS). https://doi.org/10.1109/miccis63508.2024.00047
[17] John D. Howard, & Thomas A Longstaff (1998). A common language for computer security incidents. https://doi.org/10.2172/751004
[18] George, D. (2026b). Multi-Vendor firewall strategy: IT, OT, and edge networks. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.19630402
[19] McGraw, G., Bonett, R., Figueroa, H., & McMahon, K. (2024). 23 security risks in black-box large language model foundation models. Computer, 57(4), 160-164. https://doi.org/10.1109/mc.2024.3363250
[20] George, D. (2026a). IEC 62443 Wireless Security: Deploying OT wireless controllers in industrial factory networks. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.19428491
[21] McTear, M., & Ashurkina, M. (2024). Conversational AI platforms. Transforming Conversational AI. https://doi.org/10.1007/979-8-8688-0110-5_7
[22] George, D. (2025b). Sanchar Saathi Digital Security versus Civil Liberty in India 's Smartphone Era. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.17838468
[23] Ojha, R., & Er. Siddharth (2024). Conversational AI and llms for real-time troubleshooting and decision support in asset management. Journal of Quantum Science and Technology, 1(4). https://doi.org/10.63345/jqst.v1i4.147
[24] George, D. (2025a). An exploratory study of friendship marriage and its role in redefining partnership for economic security and personal autonomy in modern society. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.17137271
[25] Thontadarya, P. S. (2025). Conversational AI for peoplesoft HCM: Elevating public sector HR. European Modern Studies Journal, 9(4), 359-368. https://doi.org/10.59573/emsj.9(4).2025.36
[26] George, D. (2026d). Data centers: critical infrastructure, global risk & geopolitical power. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.20443237
[27] George, D., Dr.S.Sagayarajan, Baskar, D., & Pandey, D. (2024). Assessing the security and privacy implications of India's DigiYatra initiative. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.14599297
[28] Sharma, D. a. K. (2026). AI-Powered Cybersecurity for OT and the IoT: A review of convergence, risk, and resilience in industrial environments. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.20156819
[29] George, D., Dr.T.Baskar, Srikaanth, P. B., & Dr.M.M.Karthikeyan. (2025). Building resilient API security through a Five-Dimensional Framework for data breach prevention in modern digital ecosystems. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.15862111
[30] Wang, W. (2024). Jailbreaking llms by suppressing rejection. 4th International Conference on AI ML, Data Science and Robotics. https://doi.org/10.51219/urforum.2024.prof-wenjie-wang
[31] George, D., Dr.T.Baskar, & Srikaanth, P. B. (2026). Beyond the perimeter Rethinking enterprise network security in an age of distributed threats. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.20329717
[32] (2003). Accounting: The language of budgeting. School District Budgeting, 67-102. https://doi.org/10.5040/9798216408161.ch-003
[33] George, D., Dr.T.Baskar, & Srikaanth, P. B. (2025). Bridging the Security Skills Gap: A comprehensive framework for developing application security competencies in modern software engineering. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.15616416
[34] (2021). Risk management of PPP model in domestic environmental protection project. Foreign Language Science and Technology Journal Database Engineering Technology. https://doi.org/10.47939/et.v2i8.230

[35] George, D., Dr.T.Baskar, & Dr.M.M.Karthikeyan. (2026). Cloud Security Architecture: A comprehensive guide to zero trust, governance, and operational resilience. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.19551592

[36] (2022). Research on risk management model of computer information system integration project. Foreign Language Science and Technology Journal Database Engineering Technology. https://doi.org/10.47939/et.v3i2.465

[37] (2024). Securing AI model weights: Preventing theft and misuse of frontier models. https://doi.org/10.7249/rra2849-1

[38] (2024). Improving response quality in conversational AI: The role of smart answer encoding in retrieval augmented generation. CONFERENCE PROCEEDING. https://doi.org/10.62919/hydg7253

[39] (2024). The five forces of customer 360. Customer 360, 23-30. https://doi.org/10.1002/9781394308668.ch4

[40] (2024). Job creation and local economic development 2024. OECD. https://doi.org/10.1787/83325127-en

[41] (2025). Supplemental material for trustworthy enough? examining trustworthiness assessments of large language model-based medical agents. Technology, Mind, and Behavior. https://doi.org/10.1037/tmb0000164.supp

[42] D. J. K. (2024). Review on solar charged batteries with inbuilt reverse current protection. International Journal For Multidisciplinary Research, 6(4). https://doi.org/10.36948/ijfmr.2024.v06i04.26640

[43] Ali, O., Shrestha, A., Chatfield, A., & Murray, P. (2020). Assessing information security risks in the cloud: A case study of australian local government authorities. Gov. Inf. Q., 37. https://doi.org/10.1016/j.giq.2019.101419

[44] Aminee, A. (2026). The agentic role of generative artificial intelligence in information systems leadership. The International Journal of Technology, Knowledge, and Society. https://doi.org/10.18848/1832-3669/cgp/a1033

[45] Bansal, A., & Suddala, S. (2024). Enhancing generative AI capabilities through retrieval-augmented generation systems and llms. Library Progress International, 44(03), 17776-17787. https://doi.org/10.36893/libpro.2024.v44n3.17776

[46] Bhatkar, P. B. (2025). Enhancing resilience posture in banking security through generative AI: Predictive, proactive, and adaptive strategies. European Journal of Computer Science and Information Technology, 13(2), 43-50. https://doi.org/10.37745/ejcsit.2013/vol13n24350

[47] Colter, J., Kinnison, M., Henderson, A., & Harbour, S. (2023). Adversarial attacks and defense on an aircraft classification model using a generative adversarial network. 2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC). https://doi.org/10.1109/dasc58513.2023.10311217

[48] Dallas, G., & Lubrano, M. (2022). Effective stewardship in practice: Monitoring, engaging, voting and reporting. Governance, Stewardship and Sustainability. https://doi.org/10.4324/9781003307082-6

[49] Gupta, N. (2025). Security risks of generative AI in financial systems: A comprehensive review. World Journal of Information Systems, 1(3), 17-24. https://doi.org/10.17013/wjis.v1i3.16

[50] Harasymchuk, O., Partyka, A., Nyemkova, E., & Sovyn, Y. (2023). AN INTEGRATED APPROACH TO CYBERSECURITY AND CYBERCRIME INVESTIGATION OF CRITICAL INFRASTRUCTURE THROUGH a RANSOMWARE INCIDENT MONITORING SYSTEM. Cybersecurity: Education, Science, Technique, 286-296. https://doi.org/10.28925/2663-4023.2023.21.286296

[51] Hollis, N. (2013). Clarity of purpose. The Meaningful Brand. https://doi.org/10.1007/978-1-137-36559-0_5

[52] Jaume, M. (2010). Security rules versus security properties. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-642-17714-9_17

[53] Leech, T. J. (2026). Defining board purpose. Mission-Critical Governance. https://doi.org/10.1201/9781003685753-13

[54] Leszczyna, R. (2019). Cost of cybersecurity management. Cybersecurity in the Electricity Sector. https://doi.org/10.1007/978-3-030-19538-0_5

[55] Malcher, M. (2019). Planning for roles and separation of duties. Oracle Cloud User Security. https://doi.org/10.1007/978-1-4842-5564-3_2

[56] Portegys, T. E. (2001). Goal-seeking behavior in a connectionist model. Artificial Intelligence Review, 16(3), 225-253. https://doi.org/10.1023/a:1011970925799

[57] Scrimshaw, N. L. (1984). Conditional statements, branching statements and more loops. At the Heart of the Mountain. https://doi.org/10.1007/978-1-4899-6795-4_18

[58] Swann, G. M. (2019). Economics of innovation. Economics as Anatomy. https://doi.org/10.4337/9781786434869.00017
[59] (2014). Asset building and livelihood rebuilding in post- disaster. Asset-Building Policies and Innovations in Asia. https://doi.org/10.4324/9781315739700-25
[60] Edelmann, A., Stumper, S., Kronstorfer, R., & Petzoldt, T. (2020). Effects of user instruction on acceptance and trust in automated driving. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). https://doi.org/10.1109/itsc45102.2020.9294511
[61] Matulionyte, R. (2023). Government automation, transparency and trade secrets. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4492382
[62] Sai, S., Gaur, A., Sai, R., Chamola, V., Guizani, M., & Rodrigues, J. (2024). Generative AI for transformative healthcare: A comprehensive study of emerging models, applications, case studies, and limitations. IEEE Access, 12, 31078-31106. https://doi.org/10.1109/ACCESS.2024.3367715
[63] Satchell, P. (2018). Current automation consequences. Innovation and Automation. https://doi.org/10.4324/9780429449888-9
[64] Teixeira, T. T., & Gatsios, R. C. (2026). Adoption of corporate governance practices by startups in the northeast. REVISTA AMBIENTE CONTÁBIL - Universidade Federal do Rio Grande do Norte - ISSN 2176-9036, 18(1). https://doi.org/10.21680/2176-9036.2026v18n1id42534
[65] (2013). Indonesian decentralization: Accountability deferred. Public Sector Reform in Developing and Transitional Countries. https://doi.org/10.4324/9780203722220-9
[66] (2024). Facing the inevitable failure of complex systems, embrace resilience. Forefront Group. https://doi.org/10.1377/forefront.20241216.100951
[67] Alberton, M. (2013). Section IV: Conclusions chapter 11 current and future trends in biodiversity and ecological connectivity protection in multi-layered systems. Toward the Protection of Biodiversity and Ecological Connectivity in Multi-Layered Systems. https://doi.org/10.5771/9783845250199-295
[68] Boyer, B. (2019). International safeguards inspections. Nuclear Safeguards, Security, and Nonproliferation. https://doi.org/10.1016/b978-0-12-803271-8.00009-6
[69] Casey, P. (2018). Models, risks, and protections (DRAFT). Oxford University Press. https://doi.org/10.1093/med/9780198786214.003.0004
[70] Farrow, S., & von Winterfeldt, D. (2020). Retrospective benefit–cost analysis of security-enhancing and cost-saving technologies. Journal of Benefit-Cost Analysis, 11(3), 479-500. https://doi.org/10.1017/bca.2020.24
[71] Franklin, D., & Michailova, S. (2024). Enhancing research impact through clear and accessible communication. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4922869
[72] Garcia, M. L. (2008). Physical protection system design. Design and Evaluation of Physical Protection Systems. https://doi.org/10.1016/b978-0-08-055428-0.50009-9
[73] King, M., Li, X., Doupe, B., & Mellish, J. (1988). Thermal protective performance of single-layer and multiple-layer fabrics exposed to electrical flashovers. Performance of Protective Clothing: Second Symposium. https://doi.org/10.1520/stp26275s
[74] Lefevre, M., Skidmore, P., & Firmin, C. (2023). Counting children and chip shops: Dilemmas and challenges in evaluating the impact of contextual safeguarding. Contextual Safeguarding. https://doi.org/10.1332/policypress/9781447366423.003.0012
[75] Seaman, J. (2021). Developing your human firewall. Protective Security. https://doi.org/10.1007/978-1-4842-6908-4_12
[76] Tahir, M., & Mazumder, S. (2008). Delay constrained optimal resource utilization of wireless networks for distributed control systems. IEEE Communications Letters, 12(4), 289-291. https://doi.org/10.1109/lcomm.2008.071975
[77] (2000). Transborder data flow contracts in the wider framework of mechanisms for privacy protection on global networks. OECD Digital Economy Papers. https://doi.org/10.1787/233311170363
[78] (2004). Benchmarks, bottlenecks, and best practices. Making Common Sense Common Practice. https://doi.org/10.1016/b978-0-7506-7821-6.50005-0
[79] (2008). Analyzing security risks. Security Software Development. https://doi.org/10.1201/9781420063813-10
[80] (2019). Understanding false positives. https://doi.org/10.4135/9781529689174
[81] (2019). Reactive instruments of social governance. https://doi.org/10.1628/978-3-16-157655-3
[82] (2025). Generative ai systems as a target of cyber threats. Generative AI Security, 171-240. https://doi.org/10.1002/9781394368532.ch5