ChatGPT

Gray Swan AI Comprehensive Analysis

Executive Summary

Gray Swan AI is an emerging **AI security and risk management platform** focused on safeguarding the deployment of advanced AI models. Founded by Carnegie Mellon researchers in 2023, the company addresses the critical problem of **unsafe or unpredictable AI behavior** in enterprise and government settings [1] [2]. Gray Swan's business model combines a **product-driven SaaS platform** (offering API-based tools and enterprise licenses) with specialized **AI safety consulting and red-teaming services**. Primary revenue streams include usage-based subscriptions for its AI "guardrail" platform (priced per million tokens) [3], on-premise software licensing for large clients [4], and contracts or partnerships with AI labs (e.g. OpenAI, Anthropic) and institutions like the UK AI Safety Institute [5]. Indirect revenue comes from **custom model safety evaluations and red-team exercises** offered to organizations on a consulting basis [6]. Gray Swan's **customer value proposition** centers on enabling organizations to **"deploy AI with confidence"** by mitigating risks such as prompt injection attacks, data leakage, policy violations, and harmful outputs [2] [7]. Its flagship tools – *Cygnal* (an AI input/output filtering layer) and *Shade* (an adversarial stress-testing suite) – provide a proactive defense and continuous monitoring around any AI model [8] [9]. These services are delivered via a cloud API (with simple integration into existing model endpoints) or through on-premise installations for sensitive environments [10] [4]. Key customer segments include **AI-forward enterprises** (e.g. in finance and healthcare) adopting generative AI, **AI vendors and labs** seeking to harden their models, and **government/national security agencies** concerned with AI safety and compliance [11] [12]. In the competitive landscape, Gray Swan faces a **growing field of AI safety and "LLM guardrail" providers**, such as Robust Intelligence, Protect AI, CalypsoAI, and HiddenLayer. All offer comparable AI-powered risk modeling or content-filtering platforms that automatically detect and block malicious or unsafe AI behaviors [13] [14]. Gray Swan differentiates itself with cutting-edge research (e.g. its unique *"circuit breaker"* method that helped its model resist nearly all jailbreak attempts [15]) and a community-driven approach (running open **"red teaming" arenas** with cash bounties to uncover new exploits) [16] [17]. However, competitors may appeal to organizations via broader toolsets (e.g. full AI supply-chain security or model validation), deeper enterprise integration (some are now backed by major firms like Cisco and F5), or simply by existing relationships in certain regulated sectors. This report analyzes Gray Swan's revenue model, customer value proposition, and competitive positioning in detail. We then evaluate options for organizations seeking AI risk mitigation and provide recommendations on leveraging Gray Swan's solutions, with practical considerations for implementation. (All information is evidence-based and sourced; see Appendices for citations.)

Problem Analysis

The rapid adoption of generative AI and autonomous AI "agents" has created a **pressing need for practical AI safety solutions**. Organizations are integrating AI into critical workflows "in just about everything," but often **lack tools to understand what could go wrong** [18]. Unlike traditional software, AI systems can produce **unpredictable and potentially harmful outputs** even from seemingly benign user inputs [19] [20]. Notable failure modes include: **prompt injection attacks** (malicious inputs that hijack model behavior) [21], the production of disallowed or dangerous content (e.g. detailed instructions for illicit activities) [16] [22], inadvertent **data leaks or misuse of sensitive info**, and **hallucinations or biased outputs** that could mislead decision-makers [23]. These risks pose serious

security, legal, and reputational threats to enterprises. For example, an AI customer service bot could be tricked into unauthorized financial transactions, or a supply-chain AI agent might be manipulated to disrupt operations. **Traditional safeguards** – such as manual content filters or relying solely on a model's built-in alignment – have proven inadequate. Researchers found that by cleverly rephrasing or obfuscating requests, adversaries can consistently bypass base model safety filters across models from OpenAI, Anthropic, Google, etc. [24] [25]. High-profile "jailbreaks" (including Gray Swan's own discovery of a universal prompt that broke many models' defenses) demonstrated how **easily safety measures can be defeated** [26]. Furthermore, even well-aligned models occasionally output harmful or false information, especially under novel or "edge-case" scenarios [27] [28]. This creates a **"gray swan" scenario** in AI deployment: events that are rare or unforeseen (e.g. a bizarre prompt causing a rogue action) but entirely plausible, with outsized impact if they occur. The **core problem** is that organizations often *lack the expertise and continuous vigilance* to secure AI systems against a constantly evolving threat landscape [2]. AI safety and adversarial robustness are highly specialized fields, and internal teams may not even be aware of all potential failure modes. There is a growing recognition that deploying AI "safely" requires more than just good intentions or basic content moderation – it demands **systematic risk assessment, stress-testing, and robust guardrails** at the model interface [6] [29]. In summary, as enterprises race to harness AI, they face the challenge of doing so responsibly: ensuring **AI models do not go rogue, leak confidential data, or violate laws and ethics**. Gray Swan AI was founded to help companies *"actually follow through"* on safe AI deployment, by providing the tools and expertise to assess and mitigate these novel risks (instead of merely trusting that AI labs have solved them) [30] [31].

Findings

Revenue Model and Business Model

Gray Swan AI employs a **hybrid revenue model** that combines subscription-based product offerings with bespoke services:

- **SaaS Subscription (Usage-Based):** The primary revenue comes from its *Cygnal* platform as a cloud service. **Clients pay based on usage (per token processed)**, similar to API pricing in the AI industry. Gray Swan publicly lists a tiered pricing structure – for instance, after a free tier of 50 million tokens, pricing starts at **$1 per million tokens** for up to 1 billion tokens, then drops to $0.70/M and $0.60/M at higher volumes [3]. This usage-based model aligns with customers' AI activity and makes adoption scalable. The **"first 50M tokens free"** indicates a product-led growth tactic: lowering the barrier for new users to try the security layer, then monetizing as they scale. Revenues from this stream are recurring and grow with the client's AI usage. Gray Swan's platform is offered as a **self-serve cloud API** (with a web dashboard and API keys for easy integration) [10], reflecting a **product-led business model** targeting developers and enterprises who can integrate safety features with minimal friction (just "change one URL" in their code to route through Cygnal's protective layer) [32]. This hints that Gray Swan's go-to-market involves attracting technical users (ML engineers, platform teams) who embed the service into AI applications.

- **Enterprise Licensing & On-Prem Deployments:** For large organizations with strict data or latency requirements, Gray Swan offers on-premise or private cloud deployments under enterprise agreements [4]. This represents a **higher-touch sales motion** typical of B2B enterprise software. Such deals likely involve an upfront or annual license fee and possibly custom integration support. The site explicitly states that on-prem "provides the ultimate in security…keeping data within your infrastructure" and encourages contacting sales for this option [4]. This suggests Gray Swan pursues **enterprise sales** for big clients (e.g., banks, defense agencies) who need custom solutions – a more traditional business model alongside its self-service SaaS. Pricing for these is not public; it would be negotiated case-by-case (the site invites contact for "higher volumes, on-premise deployments, and other specialized use cases" beyond the published token pricing) [33]. We can infer that these arrangements provide a significant revenue stream and customer lock-in via annual contracts.

- **Consulting and Custom Services:** In addition to products, Gray Swan generates revenue through **professional services** like private red-teaming events, custom model evaluations, and policy customization. The company's CEO notes that while public crowdsourced benchmarks are useful, many clients also require **"paid private evaluations"** and contracted red teams for thorough security testing [6]. Gray Swan offers a *"Private Red-Teaming"* service (as indicated on its website menu) where organizations can have their AI systems stress-tested in a controlled environment [34] [35]. This likely involves Gray Swan's team (and possibly vetted external hackers) conducting scenario-based attacks on the client's model and then reporting vulnerabilities. Revenue here would be project-based or retainer-based consulting fees. Additionally, Gray Swan provides customization such as tuning its filters to a client's specific policies or training a custom "Cygnet" variant – the site mentions **policy customization and model-specific training** available, with pricing depending on model size and infrastructure for fine-tuning [36]. Such services blur into the product offering (enhancing Cygnal for the client), but are essentially **value-add consulting** driving indirect revenue (and boosting customer retention). They position Gray Swan not just as a tool vendor but a **"white-glove" partner** for AI risk management.

- **Partnerships and Indirect Streams:** Gray Swan has also secured **notable contracts/ partnerships with major AI labs** very early in its life. Forbes reported the startup gained traction by working with **OpenAI and Anthropic** – presumably to help test or "bulletproof" those companies' models – as well as with the UK's AI Safety Institute [5]. While details are confidential, these likely took the form of **paid research or pilot contracts** (e.g., OpenAI using Gray Swan's Shade tool to audit a new model [37], or Anthropic hiring Gray Swan to improve Claude's defenses). Such relationships not only provide revenue but also serve as validation and R&D collaboration. In the government realm, Gray Swan's partnership with the UK AI Safety Institute (a public body) and involvement in U.S. AI safety initiatives hint at potential **government grants or contracts** in the future [38] [39]. We have not found public records of defense contracts as of 2025. However, Gray Swan's domain (AI security) is of high interest to defense and intelligence sectors, so SBIR grants or pilot programs could emerge (similar AI security startups have taken this route [40]).

- **Funding vs Revenue:** It's worth noting that Gray Swan is a **venture-backed startup** in its early stage. It raised **$5.5 million in seed funding** from a "nontraditional investor" and others [41], and is reportedly preparing for a Series A. This means current operations might still be funded by investor capital as the company acquires customers. Gray Swan's **business model** is thus growth-focused: proving its technology and capturing market share (especially given big competitors and consolidation in this space). Revenue figures are not disclosed (the company is private), but the presence of paying clients (OpenAI, Anthropic, etc.) suggests initial enterprise revenue. Gray Swan's strategy combines **product-led growth** (free trials, usage-based pricing to draw users) with **enterprise sales/partnerships** (to land marquee clients for larger deals). It is effectively a **hybrid model**: part cybersecurity SaaS company and part expert consulting firm. This hybrid approach is common in nascent fields – the product provides scalable income, while custom services help fund development and deepen customer relationships.

In summary, Gray Swan monetizes by selling its **AI safety platform** on a subscription basis and by **directly assisting organizations** in securing AI deployments. Its model is to embed itself as a critical layer in AI applications (driving recurring usage fees), while offering enough flexibility (on-prem, custom tuning, expert services) to satisfy enterprise demands. Over time, one can expect primary revenues to shift toward the SaaS side as the product matures, but currently both streams (product and services) are crucial. **Secondary revenue channels** like data or model licensing are not prominent at this stage – for instance, Gray Swan's proprietary hardened model (*Cygnet*) is used as a showcase and internal component, not sold as a standalone model license (no evidence of separate licensing). The company's open-source releases (e.g. the AgentHarm benchmark for LLM agents [42]) are likely marketing and community-building rather than direct revenue. We did not find any indication of Gray Swan selling customer data or analytics – in fact, their site explicitly notes they "do not and will never sell user data" [43], aligning with a security-forward business ethos. Thus, the **revenue model is straightforward:** charge for protective software and for expertise in using it. This aligns with Gray Swan's mission to be "the safety and security provider for the AI era" [44] – analogous to how a cybersecurity firm sells both a security appliance and consulting, Gray Swan provides AI "firewalls" and the know-how to deploy them effectively.

Customer Value Proposition

Gray Swan AI's value proposition lies in **making AI deployments safe, reliable, and compliant** for organizations. For enterprise customers and institutions, the company promises to **reduce the risk** that AI systems will cause harm – whether through technical failure or malicious exploitation – thereby protecting the customer's business and reputation. Key elements of this value proposition include:

- **Comprehensive AI Risk Mitigation:** Gray Swan offers tools to identify and neutralize a wide spectrum of AI failure modes. Unlike basic content filters, its solutions address **both sides of the AI interaction – inputs and outputs** – to create a bi-directional safety net [8]. For example, *Cygnal* will **block malicious inputs** (like cleverly crafted prompts containing hidden instructions or code that would hijack the model) and **filter any harmful or sensitive content** in the model's responses [45]. This two-way filtration is crucial because it stops attacks at entry and prevents unsafe outcomes from reaching end-users. Gray Swan continuously updates these defenses; Cygnal's filtering is **"continually adaptive"**, training on real-world violations to improve over time [46]. The result for customers is significantly **lower exposure to AI-related incidents** – Gray Swan cites an attack block rate of 99.98% in public evaluations, turning an otherwise vulnerable model into "by far the most-robust" system under test [47]. In practical terms, an organization using Gray Swan can trust that known attack techniques (prompt injections, toxicity prompts, etc.) will be caught in real-time, **avoiding costly scenarios** like data breaches, generation of fraudulent content, or PR disasters from offensive AI outputs.

- **Advanced Risk Assessment & "Red Team" Simulation:** Through its *Shade* platform, Gray Swan delivers **continuous AI security analysis** for the client's models [9]. Shade acts as an automated red-team: it **probes the AI with thousands of adversarial scenarios** to discover vulnerabilities and edge cases before real attackers or users do. For instance, Shade was used to **stress-test OpenAI's experimental "o1" model, finding weaknesses under worst-case conditions** [37]. The **value to customers** is proactive insight – they learn **"how your AI will stand up under the toughest conditions"** [48]. If, say, a financial chatbot is susceptible to a certain style of role-play prompt that extracts confidential data, Shade will reveal that. Armed with these findings, Gray Swan (and the client's team) can implement targeted fixes or policies (often via Cygnal's filters or model tweaks). Essentially, **Shade turns unknown risks into known, manageable ones**, giving enterprises confidence to deploy AI in critical roles. This addresses a major pain point: organizations often *don't know* what they should be worried about with AI. Gray Swan fills that gap with continuous security R&D on the client's behalf, **"staying ahead of**

the changing threat landscape" and feeding the latest protective measures into its products [49].

- **"Bulletproofing" AI Systems for Safe Deployment:** Gray Swan's entire suite collectively **hardens AI models against exploitation**, a fact recognized by industry leaders. Forbes describes Gray Swan as *"building powerful tools to mitigate risks in rapidly evolving AI landscapes"* [50] and notes it is **"leading the charge in bulletproofing AI models"** for companies like OpenAI and Anthropic [39]. For customers, this translates to a unique selling point: **access to the same expertise that top AI labs trust**. Gray Swan's founders literally helped uncover the "mother of all jailbreaks" and then built solutions to close those holes [24] [51]. By engaging Gray Swan, an enterprise essentially outsources a difficult task (AI security) to specialized experts who are at the cutting edge of the field. The value-add is not just in technology but in **peace of mind and speed** – customers can deploy AI faster and in more sensitive applications because Gray Swan has their back on safety. This can unlock use cases that were previously deemed too risky. For example, a healthcare provider might have feared using an LLM to answer patient questions due to risk of unsafe advice or privacy leaks; with rigorous safeguards from Gray Swan (including policy enforcement, output validation, and continuous monitoring), they can proceed knowing the AI's behavior is constrained within acceptable bounds.

- **Features & Services Offered:** Gray Swan's platform provides a **range of tools and delivery mechanisms** to maximize value for different user needs:

- *Cygnal:* A **secure AI middleware** that integrates via API between the user's application and the AI model [32]. It is model-agnostic and **"universally compatible"** – meaning clients can use it with OpenAI, Anthropic, open-source models, etc. by simply pointing their API calls to Cygnal [10]. This ease of integration (just adding Gray Swan's API key and endpoint) is valuable for engineering teams, as it **requires minimal code changes** to get protection. Cygnal operates with low latency overhead and supports streaming, function calling, and other modern LLM features [52] [53], ensuring it doesn't break functionality. The **precision of Cygnal's filtering** is a selling point: it reportedly outperforms big tech's own guards (Microsoft, OpenAI, etc.) with higher true positive rates and fewer false positives [54] [55]. In practice, this means customers get robust moderation without degrading user experience (less "Sorry, I can't answer that" for harmless queries). Cygnal can also be **customized to an organization's policies** – for instance, a bank might define "hateful content" differently than a social media company, and Cygnal can be tuned accordingly [36]. Delivery: via a **cloud API or on-prem appliance**, giving flexibility for different security postures.
- *Shade:* An **AI security evaluation suite** that clients use (or have Gray Swan use on their behalf) to **regularly test their AI models** [9]. It incorporates the latest adversarial techniques from research, essentially providing **"AI pen-testing as a service."** Delivered through a web interface or reports, Shade's output includes detailed findings on how the model can fail and recommendations. Gray Swan may run Shade continuously in the background ("continuous red-teaming" [56]), meaning the client is alerted to new vulnerabilities as threat tactics evolve. This is delivered as a platform feature; value-wise, it's like having an automated watchtower scanning for weaknesses 24/7, something most organizations couldn't maintain internally.
- *Arena (Community Red Teaming):* Gray Swan hosts the **Gray Swan Arena**, a platform where external participants (security researchers, enthusiasts) attempt to jailbreak or exploit models in structured challenges [57] [58]. For customers, Gray Swan can leverage this **crowdsourced security testing** in two ways. First, public competitions (with anonymized models or dummy scenarios) yield general insights that improve Gray Swan's products for all. Second, Gray Swan hints at **private arenas** for clients [59] – i.e. a company could invite Gray Swan to organize a focused red-team contest on their specific model (under NDA), harnessing the "wisdom of

crowds" safely. Either way, the customer benefits from a **broad range of attack perspectives** that no single internal team could replicate. Volunteers are motivated by learning and prizes [6], and Gray Swan curates their findings into actionable intelligence for the model owner. Essentially, Gray Swan offers **access to an entire community of AI hackers** in addition to its in-house experts – a strong value proposition for staying ahead of threats.

- *Cygnet (Secure Model):* While not a direct service to customers, Gray Swan's development of *Cygnet* (its own aligned language model) demonstrates its capabilities and potentially offers a **reference safe model**. Cygnet employs innovative *"circuit breakers"* that act as tripwires in the model's reasoning process to halt it when it veers into harmful content [15]. At Gray Swan's jailbreaking competition, Cygnet "largely withstood all attempts" whereas many popular models were broken [60]. For a customer, this showcases Gray Swan's thought leadership and may evolve into a product – e.g. offering Cygnet as a **safer alternative model** for certain use cases. Even if customers continue using third-party models, the circuit-breaker concept is incorporated into Cygnal's filtering, giving them an edge. Advisors have noted this approach as a promising direction in AI safety (Elon Musk's xAI was said to be interested in using such circuit breakers) [61]. The net value: clients get state-of-the-art defenses that aren't available elsewhere (since Gray Swan's methods are novel and proprietary).

- **Key Customer Segments and Use Cases:** Gray Swan primarily targets **enterprise and institutional clients** that deploy AI in mission-critical or sensitive contexts:

- **Large Enterprises in Regulated or High-Risk Industries:** Sectors like **finance, healthcare, and legal** have strong use cases for AI (automation of customer support, data analysis, decision support) but also severe consequences for AI errors. For example, a bank's chatbot giving fraudulent advice or leaking PII, or a medical AI offering dangerous health recommendations, could be catastrophic. Gray Swan's value for these clients is enabling them to use AI **while staying compliant with regulations (e.g., GDPR, HIPAA) and internal policies**. Its tools can enforce custom policy rules (like no financial disclosure beyond X, no medical advice without disclaimer, etc.) and provide auditable logs of AI behavior [62] [63]. We saw Gray Swan's research specifically test AI agents in domains like finance and healthcare, where agents "performed high-risk actions" when attacked [11] – highlighting the need in those fields. By adopting Gray Swan, a finance firm can avoid unauthorized trades or data leaks caused by a compromised AI agent, and a healthcare AI provider can guard against malpractice due to AI hallucinations. **Logistics and supply chain** companies (which the user asked about) similarly could use AI for planning and communication; Gray Swan would ensure those AI systems can't be tricked into chaos (e.g., a competitor injecting a prompt to disrupt an AI-driven logistics scheduler). While we found no direct case study on logistics, the general principle of **operational continuity and safety** applies.
- **Technology Companies and AI Vendors:** This includes AI model providers (OpenAI, Anthropic, etc.) and platforms integrating AI. These customers may use Gray Swan in two ways: (1) **internally**, to test and improve their models pre-release (as Anthropic and OpenAI have done in partnering with Gray Swan [5]), and (2) **embedded in their products**, to offer end-users a safer experience. For instance, a SaaS that includes an AI assistant could integrate Cygnal to filter the assistant's outputs, adding a layer of trust. Gray Swan's value here is both improving the AI product's quality (fewer bad outputs) and protecting the brand from incidents. Since Gray Swan's founders are respected AI safety researchers (one co-founder, Zico Kolter, even sits on OpenAI's safety board [64]), these tech clients gain access to top-tier expertise *without* hiring a large in-house safety team.
- **Government and National Security:** Government agencies deploying AI (for intelligence analysis, public services, etc.) have unique security concerns. Gray Swan's emphasis on

robustness against “well-resourced attackers” and worst-case scenarios is highly relevant here [65]. National security use cases might include using LLMs to analyze open-source intelligence or to interface with the public; any vulnerability (like prompt exploits) could be leveraged by adversaries or lead to misinformation. Gray Swan’s tools can enforce strict policies (e.g., prevent an AI from ever revealing sensitive methods or sources) and detect if someone is trying to manipulate an AI into, say, disclosing classified info. The partnership with the UK AI Safety Institute and involvement of US AI Safety bodies in Gray Swan’s challenges [38] indicates **governments see value in Gray Swan’s approach**. Additionally, defense contractors or military AI programs could use Gray Swan to validate the security of AI models before deployment in the field (ensuring, for example, that an autonomous AI won’t be easily tricked by enemy input signals – analogous to prompt injection in a different modality). By adopting Gray Swan, government users can **stay ahead of AI threats** that could impact national security, benefiting from the collective research Gray Swan does (the company stays “at the forefront of new developments” in AI safety) [66].

- **Mid-Size Companies and Startups Using AI:** Not to be overlooked, Gray Swan also markets to startups and smaller companies (“anyone else who needs to deploy AI with confidence” [67]). These organizations often lack specialized AI risk teams. Gray Swan provides them an **affordable safety net** (via the pay-as-you-go model) so they can adopt powerful AI models without inadvertently “shooting themselves in the foot.” For example, a startup building an AI content generator can integrate Gray Swan to avoid producing defamatory or biased content that could lead to lawsuits. The self-service nature (API and docs) means even lean teams can implement it quickly. Essentially, Gray Swan **levels the playing field** by giving smaller players access to advanced AI safety tech that only big tech would otherwise possess.

In all, Gray Swan’s customer value proposition can be summed up as: **“We handle the AI risks, so you can focus on the AI rewards.”** It adds value by **reducing fear and uncertainty** around deploying AI. Organizations get to exploit cutting-edge AI capabilities while **minimizing downside risk** – be it legal risk, security breaches, ethical lapses, or brand damage. By continuously updating its defenses (drawing from both its internal research and community findings), Gray Swan assures customers that they are protected against even newly discovered exploits, something an in-house solution would struggle to maintain. The convenience (drop-in API, ready-to-go model scanning) and credibility (backed by real metrics and expert founders) further strengthen this value proposition. In an environment where **“everyone racing to adopt AI is claiming to be doing so safely”**, Gray Swan is the partner that helps companies **actually achieve AI safety** in practice [30] [68].

Competitive Landscape

The landscape for AI-powered strategic forecasting and risk mitigation platforms – particularly those focusing on AI model safety – has grown crowded as AI adoption surges. Gray Swan AI faces **direct competition from several companies offering comparable tools for AI risk modeling, scenario testing, and guardrail enforcement**. Below we identify Gray Swan’s primary competitors and analyze differentiators, including why some organizations might choose alternatives despite Gray Swan’s strengths:

- **Robust Intelligence:** A well-established player in AI model security, recently acquired by Cisco, Robust Intelligence provides an end-to-end platform for **detecting model vulnerabilities and deploying guardrails** in production [69]. Its “AI Firewall” offers real-time protection similar to Gray Swan’s Cygnal, aiming to shield AI applications from attacks and undesired outputs [13]. Robust Intelligence has invested nearly a decade in proprietary techniques (algorithmic red-teaming, threat intelligence pipelines) to automatically generate failure examples and update its detections [70]. One differentiator is its broad **coverage of hundreds of security/safety**

categories and tight integration into enterprise workflows – it sells not just a filter, but a comprehensive **AI risk management platform** including model validation (pre-deployment testing) and ongoing monitoring [71] [72]. They also market solutions by industry (finance, insurance, government, etc.) and use-case (LLM chatbots, RAG apps) [12], which may appeal to clients wanting a vendor with domain-specific knowledge. **Why choose Robust Intelligence?** Some organizations might prefer an **all-in-one solution backed by a larger company (Cisco)** for reliability and support. If a client already uses Cisco or is a very large enterprise, Robust's integration and long track record might inspire confidence. Additionally, Robust Intelligence addresses not only prompt security but issues like **data drift, model performance monitoring, and supply-chain risks** (given its ML pipeline focus). For a customer who wants a **single platform for all ML governance (bias, robustness, security)**, Robust's broader scope could outweigh Gray Swan's narrower focus on prompt/security threats. However, Robust Intelligence's offerings may come at higher cost and less agility compared to Gray Swan. Gray Swan's nimbleness and research edge in LLM-specific exploits is a counter-advantage.

- **Protect AI:** Protect AI is another competitor, positioning itself as a **"security platform for artificial intelligence systems"** that helps organizations identify, monitor, and mitigate AI security risks [73]. It has a suite of tools like *Guardian* (for scanning models for malware/ backdoors) and *Recon* (for automated red-teaming of generative AI) [74] [75]. Protect AI's strategy is to secure the **entire AI lifecycle**: from checking model integrity before deployment (e.g., verifying that an open-source model hasn't been tampered with) [76], to simulating attack scenarios in runtime, to guiding the configuration of **cloud-specific guardrails** (they integrate with AWS Bedrock's guardrail features, for instance) [77] [78]. The company has made inroads in government as well – for example, partnering with Leidos to secure US government AI systems [79]. **Differentiators:** Protect AI provides strong **model supply-chain security** (scanning models, checking for hidden vulnerabilities before you even use them) which Gray Swan currently does not emphasize. For organizations worried about things like Trojaned models or model provenance, this could be critical. Protect AI also works closely with cloud providers (AWS) so if a client's infrastructure is heavily on AWS, Protect's solution might integrate more seamlessly (leveraging AWS's own guardrail frameworks alongside Protect's tools) [78]. Protect AI essentially combines **DevSecOps for AI** (code and model scanning) with **runtime protection**, whereas Gray Swan is mostly runtime protection and attack simulation. Customers might choose Protect AI if they need that **holistic approach** or if they are in sectors like defense that value the model validation piece (indeed, Protect AI has positioned itself in federal markets). Additionally, Protect AI's emphasis on **weekly threat updates** and a library of hundreds of attack types [80] is similar to Gray Swan, but being an older firm it might have a larger knowledge base for non-LLM models (computer vision, etc.). On the flip side, Gray Swan's specialization in LLM adversarial tactics and its community red-teaming could yield faster discovery of cutting-edge LLM exploits. Protect AI tends to work closely with large partners and might require a more involved deployment (not self-serve), which could be a factor for smaller customers who then lean toward Gray Swan's simpler model.

- **CalypsoAI:** CalypsoAI is a competitor particularly active in the **government and defense space**. They offer a GenAI security and enablement solution that provides **testing, monitoring, and real-time defense for LLMs**, quite analogous to Gray Swan's feature set [81]. CalypsoAI highlights *Independent Model Validation* and an ability to **"test and evaluate AI/ML models"** for agencies [82]. A key differentiator is Calypso's framing around **Responsible AI compliance** – ensuring AI systems are **safe, compliant, and trustworthy** through every stage of deployment [14]. They integrate with model pipelines to enforce governance, and list capabilities such as **agentic red teaming (simulating malicious prompts), real-time guardrails (blocking unsafe outputs), continuous monitoring (risk scoring), and model-agnostic deployment** [83] [84].

This sounds very much like Gray Swan's promises. CalypsoAI was recently set to be acquired by F5 (a major enterprise security firm) [85] [86], indicating its traction. **Why a customer might choose CalypsoAI:** If an organization, especially a government agency or contractor, prioritizes a proven track record in **compliance and governance**, CalypsoAI's experience since 2019 and its enterprise-grade governance features (policy enforcement, audit logs) [62] [87] could be attractive. They emphasize alignment with regulatory requirements and responsible AI standards [88]. In sectors with strict oversight (finance, government), having a solution explicitly built for **auditability and policy control** might tip the scales. CalypsoAI also offers **drift monitoring and bias testing** as part of its platform [89] [90], extending beyond just security into overall model risk management (something Gray Swan hasn't publicly covered). This one-stop-shop for **Trustworthy AI (safety + bias + explainability)** could appeal to organizations looking to satisfy all aspects of AI governance with one vendor. Moreover, being acquired by F5 suggests future deep integration with enterprise security stacks (for instance, F5 could bundle AI guardrails into its application firewalls). On the downside, CalypsoAI's government focus might mean it's less accessible to smaller private companies, and it might require more customization to fit outside environments. Gray Swan's advantage would be more agility and perhaps better performance specifically on stopping prompt-based exploits (given Calypso also does similar, the difference might come down to metrics or cost).

- **HiddenLayer:** HiddenLayer is an AI security company that takes an approach akin to cybersecurity's EDR (endpoint detection & response) but for ML. It **"protects against the full spectrum of AI attacks"** with protections rooted in frameworks like MITRE ATLAS and OWASP's Top 10 for LLMs [91]. HiddenLayer's platform includes a **Model Scanner** (for malware/backdoor detection in models) and an **AI intrusion detection system (AIDR)** for real-time monitoring of model behavior and detecting threats/anomalies [92] [93]. Essentially, HiddenLayer focuses on **ML-specific threats** including data poisoning, model theft, adversarial inputs, and it provides tools for incident response when an AI attack is detected. **Differentiators:** HiddenLayer's emphasis on **threat detection and response** (rather than just prevention) might attract organizations who want the security operations integration – e.g., alerts that feed into a SOC (Security Operations Center). They talk about **SOC integration, real-time alerts, and behavioral analytics** for AI systems [94], suggesting a product that can sit in an enterprise's security dashboard. Companies with mature security teams might prefer this approach to complement their existing defense (it's akin to having an AI-specific alarm system, whereas Gray Swan is more of a preventive shield). HiddenLayer also addresses **model IP theft and Trojan detection**, which is outside Gray Swan's current scope. Why choose HiddenLayer? If a client is very concerned about stealthy threats or wants continuous **AI threat monitoring** with the ability to investigate incidents, HiddenLayer's platform might be more suitable. Also, HiddenLayer frames itself as "the only" platform that protects ML models in a **security-first way** across training and deployment [95] (though that is marketing speak, it suggests a comprehensive approach). It's essentially an **ML security operations platform**, versus Gray Swan as an AI safety product + service. Organizations might also pick HiddenLayer if they are not only dealing with LLMs but also other types of models (CV, time-series) – HiddenLayer covers a broad range, whereas Gray Swan is very LLM/agent-focused at present. However, Gray Swan's direct LLM specialization could make it more effective specifically for language model guardrails (for example, Gray Swan's 99.98% block rate in a jailbreak test might outperform a generalist tool) [47]. Additionally, HiddenLayer's approach might generate more false positives or require security expertise to interpret alerts, while Gray Swan aims to **prevent issues automatically** without burdening the client's staff.

- **Others / Alternative Approaches:** Aside from these startups, Gray Swan competes indirectly with **big tech's in-house solutions** and open-source efforts:

- *Cloud Provider Guardrails:* Companies using Azure OpenAI, AWS Bedrock, or Google's PaLM API get some level of built-in content filtering and safety features. For some, these **default guardrails (which are essentially content moderation APIs)** might seem "good enough," especially as they improve. However, research shows these built-in filters have **widely varying effectiveness and common failure cases** – often either over-blocking innocuous inputs or letting through cleverly crafted malicious prompts [96] [28]. Gray Swan's offering is usually **more precise and robust** than generic guardrails, but a cost-conscious or less risk-sensitive customer might stick with what their model provider gives for free. The trade-off is the higher risk of successful exploits (which Gray Swan would mitigate).
- *In-House Development:* An organization could attempt to build its own AI safety layer – using open-source libraries (like OpenAI's guardrails code or custom regex/policy filters, etc.), and conduct internal red team exercises. Large tech firms like OpenAI and Anthropic themselves have internal red teams, as do some banks and big enterprises. The **decision to not use Gray Swan** might come if a company believes they can internally manage AI risk or see it as a core competence. However, given the specialized nature of adversarial AI security, it's challenging to match the breadth of Gray Swan's continuously updated attack knowledge and defenses. Still, some companies will prefer an internal solution for control or data privacy reasons (though Gray Swan counters the latter by offering on-prem deployment).

- *Traditional Consulting Firms:* For strategic forecasting and risk scenario planning (in a broader sense), companies sometimes turn to consulting firms (McKinsey, BCG) or defense think tanks. While those are not direct "AI platforms," a company might allocate budget to human analysts and scenario planners instead of a tool like Gray Swan if the perceived need is more conceptual scenario generation (e.g., geopolitical risk forecasting) rather than technical AI safety. If a customer mistakenly thought Gray Swan provided general geopolitical forecasting, they might compare it to firms like **Predata/Recorded Future** (for geopolitical risk intelligence) or **Palantir** (for scenario modeling on data). In reality, Gray Swan is solving a different problem – it forecasts *AI failure scenarios*, not world events. But it's worth noting that **some potential buyers might conflate "Gray Swan" with the idea of rare event forecasting**, and in that domain there are established tools and consultancies. Clarity in marketing is needed so Gray Swan isn't mismatched against those.

- **Competitor Comparison Summary:** All direct competitors (Robust Intelligence, Protect AI, CalypsoAI, HiddenLayer) share a common goal with Gray Swan: **make AI deployments safer through automated tools**. They each combine features like model stress-testing, policy enforcement, and monitoring, but with different emphases:

- Gray Swan stands out for its **LLM specialization, academic pedigree, and community-driven approach** (open challenges). It likely leads in cutting-edge LLM exploit defense (e.g., circuit breakers concept, high block rate proven in public). It also has a transparent, developer-friendly pricing model.
- **Robust Intelligence** is a broader platform (covering many ML risks) and now part of a major corporation (Cisco), which might assure longevity and support. It might be chosen for enterprises wanting a mature, integrated solution (especially if they have non-LLM models too).
- **Protect AI** differentiates with model file security and integration with cloud AI services. Likely favored by those deep in AWS or who need to vet open-source models for hidden threats.
- **CalypsoAI** focuses on compliance and comprehensive risk management, appealing to government and finance where reporting and governance are as important as the tech. Acquisition by F5 could make it a default for existing F5 customers.

- **HiddenLayer** focuses on detection/response and full-spectrum ML attack defense, aligning with organizations that have strong security operation centers or unique ML attack concerns (like IP theft).

In terms of **pricing and client types**, Gray Swan is relatively accessible (public pricing, free trial, catering also to startups), whereas many competitors operate on enterprise contract models (you must contact sales for any pricing, indicating likely higher cost and focus on bigger deals). This could make Gray Swan more attractive to mid-market and fast-moving tech companies that want to self-serve a solution. Conversely, extremely large enterprises might feel more comfortable with a competitor if they equate higher price and longer track record with reliability.

It’s also important to note that this space is evolving fast – we’re seeing **consolidation** (M&A: RI->Cisco, Calypso->F5) and new entrants continuously. Gray Swan, being newer, has to prove its credibility against some firms that have been around since 2019 or so. The Forbes feature and contracts with top AI labs lend it credibility [50] [5], but a risk-averse customer might still wait to see more adoption. On the other hand, **the problem is so novel** that no solution is foolproof; early adopters might trial multiple tools. In fact, some customers might use Gray Swan **in conjunction** with others – for example, using Gray Swan for intensive LLM red-teaming and filtering, while using a competitor’s tool for model supply-chain checks or for non-LLM models.

Why users might choose a competitor despite Gray Swan: - If they require **features Gray Swan lacks** (e.g., scanning for data poisoning in training, or built-in bias checking). - If they desire a **one-vendor solution** integrated with their broader IT security (Cisco’s backing of Robust or F5’s of Calypso could sway CIOs who prefer established partners). - Perceived maturity and support: Gray Swan is a young startup; some might question its long-term support or scalability and opt for a company with more enterprise deployments. - **Cost considerations:** While Gray Swan’s usage-based pricing can be economical, at very large scale it might add significant cost (e.g., millions of dollars if processing billions of tokens monthly). An enterprise might negotiate a flat license with another vendor or even rely on free open-source guardrails if budget is tight, accepting lower security as a trade-off. - **Data locality and privacy:** Although Gray Swan offers on-prem, some competitors (especially those targeting government) may already have cleared environments or credentials (FedRAMP certifications, etc.) for handling sensitive data. If Gray Swan hasn’t yet navigated those compliance hoops, a government client could require a competitor that has. That said, Gray Swan’s on-prem solution addresses many privacy concerns by not sending data to a third-party cloud.

In conclusion, Gray Swan is among the **leading innovators** in AI safety platforms but must differentiate itself in a competitive field. Its strengths are technical excellence in LLM defense and agility; its challenges are convincing risk-averse clients to trust a newer solution and broadening its feature set to match the “checklist” that some rivals boast. The competitive landscape is likely to keep evolving, and Gray Swan’s ability to stay at the frontier of research (via Shade and Arena) is a key asset – it can incorporate the **latest attack discoveries faster**, potentially giving it an edge in effectiveness even if others have more enterprise polish. Organizations evaluating options should consider both the **quantitative performance** (which solution actually blocks more attacks or finds more issues, as evidenced in independent evaluations) and qualitative factors like integration effort and vendor stability. The next section will assess these options from the perspective of a customer deciding how best to secure their AI systems.

Options Assessment

Organizations seeking to mitigate AI-related risks have several options. Below we outline and assess the main approaches, including using Gray Swan AI's platform, choosing an alternative competitor, or relying on internal measures. The goal is to evaluate each option's pros, cons, and suitability:

Option 1: Adopt Gray Swan AI's Safety Platform (Cygnal + Shade)
Description: Integrate Gray Swan's Cygnal API as a security layer in front of your AI models and use its Shade suite for ongoing risk analysis. This can be done via Gray Swan's cloud service or deployed on-premise for sensitive data. Possibly engage Gray Swan's team for initial setup, policy tuning, and optional private red-teaming exercises.
Advantages: This option provides a **state-of-the-art, specialized guardrail** for AI. Gray Swan's solution has demonstrated extremely high effectiveness in blocking attacks (only 0.02% of adversarial attempts succeeded in tests) [47] , meaning it dramatically reduces the chance of a catastrophic AI output or breach. Integration is relatively **quick and flexible** – developers can get basic protection running within minutes by redirecting API calls through Cygnal [32] . Gray Swan's continuous updates ensure you benefit from the **latest research-driven defenses** without heavy lifting on your part [9] . Shade's automated testing will keep you informed of new vulnerabilities, essentially outsourcing the R&D of "how might our AI fail?" to experts. Compared to in-house efforts, Gray Swan brings an entire community and research lab's worth of knowledge. Another pro is **customizability**: Gray Swan can tailor filters to your organization's specific rules and even train custom safety models for you [36] , yielding a solution aligned to your domain (for example, stricter filters for a healthcare AI vs. more lenient for an internal coding assistant). The **cost model** (usage-based) can be efficient if your AI usage is moderate or grows over time, and the free tier allows initial experimentation at no expense. Importantly, adopting Gray Swan signals to stakeholders (regulators, customers) that you are taking robust steps for AI governance, potentially easing compliance and trust concerns.
Risks/Disadvantages: Being a newer platform, there is some **execution risk** – you rely on Gray Swan's continued viability and support. If the company were to falter or change pricing, you might need to adjust quickly (mitigated somewhat by on-prem option giving you a lasting deployment). Another consideration is **dependency and latency**: routing all AI calls through an external filter adds a component that must stay highly available and fast. Gray Swan's design emphasizes low latency overhead [52] , but it's an added piece in your architecture that needs monitoring. In on-prem mode, you take on running the infrastructure for it. Also, **coverage limitations**: Gray Swan is excellent for LLM prompt/response safety and known exploit patterns, but if your AI stack includes other elements (like computer vision models vulnerable to adversarial images, or data preprocessing risks), Gray Swan's tools won't cover those (at least currently). You might need supplementary measures for non-LLM models. **Cost** at very high scale could become significant – e.g., a billion tokens is $1000 under current pricing [3] ; if you're doing tens of billions, this scales linearly unless volume discounts or enterprise licensing kick in. For extremely large deployments, negotiating a custom plan or even considering a flat-license competitor might be prudent. Nonetheless, for most organizations, the cost of preventing a single major AI incident (which could be millions in liability or damage) far outweighs these fees.
Assessment: Option 1 is a strong choice if your organization directly faces the LLM-related risks Gray Swan is built to mitigate and you want a **quick, high-efficacy solution**. It's particularly well-suited if you lack a large internal AI safety team – Gray Swan effectively becomes your AI safety partner. It's also a good fit if you value being on the cutting edge of security (benefiting from Gray Swan's ongoing research and community findings). Ensure that you have buy-in from IT/security to incorporate a third-party component and plan for some initial configuration to align with your needs. Overall, adopting Gray Swan can significantly **lower risk with relatively low implementation effort**, making it an attractive option for many enterprise AI projects.

Option 2: Use a Competing AI Safety Platform

Instead of Gray Swan, an organization could choose one of the alternative vendors described (Robust Intelligence, Protect AI, CalypsoAI, HiddenLayer, etc.) to secure their AI systems. The exact benefits depend on the vendor, but generally:

Advantages: Many competitors offer **broader or more integrated feature sets**. For instance, if you also need to scan models for backdoors or monitor model drift, a platform like Protect AI or CalypsoAI might cover that alongside prompt security [97] [63]. Going with an established competitor could provide **enterprise-grade support, SLAs, and integration** – e.g., Robust Intelligence (Cisco) can integrate with existing Cisco security infrastructure; CalypsoAI (F5) may soon integrate with load balancers and app firewalls, etc. If your use cases include **multi-modal AI or a lot of regulatory documentation**, some competitors might have ready-made frameworks for compliance (audit logs, reporting dashboards) that ease adoption in strict environments [98] [99]. Also, certain industries have preferred vendors (for example, defense contractors might already work with CalypsoAI or have vetted it for classified settings). Choosing a competitor might align better with internal procurement preferences or existing partnerships. In some cases, pricing negotiation for a large enterprise license might be simpler (some competitors might offer a flat annual license, which, for heavy users, could be more predictable or even cheaper than token-based billing).

Risks/Disadvantages: A potential trade-off is **effectiveness and focus**. Gray Swan's laser focus on LLM adversarial safety might mean it has an edge in blocking the newest jailbreak techniques; a broader platform might not catch everything if it's less specialized (especially if the competitor hasn't matched Gray Swan's 99.98% block rate benchmark [47] – independent evaluations would be ideal to compare). Additionally, competitors that are larger companies may move slower in updating their tools; for example, Gray Swan's small team of top researchers might implement a newly published exploit fix within days, whereas a Cisco-owned product might have longer release cycles. Another factor is **complexity**: more features can mean more complexity in deployment. Some alternatives may require deeper integration into your ML pipeline or more configuration (e.g., HiddenLayer's system might involve connecting to your logging systems, training detectors, etc.). This could increase the implementation time and require specialized staff. **Resource requirements** might also be higher – an on-prem deployment of a robust competitor suite might need significant infrastructure. Finally, each competitor has a different **scope**: if your main worry is indeed prompt and output control for LLMs, some parts of a competitor's platform (like model supply chain checks) might be overkill or not directly addressing your immediate risk, yet you pay for them. There's also the risk of vendor lock-in or acquisition changes: with industry consolidation, a competitor's product might change or be integrated into something larger (e.g., CalypsoAI into F5's offerings) which might alter its focus or pricing.

Assessment: Option 2 is viable for organizations that have **very specific needs aligning with a competitor's strengths**. For example, if you absolutely need **AI model validation and security in one** (covering training data to deployment), a competitor might serve better. Or if your company already trusts a particular vendor in this space (perhaps from a pilot or prior project), sticking with them could reduce friction. It's wise, however, to **evaluate the effectiveness** of any option: consider running a trial or proof-of-concept where you test Gray Swan vs a competitor on your own AI model with known challenges. Some research (like the Palo Alto Networks study [96]) suggests guardrail effectiveness varies widely, so empirical results should guide the choice. In summary, using a competing platform can give you a more **comprehensive but possibly less specialized** safety net. It might involve more upfront work but could satisfy a broader set of security and compliance criteria in one package. Ensure that whichever competitor you consider has proven results on the specific threat vectors you care most about (be it prompt exploits, data leakage, etc.) and weigh the cost/complexity against Gray Swan's option.

Option 3: Rely on Model Providers' Built-in Safeguards

The simplest (and often default) path is to use whatever safety features come with your AI model/API. For instance, OpenAI's GPT-4 has a built-in content moderation system, Google's Vertex AI might have safety filters, and AWS Bedrock provides *Guardrails* that can be configured [77]. Under this option, you do not add any third-party layer; you trust the model vendor's alignments and filters, possibly with some minor rules of your own.

Advantages: This is a **zero-cost (no additional fees) and zero-integration-effort** approach. You leverage the fact that top AI providers have done some alignment – e.g., models like GPT-4 are trained to refuse certain requests and come with usage policies. Many platforms have basic content filters that catch blatant toxic or illegal content. You avoid the complexity of managing another tool and keep latency to a minimum (since you're not routing through an external service). If your AI usage is relatively low-stakes or experimental, this might be acceptable initially. It's also an approach some take to see how models behave in practice *before* deciding on more heavy-duty solutions. In short, it's the **status quo** – easy and with no direct monetary cost.

Risks/Disadvantages: Relying solely on built-in safeguards is broadly considered **insufficient for serious applications** [17] [96]. Vendors' filters have known blind spots; determined adversaries often find ways around them (public jailbreaks appear almost as fast as new models come out). For example, OpenAI's own CEO has noted that no prompt filter can catch everything. Our research confirms common evasion tactics can trick these guardrails (e.g., roleplay scenarios that mask malicious intent, as highlighted in the Unit42 study [100]). So the risk is that a malicious or even an unwitting user request produces a harmful output because the built-in filter missed it – leading to outcomes like disinformation generation, hate speech, or security breaches on your watch. Additionally, the model's internal alignment (RLHF) may reduce the frequency of bad outputs but cannot eliminate them, especially as new exploits are discovered [101]. If you choose this option, you are effectively **accepting a significant residual risk** of AI misbehavior. Another drawback is lack of customization: vendor filters are one-size-fits-all and might not align perfectly with your internal policies or regional regulations. They also might be too restrictive in some cases (hurting functionality) or not restrictive enough in others, and you typically cannot fine-tune them beyond on/off or choosing broad settings. There's also minimal **transparency** – you get little insight into what was blocked or why, making compliance reporting or incident analysis hard.

Assessment: For any application where an AI mistake could cause **material harm**, this option is usually **not recommended as a standalone strategy**. It might be tolerable for very low-risk scenarios (e.g., an internal tool with human oversight at every step, or a trivial use case where errors are inconsequential). Even then, as usage grows, the chance of a "gray swan" event (rare but impactful) increases, and it's prudent to add stronger safeguards. In essence, Option 3 is what many start with by default, but the track record of AI incidents suggests it's a matter of *when*, not *if*, an unmitigated model will produce something problematic. Organizations concerned about reputation, security, or liability should view this option as **insufficient** and plan to layer on additional protection (via Option 1 or 2) sooner rather than later.

Option 4: Build an In-House AI Safety Solution

Another path is to develop your own suite of safety measures tailored to your AI use cases. This could involve assembling open-source tools, writing custom filtering rules or machine learning detectors, and running internal red team exercises with your security or AI teams. Essentially, you attempt to replicate some of what Gray Swan or others offer, but internally.

Advantages: Building in-house gives you **full control and customization**. You can design filters exactly suited to your content policies and adjust quickly as your needs change, without relying on vendor timelines. Your data stays completely internal, which might ease privacy or secrecy concerns (especially for highly sensitive domains). Over time, you develop **internal expertise** in AI safety, which could be a strategic advantage if AI is core to your business (for example, a company like Google invests heavily in its own AI alignment research rather than using external tools). Cost-wise, this option avoids direct vendor fees; it leverages existing staff and open resources. If your team is very skilled, they might innovate novel defenses that give you a competitive edge. Also, any insights or tools remain your intellectual property.

Risks/Disadvantages: The primary downside is the **significant effort and expertise required**. AI safety is a fast-moving, specialized field – attracting and retaining talent who can do this (prompt researchers, adversarial ML experts) is difficult and expensive. There's an opportunity cost: those experts could be working on your core product instead of reinventing safety tools. Without broad exposure, an internal team might miss attack techniques that are being discovered elsewhere; external platforms benefit from many clients and global research, whereas an internal effort is limited to your scope. **Maintenance burden** is high – continuous monitoring of academic literature, threat forums, etc. is needed to update your safeguards (Gray Swan literally builds its business on doing this continuously [9]). You'll also have to build infrastructure for things like logging, analytics of AI outputs, perhaps your own challenge/response red teaming environment – all that could take months of engineering. Early in-house attempts might be crude (simple regex filters that can be easily bypassed, for instance). There is a real risk of **false confidence**: thinking your custom filter is working until an incident happens that it didn't catch. Moreover, internal red teams often lack the adversarial diversity of a broader community; they might not think of the quirky exploits that a crowd or specialized firm would. In sum, building in-house can be slow and may still result in weaker protection than a dedicated vendor solution, especially in the short term.

Assessment: Option 4 tends to make sense only for organizations that either **have extreme security constraints** (can't involve any third-party, e.g., intelligence agencies working with classified data and air-gapped systems) or those that are **very large with unique needs** and have the resources to invest in an internal "AI safety department". If you choose this path, it might still be wise to use some open-source or collaborative approaches – e.g., participating in community red-team events (like those Gray Swan runs publicly) to benchmark your defenses, or using open benchmarks (HELLOT, AdvBench, etc.) to test your model. For most companies, however, the in-house route will likely be **costlier and riskier** in terms of leaving gaps, especially given the pace of new exploits. It could potentially complement Option 1 or 2 (for example, you build some unique rules in-house while also using a vendor for the heavy lifting). But as a standalone, it means committing to an ongoing investment similar to having your own mini Gray Swan internally. Only pursue this if AI safety is a core competency you're willing to develop and if no external solution can meet specific requirements you have.

After evaluating these options, many organizations find that a **hybrid approach** can work well: for instance, **using a platform like Gray Swan (Option 1)** as the backbone of defense, while also **developing some in-house policies or fail-safes (Option 4)** to address any organization-specific concerns, and continuing to utilize the model provider's improvements (Option 3) as an additional layer. Options 1 and 2 are not mutually exclusive either – some might use Gray Swan primarily but also do a one-time assessment with another vendor's tool or vice versa, especially during trial phases. The key is to achieve defense-in-depth without unnecessary redundancy.

In the next section, we provide **recommendations** based on the above analysis, assuming the goal is to robustly secure AI deployments in a cost-effective and reliable manner.

Recommendations

Based on our analysis of Gray Swan AI's offering, the customer value it provides, and the alternatives available, we recommend the following course of action for organizations aiming to ensure safe and secure AI deployments:

1. Implement a Dedicated AI Safety Layer – Preferably Gray Swan AI or an Equivalent

Adopt an advanced AI security platform rather than relying on native model safeguards alone. The evidence is clear that built-in guardrails are not sufficient to catch all malicious or harmful AI behaviors [100]. We specifically recommend **engaging Gray Swan AI's platform as the primary safety layer** for organizations whose use of generative AI involves sensitive content, external user inputs, or any high-impact domain. Gray Swan has demonstrated industry-leading effectiveness in mitigating LLM threats (e.g., virtually eliminating successful jailbreaks in tests) [47]. By integrating Cygnal and using Shade, your organization will proactively plug most of the known failure modes and continuously learn about new ones relevant to your systems. This significantly lowers the risk of an AI-related incident, which in turn protects your brand, customers, and compliance posture. If for any reason Gray Swan is not viable (say, procurement constraints or niche needs), then an **equivalent competitor's solution** (such as Robust Intelligence or CalypsoAI) should be pursued with urgency. The key recommendation is **do not deploy critical AI systems without an external, continuously updated guardrail mechanism**. The cost of these solutions is justified by the potential prevention of one major mishap (which could save you from multi-million dollar liabilities or reputation damage). It's essentially purchasing "AI insurance" in the form of a technical control. Given Gray Swan's strong focus and early successes (including trust by top AI labs) [17], it is a prime candidate to fulfill this role. Start with their free trial to evaluate fit, and move to a paid plan or enterprise deployment swiftly if it meets expectations.

2. Leverage Gray Swan's Expertise and Services for a Holistic Approach
Subscribing to the platform is step one; we further recommend **making full use of Gray Swan's ancillary services** to maximize benefits: - **Engage in a Private Red-Teaming Exercise:** Early in the deployment, consider contracting Gray Swan to run a **targeted red-team simulation on your AI application** [6]. This will produce a bespoke report of vulnerabilities (beyond the automated Shade findings) and help calibrate Cygnal's filters to your context. It's an investment in understanding your specific risk profile. Gray Swan's team can bring external red-teamers (under NDA) to try to break your model in ways your team might not envision. The output will be actionable fixes and will bolster your confidence (or reveal gaps to address) before wide release. - **Customize Policies and Models as Needed:** Work with Gray Swan to **customize the safety policies** enforced by Cygnal to match your organization's requirements [36]. For example, if you are in healthcare, define what constitutes disallowed medical advice; if in finance, define what sensitive data must never be revealed or what fraudulent transaction patterns to guard against. Gray Swan offers to train custom versions of their guardrail model for your needs – take advantage of this if your domain has unique language or criteria. This ensures the AI guardrails are not generic but tailored, reducing false positives and aligning with your compliance needs. - **Tap into Arena (Community Insights):** Even if you don't run a private contest, keep an eye on Gray Swan's **Arena challenges and results**. They often publish insights (e.g. results of UK AISI × Gray Swan agent challenge [102]) that can inform your own threat models. If appropriate, encourage your internal AI engineers or security staff to participate in these public challenges. This not only helps the community but also builds your team's skills in understanding AI exploits. Gray Swan's community approach is a differentiator – as a customer, you gain indirect benefit from the collective testing by hundreds of hackers globally. We recommend maintaining communication with Gray Swan about major findings from Arena events and requesting briefings on how those might apply to your deployments. - **Use Gray Swan's Reports for Stakeholder Assurance:** The benchmarks and reports from Gray Swan (such as safety scores, block rates, etc.) [54] [55] can be used in internal and external communications to demonstrate due diligence. For instance, if regulators or clients inquire how you're keeping your AI outputs safe, you can cite that you are using a solution that achieved 99% precision on safety benchmarks [54] and is at the forefront of AI safety research [50]. We recommend integrating Gray Swan's metrics into your governance reports. This will reassure stakeholders that you've adopted best-in-class measures, and it provides a quantifiable way to track improvements over time.

3. Continue to Monitor the Competitive Landscape and Complementary Tools
While we advise choosing Gray Swan now, also **stay informed about other tools and evolving standards** in AI safety. The field is moving quickly; new techniques (or new threats) can emerge that might be better addressed by complementary solutions. For example, if using Gray Swan for LLM

security, you might later consider a **data-centric security tool** for your training data or a bias-detection module for fairness – possibly offered by another vendor or open-source. Our recommendation is to **periodically evaluate** (say, annually) your AI risk mitigation stack against the latest offerings. This could include running bake-off tests: e.g., test a sample of adversarial queries through both Gray Swan and a competitor like HiddenLayer or an open-source filter to see if anything slips through one and not the other. If, hypothetically, a competitor demonstrates significantly better performance on a new class of attacks, be ready to either push Gray Swan for improvements or consider augmenting/switching solutions. Keep an eye on industry benchmarks or independent studies (like the Unit42 report [96]) for a neutral view of how various guardrails perform. Additionally, track acquisitions: if, say, Cisco/Robust releases a major integrated product that might suit your infrastructure, you'd want to know. Essentially, **maintain vendor agility**. The recommendation is not to bounce between solutions frequently (consistency is valuable), but to ensure you're aware of what's out there. Gray Swan itself is likely to expand capabilities; engage with them on their roadmap (ask about upcoming features like perhaps bias checks or multi-model support). Aligning with a vendor who is forward-looking is wise, but be ready to bring in supplementary tools if needed. For example, you might end up using Gray Swan for prompt security *and* an internal tool for model watermarking – that's fine, as long as it covers your threat surface. In summary: choose a primary solution now (Gray Swan), but **don't set and forget** – treat AI security as an evolving program where you will integrate new defenses as threats evolve.

4. Develop Internal Protocols and Human Oversight to Backstop the Technical Controls
Even with Gray Swan's powerful filters in place, no system is 100% foolproof (there's that residual 0.02% in their own test) [103] . We recommend establishing **internal processes to complement the automated guardrails**: - **Incident Response Plan:** Create a clear procedure for what to do if an AI output does evade safeguards and causes a problem. For instance, if a user manages to get a disallowed response, how will it be reported, who analyzes it, and how quickly can you pull in Gray Swan (or your own team) to patch the gap? Gray Swan provides logging and monitoring; ensure those logs are monitored by your security team or integrated into your SIEM (Security Info/Event Management) system so alerts are not missed [94] . Basically, treat AI incidents like security incidents with defined roles and communication channels. - **Human-in-the-Loop for High-Stakes Outputs:** For extremely high-risk tasks (e.g., an AI system making financial transactions or medical diagnoses), keep a **human review layer** in addition to Gray Swan's guardrails. The guardrails will catch known issues, but a human should still sanity-check outputs that could be life-or-death or legally binding. Gray Swan reduces the noise (so humans aren't overwhelmed by trivial issues) by filtering out most nonsense, but humans can provide the final judgment on ambiguous cases. Over time, as confidence in the AI grows, the level of human oversight can be adjusted, but initially err on the side of caution. - **Policy and Training:** Update your organization's AI usage policies to reflect the new safety layer. For example, policy can state that "All generative AI deployments must route through Gray Swan Cygnal (or an approved equivalent) and must undergo a Shade security evaluation before launch." This ensures organizational compliance and sets a baseline standard. Also, **train your staff** (developers, prompt engineers, etc.) about the capabilities and limits of Gray Swan's system. They should understand what types of prompts it will block or transform, so they are not confused during development or testing. Encourage them to *attempt* some adversarial inputs in a sandbox to see Gray Swan in action – this will both build trust in the tool and potentially surface any tuning needed for false positives. - **Regular Reviews and Updates:** Schedule periodic meetings (perhaps quarterly) with Gray Swan's customer success or engineering liaison to review performance. Use these to discuss any false positives/negatives observed, new threats, and features. Gray Swan's team likely can advise on emerging risks they see across clients. This partnership approach means you're not just a customer but an active participant in maintaining AI safety. It also keeps Gray Swan accountable to your needs (if something isn't working ideally, they can adjust or suggest changes).

5. Plan for Scalability and Integration
In implementing Gray Swan (or any solution), consider the **scalability and integration** aspects early: - Ensure your DevOps/MLOps pipeline incorporates the Gray Swan layer from development to production (e.g., your CI/CD tests could include a step where some sample prompts are run through the AI+Cygnal to catch issues before deploy). By building it in from the start, you avoid it feeling like a bolt-on or afterthought. - Monitor the **token usage and costs** as your AI usage grows. Since Gray Swan's pricing is usage-based [3], you want to avoid any budget surprises. We recommend setting up billing alerts or usage dashboards. If you foresee a massive spike (say, scaling to a new user base), proactively talk to Gray Swan about a higher volume plan or enterprise license to optimize costs [33]. They indicated openness to custom deals for high volume or on-prem cases. - If you have global operations or latency-sensitive applications, plan how the safety layer will be deployed (multiple regional instances? on-premise near your data centers?). Gray Swan on-prem might be prudent if you need sub-50ms latencies or have data sovereignty laws. Include those considerations in your implementation strategy. Gray Swan's team can assist in architecture planning for this – engage them early if needed. - Lastly, keep an **eye on user experience**. A guardrail should ideally be unobtrusive to end-users. After integrating Gray Swan, gather feedback: Are legitimate user queries ever being blocked incorrectly? If yes, adjust the policies or whitelists in Cygnal (Gray Swan can help configure thresholds to minimize false positives [54] [55]). The recommendation is to fine-tune the balance between strictness and permissiveness to suit your audience. For example, an AI for creative writing might allow more edgy content (with just illegal stuff blocked) whereas an AI for customer support might need stricter tone enforcement. Gray Swan is capable of both; you need to set those dials appropriately.

In summary, our recommendations strongly urge adopting a robust third-party AI safety solution, with Gray Swan AI being a top choice given its focused strengths. By pairing the technical solution with internal processes and continuous engagement, your organization will be in a strong position to harness AI's benefits **safely and responsibly**. This multi-layered strategy – technology, process, and people – will significantly reduce the likelihood of AI causing unwelcome surprises ("gray swan" events) and put you ahead of the curve in AI governance. The next section addresses practical considerations for implementing these recommendations, to ensure a smooth integration of Gray Swan's tools into your workflows.

Implementation Considerations

Implementing Gray Swan AI's safety platform (or a similar solution) requires thoughtful planning to ensure it delivers maximum protection with minimal disruption. Here we outline key considerations and steps for a successful rollout:

A. Integration and Deployment Planning
- **Architecture Integration:** Decide how the Gray Swan filter (Cygnal) will sit in your system architecture. Common approaches include integrating at the API gateway level or directly in the application backend that calls the AI model. For a web app, for instance, your backend would send user prompts to `api.grayswan.ai/cygnal` instead of directly to the model API [104]. Ensure that whatever microservice or component currently handles model queries is amenable to this change. If using multiple AI models, you might route all through Gray Swan for consistency. Document the flow: *User input → Gray Swan (input filter) → Model → Gray Swan (output filter) → User output* . If an on-prem deployment is chosen, set up the Gray Swan server(s) in a secure network zone and low-latency proximity to your AI compute. Gray Swan has examples and documentation – leverage those to map it onto your stack [105] [106]. - **Latency and Throughput:** Evaluate the performance impact. Gray Swan's filtering is optimized in C++/Rust (as implied by their low false positive focus) and they tout minimal added latency [52]. Still, measure end-to-end latency in a staging environment. If you have streaming

responses, confirm that Gray Swan supports streaming without waiting for full completion (their docs say streaming is supported [52]). For high-throughput systems, you might need to scale Gray Swan instances horizontally or ensure the cloud endpoint can autoscale. Check with Gray Swan about any rate limits or recommended concurrency. It's wise to start with a pilot group of users to monitor performance, and then scale up. - **Security and Access:** Treat the Gray Swan API keys or on-prem credentials as sensitive, since they control access to your safety layer. Implement proper secret management. If cloud-based, use encrypted connections and perhaps a dedicated VPN or VPC endpoint if Gray Swan offers it for enterprise clients. Also, consider outbound network rules: your app should only communicate with known Gray Swan endpoints to prevent any bypass. Conversely, ensure Gray Swan can reach your model if it's behind a firewall (though usually it intercepts the call and then calls out to the model provider if using cloud models – basically it proxies the request). - **Fallback Strategy:** Plan for what happens if the Gray Swan service is temporarily unavailable (downtime or network issues). Ideally, have a fallback mode: the AI either refuses to answer (safe fail) or uses a backup basic filter. You might, for example, cache a small local profanity filter as a last resort to avoid completely raw output. Since Gray Swan's on-prem can be made highly available (multiple instances), and their cloud is likely robust, this is a low risk, but standard resilience design applies. Document this in your runbooks.

B. Configuration and Tuning

- **Policy Settings:** Work with Gray Swan's team to configure the filtering policy to your needs. Cygnal likely comes with default settings aligned to general norms (no hate, violence incitement, etc.). Review these defaults and adjust. For instance, define what constitutes sensitive data in your context – you might want to upload a list of your company's confidential project names so if the AI ever mentions them, Gray Swan flags it (they did mention customization by organization policies [36]). If operating in multiple languages or regions, ensure the filter covers those languages (Gray Swan being CMU-born likely has multilingual capabilities, but confirm). - **Thresholds for Blocking vs. Flagging:** Gray Swan filters can either block content outright or potentially just tag it. Determine how strict each category should be. It might be acceptable to allow slightly edgy content with a warning in an internal tool, whereas a customer-facing tool should block anything remotely off-color. During initial deployment, consider a **"monitor mode"**: let Gray Swan run and flag issues but not block, just to gather data on what it would block. Then review those logs to fine-tune before enforcing. Once comfortable, switch to full enforcement mode. - **False Positive Management:** Be prepared to iterate on reducing false positives. For example, the Unit42 study noted some guardrails misclassified benign code as malicious [107]. If your AI outputs code snippets (common in developer tools), coordinate with Gray Swan on how to allow that safely (maybe they have a code-mode setting). Use the logs: Gray Swan will log blocked prompts/outputs with reasons. Regularly review these, especially in the early phase, to catch if it's blocking something that should be allowed. You can then whitelist certain phrases or adjust the strictness for that category. Gray Swan's low false positive rate (FPR95 of 16.5%, better than others [55]) is encouraging, but your actual content might differ, so tuning is expected. - **Integration with Existing Moderation:** If you already had some content moderation in place (like using OpenAI's moderation API or internal regex filters), decide whether to disable those to avoid redundancy. Running multiple filters could increase false positives or latency. Likely, Gray Swan's filter will supersede simpler ones, but do an A/B test to ensure Gray Swan covers what the old filter did. One can keep the old system as backup monitoring for a while until confidence in Gray Swan is high, then retire it to reduce complexity.

C. Testing and Validation

- **Adversarial Testing:** Before full go-live, conduct your own **penetration testing on the AI system with Gray Swan in place**. Encourage your team to try to bypass the guardrails in a controlled setting. Maybe use known jailbreak prompts from the wild (there are lists circulating on forums) and see if Gray Swan blocks them. Ideally, none should get through, but if any do, document and report to Gray Swan to see if it's a misconfiguration or a novel attack needing an update. This exercise not only validates the setup but also familiarizes your team with Gray Swan's capabilities. - **User Acceptance Testing:** Test with friendly users or employees to see if the AI behavior is still useful and coherent under the new guardrails. Sometimes adding a filter can slightly alter outputs (e.g., the model might rephrase or truncate certain answers because the filter nudged it). Make sure this doesn't break the user experience. If issues arise (like the AI response is overly cautious or refusing legitimate queries), adjust the system or instruct the model with clearer guidelines (you might need to update your prompt to the model to account for the filter's presence). Gray Swan mentions it can be configured via simple parameters or dashboard – utilize that UI to simulate and tweak until the balance is right [108]. - **Performance Benchmarking:** Keep track of any changes in model performance or cost due to the integration. Possibly the filter might cause a slight increase in token usage (if it adds tokens when calling the model? Unlikely, it probably doesn't consume model tokens itself except for scanning output, which is minor). Just ensure throughput and cost metrics are within acceptable ranges. If not, consult Gray Swan for optimizations (maybe batch processing if you have bulk inferencing, etc.).

D. Rollout Strategy
- **Phased Rollout:** It's prudent to roll out the new safeguarded AI system in phases. Start with a small subset of users or a pilot project. Monitor logs intensively and gather feedback. Then gradually ramp up to all users. This mitigates any unforeseen issues impacting everyone at once. Gray Swan's platform should handle scaling, but your integration might reveal corner cases when load increases. - **Communication to Users:** If the AI is user-facing, consider informing users that an **AI safety system is in place** for their protection. For example, if a user's query is blocked or rephrased, have a friendly message like, "Your request was modified for safety reasons" rather than a generic error. Transparency can help users understand why some queries might be disallowed. If the AI is internal, communicate to employees about the new guardrail and encourage them to not deliberately circumvent it (and to report if they somehow do). - **Feedback Loop:** Establish a channel for users (or internal testers) to give feedback if they encounter what they think is a false block or an unsafe output that wasn't caught. This feedback loop is valuable. For any unsafe output that slips through, escalate immediately with Gray Swan – as a client, you should expect quick response to patch such gaps (they might provide updated filter rules or model parameters). For false blocks, see if policy adjustment is needed or if users need clarity. Treat initial weeks as a learning period and adjust configuration accordingly.

E. Compliance and Documentation
- **Document the Controls:** Update your risk register and documentation to reflect that "AI Model X is protected by Gray Swan AI Cygnal filter as of DATE, which addresses risks A, B, C." This will be useful for audits or regulatory inquiries. Gray Swan likely can provide documentation on their methodology and results (like the precision/recall on various harmful content categories [54]). Attach that as evidence of control effectiveness. - **Privacy Assessment:** If using Gray Swan's cloud, conduct a privacy/security assessment (as you would with any vendor). Confirm what data is sent to them and how it's stored. According to Gray Swan, they don't store or sell data beyond analytics and presumably abide by strict policies [43]. Still, make sure this is covered in your DPIA (Data Protection Impact Assessment) if required. If your data is extremely sensitive and regulations disallow sharing, opt for on-prem where all data stays in-house. Gray Swan's privacy policy and any data handling terms should be reviewed by legal. - **Service Level Agreements (SLAs):** For mission-critical use, negotiate an SLA with Gray Swan (uptime, support response times). Ensure there's a clear support channel for emergencies (like a hotline or priority email if a critical flaw is discovered at 3am). Knowing how to reach their team quickly is part of incident preparedness.

F. Long-Term Maintenance
- **Regular Updates and Patches:** Gray Swan will presumably update its model and filters regularly (to incorporate new threats). If cloud-based, these come automatically. If on-prem, have a process to get updates (maybe a docker image or package). Assign someone on your team to stay on top of Gray Swan's release notes or announcements (subscribe to their newsletter or portal) [109]. We recommend

scheduling maintenance windows to apply any critical updates they provide (similar to antivirus definitions updates in traditional IT security). - **Periodic Re-evaluation:** As recommended earlier, every so often re-evaluate if Gray Swan's solution is meeting your needs or if adjustments are needed. This could involve running new test cases that have emerged in the wild (for example, if a new kind of jailbreak or bias issue is talked about in research, test against it). Maintain a relationship with Gray Swan's account manager to discuss roadmap – e.g., if you need a feature (say, support for vision models or certain compliance module), voice that; it might already be in development or they can prioritize it.

G. Contingency for Transition
- Although not expected in the near term, have a contingency plan in case you ever needed to switch away from Gray Swan (for example, if pricing changed drastically or in the unlikely event they go out of business). This could be as simple as keeping a list of alternative providers or open-source fallback. Since you now know the integration points (the filter sits at the API level), swapping to another filter vendor is feasible if needed. The idea is to avoid vendor lock-in by design – maintain some abstraction in your code where Gray Swan is implemented, so another service could be plugged in with minimal changes. This is future-proofing; you likely won't need it if Gray Swan continues to perform well and scale, but it's good practice.

By addressing the above considerations, implementation of Gray Swan AI's safety measures should be smooth and yield strong risk reduction outcomes. The focus is on being methodical: integrate carefully, tune thoughtfully, test thoroughly, and maintain actively. The Appendices provide supporting source references and additional context that informed these recommendations, ensuring that each step is grounded in best practices and documented evidence.

Appendices

Sources

1. Gray Swan AI – *Official Website (About Us)*: "Gray Swan is the safety and security provider for the AI era... dedicated to helping enterprise organizations, frontier model developers, startups... deploy AI with confidence by providing tools that assess the risks of a deployment, as well as secure models..." [110]. This establishes Gray Swan's mission and target customers in its own words.

2. Gray Swan AI – *Official Website (Product – Cygnal)*: Pricing and deployment details for Cygnal, e.g., "Your first 50M tokens are free. After that, usage is charged as follows: $1/M Token (First 1B per month), $0.70/M (1B–10B), $0.60/M (10B+)" [3]. Also notes on on-prem enterprise deployment ("ultimate in security, contact us to learn more") [4] and API integration ("understands all popular formats... change one URL in your code") [32]. These details support the revenue model (usage-based SaaS + enterprise licensing) and ease of integration.

3. Forbes (Sarah Emerson, *Forbes Australia*, Nov 8, 2024) – *"This hacker team is bulletproofing AI models..."*: An in-depth profile of Gray Swan. Key points cited: Gray Swan secured early **partnerships/contracts with OpenAI, Anthropic, and the UK AI Safety Institute** [5]; Quote from CEO on huge unmet need for practical AI risk solutions [18]; Description of Gray Swan's founding after discovering major LLM vulnerabilities [24]; Discussion of *Cygnet* model using "circuit breakers" and how it withstood jailbreaking attempts (99.98% block rate, only 0.02% attacks succeeded) [47] [15]; Funding info ($5.5M seed, prepping Series A) [41]. This source gives external validation of Gray Swan's capabilities, partnerships, and performance metrics.

4. Gray Swan AI – *Official Website (Main page)*: Outline of products and their purpose – *Cygnal:* “wraps your AI-powered applications with bi-directional security that blocks malicious inputs and filters harmful outputs” [8]; *Shade:* “comprehensive AI security and safety evaluation suite… continuously deliver insights into how your deployment will behave under worst-case conditions” [9]. Also mentions Arena as a venue for AI red teaming with community participation [58]. This corroborates Gray Swan’s feature set and value proposition.

5. TechCrunch (Kyle Wiggers, Apr 22, 2025) – *“Crowdsourced AI benchmarks have serious flaws…”*: Quote from Matt Fredrikson (Gray Swan CEO) acknowledging that public benchmarks attract volunteer red-teamers for “learning new skills” and that **public benchmarks aren’t a substitute for paid private evaluations** [6]. Indicates Gray Swan runs crowdsourced red teaming campaigns and also does private ones. This supports statements about Gray Swan’s consulting services and the need for internal benchmarks alongside public ones.

6. CSO Online (David Strom, Aug 2024) – *“5 steps for deploying agentic AI red teaming”*: Describes an academic paper led by Andy Zou (Gray Swan co-founder) where they attacked AI agents in scenarios, finding **60,000 successful prompt injection attacks out of 2 million, across domains like finance, healthcare** [11]. Highlights need for defenses. This was used to illustrate risks and Gray Swan’s involvement in cutting-edge research on multi-agent vulnerabilities.

7. CB Insights (Competitors list for Robust Intelligence): “Robust Intelligence’s top competitors include Protect AI, CalypsoAI, and LatticeFlow AI…” [111]. Used to identify relevant competitors in the AI security space.

8. Robust Intelligence – *Official Website*: Marketing snippet: “Protect generative AI applications against attacks and undesired responses. Robust Intelligence guardrails protect against security and safety threats.” [13] Also detail: “detections span hundreds of security and safety categories… automatically configured to model’s vulnerabilities identified with our AI Validation” [71]. Establishes competitor’s scope and approach.

9. Protect AI – *AWS Partner Blog (Apr 2025)*: “Protect AI is a security platform for artificial intelligence systems. It helps organizations identify, monitor, and mitigate AI security risks. The platform integrates two key tools: Guardian (scans & validates models) and Recon (automated red teaming for generative AI).” [73] [74]. Used to highlight competitor features (model scanning + attack simulation) and platform description.

10. FSD-Tech (CalypsoAI product page): “Calypso AI protects your Large Language Models from Misuse, Data Leakage and Adversarial Threats. With Testing, Monitoring and Real-Time Defense built in, it ensures Safe, Compliant and Trustworthy AI across every stage of your Deployment Lifecycle.” [14]. Supports statements about CalypsoAI’s capabilities and focus on safe/compliant AI.

11. HiddenLayer – *Official Website*: “HiddenLayer protects against the full spectrum of AI attacks. Our protections are rooted in global frameworks such as MITRE ATLAS and the OWASP Top 10 for LLMs…” [91]. Used to describe HiddenLayer’s positioning (broad AI attack protection, framework-based approach).

12. Unit42 (Palo Alto Networks, June 2025) – *Comparative study on LLM guardrails*: Exec summary notes about false positives and false negatives in guardrails, e.g., *“Highly sensitive guardrails across different systems frequently misclassified harmless queries as threats… Some prompt injection*

strategies successfully bypassed input guardrails on various platforms, and when harmful content was generated, output filters sometimes failed to intercept it.” [96] [28]. Also explanation of guardrails vs. model alignment [27]. This provided a neutral insight into limitations of typical guardrails and reinforced the need for robust solutions (and caution with built-ins).

13. Pittsburgh Post-Gazette (Aug 11, 2024) – *“Everyone racing to adopt AI… safely. This Pittsburgh startup wants to help companies follow through.”*: Paraphrased context from this piece (founders raised $5.5M, aim to save AI developers from themselves by providing tools). We couldn’t directly quote due to access issues, but it informs the narrative that companies claim to use AI “safely” and Gray Swan actually enables it. (Referenced in passing as [30] for context).

14. Forbes press release via Gray Swan site (Oct 29, 2024): “Gray Swan AI… is leading the charge in bulletproofing AI models for companies like OpenAI and Anthropic… at the forefront of AI safety, building powerful tools to mitigate risks in rapidly evolving AI landscapes.” [50]. Validates Gray Swan’s reputation and clientele, used in writing to emphasize credibility.

Each source above was used to ensure accuracy of factual claims (e.g., performance metrics, pricing, founders’ quotes) and to present an evidence-based analysis. All citations in the report follow the prescribed format and point to the relevant supporting text for verification.

---

[1] [5] [15] [16] [17] [18] [24] [25] [26] [37] [41] [51] [60] [61] [64] This Hacker Team Is Bulletproofing AI Models For Companies Like OpenAI And Anthropic
https://www.forbes.com.au/news/innovation/this-hacker-team-is-bulletproofing-ai-models-for-companies-like-openai-and-anthropic/

[2] [44] [59] [65] [67] [109] [110] About Gray Swan
https://www.grayswan.ai/about

[3] [4] [10] [32] [33] [36] [46] [47] [52] [53] [54] [55] [56] [103] [104] [105] [106] [108] Cygnal
https://www.grayswan.ai/product/cygnal

[6] [31] Crowdsourced AI benchmarks have serious flaws, some experts say | TechCrunch
https://techcrunch.com/2025/04/22/crowdsourced-ai-benchmarks-have-serious-flaws-some-experts-say/

[7] [8] [9] [19] [20] [21] [22] [23] [43] [45] [48] [49] [57] [58] [66] Gray Swan AI: Enterprise Security for AI-Powered Applications
https://www.grayswan.ai/

[11] 5 steps for deploying agentic AI red teaming | CSO Online
https://www.csoonline.com/article/4055224/5-steps-for-deploying-agentic-ai-red-teaming.html

[12] [13] [69] [70] [71] [72] Protect your AI applications in real time — Robust Intelligence
https://www.robustintelligence.com/platform/ai-firewall-guardrails

[14] [62] [63] [81] [83] [84] [87] [88] [89] [90] [94] [98] [99] FSD-Tech | CalypsoAI Trusted AI Validation Tools
https://fsd-tech.com/products-calypsoai

[27] [28] [29] [96] [100] [101] [107] How Good Are the LLM Guardrails on the Market? A Comparative Study on the Effectiveness of LLM Content Filtering Across Major GenAI Platforms
https://unit42.paloaltonetworks.com/comparing-llm-guardrails-across-genai-platforms/

[30] Everyone racing to adopt AI is claiming to be doing so 'safely.' This ...
https://www.post-gazette.com/business/tech-news/2024/08/11/companies-ai-pittsburgh-startup-graw-swan/stories/202408110042

[34] [35] [38] [39] [42] [50] [68] [102] Gray Swan News
https://www.grayswan.ai/news

[40] [82] CalypsoAI & Carahsoft Drive ML/AI Adoption for Government
https://www.carahsoft.com/news/calypsoAI-and-carahsoft-partner-to-deliver-AI-ML-technology

[73] [74] [75] [76] [77] [78] [80] [97] Protect DeepSeek model deployments with Protect AI and Amazon Bedrock | AWS Partner Network (APN) Blog
https://aws.amazon.com/blogs/apn/protect-deepseek-model-deployments-with-protect-ai-and-amazon-bedrock/

[79] Protect AI and Leidos to Secure AI Across U.S. Government Systems
https://protectai.com/newsroom/protect-ai-and-leidos-secure-government-systems

[85] F5 to acquire CalypsoAI to bring advanced AI guardrails to large ...
https://www.f5.com/company/news/press-releases/f5-to-acquire-calypsoai-to-bring-advanced-ai-guardrails-to-large-enterprises

[86] Why Paladin invested in CalypsoAI
https://www.paladincapgroup.com/enabling-safe-and-confident-use-of-large-language-models-and-generative-ai-within-an-enterprise-why-paladin-invested-in-calypsoai/

[91] HiddenLayer | Security for AI
https://hiddenlayer.com/

[92] [93] HiddenLayer AI Security Platform: Complete Protection for Your AI ...
https://www.youtube.com/watch?v=gcj5ZcAV02A

[95] The AI Security Playbook - HiddenLayer
https://hiddenlayer.com/innovation-hub/the-ai-security-playbook/

[111] Top Robust Intelligence Alternatives, Competitors - CB Insights
https://www.cbinsights.com/company/robust-intelligence/alternatives-competitors