IntuitionLabs

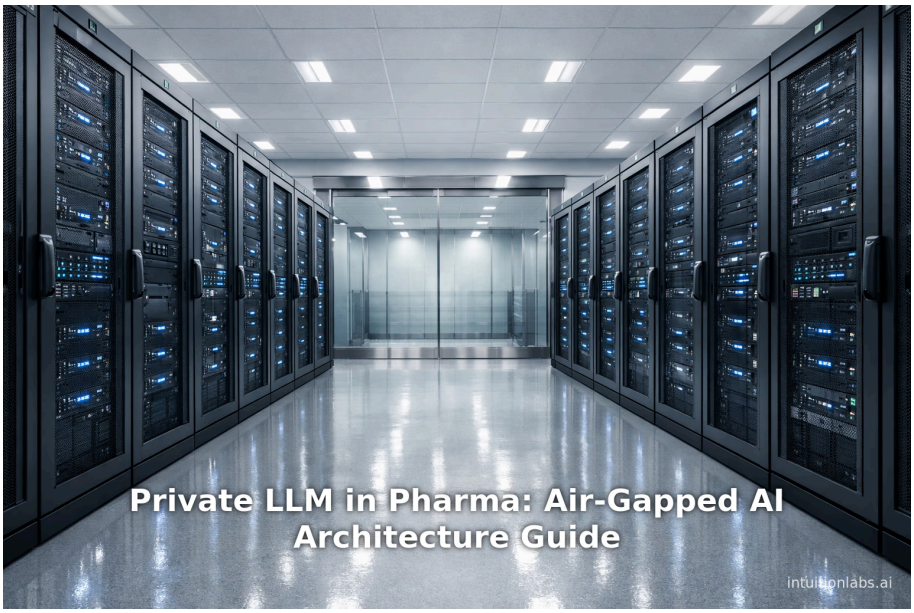
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Private LLM in Pharma: Air-Gapped AI Architecture Guide

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-04-09 · private llm · air-gapped ai · pharma ai · data sovereignty · ai compliance · healthcare llm · ai architecture · on-premise llm

Original article: https://intuitionlabs.ai/articles/private-llm-pharma-air-gapped-architecture



In this article

1. Executive Summary
2. Introduction and Background
3. Regulatory and Compliance Landscape
4. Private LLM Deployment Architectures
5. Security Controls and Threat Mitigation
6. 1. Access Control and Network Segmentation
7. 2. Monitoring and Audit Trails
8. 3. Input Validation and Prompt Filtering
9. 4. Output Control and Validation
10. 5. Monitoring for LLM-Specific Attacks
11. 6. Integrating with Enterprise Security
12. Data Management and RAG Considerations
13. Case Studies and Real-World Examples
14. Implications and Future Directions
15. Conclusion

Executive Summary

Pharmaceutical companies are racing to integrate generative AI into research, development, and operations, but must do so within a vastly more constrained environment than most industries. Pharma's massive confidential data stores (clinical trial data, proprietary research, patient records, etc.) and stringent regulatory demands (HIPAA, 21 CFR Part 11/GxP, GDPR, the EU AI Act (generally applicable since 2 August 2026, with exceptions; European Commission), etc.) mean that simply calling a public AI API is often impossible. Instead, firms are increasingly turning to **private LLM deployments** – on-premise or isolated environments where models and data remain under the company's direct control. In these settings, **air-gapped architecture** (no physical connection and only manually controlled transfers; NIST) and strict **data sovereignty** controls are critical.

This report examines how pharmaceutical organizations can architect, secure, and govern private LLMs to both harness their power and satisfy regulatory/compliance requirements. We draw on industry analyses, technical research, and real-world deployments to show best practices and pitfalls. Key findings include:

- **Value of AI in Pharma.** Generative AI may unlock tens of billions of dollars per year in value for pharma by speeding drug discovery, optimizing clinical trials, and improving operations (www.mckinsey.com) (www.scilife.io). However, these use cases inevitably involve highly sensitive data.
- **Regulatory Imperatives.** Pharma is among the *most heavily regulated* sectors. US rules (HIPAA, FDA 21 CFR Part 11/210, GxP guidelines) and EU rules (GDPR, AI Act, medical device regulations) all impose rigorous data protection, auditability, and validation requirements (www.atlas-compliance.ai). For instance, determine validation, audit-trail, and access-control needs from intended use, applicable

predicate rules, and the regulated records affected (FDA guidance); assess data integrity (ALCOA+ compliance) within that scope (www.atlas-compliance.ai) (www.techradar.com). Failure to comply can mean major fines or halted drug approvals.

- **Architecture Strategies.** Pharma firms can evaluate private on-premises environments or secured cloud/VPC deployments. Distinguish a physical air gap, with manually controlled transfers and no physical connection between systems, from an enclave with a hardware-enforced one-way gateway (NIST). Controlled cross-domain connectivity needs an assessed design with appropriate filtering and security enforcement; software forwarding alone does not provide the assurance of a hardware diode (ASD).
- **Security Controls.** Beyond basic network isolation, private LLM deployments must guard against novel AI-specific attacks. Prompt injection, Trojaned RAG knowledge, illicit data exfiltration, and model-inversion threats all arise in LLMs. Defense-in-depth is needed: input sanitization and dynamic filters, continual monitoring, and specialized "LLM agent" classifiers for malicious prompts (cloudsecurityalliance.org) (arxiv.org). Use layered controls and restrict tool permissions: prompt-injection mitigations cannot eliminate the vulnerability entirely (OWASP).
- **Case Studies.** In practice, defensive architecture varies by use case. U.S. national labs and the military run self-hosted "Ask Sage" LLMs inside classified enclaves (blog.dreamfactory.com), while TrueFoundry reports that a Fortune 100 healthcare company deployed models separately across regions in response to data-movement regulations (TrueFoundry case study). Technology vendors are responding: AWS now offers on-prem "AI Factory" cabinets to satisfy sovereignty requirements (www.techradar.com), and HPE's Private Cloud AI can run in fully disconnected (air-gapped) mode (www.itpro.com).
- **Future Outlook.** The EU AI Act became generally applicable on 2 August 2026, with exceptions; its high-risk rules apply from 2 December 2027 for Annex III uses and 2 August 2028 for high-risk systems embedded in regulated products (European Commission). Pharma companies that invest now in robust private LLM infrastructure – embedding security, auditability, and compliance by design – can not only avoid legal peril but gain first-mover advantage. However, these systems demand a new operational model (specialized hardware, disciplined data governance, and security expertise) that broad training and modern governance tools must support.

This report dives into each of these areas: we analyze regulatory requirements, architectural options, security controls, and emerging solutions. We draw on technical research and expert commentary to provide evidence-backed recommendations for pharma organizations moving into private LLM deployment.

Introduction and Background

Generative AI (and Large Language Models, LLMs) have rapidly evolved from an experimental novelty to critical tools across industries. In healthcare and life sciences, **the potential impacts are enormous**. Recent analyses estimate that AI could create **hundreds of billions of dollars in annual value by 2025** in the pharmaceutical sector (www.scilife.io) (www.mckinsey.com). Benefits include accelerating drug discovery (e.g. suggesting new compounds or interpreting complex genetic data), optimizing trial management (targeting

patients, minimizing delays), automating regulatory writing, and personalizing patient engagement.For example, McKinsey reports that generative AI might boost productivity to unlock roughly **$60–110 billion per year** in pharma by speeding discovery and development processes (www.mckinsey.com). These opportunities are especially vital in pharma, where bringing a single drug to market can cost over a billion dollars and take a decade; any acceleration or insight provided by AI can greatly improve ROI.

At the same time, **pharma's use cases are data-intensive and highly sensitive**. Pharmaceutical companies generate and consume massive data volumes: clinical trial datasets, patient health records, proprietary research manuscripts, and milling pharmacovigilance logs, among others. Much of this data is regulated (patient data is protected by HIPAA in the U.S. and GDPR in Europe), and much of it is also core intellectual property (unpublished research hypotheses, novel molecular structures) that companies cannot afford to leak. This creates a stark tension: on one hand, companies want to use AI to *analyze* this data; on the other hand, *feeding it into external services* (like public cloud APIs) risks exposure.

Indeed, many executives now explicitly ask: *"How can we use AI to create value **without losing control** of our sensitive data, intellectual property, or compliance posture?"* (www.techradar.com). The notion of **"private AI"** arises from this demand. Private AI refers to deploying LLMs in controlled environments where all data processing happens under the organization's direct supervision (www.techradar.com). Practically, this often means running LLMs on-premises (in the company's own data centers or private clouds) with strict network and access controls, rather than sending data to third-party servers. The goal is to keep **"data and intellectual property within an organization's infrastructure,"** eliminating the inadvertent leakage that can happen with public AI models (www.techradar.com).

This shift is already underway. In one large healthcare company, executives identified 30+ enterprise LLM use cases by 2025, but had to build a *regionalized architecture* because "data movement regulations forced the company to deploy models separately in each region of presence" (www.truefoundry.com). Likewise, technology providers are moving to meet demand: AWS's new "AI Factories" bring managed AI hardware into the customer's facilities (www.techradar.com), and HPE announced an air-gapped mode for its Private Cloud AI to satisfy strict sovereignty requirements (www.itpro.com). These developments underscore that for regulated industries, **private LLM deployments are becoming a strategic imperative**, not an optional convenience.

In summary, **the pharmaceutical industry stands at the intersection of immense AI opportunity and stringent data constraints**. This report surveys the landscape of deploying LLMs under these constraints. We review the relevant regulations and compliance demands (Section 3), outline architectural patterns for private LLMs and air-gapped systems (Section 4), discuss risk vectors and security controls (Section 5), and present real-world examples (Section 6). Throughout, we ground our discussion in data and expert insight, emphasizing how to align LLM innovations with the sector's rigorous risk tolerances.

Regulatory and Compliance Landscape

Pharma companies operate in a **highly regulated environment**. Before deploying any LLM on pharmaceutical data, organizations must ensure compliance with multiple overlapping legal and industry standards. Key frameworks include:

- **Health Information Privacy (HIPAA).** HIPAA obligations apply to covered entities and business associates handling PHI, rather than to every use of health-related data. Cloud processing of ePHI is permitted when the service relationship has the required HIPAA-compliant BAA and the parties otherwise comply with the HIPAA Rules (HHS cloud guidance). Strongly prefer encryption at rest and in transit. Under the Security Rule, encryption is an addressable implementation specification: make a documented risk-based determination, implement it when reasonable and appropriate, and document any justified alternative or rationale (HHS encryption guidance). (blog.dreamfactory.com).
- **FDA 21 CFR Part 11 and GxP.** Determine Part 11 applicability from the electronic records and signatures used to meet applicable predicate rules or FDA submission requirements. Document scope decisions and base validation and audit-trail controls on the system's intended use, applicable requirements, and a justified risk assessment (FDA guidance). Likewise, Good Laboratory/Manufacturing/Clinical Practice (GLP/GMP/GCP) guidelines demand rigorous documentation and data integrity (the ALCOA+ principles: Attributable, Legible, Contemporaneous, Original, Accurate, plus Complete/Consistent/Enduring/Available (www.atlas-compliance.ai)). An AI system used for quality decisions (e.g. QA oversight or clinical analysis) would need to produce reliable records that inspectors can follow from input through output. As one compliance analysis notes, LLM-powered tools *can* be made Part 11 compliant **only if treated like any other validated system** – with defined intended use, controlled inputs/outputs, rigorous audit trails, and documented human oversight (www.atlas-compliance.ai). In practice, this often means developing an LLM within a quality system (QMS) framework, including pre-deployment validation and continued Performance Qualification.
- **Data Protection (GDPR, regional laws).** Where GDPR applies, identify an Article 6 lawful basis for processing personal data. Processing sensitive clinical records also requires an applicable Article 9 condition; legitimate interests alone do not authorize processing health data (CNIL). For international transfers, assess the applicable Chapter V mechanism: an adequacy decision, appropriate safeguards, or a narrowly applicable derogation. Where relying on standard contractual clauses, document the transfer impact assessment and any necessary additional safeguards (EDPB). Distinguish these transfer conditions from any additional residency policies or obligations applicable to the deployment.
- **EU Artificial Intelligence Act.** Determine each system's risk classification from its intended purpose and the applicable product or listed-use provisions, rather than assuming that all healthcare or scientific uses are high-risk. Assess obligations separately for AI-system providers, deployers, and general-purpose AI-model providers (European Commission guidance).
- **Other Compliance Standards.** Pharma firms must often adhere to industry best practices and certifications too. For example, SOC 2 or ISO 27001 for information security; FDA's Data Integrity guidelines; and, in defense-related aerospace projects, ITAR export controls. Air-gapped LLMs especially often need to comply with strict export controls (ITAR/EAR) and ensure absolute non-exfiltration of any data (blog.dreamfactory.com). Notably, regulations like HIPAA and GDPR now treat AI-generated insights no differently than traditional data processing: any leakage of underlying PHI or personal health content through an AI model could be a violation.

These regulatory demands converge on several **common requirements for Pharma AI**:

- **Data Residency and Sovereignty.** Assess international transfers under the applicable legal framework. Under GDPR, transfers outside the EEA may rely on adequacy decisions or appropriate safeguards; derogations are limited to specified situations (EDPB). Document any additional national, sectoral, contractual, or organisational residency requirements separately before selecting deployment regions.
- **Access Controls and Auditability.** Define security logging according to the deployment's risks, privacy constraints, and intended use. Determine which electronic records fall under Part 11 using applicable predicate rules. FDA recommends basing audit-trail decisions on those requirements and a documented assessment of effects on product quality, safety, and record integrity (FDA guidance). Restrict access to authorised users and require review for consequential actions.
- **Privacy by Design.** Organizations must minimize sensitive data exposure. PHI or confidential research documents should be pre-processed (de-identified or summarized) before ingestion, and all results should be checked for inadvertent memorization of personal details. Under GDPR and HIPAA, special caution is required when using models that can "regurgitate" training data (dataprotection.ie).
- **Validation and Risk Management.** LLMs should be treated like any critical system: with formal validation studies showing accuracy, testing for failure modes, and risk controls (e.g. human-in-the-loop for critical decisions). As noted by compliance advisors, it may be necessary to accept that LLMs are probabilistic ("opaque") and thus *constrain* their use cases (e.g. not letting them autonomously issue regulatory filings). Any LLM-based tool in quality operations should be thoroughly validated and documented beforehand (www.atlas-compliance.ai).

In practice, meeting these requirements often **dictates the architecture**. For example, many pharma companies simply do not allow wall-to-wall internet connectivity in labs or wings handling sensitive data. Instead, they create segregated networks or even physically isolated facilities for AI. Network isolation should not be treated as proof of compliance. Assess GDPR obligations for the particular AI product and processing activity (Data Protection Commission). We explore these architectures in the next section.

Table 1 (below) summarizes select regulations pertinent to Pharma LLM deployment.

Regulation/Standard	Scope	Key Requirements	Implications for LLMs
HIPAA (US)	PHI handled by covered entities and business associates	Privacy and Security Rule safeguards; risk analysis and appropriate controls	Prefer encryption at rest and in transit; document the risk-based determination for this addressable specification (HHS). (blog.dreamfactory.com).
GDPR (EU)	Personal data within GDPR's scope	Lawful processing, data minimisation, data-subject rights, and Chapter V transfer conditions	GDPR does not impose a blanket EU-only residency rule. Transfers outside the EEA can use an adequacy decision or appropriate safeguards; limited derogations may apply subject to their conditions. Assess the applicable transfer mechanism (EDPB).
21 CFR Part 11 (US)	FDA-regulated electronic records	System validation; ALCOA+ data integrity; audit trails; user accountability	LLM-based tools in quality systems must be validated; outputs must be auditable and include date/user stamps (www.atlas-compliance.ai).

Regulation/Standard	Scope	Key Requirements	Implications for LLMs
EU AI Act	AI systems and general-purpose AI models within scope	Obligations vary by risk classification and operator role	Most rules became applicable on 2 August 2026. High-risk rules apply from 2 December 2027 for Annex III uses and 2 August 2028 for high-risk systems embedded in regulated products; determine the applicable category and operator duties (European Commission).
GxP	Pharmaceutical research and manufacturing	Detailed documentation (batch records), controlled processes	For drug CGMP activities, FDA expects all data to be reliable and accurate (FDA data-integrity guidance).
ITAR/EAR (US)	Defense-related tech/data	Export controls on technical data and software	LLMs containing controlled technical info (e.g. biological agents) may require internal-only deployment; shared on secured systems only (blog.dreamfactory.com).

Each of these regimes has severe penalties for breach – e.g. GDPR fines up to 4% of global turnover (blog.dreamfactory.com). Consequently, **private deployment** is often the de facto choice in regulated settings, as it allows pharmaceutical firms to implement necessary controls (encryption, logging, local archiving) themselves. In the sections ahead, we will show how to design these architectures so that compliance is built-in from the ground up.

Private LLM Deployment Architectures

Deploying an LLM privately entails moving the model and inference infrastructure behind an organization's secure perimeter. There are several architectural modes, each offering different trade-offs for performance, manageability, and control:

1. **On-Premises Servers/Private Cloud.** In the simplest private setup, the company builds or rents GPU clusters within its own data centers or a dedicated private cloud. The LLM (often a large open-source model like Llama 3, Mistral, or Qwen) runs on these servers behind the corporate firewall. All data inputs and outputs stay on-prem. This setup provides the highest level of control: administrators can enforce network segmentation, strict IAM policies, and hardware access restrictions. It incurs capital and operational cost (buying GPUs, hiring ops staff), but many pharma R&D centers already run HPC clusters for molecular simulation and can repurpose those for AI.

An example is **AWS AI Factory**: AWS installs identical hardware stacks (GPUs, servers) inside the customer's data center and manages them as a dedicated "private AWS region" (www.techradar.com). This gives organizations the convenience of a managed service while keeping all data physically local. As TechRadar noted, in this model *"AWS says it can deploy these systems in months… With AWS managing the entire AI environment exclusively for the one customer, data stays local and hardware will not be shared with others."* (www.techradar.com). This "on-prem plus managed" approach (-AWS AI Factory, Azure Stack AI, etc-) is attractive for large enterprises that need both control and scalability.

2. **Virtual Private Clouds (VPCs) with Isolation.** Some companies opt for cloud-based hardware but with strong isolation. In this model, the LLM runs in a customer's private VPC or on-cluster (often near an air-gapped data lake), and strict security groups/firewalls limit external access. All model inferences might be served through private APIs, and the chosen cloud provider must guarantee data never leaves tenancy. For instance, HPE's Private Cloud AI can run in a mode that is completely air-gapped from external networks (www.itpro.com). The advantage is flexibility and the ability to leverage cloud elasticity; the key requirement is certifying that the cloud provider's control plane and side channels are not allowed to exfiltrate data.
3. **Air-Gapped Environments.** A physical air gap has no physical connection between systems and permits logical transfers only manually, under human control (NIST definition). A hardware-enforced unidirectional gateway is a different architecture: it provides controlled connectivity in one direction. Ordinary network segmentation, software filtering, or syslog forwarding should not be described as a physical air gap or treated as equivalent to a hardware diode.

A cross-domain solution requires more than installing a diode: define the permitted transfer direction, content filters, failure behaviour, and residual egress risk, and assess the assurance of the complete design (ASD guidance).

Figure 1. *Example of an isolated LLM enclave with controlled one-way export, rather than a physical air gap. Data enters through approved physical-media procedures; a hardware-enforced unidirectional gateway exports approved logs or results. Assess filtering and residual disclosure risk before release.*

In this design, any model updates or new datasets must also be transferred manually via encrypted USB drives or the equivalent. As DreamFactory's Terence Bennett notes, this *"offline data transfer… must follow strict chain-of-custody protocols"* (blog.dreamfactory.com). Thus, governance of physical media is critical (virus scanning, limited personnel).

4. **Hybrid and Edge-Assisted Models.** Some hybrid strategies are emerging. For example, key AI models may run on-prem, but less-sensitive components (like the user interface or non-confidential prompts) might connect to lower-trust cloud services for efficiency. Another pattern is to use small "edge" LLM clients that handle lightweight tasks locally and only send batched, anonymized data into a secure AI core for heavy reasoning. However, in most **strictly regulated pharma contexts**, the safe default is to minimize all egress: that often leads organizations to a *"bring compute to data"* mentality rather than vice versa (www.techradar.com). In practice, this means preferring on-site model execution over any cross-network processing.

These architectural choices directly impact **operational considerations**:

- **Connectivity and Updates.** Private LLM clusters cannot rely on automatic cloud updates or external APIs. All software dependencies (model weights, libraries, vendor patches) must be curated offline. Feature branches of LLM frameworks may need containerization and robust version control to ensure reproducibility. Some organizations set up dedicated "update stations" within their DMZs to pull approved packages via a one-way bridge.

- **Hardware and Scale.** LLMs are resource-intensive – e.g. a 70B-parameter model may require multiple high-memory GPUs (A100/H100) to run efficiently. Building such capacity on-prem is costly: companies may invest millions in GPU clusters. This is why offerings like AWS AI Factories or HPE's scaled racks (now double to 128 GPUs) are being marketed to enterprise customers (www.itpro.com). The architecture must allow horizontal scaling (adding GPUs/nodes) while maintaining isolation.
- **Software Stack.** Common inference frameworks for private LLM include vLLM, Ollama, or vendor platforms that can run containerized Hugging Face or custom models on-prem. These platforms must be audited and secured. Companies often apply their own hardened Linux distributions, disable unnecessary telemetry, and implement container security scanning. Multi-tenancy (sharing a cluster across departments) is possible but requires careful namespace and GPU scheduling isolation.

In short, private LLM deployment in pharma looks **much more like a traditional on-prem IT project than a simple cloud SaaS integration**. It demands in-house GPU infrastructure, stringent networking, and enterprise-grade orchestration. The next section will discuss how to layer security controls onto this architecture to prevent both traditional cyber threats and AI-specific attacks.

Security Controls and Threat Mitigation

Running an LLM privately significantly reduces some risks (e.g. direct cloud breaches), but it introduces others and does not obviate core security needs. Notably, **the AI introduces new threat vectors and requires specialized defenses**. We group the key control areas as follows:

1. Access Control and Network Segmentation

Standard IT security best practices apply: treat the LLM infrastructure as highly sensitive. This means strict network isolation (as described), firewalled segment, and minimal external interfaces. Role-Based Access Control (RBAC) should ensure that only specific users or service accounts can query the model. Administrative access to the GPU servers themselves must use multi-factor authentication and logging. Network appliances (if any) in the air-gapped zone must also be locked down – for instance, any NFS mounts or databases feeding the RAG system should be opening to only necessary IPs.

For example, DreamFactory's platform enforces "identity passthrough" from corporate IAM into the LLM API layer, so that all data requests carry user identity meta-data. This ensures accountability: any query in the LLM logs can be traced back to a logged-in individual (a requirement under many compliance regimes (www.atlas-compliance.ai)). Dedicated API gateways or service meshes can mediate all LLM calls, checking tokens against an identity provider and enforcing policies about allowable prompts.

2. Monitoring and Audit Trails

Audit logging is critical both for security and compliance. Every LLM prompt and response ideally should be logged (with redaction of any sensitive output, per privacy rules). The logs must be sent through the data diode (or DMZ collector) to an enterprise SIEM system. As Hannecke describes for sovereign deployments: *"your SIEM needs AI-specific event schemas, your SOC needs runbooks for prompt-injection alerts, and your*

approval workflows need to handle AI-specific escalations" (medium.com). In practice, this means defining new log types (e.g. `PROMPT_REQUEST`, `PROMPT_INJECTION_BLOCKED`, `PII_DETECTED_IN_RESPONSE`) and ensuring all security tools ingest them.

Tamper-proof audit is a must. Techniques might include write-once log files, remote logging by data diode, and cryptographic signing of events. Insider threats are a concern: anyone with admin rights could conceptually turn off logging, so strong separation of duties and periodic audits (even manual reviews) are recommended.

3. Input Validation and Prompt Filtering

AI models are vulnerable to malicious inputs. The industry's OWASP Top 10 for LLMs identifies dual categories: *direct injection* (malicious user prompt) and *indirect injection* (poisoned context from RAG documents) (medium.com). Preventing simple exploits requires the same "stage 1, 2, 3" guardrails used for cloud LLMs, but all running locally. This includes:

- **Whitelist/Blacklist Filters:** Block known disallowed content in queries (e.g. attempts to reveal PII or proprietary data) and in RAG content. For example, any user question that looks like *"Ignore previous instructions…"* could indicate a prompt-injection attempt. A local filter might block or query a security LLM.
- **Tiered Inspection:** Major vendors use a multi-stage approach: Stage 1 syntactic filters, Stage 2 ML classifiers, Stage 3 semantic checks via another LLM. In an air-gapped setup, Stage 3 must use a local model (e.g. Llama 3 or Qwen). Hannecke's January 2025 model-selection discussion compares local models with GPT-4 for prompt-injection classification. The author describes internal tests against public injection datasets, rather than published benchmarks; treat the results as limited evidence and evaluate the deployed configuration separately (Hannecke). To compensate, one can lower the escalation thresholds (treat more inputs as suspicious) and tighten Stage 2. For example, if GPT-4 would only escalate an input with risk score ≥0.7, a sovereign system might autoclassify anything ≥0.5 as suspicious and send it to Stage 3 (medium.com).
- **Chat Hygiene:** Require that all user prompts pass through sanitization. Some deployments use a neural network or ruleset to strip out content like SSNs, credit cards, or API keys before passing text to the model. When a user is allowed to upload files (for RAG), each document must be scanned for hidden scripts or exfiltration commands. In the context of pharma, this might include redacting patient identifiers or manual curation to ensure trade secrets aren't inadvertently exposed.
- **Context Checking:** In Retrieval-Augmented Generation (RAG), the model's context is drawn from internal document stores. CSA recommends validating *each retrieval result* to ensure it contains no malicious payloads (cloudsecurityalliance.org). For instance, before a document's text is put into a run, it could be screened by an LLM or rule-based system for suspicious instructions. This adds latency but is prudent when RAG references sensitive knowledge graphs (see sidebar).

4. Output Control and Validation

Just as inputs need filtering, outputs from the LLM must be vetted. Outbound content may inadvertently reveal PHI or IP. Controls include:

- **Sensitive Data Detection:** Scan outputs for patient identifiers and confidential content, with redaction or blocking and alerts for detected disclosures. Presidio offers PII identification and anonymization with predefined or custom recognizers; configure and validate recognition for the data in this deployment. For proprietary compound information, evaluate dedicated rules or models rather than assume that generic PII tools cover it. Treat Detoxify as a toxic-comment classifier, separately from PII/IP disclosure detection. Azure AI Content Safety detects harmful content; its disconnected containers and embedded SDK are limited-access features requiring registration and restricted use cases (Microsoft access requirements). Containers support only a subset of features and are in public preview; verify the supported features and offline approval before selecting them (container documentation).
- **Approval Workflows:** A common pattern is Human-in-the-Loop (HITL) for high-risk outputs. For example, if an LLM generates a summary intended for external publication or regulatory submission, it might be routed as a “draft” to a human reviewer before finalization. This defers ultimate trust to an expert.
- **Versioning and Reproducibility:** Each LLM response should be tied to a specific model version and prompt. If a model or its data is updated, all outputs should include version identifiers. For regulated records, one may need the ability to “freeze” an LLM’s weights for a given submission.

5. Monitoring for LLM-Specific Attacks

LLMs introduce new attack surfaces:

- **Prompt Injection Attacks:** As detailed above, these can subvert the model’s intent. We have already mentioned defenses (multi-stage filters and local injection-detection models discussed in Hannecke’s guide (medium.com) (arxiv.org)). Research shows that specialized pipelines of LLM “security agents” can even drive attack success rate effectively to zero in testing (arxiv.org). Implementing such frameworks internally (sometimes called “self-defending AI”) may become best practice.
- **Knowledge Base Poisoning:** Do not seed operational pharmaceutical knowledge sources with false entries. Protect source-data integrity: FDA expects CGMP data to be reliable and accurate (FDA guidance).
- **Model Extraction and Exfiltration:** Attackers might try to steal the LLM itself (e.g. copying weight files) or extract proprietary training data from it. Countermeasures include: encrypting model files at rest, limiting file system permissions, and disabling copier devices on the GPUs. Some deploy model watermarking (embedding subtle patterns) so that if someone leaked a copy, it could be traced. More proactively, watermarking is also studied for keeping generated outputs secret (most watermark methods allow licensed users to verify origins). We do not have formal pharma references here yet, but the principle is clear: treat LLMs as highly sensitive IT assets, not as disposable algorithms.

- **Side-Channels and Leakage:** Even air-gapped systems are not immune to side-channel leaks. Researchers have demonstrated that deep learning hardware can emit electromagnetic patterns during computation that encode neural activations. High-security deployments may need EM shielding or power-line noise filtering. Such measures are more common in military settings, but awareness is growing. For our purposes, simply note that **every measure possible should be taken** – assume an adversary is highly motivated.

6. Integrating with Enterprise Security

Finally, private LLMs must plug into the organization's overall security stack:

- **SIEM/SOAR Integration:** As noted, all AI-related events should flow into the Security Information & Event Management system. This allows correlation with other alerts (e.g., if a phishing campaign is underway concurrently). Security Orchestration toolruns can automatically quarantine suspect behavior: for instance, if a user suddenly runs many large queries *and* tries to export results, SOC rules could shut down their session and require re-authentication.
- **Incident Response Playbooks:** Security teams should have specific runbooks for AI incidents. For example, if a prompt injection attempt is flagged as malicious, a defined process should exist: lock the session, escalate to an AI security expert, annotate the attempt, and consider retraining models to recognize that exploit. Without these playbooks, AI-induced anomalies may be dismissed as "false positives" or ignored.
- **Continuous Testing:** Because threat models are new, regular adversarial testing is necessary. Red-team engagements might try sophisticated prompt injections, context poisoning, or large retrieval queries to stress the RAG system. These tests should simulate scenarios like an internal user trying to steal R&D secrets via clever prompts, or an uploaded patient chart containing hidden commands. The insights feed back into tightening the defenses.

Together, these measures form a **defense-in-depth** architecture for private LLMs. There is no single silver bullet; rather, one layers traditional IT security (encryption, IAM, patching) with AI-specific monitoring and human oversight. The result is an environment where LLM capabilities are unlocked **under controlled and auditable conditions**.

Table 2 (below) compares key architectural options along several dimensions relevant to pharma use cases.

Deployment Type	Connectivity	Use Cases	Strengths	Drawbacks
Public Cloud LLM (API)	Internet-accessible enterprise API or consumer service	Assess suitability for the particular workload	Managed infrastructure; low operational overhead	Evaluate the particular service's confidentiality, retention/training terms, security and transfer conditions. PHI processing requires applicable HIPAA safeguards and a BAA where required; HHS permits compliant cloud processing (HHS).

Deployment Type	Connectivity	Use Cases	Strengths	Drawbacks
Private Cloud / VPC	Network-isolated, possibly limited egress	Core R&D analysis, internal document search	Good control (logs, encryption), elastic scale	Still reliant on provider compliance; some remote management overhead.
On-Prem GPU Cluster	Fully on-site (intranet only)	High-sensitivity research, GxP workflows	Maximum data control; finely tuned for compliance	High capex/maintenance; slower feature updates; limited scalability vs cloud.
Air-Gapped Enclave	No physical network connection; transfers manually controlled (NIST)	Workloads requiring physical network isolation	Reduces network connectivity	Manual transfer and update procedures; assess physical-media risks. A one-way gateway is a separate controlled-connectivity architecture.
Hybrid/Edge-Assisted	Mixed (local inference + selective upload)	Some scalable tasks with partial cloud assist	Balances performance and control (when designed well)	Complex to implement; may still expose some data for cloud tasks.

Data Management and RAG Considerations

A crucial component of enterprise LLM use is managing the **knowledge base** – the data fed into or referenced by the model. This includes:

- **Data Classification.** Not all data can co-mingle. Before any processing, organizations should classify data by sensitivity (e.g. “PHI”, “Company Confidential IP”, “Public/NFS”). Only appropriate data should enter the LLM context. For instance, patient health records should be anonymized or strictly access-controlled; trademark or patent data might be kept completely offline. Classification tools and policies should prevent accidental inclusion of highly regulated content in model ingest.
- **Data Preprocessing.** Proprietary scientific data often needs tokenization or summarization. For example, crystal structure files or genomic sequences should be transformed into text with least necessary detail. Custom rules or scrubbers should remove legal or personally-identifiable information. Given GDPR and HIPAA risks, even LLM embeddings of patient data are considered processing, so best practice is to encrypt or hash any patient ID fields before model ingestion.
- **Fine-Tuning vs. Retrieval.** Pharma orgs often debate whether to fine-tune an LLM on private data or to use RAG. **Fine-tuning** (retraining the base model on company research papers, data) embeds knowledge permanently in a new model. This can improve output relevance but comes with risk: it implicitly “copies” possibly copyrighted or sensitive information into model weights. Because fine-tuned models can’t easily prove where knowledge came from, regulators may push back. On the other hand, **RAG** keeps a base LLM generic and retrieves answers from a separate document store. In RAG, the AI

model references papers or data only at query time, which can be better for provenance. Pharma companies are exploring both approaches, but often lean toward RAG for mutable knowledge (since drug data evolves constantly) and reserve fine-tuning for general domain alignment (e.g. tele-pharma jargon).

- **Vector Databases and Search.** RAG systems rely on vector databases to store document embeddings. Security of this component is vital. The Cloud Security Alliance points out that *"vector databases that store knowledge sources… are critical infrastructure"* (cloudsecurityalliance.org). We must secure the database at rest (encryption) and in transit (TLS), apply strict IAM, and monitor queries as noted earlier. Importantly, **the similarity search capability itself can leak information**. An attacker could, with enough queries, infer relationships between embedded documents (e.g. "which documents are semantically closest to Patent X?"), potentially revealing hidden connections. The CSA blog warns of "data leakage through similarity queries" where clever attackers retrieve semantically similar sensitive data (cloudsecurityalliance.org). Defenses include limiting search granularity, rate-limiting queries, and auditing any large-volume or outlier search patterns.
- **Knowledge Poisoning Risks.** When using RAG, one must assume adversaries might attempt to poison the knowledge base. For example, if a researcher downloads internal docs to feed into a bot, an attacker could slip in fake versions first. Once inside, the LLM would use the poisoned data to generate illusory answers. Pharma companies should vet all RAG inputs: ensure only approved, versioned documentation is embedded. Periodic integrity checks (hash comparisons) of the vector DB against a known-good snapshot can detect tampering. In high-security scenarios, one might even obscure retrieval (e.g. use secure enclaves) so that raw documents never reside on a digital drive concurrently with the running system.
- **Governance and Lineage.** Any output from a pharma LLM may need explaining back to its source. If the model cites an internal study, record that citation. Systems should track data lineage: which documents contributed to a given answer. This is not trivial with neural LLMs, but techniques like tool-augmented LLM (having the model quote sources) or chain-of-thought logging can help. At minimum, logs should show which retrieval results were passed to the model for generating each answer.

These data-management practices should be assessed alongside the GDPR obligations of the particular AI product and processing activity; they do not establish compliance by themselves (Data Protection Commission).

Case Studies and Real-World Examples

Pharma is still early in LLM deployment, but we see emerging examples across sectors that illustrate best practices:

- **Defence and National Labs:** Organizations that have long used air-gapped computing (e.g. defense contractors, nuclear labs) are among the most advanced in air-gapped AI. For instance, Los Alamos National Laboratory reportedly operates its *own self-hosted LLMs* for intelligence tasks (blog.dreamfactory.com). Similarly, the US Army's "Ask Sage" workspace provides classified AI-powered

analysis in an isolated environment. These examples show that even highly sophisticated agencies accept the operational complexity of air gaps to secure AI. Key takeaways: leverage dedicated hardware with no external links, enforce strict software whitelisting, and train SOC teams on AI-specific alerts (blog.dreamfactory.com).

- **Large Healthcare Company (TrueFoundry Case).** A Fortune 100 healthcare organization (50,000+ employees, global footprint) partnered with TrueFoundry to deploy 30+ LLM use cases internally (www.truefoundry.com). The company's AI team pursued rapid ROI (projected $500M+ annual impact) across research, operations, and patient support. However, they faced a significant governance hurdle: **data residency laws**. "Data movement regulations forced the company to deploy models separately in each region of presence," creating *"a management nightmare"* (www.truefoundry.com). According to TrueFoundry, the company managed models across regional clusters and achieved a *60–80% reduction in time-to-value* for LLM use cases using reusable templates and simplified deployment (TrueFoundry case study). Lessons for pharma: multi-region deployments may be needed (e.g. EU and US clusters), and investment in orchestration tools pays off.
- **Technology Vendors.** Major compute providers are launching products tailored to these needs. As reported by ITPro in March 2026, **HPE's Private Cloud AI** now includes a fully air-gapped configuration (www.itpro.com). HPE's AI solution can operate "completely disconnected from the Internet," addressing customers who demand both virtualization flexibility and maximal isolation (www.itpro.com). Similarly, Amazon Web Services' new **AI Factories** take this further: AWS places a rack of AI servers (with the latest Nvidia Blackwell GPUs) behind a customer's firewall, and then *operates it remotely as a single-tenant cloud* (www.techradar.com). In AWS's own words, this setup ensures that "data stays local and hardware will not be shared with others" (www.techradar.com). These hybrid offerings reflect recognition that enterprises want cloud-like AI without the cloud's external risks.
- **AI Governance Initiatives.** Internally, companies are formalizing AI governance groups. For example, guidance on LinkedIn and industry blogs emphasizes creating dedicated AI compliance teams that involve legal, security, and IT experts. While not pharma-specific, this aligns with the advice to treat AI with the same rigor as clinical systems. Some early adopters have instituted "model risk management" akin to what banks do with financial models, requiring documentation (model card, data card) for each LLM deployment. (www.atlas-compliance.ai) (Atlas Compliance suggests assembling auditors' records for LLM inspections, an approach pharma could emulate).
- **Security Research.** In academia and industry labs, experiments underscore the vulnerabilities and defenders' progress. For instance, a 2025 study by Hossain et al. demonstrated a **multi-agent defense pipeline** that neutralized 400 prompt-injection attacks across different LLMs (reducing attack success to 0%) (arxiv.org). These experimental results do not establish complete protection against prompt injection. OWASP states that mitigations cannot eliminate it entirely (OWASP).

Collectively, these examples show that private LLM deployment is already feasible at scale, but requires holistic planning. The companies and agencies succeeding are those that align architecture, compliance, and security from the outset – not retrofitting AI onto loose controls.

Implications and Future Directions

Looking ahead, several trends will shape how pharmaceutical enterprises exploit private LLMs:

- **Evolving Regulations.** The AI regulatory landscape will tighten. The EU AI Act classifies systems by intended purpose and the applicable product or listed-use provisions; medical or scientific use alone does not establish high-risk status (European Commission). The U.S. is also moving (FDA has piloted its own GenAI tools for internal review (www.axios.com), and is likely to update guidance on software in medical devices). Globally, nations are crafting data localization laws (India's DPDP Act, Brazil's LGPD, etc.). Pharma firms must prepare by designing future-proof systems: assume no data transfer overseas without permission, maintain explainable audit logs, and plan for third-party compliance audits of their AI pipelines. Technology trends like **"AI compliance by design"** (built on policy-as-code frameworks) are becoming mainstream (peerobyte.com).
- **Sovereign AI Infrastructure.** Europe and other regions are actively funding domestic AI hardware initiatives (the notion of "digital sovereignty"), because reliance on foreign cloud providers is viewed as strategic vulnerability (www.techradar.com) (www.itpro.com). For pharma companies with global operations (e.g. a US parent with EU R&D centers), this means we can expect growth in regional on-prem AI clouds. CTOs should architect AI workloads to run on cloud stacks that satisfy any future regional security certification.
- **Better Local Models.** The quality of open-source LLMs is rapidly improving. Within a few years, local models (70–200B parameters) may match proprietary APIs on many tasks. Already, companies can run "GPT-4-like" reasoning on-site using models like Llama 3, GPT-NeoX, Guanaco or the new Chinese models (Qwen, InternLM) on suitably beefed-up hardware (which is what HPE and NVIDIA are enabling (www.itpro.com)). For each deployment, evaluate toxicity, hallucination, and prompt-injection detection as separate tasks under representative operating conditions; document the metrics and test configuration (NIST AI RMF Playbook). It also means pharma can avoid vendor lock-in – but it should validate (per Part 11) that the chosen open model is fit for their intended use and does not contain unauthorized licenses.
- **Integrated Digital Workflows.** Private AI will be embedded into broader digital transformation. For example, AI-driven *digital twins* of biomanufacturing processes (virtual plants simulated by LLMs) are now being prototyped. Systems may span from on-site LLMs to field devices measuring equipment, all under a unified security envelope. Ensuring consistency of AI governance across these layers will be crucial. In the near term, expect to see more **MLOps** platforms tailored for private LLMs: offerings that handle data versioning, model lineage, and audit logging in a unified way under regulatory constraints (similar to how MLFlow tracks experiments but with biotech-aware extensions).
- **Emerging Threats and Mitigations.** As adversaries adapt, new threats will arise. For example, attackers might try to exploit large-but-not-air-gapped networks (i.e. one organization's internal network) as the next target, knowing they can't hit an internet-facing service. Defensive researchers will continue to innovate. Promising directions include hardware enclaves (like AMD SEV or Intel SGX) to run LLM inference in tamper-resistant zones, and **differential privacy** techniques to train models without risking

any individual's data. Also, **explainability tools** will improve, helping auditors understand LLM decisions in high-stakes contexts. Although generative AI is still young, the pace of innovation in security tools (some of which we've cited) suggests pharma can stay ahead if it commits resources.

In conclusion, **private LLM deployments in pharma represent a transformational but challenging frontier**. With suitable architecture and governance, companies can unlock AI's benefits while preserving trust and compliance. The next few years will see a convergence: regulators demanding strict accountability, enterprises demanding sovereign AI platforms, and technology vendors delivering the hardware/software building blocks. Forward-looking organizations will treat this not as a one-off project, but as a foundational capability – investing now in people, processes, and infrastructure to make AI a safe and sustainable part of pharmaceutical innovation.

Conclusion

Large language models hold tremendous promise for the pharmaceutical industry, from speeding up research to enhancing patient care. But the imperative to protect sensitive data and comply with strict regulations forces a departure from typical cloud-first AI strategies. **Air-gapped, private deployments** – carefully engineered with robust security – are emerging as the practical solution.

This guide has shown that such deployments demand expertise across multiple domains: IT architecture, cybersecurity, regulatory law, and ML operations. Key recommendations include:

- **Design for sovereignty:** Architect AI so that data never leaves approved boundaries, through on-premises or fully isolated private cloud setups (peerobyte.com) (www.itpro.com).
- **Enforce deep security:** Apply defense-in-depth specific to LLMs: from one-way data diodes and strict logging (as in classified defense applications (medium.com)) to multi-stage prompt filtering and RAG validation (cloudsecurityalliance.org) (arxiv.org).
- **Embed compliance by default:** Treat AI tools like any other regulated IT system. Validate them, keep audit trails (ALCOA+), and document every step (www.atlas-compliance.ai).
- **Leverage emerging platforms:** Use vendor solutions (AWS AI Factories, HPE Private AI) or mature open-source tools that support offline operation. Engage with tech partners who understand pharma security needs.
- **Train and govern:** Educate users and data scientists on safe AI usage; establish an AI governance board involving compliance, legal, and technical stakeholders.

By balancing innovation with careful adherence to sovereignty and security, pharma companies can reap AI's rewards without sacrificing trust. Ultimately, maintaining **data ownership and strict oversight** will not only prevent breaches and compliance failures, but also sustain *pharmaceutical innovation itself* – ensuring that new drugs and treatments are guided by AI insights while protected by the highest standards of confidentiality and integrity.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
2. https://csrc.nist.gov/glossary/term/air_gap
3. https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality
4. https://www.scilife.io/blog/ai-pharma-innovation-challenges
5. https://www.atlas-compliance.ai/blog/can-llm-based-quality-systems-pass-an-fda-part-11-audit
6. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/part-11-electronic-records-electronic-signatures-scope-and-application
7. https://www.techradar.com/pro/why-data-sovereignty-is-essential-to-help-businesses-prepare-for-impending-ai-regulation
8. https://www.cyber.gov.au/publication/fundamentals-of-cross-domain-solutions
9. https://cloudsecurityalliance.org/blog/2023/11/22/mitigating-security-risks-in-retrieval-augmented-generation-rag-llm-applications
10. https://arxiv.org/abs/2509.14285
11. https://genai.owasp.org/glossary/prompt-injection
12. https://blog.dreamfactory.com/government-and-defense-air-gapped-llm-data-access-dreamfactory
13. https://www.truefoundry.com/case-studies/fortune-100-healthcare-ships-llm-use-cases-truefoundry
14. https://www.techradar.com/pro/amazon-is-testing-out-private-on-premises-ai-factories
15. https://www.itpro.com/cloud/private-cloud/were-meeting-customers-where-they-are-hpe-expands-private-cloud-ai-service-with-new-sovereignty-controls-air-gapped-features
16. https://www.techradar.com/pro/building-private-ai-control-compliance-and-competitive-edge
17. https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html
18. https://www.hhs.gov/hipaa/for-professionals/faq/is-the-use-of-encryption-mandatory-in-the-security-rule/index.html
19. https://www.cnil.fr/en/sheet-ndeg15-take-account-legal-basis-technical-implementation
20. https://www.edpb.europa.eu/sme/be-compliant/international-data-transfers_en
21. https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act
22. https://dataprotection.ie/en/dpc-guidance/blogs/AI-LLMs-and-Data-Protection
23. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-integrity-and-compliance-drug-cgmp-questions-and-answers
24. https://medium.com/%40michael.hannecke/sovereign-ai-agent-security-air-gapped-deployments-and-enterprise-integration-efc770879cf8
25. https://medium.com/%40michael.hannecke/securing-ai-agents-monitoring-for-threats-you-cant-unit-test-0674d4a3c762
26. https://presidio.dataprivacystack.org
27. https://github.com/unitaryai/detoxify
28. https://learn.microsoft.com/en-us/azure/ai-services/content-safety/overview
29. https://learn.microsoft.com/en-us/azure/ai-services/content-safety/limited-access
30. https://learn.microsoft.com/en-us/azure/ai-services/content-safety/how-to/containers/container-overview

31. https://www.axios.com/2025/05/12/fda-ai-drugs-company-data-questions
32. https://peerobyte.com/blog/eu-ai-act-for-cloud-infrastructure-what-you-need-to-get-done-by-2025-2026-and-how-it-affects-architecture-and-logs
33. https://airc.nist.gov/airmf-resources/playbook/measure

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.