# NC ASK – AI Security & Governance Guiding Principles (v0.1)

**Owner:** NC ASK Program (Product Owner + Security/Privacy Leads)  
**Audience:** Builders, researchers, and stakeholders working on NC ASK (web/mobile)  
**Scope:** Educational, non‑diagnostic AI features; data, models, infrastructure; research metadata

## 0) One‑page TL;DR

**Purpose.** Keep families’ and providers’ information safe; provide transparent, equitable, and accountable AI; and align with UNC expectations and applicable laws (HIPAA/FERPA where in scope).  
**What this means in practice.** 1. **No sensitive data in public LLMs.** PHI/FERPA‑covered or confidential data **never** leaves approved private environments.  
2. **Human accountability.** People, not models, remain responsible for decisions; NC ASK is an educational tool—not medical or legal advice.  
3. **Grounded answers only.** Responses are restricted to approved sources with citations; avoid speculation.  
4. **Least‑privilege by default.** Tight identity/access, network isolation, secrets management, and audit logging.  
5. **LLM‑specific safeguards.** Prevent prompt injection, insecure output handling, and data leakage; red team regularly.  
6. **Govern by lifecycle.** Inventory every dataset/model, document it, evaluate it, monitor it, and retire it safely.  
7. **Equity & accessibility.** Design for diverse users and guard against bias; follow WCAG; offer plain‑language options.  
8. **Change control.** Version everything (prompts, models, data, UI guardrails) and review before release.

## 1) Purpose & Principles

* **Safety & Privacy First.** Protect users’ privacy and confidentiality; minimize data collection; default to de‑identified processing.
* **Transparency & Disclosure.** Clearly disclose AI use, data sources, limitations, and intended use (education, not clinical/legal advice).
* **Accountability.** Assign clear human owners for data, models, prompts, and releases; maintain decision trails.
* **Equity, Inclusion & Accessibility.** Mitigate representational harms; support multiple reading levels; adhere to WCAG 2.2 AA.
* **Security‑by‑Design.** Embed controls across the lifecycle: identity, network, code, data, and model safeguards.
* **Reliability & Evaluations.** Ground outputs in approved sources; continuously test quality, bias, and safety.
* **Sustainability.** Prefer efficient compute, caching, and right‑sized models; publish environmental considerations.

### 1A) Risk‑Scaled Deployment Tiers (Research vs Production)

Adopt a **two‑tier governance model** so controls scale with risk and intended audience.

**Tier R – Research / Exploration**  
Purpose: internal prototyping, feasibility studies, model/prompt evaluation.  
Data: no PHI/FERPA/regulated data; de‑identified or synthetic only.  
Environment: private, access‑controlled workspace; egress‑restricted; logging enabled.  
Controls: lightweight change control; preliminary red‑team; evals before external sharing.  
Approvals: Product Owner + Security/Privacy Lead sign‑off for data onboarding.  
UI: clearly watermarked *Research / Not for public use*; no public endpoints.

**Tier P – Production / Public‑Facing**  
Purpose: end‑user features (families/providers) and published tools.  
Data: curated, documented corpora; zero regulated user input by default; DLP on telemetry.  
Environment: hardened network paths, WAF/API gateway, secret rotation, SBOMs, signed images.  
Controls: formal change control; full adversarial testing; uptime/SLOs; rollback; incident runbooks.  
Approvals: Executive Sponsor + Security/Privacy Lead; compliance check (e.g., UNC guidance, IRB where applicable).  
UI: educational‑use disclaimers, source transparency, and report‑harm channel.

## 2) Roles & Responsibilities

* **Executive Sponsor:** Approves risk appetite and resourcing; signs off on major policy changes.
* **Product Owner (NC ASK):** Owns roadmap, scope, and risk acceptance; coordinates with Privacy/IRB/Compliance.
* **Security Lead:** Designs and validates security controls, incident response, and red‑team program.
* **Privacy/Compliance Lead:** Oversees HIPAA/FERPA applicability, consent/notice language, data retention, and IRB alignment.
* **Data Steward(s):** Own data catalog, classification, provenance, quality checks, and data‑sharing agreements.
* **Model Owner(s):** Maintain model cards, prompt libraries, eval suites, drift monitors, and release notes.
* **Engineering Lead:** Implements controls (identity, network, secrets, CI/CD), code reviews, and dependency hygiene.
* **Clinical & Community Advisors:** Review content safety, bias, and usability; help set red‑flags/escalation rules.

## 3) Data Governance & Classification

* **Classify all inputs/outputs.** PHI, FERPA, confidential, internal, public. Default treat as confidential unless proven otherwise.
* **Data minimization.** Collect the least necessary; prefer session‑only or anonymized analytics; make metadata opt‑in where feasible.
* **Approved storage/processing.** Restricted to approved private cloud resources; no copying to personal devices or public tools.
* **Provenance & licenses.** Track source, license, and update cadence for each document; record last‑review dates.
* **Retention & deletion.** Time‑bound retention with documented purging; per‑category schedules; secure deletion on retirement.

**Consent & Notices**  
- Plain‑language consent/notice at first use; link to privacy policy; flag research metadata collection; provide contact for questions.  
- Provide opt‑out where feasible without degrading safety; disclose third‑party processing.

## 4) Model & Prompt Lifecycle Governance

* **Inventory & model cards.** For every model (base, fine‑tuned, or retrieval‑augmented), maintain purpose, data scope, limitations, and known failure modes.
* **Prompt library & versioning.** Treat system prompts and RAG instructions as code; review, test, and version‑control them.
* **Evaluation gates.** Ship only after passing accuracy, harmlessness, bias/fairness, accessibility, and latency SLOs.
* **Monitoring & drift.** Watch for content regressions, hallucinations, prompt‑leak signs, and source coverage gaps; roll back if needed.
* **Red teaming.** Periodically test jailbreaks, prompt injection, data exfiltration, toxic content, and social‑engineering vectors.
* **Decommissioning.** Archive artifacts, revoke secrets, and delete cached embeddings and derived datasets.

## 5) LLM‑Specific Security Guardrails

* **Trust boundaries for RAG.**
  + Separate untrusted user content from trusted corpora.
  + Sanitize/normalize documents before indexing; strip active content.
  + Annotate chunks with provenance; restrict what can be cited.
* **Prompt‑injection defenses.**
  + Layered instructions: immutable system prompt + retrieval‑only mode where possible.
  + Pre‑/post‑processing to detect and neutralize “ignore previous” and exfiltration attempts.
  + Disallow model‑executed code/links unless explicitly sandboxed and approved.
* **Insecure output handling.**
  + Treat LLM output as **untrusted**; sanitize before rendering; never auto‑execute commands or HTML/JS.
  + Require human confirmation for any action with side effects (emails, forms, API calls).
* **Data leakage controls.**
  + No training/fine‑tuning on sensitive logs by default; use de‑identified corpora with DLP scanning.
  + Mask/replace identifiers in telemetry; segregate analytics from content stores.
* **Safety filters & content policy.**
  + Enable toxicity/self‑harm filters; block medical/legal instruction beyond educational scope; embed escalation guidance.

## 6) Platform & Engineering Controls (reference architecture‑agnostic)

* **Identity & Access.** SSO, MFA, RBAC/ABAC; least privilege; time‑bounded tokens; just‑in‑time elevation; break‑glass accounts.
* **Network & Egress.** Private networking, egress allow‑lists, VPC/VNet peering, WAF, API gateways; deny default outbound where possible.
* **Secrets & Keys.** Centralized secret manager; key rotation; no secrets in code or prompts; signed releases.
* **Supply Chain.** SBOMs, dependency pinning, vulnerability scanning, and license checks; container image signing.
* **Logging & Audit.** Immutable logs for admin, data access, and model calls; retention aligned to policy; tamper‑evident storage.
* **CI/CD & Testing.** Static/dynamic scans, policy‑as‑code, unit/integration/e2e + adversarial tests; gated deployments.
* **Resilience.** Backups, runbooks, chaos/fire‑drill tests, and RTO/RPO objectives; health probes and circuit breakers for models.

## 7) Safety, Ethics, and Equity

* **Bias & representation.** Evaluate for disparate impact across race/ethnicity, language, disability, and county/rural status; document mitigation.
* **Inclusive design.** Offer plain‑language modes, multilingual support where possible, and visual aids; avoid stigmatizing language.
* **Human in the loop.** Provide escalation pathways to clinicians, advocates, or help lines; flag red‑line situations (e.g., safety concerns).
* **Documentation & disclosure.** Publish model/system cards and data source lists; note known limitations and update cadence.

## 8) Incident Response & Reporting

* **Triggers.** Data exposure, model behavior causing harm, integrity breaches, unauthorized access, or content policy violations.
* **Process.** Contain, eradicate, recover; notify stakeholders; document lessons; update controls; communicate externally per policy.
* **User remedies.** Provide a mechanism for users to report harms, request corrections, or delete their data where applicable.

## 9) Compliance Mapping (for governance alignment)

* **Risk Management frameworks.** Map controls to recognized frameworks (e.g., GOVERN/MAP/MEASURE/MANAGE lifecycle).
* **LLM security posture.** Track mitigations against common LLM risks (e.g., prompt injection, insecure output handling, training data poisoning, excessive agency).
* **Institutional policies.** Align disclosures with University research and operations guidance; confirm IRB status for metadata studies.

## 10) Versioning & Change Control

* Maintain a changelog; major changes require Security/Privacy sign‑off and stakeholder review.
* Version prompts, models, datasets, guardrails, and UI notices; archive superseded versions for audit.

## 11) Public‑Facing Disclaimers (UI language starter)

* **Educational use only.** NC ASK provides educational information and resource navigation; it is **not** medical or legal advice and does not replace clinical care.
* **Source transparency.** Responses summarize approved sources and may be incomplete; verify with your care team or official agencies.
* **Privacy note.** Do not include personal or identifying information in free‑text questions.

## 12) Appendices (to be added as artifacts mature)

* A. Data catalog & classification register
* B. Model and prompt cards
* C. Evaluation suite & thresholds
* D. Red‑team playbook
* E. Incident response runbook
* F. Accessibility and plain‑language checklists
* G. Compliance crosswalk (institutional + regulatory)
* **H. OWASP LLM Top‑10 → NC ASK Controls Crosswalk (one‑page)**
* **I. UNC Reviewer Packet (cover letter + framework table)**

## 13) OWASP LLM Top‑10 → NC ASK Controls Crosswalk (one‑page)

| OWASP Risk (2025) | Example failure mode | Primary controls in this doc | Control owner(s) | Evidence/artifact |
| --- | --- | --- | --- | --- |
| **LLM01 Prompt Injection** | Indirect prompt in retrieved doc alters behavior | §5 Prompt‑injection defenses; §4 Eval gates; §6 CI/CD | Security Lead, Model Owner, Eng Lead | Red‑team logs; blocked patterns; eval results |
| **LLM02 Sensitive Information Disclosure** | Model echoes secrets/PHI from context | §3 Data classification; §5 Data‑leakage controls; §6 Logging | Privacy Lead, Data Steward, Security Lead | DLP scans; masked telemetry; access audits |
| **LLM03 Supply Chain** | Compromised model/container or dependency | §6 Supply chain (SBOM, signing); §6 CI/CD | Eng Lead, Security Lead | SBOM; signature attestation; vuln scans |
| **LLM04 Data & Model Poisoning** | Poisoned fine‑tune or embeddings | §3 Provenance; §4 Model cards; §5 Trust boundaries | Data Steward, Model Owner | Source registry; dataset checksums; review log |
| **LLM05 Improper Output Handling** | XSS/HTML injection via model output | §5 Insecure output handling; §6 WAF | Eng Lead, Security Lead | Output sanitizer tests; e2e security tests |
| **LLM06 Excessive Agency** | Unchecked tool use causes side‑effects | §5 Human‑in‑the‑loop; sandboxing; confirmations | Product Owner, Eng Lead | Action confirmation UX; sandbox logs |
| **LLM07 System Prompt Leakage** | Prompt revealed via jailbreak or error | §4 Prompt versioning; §5 Layered instructions | Model Owner, Security Lead | Prompt diffs; leak tests; error‑message review |
| **LLM08 Vector/Embedding Weaknesses** | Access‑control bypass via embeddings | §5 Trust boundaries; §6 Network/Egress | Eng Lead, Security Lead | ACL tests; retrieval audit; index ACLs |
| **LLM09 Misinformation** | Hallucinated citations/resources | §1 Transparency; §4 Evaluation gates; RAG source pinning | Model Owner, Clinical Advisors | Accuracy evals; citation checks |
| **LLM10 Unbounded Consumption** | Token/compute exhaustion (DoS) | §6 Resilience; rate limits; circuit breakers | Eng Lead, Security Lead | Rate‑limit config; chaos drills |

*Reference:* OWASP **Top 10 for LLMs 2025** (Prompt Injection, Sensitive Information Disclosure, Supply Chain, Data & Model Poisoning, Improper Output Handling, Excessive Agency, System Prompt Leakage, Vector & Embedding Weaknesses, Misinformation, Unbounded Consumption).

## 14) UNC Reviewer Packet (for institutional review)

### A) Cover Letter Template (paste to pdf/email)

**Subject:** NC ASK – Security & Governance Summary and Request for Review  
**To:** UNC Privacy/Compliance, Research Integrity, IT Security (as applicable)  
**From:** [Executive Sponsor], [Product Owner], [Security Lead], [Privacy Lead]  
**Summary:** NC ASK is an educational, non‑diagnostic digital platform supporting families and providers. This packet summarizes data classification, model governance, security controls, evaluation results, and incident response. We request review/acknowledgment prior to moving from Tier R (Research) to Tier P (Production).

**Scope & Data:** Data classes in scope (public, internal); explicit exclusions (PHI/FERPA). Provenance and license registry attached.  
**Controls:** Identity/least‑privilege, network egress allow‑lists, secrets management, SBOM/signing, logging/audit, eval gates, red‑team cadence.  
**LLM Safeguards:** Prompt‑injection defenses, retrieval trust boundaries, output sanitization, tool‑use confirmation, leakage/DLP, model cards.  
**Risk & Frameworks:** Controls mapped to NIST AI RMF functions and OWASP LLM Top‑10 (see tables).  
**Contact & Approvals Needed:** Names/roles for sign‑off; planned go‑live date; rollback plan.

### B) Framework Table (excerpt)

| Area | NC ASK control | NIST AI RMF | UNC guidance linkage |
| --- | --- | --- | --- |
| Data minimization & classification | §3 Data governance | **GOVERN/MAP** | Research & Staff guidance: privacy, sensitive data handling |
| Transparent AI use / disclaimers | §11 UI language | **GOVERN** | Research/Teaching: disclosure & authorship expectations |
| Evaluation gates (accuracy, safety, bias) | §4 Eval gates | **MEASURE/MANAGE** | Research: quality & reproducibility |
| Prompt‑injection & output sanitization | §5 LLM guardrails | **MANAGE** | Operations/Staff: security safeguards |
| Change control & versioning | §10 Versioning | **GOVERN/MANAGE** | Operations: change mgmt & accountability |
| Incident response & user remedies | §8 IR plan | **MANAGE** | Institutional incident reporting norms |

Attachments suggested: model card(s), red‑team summary, eval results, SBOM attestation, data catalog excerpt.