SPECIFIC STANDARDS

FOR DIAGNOSTIC APPLICATIONS LEVERAGING LARGE LANGUAGE MODELS (LLMS) IN THE FREE HEALTH CHECK PROGRAMME

Table of Contents

1. Background

Health data are at the core of Indonesia's digital health transformation agenda under the Digital Health Transformation Strategy (DHTS) 2.0, which seeks to build an integrated, interoperable, and secure health data ecosystem. The Framework for AI-assisted Diagnostic Data Management in Indonesia defines the overarching governance, quality, and ethical foundations for how diagnostic data are collected, curated, stored, shared, and monitored. It aligns with the Personal Data Protection Law (UU PDP 2022), Ministry of Health data governance policies, and international guidance such as the WHO Ethics and Governance of AI for Health, as well as FAIR and CARE principles. Within this broader policy and governance environment, large language models (LLMs) are recognized as powerful tools that can support clinical decision-making, but only when grounded in robust data management and clear accountability mechanisms.

At the same time, Indonesia's diagnostic data remain fragmented across laboratory systems, imaging platforms, and program-specific registries, including those used for population-based screening initiatives. This fragmentation can undermine AI performance by limiting data completeness, interoperability, and representativeness, and by increasing the risk of bias, inconsistent outputs, and unsafe recommendations. The framework responds to these challenges by setting expectations across the entire diagnostic data lifecycle—collection, cleaning and validation, annotation and structuring, storage and versioning, sharing and reuse, and continuous monitoring and feedback. It clarifies the roles of the Ministry of Health, Pusdatin, program managers, health facilities, and system developers in ensuring that diagnostic data used for AI applications are accurate at the source, properly standardized, and used ethically and transparently.

Within this context, the Cek Kesehatan Gratis (Free Health Check) programme is a critical use case for operationalizing AI-assisted diagnostics at scale. The programme spans the full life course—from newborn and

infant screening through school-age children, adolescents, adults in community settings (with emphasis on men), to geriatric screening for older adults—with diverse data types ranging from structured measurements (e.g., anthropometry, blood pressure, blood sugar) to semi-structured questionnaires (e.g., lifestyle, mental health, reproductive health) and clinical observations. Integrating LLMs into this ecosystem offers significant potential to synthesize these heterogeneous data, flag risks, surface missed patterns, and support health workers with context-aware recommendations. However, it also amplifies the need for clear, technical standards that specify what data must be supported, how LLMs may operate on them, and what minimum safeguards are required before use in real-world screening workflows.

The LLM Diagnostic Standards for the Free Health Check programme translate the high-level principles of the framework into concrete, testable technical requirements. They define, in normative language (SHALL/SHOULD/MAY), the minimum set of data elements and formats that LLM-enabled applications must ingest; the allowed diagnostic support functions (such as risk stratification, summarization of findings, and generation of follow-up suggestions); the required structure and explainability of outputs; and the safety and governance controls that must be in place, including auditability, human oversight, and mechanisms for error reporting and correction. These standards also address interoperability with SATUSEHAT and other national systems, version control for models and prompts, and alignment with privacy-by-design principles to ensure that LLM operations on CKG data remain compliant with UU PDP 2022.

By anchoring technical requirements in the existing framework and national policy environment, this standard creates a bridge between governance and implementation. It provides programme owners, clinical leaders, digital health and interoperability teams, AI/LLM developers, and vendors with a common reference for designing, procuring, and evaluating LLM-enabled diagnostic tools. At the same time, it offers regulators,

monitoring and evaluation teams, and other oversight bodies a clear set of criteria against which to assess compliance, safety, and performance over time. In doing so, the standard helps ensure that the use of LLMs in the Free Health Check programme not only accelerates innovation, but also strengthens public trust, clinical quality, and long-term sustainability of Indonesia's digital health system.

2. Purpose, Scope, and Application

2.1 Purpose

This document defines specific, minimum standards that any diagnostic application using large language models (LLMs) must meet when operating on data from the Cek Kesehatan Gratis/ Free Health Check programme. It translates the broader technical guidance into clear, testable requirements so that implementers, vendors, and internal entities know exactly what data must be supported, what LLM functions are allowed, how outputs must be structured, and what safety and governance controls are mandatory before such systems can be used in real screening workflows.

2.2 Scope

These standards apply to LLM-enabled diagnostic support across the full Free Health Check lifecycle and settings, including newborn and infant screening (growth, congenital heart, SHK/ G6PD/ HAK, bile duct abnormalities), toddlers and preschool/ early school-age children (growth, development, eye/ ear, dental, TB and thalassemia), school-age and high school adolescents (demographics, lifestyle, mental health, reproductive health, NTD, liver, nutrition, blood pressure, blood sugar, anemia, hearing, vision, dental, fitness), adults in community settings with emphasis on men (smoking, physical activity, TB, liver, cardio-metabolic, respiratory, sensory, dental, and cancer screening), and older adults through geriatric screening (cognition, mobility, malnutrition, depression, functional status, and relevant cancer or chronic disease screening).

2.3 Intended Audience

The primary audience for this document is programme owners and clinical leaders responsible for Free Health Check design and quality, along with digital health and interoperability teams, data stewards, AI/ LLM developers, and vendors who design, build, or procure diagnostic applications. It is also relevant to monitoring and evaluation teams, regulators, and governance bodies who need a concise reference to assess whether proposed or existing LLM-based solutions comply with national expectations for data handling, diagnostic support functions, and safety controls.

2.4 Use of Normative Language (SHALL/ SHOULD/ MAY)

Throughout this document, requirements are expressed using standard normative terms: “SHALL” indicates a mandatory requirement that must be met for an LLM-enabled diagnostic application to be considered compliant; “SHOULD” indicates a strong recommendation that applies in most situations but may be deviated from with clear justification; and “MAY” indicates an optional feature or behaviour. Implementers are expected to treat “SHALL” clauses as non-negotiable, document any deviations from “SHOULD” clauses, and use “MAY” clauses as guidance for future enhancement rather than minimum compliance.

3. Data and Coding Standards for Free Health Check LLM Use

This section defines the minimum data elements and coding practices that any LLM-enabled diagnostic application must support when working with Free Health Check data. The goal is to ensure that all age-, sex- and group-specific forms are represented in a consistent, machine-readable way so the LLM can reliably interpret growth, development, mental health, infectious disease, cardio-metabolic, cancer, and geriatric information.

3.1 Identity and Context Data

The application shall correctly capture and process identity and context data so that each screening encounter is uniquely identifiable, linkable over time, and associated with the correct individual, facility, and programme.

- Health facility identity and code (Identitas Faskes, Kode Faskes)
- School identity where applicable (Identitas Sekolah)
- Participant identity by life stage (Identitas Balita, Anak Sekolah, Dewasa, Lansia)
- Guardian/ relative identity where applicable (Identitas Wali)
- National Identification Number (NIK)
- Date of birth (Tanggal Lahir) and derived age/ age group
- Sex (Jenis Kelamin)
- Screening date (Tanggal Pemeriksaan) using DD/ MM/ YYYY format

3.2 Minimum Clinical Data Sets by Life Stage

The application shall support a core set of clinical variables for each life stage so that key screening programmes can be run end-to-end. Not every optional field must be used in every implementation, but these groups

define the minimum expected data for safe and meaningful LLM-based support.

3.2.1 Newborn and Infant Screening

For newborns and infants, the focus is on early detection of growth problems, congenital and metabolic disorders, and selected infectious risks.

- Growth and anthropometry: weight, length/ height, head circumference, basic growth classification (e.g., BB/ U, PB/ U or TB/ U, BB/ PB or BB/ TB)
- Congenital and metabolic screening: congenital heart screening; SHK, G6PD, HAK screening and confirmation tests when positive
- Bile duct abnormalities: bile duct screening at 0–2 days, 14–28 days, and 2–4 months
- Infectious disease (where present in forms): TB-related items for infants

3.2.2 Toddlers, Preschool, and Early School-Age Children

In younger children, the minimum dataset covers growth, development and behaviour, basic sensory and dental health, and early TB and blood-related problems.

- Growth: weight, height/ length, head circumference, derived growth indicators and nutritional status
- Development and behaviour: KPSP result; behavioural/ emotional questions (tantrums, aggression, emotional changes); KMPE and GPPH results where indicated
- Eye and ear: hearing test result; white pupil examination; basic vision screening where used
- Dental: dental screening result

- Tuberculosis: TB in infants and preschool children; TB signs (e.g., lymph node enlargement, bone/ joint swelling); Mantoux test induration and result; TB scoring when Mantoux is done
- Hematology and thalassemia: random blood sugar (if present); hemoglobin (Hb) result; thalassemia screening (rapid Hb) and follow-up
- Immunisation: general immunisation status where recorded

3.2.3 School-Age and High School Adolescents

For adolescents, the dataset adds mental health, reproductive health, NTDs, lifestyle risks, and early cardio-metabolic parameters on top of growth and sensory/ dental items.

- Demographics and lifestyle: student identity and context; smoking history/ behaviour; physical activity level; fitness tests (*kebugaran*)
- Mental health: Mini MINDHEAR Youth Scale A (anxiety) and B (depression)
- Reproductive health: reproductive health for girls and boys; HPV immunisation history
- Infectious disease and NTD: TB history; history of hepatitis B and C; family history of thalassemia; NTD variables (*kusta*, *skabies*, malaria); malaria risk factors and RDT malaria results
- Growth, nutrition, and cardio-metabolic: weight, height and growth indicators; blood pressure; random blood sugar; anemia checks by grade/ sex as specified
- Sensory and dental: hearing acuity examination; visual acuity examination; dental examination
- Immunisation: general immunisation status where recorded

3.2.4 Adults in Community (Emphasis on Men)

For adults, especially men, the minimum dataset supports integrated assessment of lifestyle risks, infectious and liver disease, cardio-metabolic risk, respiratory risk, sensory and dental health, and key cancers.

- Identity and context: adult identity and demographics, including linkage to facility and programme
- Risk factors and history: smoking behaviour; physical activity level; TB history and examination (including sputum when present); history of diabetes mellitus
- Liver and infectious disease: hepatitis B and C status; cirrhosis indicators; hepatitis and fibrosis/ cirrhosis lab results; syphilis and HIV rapid test results where used (e.g., premarital testing)
- Cardio-metabolic: nutrition (weight, height, waist circumference); blood pressure; blood sugar (initial and follow-up: random, GDP, GD 2PP, HbA1c); kidney function tests for relevant risk groups; lipid profile for HT/ DM; variables required for heart and stroke risk prediction
- Respiratory: COPD/ PUMA screening items (age, smoking, symptoms, activity)
- Sensory and dental: ear and eye screening by age band (including pupil exam, visual acuity tests); dental screening by age band
- Cancer screening: colorectal cancer screening and follow-up; lung cancer screening; cervical cancer screening (IVA, HPV DNA) for eligible women; breast cancer screening (breast exam, SADANIS, breast ultrasound)
- Mental health: adult mental health screening where included in Free Health Check forms

3.2.5 Older Adults and Geriatric Screening

In older adults (age 60 and above), the minimum dataset focuses on geriatric syndromes and functional decline, alongside any ongoing cancer or chronic disease screening.

- Cognition: cognitive decline screening; Mini-Cog/ clock draw follow-up; AD-8 INA follow-up
- Mobility: mobility limitation; SPBB follow-up examination
- Nutrition: malnutrition screening; MNSF follow-up examination
- Mood: depression symptom screening; follow-up depression assessment
- Function: Barthel Index for functional status
- Additional cancer and chronic disease screening variables included in the geriatric form

3.3 Derived Indicators and Scores

The application shall either calculate or correctly ingest a set of derived indicators and scores that turn raw measurements and questionnaire responses into clinically meaningful categories for the LLM to use.

- Growth indices (e.g., weight-for-age, height-for-age, weight-for-height or weight-for-length)
- BMI and waist-based indicators for adolescents and adults
- Questionnaire and scale scores (KPSP, KMPE, GPPH, Mini MINDHEAR, geriatric tools such as Mini-Cog, AD-8 INA, SPBB, MNSF, Barthel Index)
- TB scoring (including Mantoux-based score where applicable)
- COPD/ PUMA scores
- Composite heart and stroke risk scores and other cardio-metabolic indices, where implemented

3.4 Data Quality and Validation Requirements

To support safe diagnostic support, all data used by the LLM must be captured and stored with consistent structure, explicit units, and basic validation. The application shall enforce or check these rules as part of its normal operation.

- All numeric measurements (e.g., anthropometrics, blood pressure, blood sugar, Hb, kidney and liver function, lipids) must have explicit units and be constrained to plausible ranges.
- Categorical results (e.g., positive/ negative, yes/ no, normal/ abnormal, present/ absent) must use controlled vocabularies rather than free text.
- Key fields for identity, dates, life stage, and sex must be complete for each screening encounter, or the system must flag the record as incomplete.
- Missing, inconsistent, or implausible values must be detected and flagged so they are not silently treated as valid inputs in LLM-based decision support.

4. Functional Standards for LLM-Based Diagnostic Support

This section defines how LLMs are allowed to behave when working with Free Health Check data. It specifies the roles the LLM may play, how it must use deterministic tools and calculators, and what kinds of outputs it must produce so that screening and diagnostic support remain consistent, explainable, and under human control across all age groups and programme areas.

4.1 Permitted Roles of the LLM

The LLM is a decision-support component that works only on approved data and rules; it does not replace clinical judgement or national guidelines.

The application SHALL ensure that the LLM:

- Uses only data provided via the Health Data Graph (HDG) or approved APIs for the current participant and encounter.
- Summarises screening data into concise, clinically meaningful snapshots (e.g., main growth, development, infectious, cardio-metabolic, cancer, or geriatric findings).
- Interprets findings strictly according to approved rules, calculators, and thresholds for each domain and life stage.
- Proposes triage and follow-up actions as recommendations for health worker confirmation, not as automatic orders.
- Provides clear explanations of why a case has been classified as normal, at risk, high risk, or urgent, referencing key variables or scores.

The application SHALL ensure that the LLM does NOT:

- Invent new diagnostic criteria, thresholds, or screening pathways.
- Provide final diagnostic labels or prescribe medication.
- Ignore missing or clearly inconsistent critical data without flagging this limitation in its outputs.

4.2 Use of Tools and Calculators

Deterministic tools and calculators implement the formal logic of growth charts, scores, and screening rules; the LLM must orchestrate these tools, not replace them.

The application SHALL:

- Implement or integrate deterministic modules for:
    - Growth and nutritional assessment (e.g., W/ A, H/ A, W/ H or W/ L, BMI, waist-based indices).
    - Developmental and behavioural tools (e.g., KPSP, KMPE, GPPH).
    - Mental health scales (e.g., Mini MINDHEAR, geriatric depression tools).
    - Disease-specific screening logic (e.g., TB scoring including Mantoux, PUMA/ COPD, liver risk, thalassemia follow-up).
    - Cardio-metabolic and cardiovascular risk calculators (heart and stroke risk prediction).
    - Cancer screening pathways (colorectal, lung, cervical, breast) according to programme rules.
    - Geriatric assessment tools (Mini-Cog, AD-8 INA, SPBB, MNSF, Barthel Index).
- Treat these tools as the authoritative source of numeric scores and categorical risk outputs for their respective domains.

The application SHALL ensure that the LLM:

- Selects and calls appropriate tools based on age group, sex, screening form type, and available data.
- Passes clean, validated input data from the CKG into tools and uses returned scores and categories unchanged.
- Builds narrative explanations and recommendations on top of tool outputs, without altering underlying numeric results or thresholds.

4.3 Standard Output Types

LLM-supported applications must produce predictable, structured outputs that can be stored, audited, and integrated into Free Health Check workflows and national systems. Additionally, the output should include information about the LLM version used, to ensure reproducibility and consistency across different executions.

For each screening encounter, the application SHALL output at minimum:

- **Structured risk categories per domain**
    - Domain-specific risk labels (e.g., growth/ nutrition, development/ behaviour, mental health, infectious disease, cardio-metabolic, cancer, geriatric) using a standard scale such as:
        - Normal/ within expected range
        - At risk/ needs monitoring
        - High risk/ needs further evaluation
        - Urgent/ requires immediate attention (for defined red-flag situations only).
- **Recommended actions**
    - A list of proposed next steps, each with:
        - Type (e.g., repeat measurement, confirmatory test, referral, health education).
        - Priority (e.g., routine, soon, urgent).
        - Suggested timeframe (where applicable).
- **Two levels of narrative summary**
    - A clinician-facing summary explaining key findings, relevant scores, risk categories, and the rationale for recommendations.
    - Where applicable, a participant-facing explanation in simpler language, aligned with approved health education messages.
- **Machine-readable structure**

- All risk categories, flags, and recommendations encoded in a structured format that can be linked back to the participant, encounter, and underlying data, and stored in the CKG and connected systems for audit, monitoring, and future analysis.

5. Evaluation and Safety Standards

This section defines how LLM-enabled diagnostic applications must be tested and operated to ensure that they are accurate, safe, and fair when used with Free Health Check data. It covers performance and calibration requirements, safety and bias controls, and how incidents are reported and handled across all life stages, from newborn screening to geriatric assessments.

5.1 Performance and Calibration Requirements

The application shall be evaluated and calibrated so that its risk classifications and recommendations are reliable for the populations and screening programmes in which it is used.

The application SHALL:

- **Separate evaluation of components**
    - Evaluate deterministic tools and calculators (e.g., growth indices, KPSP/ KMPE/ GPPH, Mini MINDHEAR, TB scoring including Mantoux, PUMA, cardio-metabolic and cancer risk tools, geriatric scales) independently to confirm correct implementation of clinical logic.
    - Evaluate the LLM layer on its use of these tools, its consistency with thresholds, and the correctness of its risk categories and recommendations.
- **Define domain-specific performance targets**
    - Agree minimum acceptable levels of sensitivity, specificity, and (where applicable) positive and negative predictive values for key flags in each domain (e.g., TB suspicion, high cardio-metabolic risk, severe depression, suspected cancer, geriatric impairment).

    - Document these targets for each life stage and screening programme where the LLM is deployed.
- **Conduct pre-deployment testing**
    - Run the full system on representative historical or pilot data from Free Health Check forms (covering newborn, child, adolescent, adult, and geriatric cohorts).
    - Compare system outputs with reference decisions (e.g., expert review, existing protocols) to confirm performance against agreed targets before any live use.
- **Check calibration where outcome data are available**
    - For risk scores that predict events or conditions (e.g., heart and stroke risk, TB suspicion, severe malnutrition), assess whether predicted risk categories align with observed outcomes over time in the pilot sites.
    - Adjust thresholds or calibration parameters through a controlled process if systematic under- or over-estimation is observed.
- **Use clear acceptance criteria for rollout**
    - Define in advance the criteria under which a domain or life-stage configuration may move from pilot to wider deployment (e.g., minimum performance metrics, absence of critical safety issues, acceptable user feedback).
    - Document and obtain governance approval before progressing.

5.2 Safety, Bias, and Monitoring

The application shall enforce safety constraints and be monitored for potential bias or drift, ensuring that no group is systematically harmed or disadvantaged by LLM-supported decisions.

The application SHALL:

- **Apply conservative behaviour with incomplete or inconsistent data**
    - Detect missing or contradictory critical inputs (e.g., key growth measures, blood pressure, blood sugar, TB test results, major questionnaire items) and avoid strong conclusions when these are absent.
    - Favour outputs such as “insufficient data – complete screening items” rather than assigning false reassurance or unwarranted high-risk labels.
- **Implement red-flag safety rules**
    - Encode and enforce predefined red-flag patterns (e.g., very high blood pressure, extremely low Hb, positive TB tests with compatible symptoms, high Mini MINDHEAR scores, strong cancer suspicion, severe geriatric impairment).
    - Ensure that red flags always trigger clear urgent or high-priority recommendations and explicit prompts for immediate human review.
- **Monitor for bias and inequity**
    - Periodically analyse outputs by age group, sex, facility type, region, and other relevant factors to identify systematic differences in risk categorisation, referral rates, or recommended actions.
    - Investigate whether observed differences are clinically justified (e.g., due to underlying prevalence) or indicate bias in data, tools, or LLM behaviour.
- **Provide health workers with override capability**
    - Allow health workers to accept, modify, or reject LLM-supported recommendations for each case.
    - Record overrides in a structured way to support later analysis of patterns (e.g., frequent rejection of certain recommendations in specific domains).
- **Monitor model and system behaviour in routine use**

- Track distributions of risk categories, number and type of recommendations, and changes in these patterns over time and across sites.
- Establish thresholds for abnormal patterns (e.g., sudden drop in high-risk TB flags or surge in urgent cardio-metabolic alerts) that prompt technical and clinical review.

5.3 Incident Management

There shall be a clear and documented process to handle cases where the LLM or associated tools produce unsafe, clearly inappropriate, or unexpected outputs, so that problems are corrected quickly and lessons inform future improvements.

The application and its operating environment SHALL:

- **Enable easy reporting of incidents**
  - Provide mechanisms for health workers and programme staff to report cases where recommendations are clearly unsafe, contradict obvious clinical facts, or conflict with programme guidelines.
  - Allow these reports to include identifiers for the participant and encounter, a description of the issue, and the user's corrective action.
- **Log and preserve full context for investigation**
  - For reported incidents, ensure that relevant logs are available, including:
    - Input data from the CKG used in the decision,
    - Tools and calculators invoked and their outputs,
    - LLM version and configuration,
    - Final risk categories and recommendations shown to the user,
    - Any overrides or manual actions taken.

- **Conduct structured root-cause analysis**
    - Review reported incidents with both technical and clinical stakeholders to determine whether the cause lies in data quality, rule implementation, tool configuration, LLM prompting, or user misunderstanding.
    - Classify incidents by severity and domain to prioritise corrective actions.
- **Implement corrective and preventive actions**
    - Apply fixes such as data validation tightening, tool or rule corrections, prompt adjustments, updated thresholds, or targeted training for users.
    - For serious issues, temporarily disable affected functionalities (e.g., a specific domain or life stage module) until a fix is validated.
- **Feed lessons back into governance and monitoring**
    - Summarise incident patterns regularly for the governance body, along with implemented corrective actions.
    - Use these insights to refine evaluation procedures, update guidance, and improve future versions of LLM-enabled diagnostic applications.

6. Governance, Roles, and Deployment Requirements

This section defines how LLM-enabled diagnostic applications for the Free Health Check must be governed, who is responsible for what, how audit trails are maintained, and how deployment and changes are controlled. The aim is to ensure that any system used in practice is not just technically sound but also accountable, transparent, and aligned with programme and regulatory expectations.

6.1 Governance and Approval

A clear governance structure is required so that clinical, technical, and operational decisions about LLM-enabled diagnostics are made in a coordinated and accountable way.

The programme SHALL:

- Establish a formal governance body (or designate an existing one) responsible for oversight of LLM-based diagnostic support in the Free Health Check.
- Ensure this body includes representation from at least:
    - public health and programme leadership,
    - relevant clinical disciplines (e.g., paediatrics, internal medicine, mental health, geriatrics, oncology),
    - digital health and interoperability teams,
    - data and AI/ LLM specialists,
    - monitoring and evaluation, and, where applicable,
    - ethics/ legal.
- Require governance approval for:
    - introduction of new screening logic, calculators, or questionnaires,
    - changes to thresholds and triage rules,

    - major modifications to LLM prompts, guardrails, or configurations,
    - progression from pilot to broader deployment in any domain or life stage.

The governance body SHOULD meet regularly to review performance, incidents, and proposed changes, and MAY issue updated standards as experience accumulates.

6.2 Roles and Responsibilities

Clearly defined roles prevent gaps and overlaps in accountability, especially when multiple organisations or vendors are involved.

At minimum, the following roles SHALL be defined and assigned:

- **Clinical leads**
    - Own clinical pathways, screening rules, and threshold definitions.
    - Review performance and safety results from a clinical perspective.
    - Sign off on any clinical changes affecting how the LLM interprets data or recommends follow-up.
- **Data stewards and interoperability leads**
    - Maintain data dictionaries, mappings from Free Health Check forms to the Health Data Graph, and alignment with national standards.
    - Define and enforce data quality rules and validation checks.
    - Coordinate changes when forms, codes, or data models are updated.
- **AI/ LLM engineers and tool developers**
    - Implement and maintain the LLM orchestration, prompts, and tool integrations in accordance with approved specifications.

    - Ensure that deterministic calculators, scoring modules, and decision rules are implemented correctly and versioned.
    - Support evaluation, monitoring, and incident investigations from a technical point of view.
- **System operators and infrastructure teams**
    - Ensure reliable operation of data pipelines, CKG, LLM services, and user interfaces.
    - Manage access control, security, backups, and disaster recovery.
- **Programme monitoring and evaluation teams**
    - Use structured outputs (risk flags, recommendations, completion rates) for routine reporting and quality improvement.
    - Provide feedback about how the system affects workload, equity, and programme outcomes.

Each implementing organisation SHOULD document how these roles are distributed (e.g., which unit or vendor holds which responsibilities) and MAY designate additional roles as needed.

6.3 Audit and Traceability

To maintain trust and support investigation of issues, every LLM-supported output must be traceable back to the data and configuration that produced it.

The application and its surrounding systems SHALL:

- Log, for each LLM-supported recommendation or report:
    - participant and encounter identifiers,
    - key input data elements used (or references to them in the CKG),
    - tools and calculators invoked and their outputs,
    - LLM model or configuration version,

    - final risk categories, flags, and recommendations presented to users,
    - health worker actions (e.g., accepted, modified, rejected, ignored).
- Ensure that logs are time-stamped, tamper-resistant, and retained for at least the minimum period defined by programme and regulatory policy.
- Provide authorised users (e.g., governance, audit, and investigation teams) with a way to reconstruct the decision path for any given case, including the interplay of data, tools, and LLM reasoning.

Implementations SHOULD avoid logging sensitive free text unnecessarily and SHOULD apply appropriate de-identification or minimisation practices when logs are used for training or research.

6.4 Deployment and Change Management

Deployment and changes to LLM-enabled diagnostic applications must be deliberate and controlled, recognising that even small changes can alter clinical behaviour.

The programme and implementers SHALL:

- Follow a standard deployment pattern:
    - design and internal testing,
    - pilot deployment in selected sites,
    - evaluation against predefined criteria,
    - governance approval,
    - staged scale-up with monitoring.
- Maintain explicit versioning for:
    - LLM prompts and configuration,
    - deterministic tools and calculators,
    - thresholds and rules,

    - data mappings and form definitions.
- Document and communicate each release, including:
    - what changed,
    - which domains and life stages are affected,
    - any revised usage guidance for health workers.

Any substantial change (e.g., a new questionnaire, new or modified risk calculator, updated thresholds, or significant prompt revision) SHALL be treated as a change request, with impact analysis, testing, and governance approval before rollout. For high-risk domains or life stages, implementers SHOULD repeat a limited pilot and validation phase after major changes.

By adhering to these governance, audit, and change management standards, the Free Health Check programme can introduce and evolve LLM-enabled diagnostic applications in a way that is controlled, transparent, and accountable to both clinical and public health objectives.

7. Final Provisions

These standards establish the minimum, non-negotiable expectations for any LLM-enabled diagnostic application operating within the Free Health Check programme. By defining the required data elements, permitted LLM functions, output formats, and safety and governance controls, this document is intended to ensure that the use of LLMs strengthens – rather than compromises – clinical quality, patient safety, data protection, and equity across all screening populations and settings.

Compliance with all applicable SHALL requirements in this document is a prerequisite for piloting, procuring, or deploying any LLM-based diagnostic support tool in real screening workflows. Programme owners, clinical leaders, digital health teams, vendors, and oversight bodies are jointly responsible for ensuring that solutions are designed, implemented, monitored, and iteratively improved in line with these standards, and that deviations from SHOULD requirements are explicitly justified and documented.

Given the rapid evolution of AI and LLM technologies, this document is a living instrument. It SHALL be periodically reviewed and updated to reflect new evidence, regulatory developments, and lessons learned from implementation in the Free Health Check programme. Future revisions SHOULD continue to centre on safe, explainable, and context-appropriate use of LLMs that supports, not replaces, professional clinical judgement, promotes trust among communities, and contributes to better health outcomes across the life course.