TECHNICAL GUIDANCE

FOR DIAGNOSTIC APPLICATIONS LEVERAGING LARGE LANGUAGE MODELS (LLMS)

Table of Contents

Table of Contents

1. Background

Indonesia's digital health transformation under the Digital Health Transformation Strategy (DHTS) 2.0 places population-based screening and early detection at the centre of efforts to improve health outcomes across the life course. The Cek Kesehatan Gratis (CKG) / Free Health Check programme is a flagship implementation of this agenda, providing systematic screening for newborns, children, adolescents, adults, and older persons in both facility and community settings. Across these age groups and programmes, large volumes of structured and semi-structured data are collected through standard forms, clinical measurements, and screening tools, and increasingly integrated into shared data platforms and clinical knowledge representations. As digital systems mature, there is a growing need to ensure that these data are not only captured consistently, but also interpreted in a way that is clinically sound, explainable, and aligned with national guidelines.

Building on the Framework for AI-assisted Diagnostic Data Management and the accompanying standards for LLM-enabled diagnostic applications, the Ministry of Health is beginning to introduce large language models (LLMs) as a layer on top of CKG data to support screening workflows. In this model, LLMs do not replace clinical judgement, but help synthesize information from multiple screening domains, map findings to clinical guidelines, orchestrate risk calculators, and generate structured outputs such as triage suggestions and follow-up recommendations. This approach can increase the consistency of decision support across programmes, help front-line health workers navigate complex screening algorithms, and reduce missed opportunities for early detection. However, it also introduces new technical and operational risks if data representations, prompts, safety checks, and integration patterns are not clearly defined.

The technical guideline responds to this need by providing practical, implementation-focused direction on how LLMs should be embedded

within CKG-based screening systems. While the framework and standards define what is expected in terms of governance, data quality, and safety (for example, compliance with UU PDP 2022, alignment with SATUSEHAT interoperability requirements, and use of normative “SHALL/SHOULD/MAY” language), this guideline focuses on how to realise those expectations in concrete system designs. It describes how screening programme data are modelled in the Health Data Graph, how LLM components interface with risk scores and clinical decision support systems, and how outputs are passed to downstream systems such as electronic medical records or national health platforms.

By grounding LLM use in the realities of clinical screening programmes, the technical guideline aims to create a consistent, repeatable blueprint for implementers across different provinces, vendors, and deployment environments. It is intended to reduce ambiguity for technical teams, ensure that clinical and programme owners can trace how recommendations are generated, and provide regulators and monitoring teams with a transparent view of where and how LLMs operate in the screening pipeline. In doing so, the guideline helps ensure that AI-assisted diagnostic support within the Free Health Check programme is not only innovative, but also safe, interoperable, and sustainable as part of Indonesia’s broader digital health ecosystem.

2. Purpose and Scope

2.1 Background and Context

This guidance describes how large language models (LLMs) are used to support diagnostic decision-making in the context of the Cek Kesehatan Gratis (CKG) / Free Health Check programme. The programme collects structured and semi-structured data across multiple screening domains (demographics, mental health, reproductive health, nutrition and

cardiometabolic risk, liver health, geriatric assessments, heart screening, and self-reported questionnaires) and stores them in a Health Data Graph (HDG). LLMs are introduced on top of this CKG to interpret findings against guidelines, orchestrate risk calculators, generate explanations, and surface triage and referral recommendations while remaining grounded in coded clinical data and established standards.

2.2 Objectives of This Guidance

The primary objective of this document is to provide implementers with a practical, end-to-end view of how diagnostic applications leveraging LLMs should be designed for CKG-based screening. It explains which screening programmes and data types are in scope, how they are represented in the CKG, and how LLM components interact with risk scores, clinical decision support systems (CDSS), and downstream systems such as EMRs or national health platforms. A second objective is to create a shared technical baseline that can later be turned into normative “specific standards” (for example, formal requirements on data formats, calibration practices, or safety monitoring) without rewriting the overall architecture and clinical mapping.

2.3 In-Scope Diagnostic Use Cases and Users

This guidance covers diagnostic support for the Cek Kesehatan Gratis / Free Health Check programme across the life course, from newborns to older adults, and for both male and female participants. In scope are workflows that start from structured manual forms and proceed to risk assessment, early detection, and follow-up recommendations based on the collected data.

For newborns and young children, the scope includes growth and nutritional assessment, congenital heart screening, SHK/G6PD/HAK newborn screening, bile duct abnormality screening at different age windows (0–2 days, 14–28 days, 2–4 months), and early tuberculosis (TB) risk screening. For preschool and school-age children, the scope extends to

growth and developmental screening (KPSP, KMPE, GPPH and related items), eye and ear screening, dental screening, TB screening and scoring, and thalassemia screening and follow-up.

For adolescents and high school students, the guidance covers demographic and lifestyle profiling, mental health screening using structured scales (for example, Mini MINDHEAR Youth Scale A and B for anxiety and depression symptoms), reproductive health for girls and boys, history of TB, hepatitis B and C, smoking and physical activity, neglected tropical diseases (kusta, skabies, malaria), liver screening, and cardio-metabolic variables such as anthropometrics, blood pressure, blood sugar and anemia checks, visual and hearing acuity, and fitness tests.

For adults in the community, the scope includes male demographic screening, behavioural risk profiling (smoking, physical activity), TB examination (including sputum), liver disease and cirrhosis risks, nutrition, blood pressure and blood sugar (including follow-up tests such as fasting blood sugar, 2-hour post-prandial, and HbA1c), kidney and liver function tests, lipid profiles, COPD screening (PUMA), ear and eye screening by age band, dental screening, and integrated cardiovascular risk assessment and heart screening (including ECG for people with hypertension). Cancer-related use cases include colorectal cancer screening and follow-up, lung cancer screening, cervical cancer screening (IVA, DNA HPV and related steps), and breast cancer screening (clinical breast exam, ultrasound and related findings).

For older adults and geriatric screening, the guidance includes cognitive decline screening and follow-up (Mini-Cog/clock draw, AD-8 INA), mobility limitation and follow-up (SPBB), malnutrition screening and follow-up (MNSF), depressive symptoms, and functional status using tools such as the Barthel Index.

Typical users of these diagnostic applications include front-line health workers at facilities and in the community, school health teams, programme staff running the Free Health Check, and clinical supervisors. On the technical side, users include data and interoperability teams, CKG

and database administrators, LLM and application developers, and monitoring and quality teams who rely on consistent data and outputs. The guidance assumes that LLMs support these users by interpreting structured data, highlighting risks, and proposing follow-up actions, while final clinical judgement remains with qualified health professionals.

3. Clinical Screening Programs and Data

3.1 Overview of Screening Programs

The Free Health Check is organised as a set of age- and sex-specific screening programmes, each represented by a group of manual form headers.

For newborns and infants, the main programme components include growth screening, congenital heart screening, newborn screening for SHK/G6PD/HAK, and bile duct abnormality screening at the specified age intervals (0–2 days, 14–28 days, 2–4 months). These forms focus on early detection of congenital and metabolic conditions, jaundice and biliary problems, and early growth issues.

For toddlers, preschool and early school-age children, programme components include growth screening, developmental screening, eye and ear screening, dental screening, and tuberculosis screening. TB scoring is performed when a Mantoux test is done, and behavioural and emotional questions are used when indicated to identify developmental or behavioural problems that may require follow-up.

For school-age and high school adolescents, programme headers cover student demographics, mental health, reproductive health (separately for girls and boys), neglected tropical diseases, liver screening, and nutrition / blood pressure / blood sugar screening. These programmes bring together lifestyle and risk-factor questions (smoking, physical activity), infectious disease history (TB, hepatitis, malaria), and basic measurements (anthropometrics, blood pressure, random blood sugar, anemia, hearing, vision, dental health, fitness).

For adults in the community, programme headers include adult male demographics, liver screening, nutrition and cardio-metabolic screening

(blood pressure and blood sugar), tuberculosis examinations, ear and eye screening by age group, dental screening, hepatitis and fibrosis/cirrhosis laboratory panels, premarital rapid tests (HIV, syphilis), COPD screening (PUMA), kidney function and lipid profiles for higher-risk individuals, integrated cardiovascular risk prediction, heart screening with ECG where appropriate, colorectal and lung cancer screening, and geriatric screening.

For older adults, the geriatric screening programme consolidates cognitive, mobility, nutritional, mood, and functional assessments into a structured workflow that can be reused for monitoring and follow-up. Separate cancer screening flows for cervical and breast cancer are layered on top of this age-stratified structure.

Table Health Screening and Examination Categories by Age Group

Newborn	Toddler and Preschool Children	Adults and Elderly
• Thyroid hormone deficiency (Congenital Hypothyroidism) • Enzyme deficiency for red blood cell protection (G6PD Deficiency) • Adrenal hormone deficiency (Congenital Adrenal Insufficiency) • Critical Congenital Heart Disease • Biliary Atresia and Growth (Weight, Height)	• Growth and Development • Tuberculosis (TB) • Ears • Eyes • Teeth • Thalassemia (Blood test at 2 years old) • Blood Glucose (Blood test at 2 years old)	• Smoking • Physical Activity Level • Nutritional Status • Teeth • Blood Pressure • Blood Glucose • Stroke Risk • Heart Disease Risk (≥ 40 years old) • Kidney Function (≥ 40 years old) • Tuberculosis (TB) • Chronic Obstructive Pulmonary Disease (COPD) • Breast Cancer (≥ 30 years old) • Cervical Cancer (≥ 30 years old) • Parasitic Heart Disease (PKG) • Lung Cancer (≥ 45 years old) • Colorectal Cancer (≥ 50 years old) • Eyes • Ears • Mental Health • Liver (Hepatitis B, C, Cirrhosis) • Pre-marriage Checkups (Anemia, Syphilis, HIV) • Geriatrics (≥ 60 years old)

Table. Health Screening and Examination Categories by Educational Level

Examination Type	SD (7-12 years)	SMP (13-15 years)	SMA (16-17 years)
Nutritional Status	✓	✓	✓
Smoking		✓	✓
Physical Activity Level	✓	✓	✓
Blood Pressure	✓	✓	✓
Blood Glucose	✓	✓	✓
Tuberculosis	✓	✓	✓
Thalassemia		✓ (Grade 7)	
Anemia (Teenage Girls)		✓ (Grade 7)	✓ (Grade 10)
Ears	✓	✓	✓
Eyes	✓	✓	✓
Teeth	✓	✓	✓
Mental Health	✓		✓
Liver (Hepatitis B)	✓	✓	✓
Liver (Hepatitis C)		✓	✓

3.2 Key Data Elements and Scores (PHQ-9, ASCVD, ALBI, etc.)

Across these programmes, the forms consistently capture identity and context variables (health facility identification, school or community context, child/student/adult/guardian identity, national ID where applicable, date of screening) to support linkage across visits and services.

For growth and nutrition, key data elements include weight, length/height, head circumference, and derived indicators such as weight-for-age, length/height-for-age, and weight-for-length/height. These indicators are used to classify nutritional status and growth patterns in newborns, young children and older age groups, and they provide essential input for any automated growth or nutrition assessment logic.

For development and behaviour, the forms collect KPSP results, behavioural and emotional questionnaires (such as KMPE), and ADHD-related assessments (GPPH) when indicated. In adolescents, mental health is captured using structured scales such as the Mini MINDHEAR Youth Scale A (anxiety symptoms) and B (depression symptoms), while in older adults, depressive symptoms are captured in geriatric mental health sections.

For infectious and chronic diseases, variables include history and symptoms of tuberculosis (for infants, children and adults), lymph node and bone/joint signs, Mantoux test induration and result, hepatitis B and C status, syphilis and HIV rapid test results, malaria screening and risk factors, and thalassemia screening and follow-up results. These elements are complemented by vaccination and immunisation data such as HPV immunisation history and general immunisation status where relevant.

For sensory, dental and organ-specific screenings, the forms capture hearing test results, ear examinations (including follow-up with otoscope and tuning fork), visual acuity tests using E-Tumbling and Snellen charts with pinhole, pupil examinations, and detailed dental screening items across age bands.

For cardio-metabolic and respiratory risk, key elements include blood pressure readings, random and fasting blood sugar, 2-hour post-prandial values, HbA1c, kidney function parameters for selected high-risk individuals, lipid profiles, anthropometric measures and waist circumference, smoking behaviour, physical activity level, COPD/PUMA screening fields, and any variables required to compute composite heart and stroke risk predictions. Heart screening with ECG yields structured findings and free-text descriptions of abnormalities where applicable.

For cancer and geriatric screening, the forms record colorectal cancer screening results and follow-up, lung cancer screening, cervical cancer screening steps (examination with VIA, HPV DNA tests), breast examination findings (including SADANIS and ultrasound results), as well as geriatric cognitive, mobility, nutritional, depressive and functional status variables (Mini-Cog, AD-8 INA, SPBB, MNSF, Barthel Index and related fields).

These elements can be directly mapped to machine-readable representations and, where applicable, to derived indices or scores that an LLM-enabled diagnostic application will use when generating risk stratifications, recommendations, and follow-up plans.

3.3 Data Quality, Coding, and Standards

To support safe diagnostic decision support, all of the above data must be captured with consistent structure, units, and value sets across age groups and forms. Identity and context fields (facility, school, individual, guardian, date of screening) should follow a canonical format so that records can be linked longitudinally and across programmes. Dates should use a consistent

day/month/year format, and unique identifiers such as national ID and facility codes should be validated where possible.

For numeric clinical measurements such as weight, height, head circumference, blood pressure, blood sugar, hemoglobin, kidney and liver function parameters, and lipid levels, units should be explicitly recorded and constrained to approved options. Value ranges should be checked at the point of entry to detect obvious errors (for example, biologically implausible heights or blood pressures). For derived indicators (for example weight-for-age, weight-for-height, growth classifications, and computed risk scores), the underlying raw values should also be stored so that calculations can be verified and re-run if algorithms or reference standards change.

For categorical and Boolean fields—such as screening outcomes, test results (positive/negative), symptom presence/absence, and risk-factor checklists—controlled vocabularies should be used instead of free text. This includes standardised options for TB signs, NTD findings, mental health symptoms, reproductive health factors, immunisation status, and cancer screening results. Where follow-up examinations are present (for example, SPBB, MNSF, follow-up blood sugar tests), each stage should be explicitly tagged as initial or follow-up to support longitudinal analysis.

Although the precise choice of interoperability standards is documented elsewhere in the broader architecture, this guidance assumes that data quality practices will enable mapping to common health data models and code systems. That includes keeping question-level data for questionnaires, recording both raw test results and their interpretations, and ensuring that variables are named and structured in a way that is stable over time. High, consistent data quality at this level is a prerequisite for reliable transformation into the Health Data Graph and for the LLM components to reason accurately over the Free Health Check data.

4. Data Model and System Architecture

4.1 Health Data Graph (HDG) and Core Entities

The Health Data Graph (HDG) is the central data structure used to represent all information collected through the Free Health Check across age groups and screening categories. Instead of storing data only as flat tables or isolated records, the CKG links people, encounters, measurements, questionnaire responses, conditions, and risk assessments as nodes and relationships. This structure allows LLM-based tools to retrieve context-rich information and reason over complete clinical snapshots rather than isolated variables.

At a minimum, the CKG includes the following core entity types:

- Person and Identity Entities
    - Individual participants: newborns, children, adolescents, adults, and older adults.
    - Linked identity attributes: national identification number, date of birth, sex, and age group derived from date of screening.
    - Related identity nodes such as guardian/parent for infants and school-age children, and facility/school for site-level attributes.
- Organization and Location Entities
    - Health facilities (Puskesmas, hospitals, clinics) and their identifiers.
    - Schools and community locations involved in CKG Sekolah and CKG Komunitas.
    - Links between encounters and the organization where screening took place.
- Encounter / Screening Episode Entities
    - A screening encounter represents a single Free Health Check contact, categorised by life stage (newborn, toddler/preschool,

school-age, adolescent, adult, geriatric) and by CKG category (CKG Ulang Tahun, CKG Sekolah, CKG Komunitas).
    - Each encounter node is time-stamped and linked to all observations, questionnaires, and procedures performed during that visit.
- Observation Entities
    - Structured measurements: weight, length/height, head circumference, waist circumference, blood pressure, blood sugar (and follow-up values), hemoglobin, kidney and liver function parameters, lipid profile, and other numeric or categorical lab results.
    - Physical examination findings: lymph node enlargement, joint/bone swelling, pupil examination, hearing and visual acuity results, dental findings.
    - Each observation includes value, unit, reference range, collection time, and method.
- Questionnaire and Response Entities
    - Child development and behaviour tools (KPSP, KMPE, GPPH, behavioural questions).
    - Youth mental health tools such as Mini MINDHEAR Youth Scale A and B.
    - Geriatric instruments for cognition, depression, mobility, malnutrition, and functional status (Mini-Cog, AD-8 INA, SPBB, MNSF, Barthel Index, and related questions).
    - Each questionnaire is represented as a structured object linked to item-level responses and scoring logic.
- Condition, Diagnosis, and Problem Entities
    - Confirmed or suspected conditions such as tuberculosis, hepatitis B/C, cirrhosis, COPD, colorectal cancer, lung cancer, cervical cancer, breast cancer, NTDs (kusta, skabies, malaria), and nutritional disorders.

    - These entities are connected to supporting observations, screening results, and follow-up investigations.
- Procedure and Test Entities
    - Procedures such as Mantoux testing, sputum examination, ECG, ultrasound of the breast, cervical VIA, HPV DNA testing, dental interventions, and follow-up imaging or lab investigations.
    - Linked to both encounters and resulting observations or diagnostic reports.
- Risk Assessment and Score Entities
    - Computed outputs such as growth classifications (e.g., weight-for-age categories), nutritional risk levels, tuberculosis risk, COPD/PUMA risk, risk of heart attack and stroke, and geriatric risk stratifications.
    - In later phases, composite cardiovascular and metabolic risk scores or mental health severity scores can be stored here when implemented.
- Report and Recommendation Entities
    - Structured summaries generated for clinicians or participants (e.g., “Free Health Check result summary”, “Geriatric screening profile”).
    - LLM-generated interpretations and suggested follow-up actions are represented as linked report nodes, with provenance information (model version, timestamp).

Relations in the CKG (e.g., *“person–hasEncounter–screening visit”*, *“encounter–hasObservation–blood pressure”*, *“observation–supports–condition TB”*, *“encounter–hasQuestionnaireResponse–Mini MINDHEAR”*) ensure that each data point is traceable to its origin and clinical context. This structure supports both graph-based retrieval for LLM prompts and auditable analytics for programme monitoring.

4.2 Integration with Forms, EMR, and External Systems

The CKG does not replace existing digital health systems; it sits on top of them as a semantic layer that unifies data coming from multiple sources. Integration is organised around national digital health infrastructure and technical guidelines, ensuring that data flow from manual forms into interoperable, standards-based representations.

- Integration with Manual and Digital Forms
    - Age- and sex-specific Free Health Check forms (newborn, child, adolescent, adult, geriatric; male/female; programme-specific headers) remain the primary tools used by health workers and school teams.
    - Data from these forms are captured through SATUSEHAT Mobile, ASIK, EMR interfaces, or other approved digital tools aligned with Juknis CKG.
    - Each form field is mapped to a standardised data model (e.g., FHIR Observation, QuestionnaireResponse, Condition, Procedure) and then transformed into CKG entities and relationships.
- Integration with EMR Systems
    - Health facilities continue to use their own electronic medical record (EMR) systems for routine care. The Free Health Check screening episodes can be:
        - documented directly in the EMR and pushed to SATUSEHAT, or
        - recorded in programme tools (ASIK, SATUSEHAT Mobile) and then exposed to EMRs as structured summaries.
    - The CKG uses the same underlying identifiers and codes as SATUSEHAT, allowing it to link screening episodes with later clinical encounters, laboratory results, and diagnoses in facility EMRs.

- Integration with Laboratory, Imaging, and Specialised Systems
    - Laboratory results for hepatitis, Hb, kidney and liver function, lipid profile, HbA1c and related tests are ingested from laboratory information systems or from manual entry in line with national data exchange formats.
    - Imaging-related data (e.g., ECG findings, breast ultrasound reports) are either stored as structured fields or linked to external systems using references; relevant summary fields are represented as observations or diagnostic reports in the CKG.
    - Screening programmes that rely on separate registries (e.g., TB, cancer) can expose their core variables via SATUSEHAT-compatible APIs, enabling them to be linked to CKG encounters and conditions.
- Integration with National Platforms (SATUSEHAT and ASIK)
    - SATUSEHAT acts as the primary data exchange and interoperability layer. The CKG is logically aligned with SATUSEHAT data models so that:
        - incoming data from SATUSEHAT can be converted into graph entities with minimal transformation, and
        - outputs from the LLM-based diagnostic application (e.g., risk categories, follow-up recommendations) can be shared back as FHIR resources or programme-specific messages.
    - ASIK and other community-level reporting tools provide entry points for Posyandu and other screening sites; their data flows through standard pipelines before being mapped into the CKG.

Through this integration approach, the CKG becomes a federated semantic layer that can be populated from multiple systems yet remains consistent with national interoperability standards and data governance policies.

4.3 High-Level Architecture and Orchestration Flow

The system architecture is designed so that LLM-based diagnostic support is modular, auditable, and tightly anchored to structured data. At a high level, the architecture consists of six layers: data capture, ingestion and quality, CKG storage and query, LLM and tools, application interfaces, and monitoring.

1. Data Capture Layer
    - Health workers, school teams, and community staff collect data using Juknis-aligned forms through SATUSEHAT Mobile, ASIK, EMR modules, or web-based interfaces.
    - Participants (or caregivers) may also provide self-reported data for certain age groups, especially adolescents and adults, via digital channels such as mobile apps or chatbots.
    - Each record is tagged with participant identity, age group, screening category (newborn, preschool, school, adolescent, adult, elderly), and screening programme type (CKG Sekolah, CKG Komunitas, CKG Ulang Tahun).
2. Ingestion, ETL, and Data Quality Layer
    - Data from capture systems are ingested into a central processing pipeline where ETL (Extract, Transform, Load) operations normalise formats, units, and codes.
    - Automated validation rules check for completeness, plausibility, duplicates, and alignment with reference standards (e.g., valid ranges for anthropometrics and labs, consistency between age and form type).
    - Cleaned, validated data are persisted in a structured store (e.g., FHIR-based data lake or warehouse) that acts as the source for the CKG.
3. CKG Storage and Query Layer

- The CKG is built on top of the structured store, transforming FHIR resources and related data into a graph of patients, encounters, observations, questionnaires, conditions, procedures, risk assessments, and reports.
- A graph database or equivalent technology provides APIs for graph queries such as “all screening results and risk factors for this adolescent in the last 12 months” or “all conditions associated with abnormal liver function tests in this facility.”
- Versioning and provenance metadata ensure that every node and relationship can be traced to its source system, data collection time, and, where relevant, previous versions.

4. LLM and Tools Layer
    - An LLM service interacts with the CKG through a retrieval and tools interface. For a given participant or cohort, the system:
        - queries the CKG for relevant context (age, sex, screening category, key observations, questionnaire results, existing conditions),
        - optionally calls calculation tools (e.g., risk score or classification functions) to derive indicators such as growth status, nutritional risk, or cardiovascular risk from raw measurements, and
        - uses retrieved facts and tool outputs as inputs for prompt construction.
    - The LLM then generates structured outputs (risk levels, flags, suggested follow-up actions) and narrative explanations aligned with Juknis and clinical protocols. Calibration and thresholding logic, as described in later sections, translates model scores into consistent “low/medium/high risk” categories.
5. Application and User Interface Layer

- Outputs from the LLM are returned to user-facing systems: SATUSEHAT Mobile, EMRs, dashboards for programme managers, and any dedicated CKG interfaces.
- For health workers, the interface prioritises clear risk categories, key reasons, and recommended next steps (for example, repeat screening, lifestyle advice, referral to a specific service, or confirmatory testing).
- For participants or caregivers, the interface provides simplified explanations and actionable advice while maintaining consistency with what clinicians see.

6. Monitoring, Logging, and Governance Layer
    - All key events are logged: data ingestion errors, CKG operations, LLM queries and responses, tool calls, and user actions on recommendations.
    - Monitoring dashboards track data completeness and quality, system performance, LLM output patterns, and calibration stability across age groups and facilities.
    - Governance controls (role-based access, audit trails, consent status, and data-use policies) are applied consistently across layers, aligning with national regulations and the broader diagnostic data framework.

Together, these layers ensure that Free Health Check data move from forms to actionable diagnostic support in a transparent, standards-aligned way. The architecture enables future expansion (e.g., additional risk calculators or more advanced LLM capabilities) without disrupting core data flows or governance arrangements.

5. LLM Functions for Diagnostic Support

This section describes how large language models (LLMs) are used in the Free Health Check context to support screening and diagnostic decision-making. The focus is on how the LLM behaves, not on its internal architecture: what information it can see from the Health Data Graph (HDG), which tools and calculators it can call, and what types of outputs it is allowed to produce. The same general patterns apply across age groups and programmes—from newborn growth and congenital screening, through child development, adolescent mental health and reproductive health, adult cardio-metabolic and cancer screening, to geriatric assessments—while respecting differences in form content and clinical pathways.

5.1 Roles of LLMs in Screening and Triage

Within the Free Health Check ecosystem, the LLM plays several clearly defined roles, all of which are supportive to health workers and do not replace clinical judgement:

1. Contextual summariser of screening data
    - The LLM receives a structured snapshot from the CKG for a given participant: age, sex, screening category (newborn, child, adolescent, adult, elderly), relevant encounters, key measurements, questionnaire results, and any existing conditions.
    - It summarises this into a concise clinical picture, highlighting normal and abnormal findings, trends (e.g., repeated elevated blood pressure or progressive weight loss), and important risk factors (e.g., smoking and low physical activity in adults, or recurrent TB history).
2. Guideline-aligned interpreter of findings

- Using structured logic and prompts aligned with national technical guidelines (Juknis) and programme protocols, the LLM interprets combinations of data points.
- Examples:
  - In newborns, it correlates growth parameters, congenital heart screening results, SHK/G6PD/HAK findings, and bile duct screening results with expected normal ranges and red-flag patterns.
  - In children, it interprets KPSP, KMPE, GPPH, eye/ear screening, and TB risk signs to identify possible developmental delay, behavioural problems, hearing or vision impairment, or TB suspicion.
  - In adolescents, it relates Mini MINDHEAR scores, reproductive health responses, NTD variables, liver history, anthropometrics, blood pressure, blood sugar and anemia to mental health risk, reproductive risk, and early metabolic risk.
  - In adults, it integrates smoking, physical activity, TB, liver, cardio-metabolic, respiratory (PUMA), and cancer-related items to form a coherent risk profile.
  - In older adults, it interprets cognitive, mobility, nutrition, depression and functional status in terms of geriatric syndromes and needs for follow-up.

3. Triage and follow-up support
   - Based on interpreted risk and pre-defined rules, the LLM assists with triage by proposing appropriate follow-up actions such as:
     - repeat measurement or confirmatory test (e.g., confirmatory blood sugar or HbA1c, follow-up TB tests),
     - lifestyle advice and routine follow-up at primary care level,
     - non-urgent referral (e.g., to mental health, nutrition, physiotherapy, eye/ear services),

        - urgent referral or escalation (e.g., suspected congenital heart disease, severe malnutrition, possible cancer or severe depression).
    - The LLM does not autonomously schedule or execute these actions; it proposes them for health worker confirmation.
4. Communication and explanation
    - The LLM transforms technical findings into clear language for two audiences:
        - Health workers: structured explanation of why a case is classified as low, moderate or high risk, including which variables or scores contributed most.
        - Participants and caregivers: simple explanations of results (for example, “your blood pressure is above the recommended range,” “this screening suggests you may benefit from a more detailed eye examination”) and practical advice consistent with programme materials.

By constraining the LLM to these roles and grounding every decision in CKG data and explicit rules, the system aims to maximise usefulness while minimising hallucinations and unsafe behaviours.

5.2 Use of Tools and Calculators (Risk Scores, CDSS Modules)

The LLM does not perform all calculations internally. Instead, it orchestrates a set of deterministic tools and calculators that implement well-defined clinical logic. This separation ensures that numeric results and classifications are reproducible, testable, and easier to govern.

1. Growth and nutritional assessment tools
    - For newborns, infants, and children, tools compute indicators such as weight-for-age, length/height-for-age, weight-for-

length/height, and classify nutritional status based on reference standards.

- For adolescents and adults, tools use weight, height and waist circumference to derive body mass index (BMI) and other anthropometric indices used in nutrition and cardio-metabolic risk assessment.
- The LLM calls these tools with raw measurements extracted from the CKG and receives back categorical outputs (e.g., normal, stunted, wasted, obese) and z-scores or percentiles where relevant.

2. Developmental, behavioural, and mental health scoring tools
    - Questionnaires such as KPSP, KMPE, GPPH and Mini MINDHEAR are implemented as scoring modules that:
        - sum or weight item-level responses,
        - apply cut-offs for “normal”, “needs monitoring”, or “needs further evaluation”.
    - The LLM queries the CKG for questionnaire responses, passes them into these modules, and uses the resulting scores and categories when generating interpretations and follow-up recommendations.
3. Disease-specific screening and risk calculators
    - For TB, tools may combine symptom signs, Mantoux results, and other factors into a risk or scoring scheme consistent with programme guidance.
    - For respiratory conditions such as COPD (PUMA screening), tools evaluate age, smoking behaviour and respiratory symptoms to classify risk level and indicate need for spirometry or specialist review.
    - For liver disease, tools evaluate hepatitis B/C status, liver function tests and cirrhosis markers as appropriate to classify risk and suggest follow-up.

- For cardio-metabolic and cardiovascular risk, tools combine blood pressure, blood sugar, lipid profile, anthropometrics, smoking and other risk factors into composite risk estimates according to defined algorithms.

4. Cancer screening decision modules
    - For colorectal, lung, cervical and breast cancer screening, tools encode the logic of:
        - who is eligible for which screening,
        - how to interpret screening and confirmatory test results,
        - what follow-up steps are recommended after a given result pattern.
    - The LLM consults these tools rather than inferring patterns by itself, ensuring that the pathway from screening result to recommended action is aligned with programme algorithms.
5. Geriatric assessment modules
    - Tools implement scoring for Mini-Cog, AD-8 INA, SPBB, MNSF, Barthel Index and any other instruments included in the geriatric forms.
    - They classify participants into categories such as “no impairment”, “possible impairment – monitor”, or “needs further geriatric assessment”, which the LLM then explains and situates in context (e.g., combined with nutrition or mood findings).

In practice, the LLM acts as an orchestrator: it decides which tools to call based on age group, form type and available data; it passes clean input data from the CKG into those tools; and it uses their outputs to build its narrative and structured recommendations. Where no formal calculator exists yet, simple rule-based modules can be used, and the LLM’s role is limited to explanation rather than invention of new thresholds.

5.3 Output Types: Recommendations, Flags, and Reports

LLM outputs are designed to be structured, predictable, and easy to integrate into health worker workflows and national systems. For each screening encounter, the LLM may produce three main classes of outputs.

1. Structured risk categories and flags
    - For each relevant domain (growth and nutrition, development, mental health, infectious disease, cardio-metabolic risk, cancer, geriatric status), the LLM returns categorical risk labels such as:
        - Normal / within target,
        - At risk / needs monitoring,
        - High risk / needs further evaluation,
        - Urgent / requires immediate attention (in rare, clearly-defined cases).
    - It may also produce specific flags, for example:
        - “Possible congenital heart disease – abnormal screening result”
        - “Suspected TB – positive screening signs and test result”
        - “High cardiovascular risk based on composite risk tool”
        - “Cognitive impairment suspected – refer for geriatric assessment”.
    - These flags are output in a machine-readable structure that can be stored as risk assessments or alerts in the CKG and in connected systems.
2. Actionable recommendations
    - For each flag or risk category, the LLM proposes recommended next steps mapped to programme protocols, such as:
        - Repeat or confirmatory tests (e.g., repeat blood pressure, fasting blood sugar, HbA1c, imaging, additional labs).

        - Referral (e.g., to primary care doctor, mental health services, nutrition counselling, TB clinic, eye/ear clinic, cancer diagnostic centre, geriatric service).
        - Follow-up timeline (e.g., “re-screen in 6 months”, “schedule visit within 1 month”, “same day referral”).
        - Health education and lifestyle advice consistent with approved messages.
    - Recommendations are produced in a structured format (e.g., a list of actions with priority and suggested timing) and can be rendered in user interfaces or exported into other workflow tools. Final decisions remain with health workers, who can accept, modify or reject recommendations.
3. Reports and explanations
    - Clinical summary report:
        - A concise, structured narrative aimed at clinicians and programme staff, describing the participant’s key findings, scores, risk categories, and rationale for recommendations.
        - This report can be attached to the participant’s record in the CKG and shared with EMR systems as a screening summary.
    - Participant-facing explanation:
        - A simplified version of the report that uses non-technical language, explains what the screening did and did not find, and emphasises next steps in a supportive way.
        - It is intended for delivery via printed output, SMS, app notifications, or counselling sessions, depending on programme design.
    - Programme and monitoring views:
        - Aggregated outputs (e.g., counts of high-risk flags by age group or facility) can be used to feed dashboards and monitoring tools; the LLM itself does not compute those

aggregates but provides standardised, structured case-level outputs that make aggregation straightforward.

By constraining outputs to these defined types, the system ensures that LLM behaviour remains consistent, reviewable, and aligned with the overall diagnostic framework. The next sections (Evaluation, Safety and Governance, Implementation Roadmap) will specify how these outputs are tested, calibrated, and governed before and after deployment.

6. Evaluation, Safety, and Governance

This section describes how the Free Health Check LLM components are evaluated before deployment, monitored during use, and governed over time. The goal is to ensure that any diagnostic support provided by the system is accurate, safe, fair, and auditable, across all age groups and programmes (from newborn screening to geriatric care).

6.1 Calibration, Thresholds, and Performance Metrics

Evaluation and calibration for the Free Health Check LLM system are organised by domain (e.g., growth and nutrition, TB, mental health, cardio-metabolic risk, cancer, geriatric screening) and by age group. Each combination has its own performance targets and thresholds, based on local epidemiology and programme priorities.

1. Separation of components for evaluation
    - Deterministic tools and calculators (e.g., growth charts, TB scores, COPD/PUMA rules, cardiovascular risk estimators, questionnaire scoring modules) are evaluated independently to ensure that they implement clinical logic correctly.
    - The LLM orchestration and explanation layer is evaluated on how accurately it uses these tools, how consistently it interprets their outputs, and how reliably it follows decision rules for recommendations and triage.
2. Thresholds and risk categories
    - For each calculator or decision module, thresholds define what counts as “normal”, “at risk / needs monitoring”, “high risk / needs further evaluation”, and “urgent”.
    - These thresholds are set and approved by clinical leads and programme managers, not by the LLM. For example:
        - growth z-score cut-offs for moderate/severe undernutrition,

        - symptom combinations and test thresholds for TB suspicion,
        - cut-offs for mental health questionnaire scores indicating minimal, mild, moderate, or severe symptoms,
        - cardiovascular risk bands defined by absolute 10-year risk where such tools are available.
    - The LLM must treat these thresholds as fixed constraints and may not "invent" new categories.
3. Performance metrics
    - For classification-type outputs (e.g., high vs low risk, refer vs no refer), metrics such as sensitivity, specificity, positive predictive value, and negative predictive value are used. The relative importance of these metrics depends on the programme; for example, TB and cancer screening may prioritise sensitivity, whereas some low-risk lifestyle advice may tolerate lower sensitivity.
    - For calibration, the system checks whether predicted risk categories match observed outcomes over time. For example, among participants flagged as "high cardio-metabolic risk", the proportion who later show elevated blood pressure or blood sugar in routine care should align with expectations.
    - For process metrics, the system tracks whether recommendations are practical and acceptable: referral rates, completion of confirmatory tests, and follow-up within the recommended timeframe.
4. Evaluation phases
    - Offline evaluation: Run the system on historical or pilot data from Free Health Check forms to measure performance before any live use. Adjust thresholds and prompts to meet pre-defined performance criteria.
    - Limited pilot evaluation: Deploy in selected facilities or schools with close supervision. Health workers can mark outputs as

"appropriate" or "not appropriate", and discrepancies are reviewed.
    - Scale-up with ongoing monitoring: After performance is acceptable in pilot sites, the system may be expanded, but with continuous data collection to detect drift or degradation.
5. Recalibration
    - As population characteristics, prevalence patterns, or screening protocols change, thresholds and risk bands may need adjustment.
    - Recalibration is treated as a controlled change: proposals are tested offline, validated by clinical governance, and deployed with clear versioning and documentation.

6.2 Safety, Bias, and Monitoring Requirements

Safety and equity are central to the use of LLMs in diagnostic contexts. The system is designed to fail safely, avoid reinforcing existing inequities, and provide mechanisms to detect and correct problems early.

1. Defined safety boundaries for the LLM
    - The LLM is not permitted to override clearly defined programme rules or invent new diagnostic criteria. It operates within a constrained set of tools, thresholds, and risk categories.
    - Where input data are incomplete or inconsistent (e.g., missing critical measurements or conflicting history), the LLM must favour conservative outputs such as "insufficient data" or "advise completion of screening items" rather than making strong conclusions.
    - For high-risk recommendations (e.g., urgent referral, suspected serious disease), the system always prompts for human review and confirmation.
2. Clinical safety checks

- For each domain (e.g., newborn screening, TB, mental health, cardiovascular risk), specific “red flag” patterns are defined. If these appear in the CKG (e.g., very low hemoglobin, dangerously high blood pressure, positive TB tests, severe depression scores), the system triggers standardised safety messages and escalation prompts.
- Safety cases are documented, showing how the system behaves for typical scenarios and worst-case scenarios, and these are reviewed during evaluation.

3. Bias and fairness considerations
    - Outputs are periodically analysed for systematic differences by age group, sex, facility type, region, socio-economic indicators where available, and other relevant factors.
    - The goal is to detect patterns such as:
        - under-identification of risk in certain subgroups,
        - systematic over-referral from certain facilities or populations,
        - inconsistent application of thresholds across demographic groups.
    - Where such patterns are found, root causes are investigated (data quality, model behaviour, or programme design), and mitigation measures are designed.
4. Monitoring requirements
    - Data-level monitoring: Completeness and quality of key variables (e.g., weight, height, blood pressure, questionnaire items) are tracked. High rates of missing or implausible values are flagged for feedback to data collectors and system owners.
    - Model-level monitoring: Frequencies of risk categories and recommendations, changes in output distributions over time, and patterns in overrides by health workers (e.g., frequent rejection of certain recommendations) are monitored. Significant shifts trigger review.

- Safety incident handling: Mechanisms are in place for health workers to report hazardous or clearly inappropriate recommendations. Such incidents are logged, investigated, and used to improve prompts, thresholds, or tooling.

5. Human oversight
    - Health workers remain responsible for clinical decisions. The interface should explicitly remind users that LLM recommendations are decision-support, not automatic orders.
    - Training and communication emphasise that users should feel empowered to override or ignore recommendations when they are not clinically appropriate, and that such overrides are valuable feedback for system improvement.

6.3 Governance, Roles, and Auditability

Clear governance is required to manage responsibilities, approve changes, and ensure that the Free Health Check LLM system remains trustworthy and aligned with national policies and programme goals.

1. Governance structure
    - A programme governance group oversees the overall use of LLM-based diagnostic support within the Free Health Check. It typically includes representatives from:
        - public health and programme management,
        - clinical experts (e.g., paediatrics, internal medicine, mental health, geriatrics),
        - digital health and interoperability teams,
        - data and AI specialists,
        - ethics and legal/compliance functions where relevant.
    - This group approves major design choices, thresholds, evaluation plans, and deployment phases.
2. Operational roles and responsibilities

- Clinical leads: define and update clinical pathways, thresholds, and safety rules; interpret evaluation results from a clinical perspective.
- Data stewards and interoperability leads: manage data dictionaries, mapping from forms to the CKG, data quality rules, and alignment with national standards.
- AI and LLM engineers: maintain the LLM prompts, tool integrations, and orchestration logic; implement changes approved by governance.
- System operators and support teams: monitor infrastructure, handle incidents, manage access control and security.
- Programme monitoring teams: use aggregated outputs for reporting and quality improvement, and provide feedback on the system's usefulness.

3. Change management
    - Any substantial changes—such as new questionnaires, new risk calculators, updated thresholds, or significant prompt changes—follow a defined change process: proposal, impact analysis, testing, governance approval, and controlled rollout.
    - Each change is associated with a new version of the rules or model configuration so that outputs can always be interpreted against the correct context.
4. Auditability and traceability
    - For every LLM-generated recommendation or report, the system logs at least:
        - the participant and encounter identifiers,
        - the data retrieved from the CKG (or references to it),
        - the tools and calculators called and their outputs,
        - the LLM configuration or model version used,
        - the final risk categories and recommendations presented to the user,
        - any user actions taken (accepted, modified, rejected).

- These logs allow investigators to reconstruct how a particular output was generated and to assess whether the system behaved according to approved rules.

5. Compliance and data protection
    - Governance processes ensure that use of LLMs respects applicable laws and policies on data protection, consent, and data sharing.
    - Access to identifiable data is role-based and limited to what is necessary for care and operations.
    - When LLM components are trained or tuned on Free Health Check data, appropriate de-identification, minimisation, and security controls are applied, and such training activity is itself subject to governance approval.

By combining clear governance, defined roles, robust audit trails, and ongoing monitoring, the Free Health Check LLM system can evolve and improve while maintaining trust, safety, and accountability.

7. Implementation Roadmap

This section outlines how the Free Health Check LLM-based diagnostic support should be implemented in stages. The roadmap is designed to be incremental and reversible: start simple, prove value and safety on a small scale, then gradually expand across age groups, programmes, and regions.

7.1 Phased Implementation Approach

The implementation is divided into four main phases. Each phase can be run in selected sites and then expanded once minimum success criteria are met.

Phase 0 – Foundations: Forms, Data and CKG Readiness

- Finalise and stabilise the manual and digital Free Health Check forms for all age groups and sex-specific flows (newborn–infant, preschool/child, adolescent, adult, geriatric; male/female; school and community).
- Map every form field to a standard data model and to its representation in the Health Data Graph (HDG).
- Set up the basic data pipeline: capture through SATUSEHAT Mobile, ASIK, EMRs or web forms; ETL processes for cleaning and validation; storage in a structured repository; and transformation into the CKG.
- Establish data quality dashboards and routines, so implementation teams can see completeness, common errors, and progress by site.
- No LLM functions are exposed to users in this phase; the goal is to ensure data are clean, consistently structured, and traceable.

Phase 1 – CKG-based Summaries and Static Decision Support

- Deploy the CKG in a small number of pilot facilities or schools with reliable data capture.

- Enable a first version of LLM support that focuses on summarisation and static decision support:
    - summarising screening data into structured, human-readable clinical snapshots,
    - highlighting obvious red flags based on simple, transparent rules,
    - generating draft screening summary reports for health workers to review.
- During Phase 1, LLM behaviour is conservative and heavily constrained; it relies on existing thresholds and simple rules, with health workers required to confirm all suggestions.
- Collect feedback from users on the clarity, relevance, and workload impact of the summaries and reports.

Phase 2 – Tool-Orchestrated Risk Assessment and Triage Support

- Introduce and integrate deterministic tools and calculators for growth, nutrition, development, TB, mental health, cardio-metabolic and respiratory risk, cancer eligibility and follow-up, and geriatric assessments.
- Extend LLM prompts and orchestration logic so the LLM can:
    - call these tools based on age, form type and available data,
    - combine their outputs into multi-domain risk profiles,
    - propose structured triage and follow-up recommendations.
- Conduct intensive offline and live evaluation in pilot sites, including calibration checks, sensitivity/specificity measurement, and user feedback on recommendations.
- In this phase, LLM support is still labelled as “pilot” and deployment is restricted to settings with strong supervision and monitoring.

Phase 3 – Expanded Deployment and Programme Integration

- Once Phase 2 results meet agreed criteria, expand deployment to additional facilities, schools and community sites.
- Integrate LLM outputs more deeply into existing workflows:
    - embed risk categories and recommendations into EMR and SATUSEHAT views,
    - feed structured flags into programme dashboards,
    - provide participant-facing explanations through standard communication channels.
- Introduce governance-controlled updates based on field experience, including adjusted thresholds, additional calculators, or more refined prompts, always following the evaluation and approval process defined in the governance section.

Each phase has clear entry and exit criteria, and rollout can be staggered by region or programme area. Importantly, any step can be paused or rolled back if safety concerns, data quality issues, or operational challenges arise.

7.2 Pilot and Validation Activities

Pilot and validation activities ensure that the system works in real-world conditions before it is widely used. They should be planned and documented in detail, with clear roles and timelines.

1. Site selection and preparation
    - Choose a small, diverse set of pilot sites (e.g., one or two schools, several community health facilities, at least one setting with geriatric services).
    - Confirm that these sites have stable data capture processes, trained staff, and basic digital infrastructure.
    - Provide orientation about the LLM system, emphasizing that it is a support tool under evaluation, not a replacement for clinical judgement.
2. Data and workflow validation

- Run “shadow mode” tests where data from forms flow through the pipeline and LLM, but outputs are not yet used for decisions. Compare system outputs with existing manual processes or expert assessments to identify inconsistencies.
- Validate mapping from forms to the CKG, ensuring that every variable appears correctly in the graph and that queries return expected results.

3. User-facing pilot
    - Activate LLM outputs for a subset of participants at pilot sites. Health workers see and can act on the system’s summaries, flags and recommendations, while recording whether they accepted, modified, or rejected them.
    - Provide simple in-app or paper forms for recording quick feedback: “useful / partly useful / not useful”, reasons for disagreement, and suggestions for improvement.
4. Quantitative and qualitative evaluation
    - Quantitative: evaluate sensitivity, specificity and calibration for key risk categories using pilot data; measure impact on referral patterns, screening completion, and follow-up rates where feasible.
    - Qualitative: organise interviews or focus groups with health workers, programme managers, and possibly participants, to understand usability, trust, clarity of language, and perceived value.
5. Decision on progression
    - Summarise results and bring them to the governance group.
    - Decide whether to:
        - remain in pilot with adjustments,
        - expand pilots to more sites,
        - or proceed to broader deployment for specific age groups or programmes only (for example, roll out child and

adolescent modules first, keep certain cancer modules in pilot).

Pilot and validation activities are not a one-time exercise; any major new feature or significant change in thresholds should trigger a scaled-down version of this process before being adopted widely.

7.3 Scale-Up, Maintenance, and Continuous Improvement

Once the system has proved safe and useful in pilots, the focus shifts to scale, reliability, and ongoing learning.

1. Planned scale-up
    - Gradually onboard new districts, facilities and schools in waves, rather than all at once.
    - For each wave, repeat basic readiness checks: data infrastructure, staff training, confirmed mappings from local workflows to standard forms and codes.
    - Monitor early usage and outputs in newly onboarded sites to catch configuration issues quickly.
2. Operational maintenance
    - Establish clear ownership for day-to-day operations of the data pipeline, CKG, LLM services, and user interfaces.
    - Set service-level expectations (uptime, response times) and procedures for handling technical incidents.
    - Maintain and update reference tables (e.g., facility lists, code sets, questionnaire versions) as part of normal operations, with proper version control.
3. Routine monitoring and feedback loops

- Use dashboards to monitor data completeness, number and distribution of risk flags, referral rates, and patterns of health worker overrides.
- Implement a structured channel (e.g., periodic surveys, helplines, in-app feedback) for users to report issues and ideas.
- Regularly review safety incidents or atypical outputs as part of governance meetings and prioritise improvements accordingly.

4. Model and rule updates
    - Over time, clinical guidelines, screening protocols, or local patterns may change. Updates might include:
        - new or revised questionnaires,
        - new risk calculators or updated algorithms,
        - adjusted thresholds and triage rules,
        - refined LLM prompts or guardrails.
    - Each update follows the change management process: design, offline testing, pilot testing where needed, governance approval, and controlled rollout with versioning and documentation.
5. Capacity building and institutionalisation
    - Provide ongoing training for new staff and refresher sessions for existing users, focusing on correct use, limitations, and interpretation of LLM outputs.
    - Strengthen in-country capacity for CKG maintenance, data engineering, LLM prompt engineering, monitoring and evaluation, so reliance on external expertise is reduced over time.
    - Document lessons learned and feed them back into broader digital health and diagnostic policies, so that the Free Health Check experience can inform other programmes.
6. Long-term evolution
    - As confidence grows and data accumulates, the system can support more advanced use cases, such as:

        - refined risk stratification that combines multiple domains,
        - population-level analyses to optimise screening strategies,
        - integration with additional digital services (e.g., follow-up appointment scheduling, remote counselling).
    - Any expansion remains anchored in the same principles: transparent data structures, clear clinical rules, robust evaluation, and strong governance.

Through this phased and iterative roadmap, the Free Health Check LLM system can move from concept to an integral part of routine screening, while maintaining a clear focus on safety, equity, and continuous learning.

8. Closing Remarks

This technical guideline provides a practical blueprint for how large language models (LLMs) should be designed, integrated, and governed within the Cek Kesehatan Gratis (CKG) / Free Health Check programme. It operationalises the broader Framework for AI-assisted Diagnostic Data Management and the accompanying LLM standards by translating high-level principles into concrete design patterns, data flows, and interaction models with the Health Data Graph (HDG), risk calculators, and clinical decision support systems. In doing so, it aims to ensure that AI-assisted diagnostic functions are consistently safe, explainable, and aligned with national policies, including the Digital Health Transformation Strategy (DHTS) 2.0 and the Personal Data Protection Law (UU PDP 2022).

Successful implementation of this guideline depends on coordinated action across multiple stakeholders. Programme owners and clinical leaders are expected to anchor LLM use in evidence-based screening protocols and maintain clinical oversight over final decisions. Digital health and interoperability teams, CKG administrators, and LLM developers are responsible for ensuring that data models, integrations, and model behaviours follow the requirements and controls outlined here, including versioning, monitoring, and auditability. Monitoring and evaluation teams, regulators, and governance bodies play a key role in reviewing performance, equity, and safety outcomes and in enforcing adherence to both the technical guideline and the underlying standards.

This document should be treated as a living reference rather than a static specification. As screening programmes evolve, new data sources become available, and evidence on AI and LLM performance in Indonesian settings grows, the guideline SHOULD be periodically reviewed and updated. Future iterations MAY expand to cover additional screening domains, new model architectures, or more detailed requirements for calibration, localisation, and safety monitoring. Feedback from front-line health workers,

programme teams, and technical implementers is essential to refine the guidance and ensure that it remains realistic, usable, and responsive to implementation challenges.

Ultimately, the aim of this guideline is to ensure that the introduction of LLMs into the Free Health Check programme strengthens—not replaces—clinical judgement, public trust, and health system resilience. By following the technical patterns and safeguards described here, implementers can help ensure that AI-assisted diagnostic support contributes to earlier detection, better continuity of care, and more equitable health outcomes across Indonesia, while maintaining accountability, data protection, and ethical integrity at every step.