The Codebook for GCC AI Strategy Document Analysis (v2)

A1.1 Purpose and Scope

This codebook specifies how references to capacity building and enabling conditions are identified and coded in the six GCC National AI Strategies (NASs). It combines qualitative coding (to assign thematic categories) with term-salience scoring (TF–IDF) to quantify emphasis on specific AI capabilities. The categories are aligned to the constructs used in the survey and model:

- **Technical Infrastructure (TI)** - Q2–Q4 *(reflective)*
- **Organizational Readiness (OR)** - Q5–Q7 *(reflective)*
- **Governance Environment (GE)** - Q8–Q9 *(formative)*
- **AI Outcomes (AIO)** - Q10–Q12 *(reflective; used for outcome tagging rather than NAS coding)*

Unit of analysis: paragraph or numbered policy bullet.
Document scope: only sections that mention capacity building, skills, training, resources/infrastructure, policy/ethics/regulation, or programs that enable AI adoption.

A1.2 Construct Definitions (for consistency across the paper)

- **Technical Infrastructure (TI):** Availability and quality of ICT and data foundations that enable AI (compute, networks/5G, cloud, platforms, datasets, tooling).
- **Organizational Readiness (OR):** Skills, leadership commitment, processes, and culture that allow agencies to leverage and manage AI resources.
- **Governance Environment (GE):** Policy and regulatory clarity, ethical guidance, oversight, and data-governance provisions that shape responsible AI use.
- **AI Outcomes (AIO):** Intended/observed public-sector results (service efficiency, citizen satisfaction, trust, KPIs).

A1.3 Inclusion / Exclusion Rules

Include passages that:

- Describe investments, programs, reforms, or directives that build TI/OR/GE.
- Mention training, upskilling, AI labs, data platforms, ethics guidelines, PDPLs, standards, oversight, leadership initiatives, or operational processes.

- Name specific AI capabilities (for example, NLP, CV, cybersecurity) in a capacity-building context.

Exclude passages that:

- Only describe vision statements with no clear capacity-building element.
- Report achieved outcomes without referencing the enabling inputs (code as AIO-tags only).
- Are press or foreword material lacking substantive policy content.

A1.4 Code Families and Codes

Use one primary code (TI / OR / GE) for each eligible passage. Apply secondary tags from AIO and Context Tags when relevant.
For each code, we list a definition, inclusion cues, exclusions, example keywords (EN/AR), and the survey linkage.

A. Technical Infrastructure (TI)

Code ID	Name	Definition & Inclusion Cues	Exclusions	Examples (EN / AR)	Survey Link
TI1	Connectivity & Compute	Networks, broadband, 5G, data centers, HPC/GPUs, edge	General "digital vision" with no infra content	*5G, cloud region, data center, compute capacity* / بنية تحتية، 5G، مركز بيانات، قدرات الحوسبة	Q2–Q4
TI2	Cloud & Platforms	Gov cloud, PaaS, MLOps, API gateways, shared AI platforms	Pure procurement policy without tech specifics	*government cloud, platform services, MLOps* / سحابة حكومية، منصات، تشغيل النماذج	Q2–Q4
TI3	Data Foundations	Data lakes, catalogs, interoperability, data sharing	Privacy/ethics rules (use GE3)	*data lake, data pipeline, interoperability* / بحيرة بيانات، مشاركة البيانات	Q2–Q4
TI4	Tools & Labs	AI sandboxes, testbeds, model repos, benchmarking	Academic research not tied to capacity building	*AI lab, sandbox, benchmark suite* / مختبرات الذكاء الاصطناعي، بيئات اختبار	Q2–Q4
TI5	Investment in TI	Budget lines, PPPs, incentives for infra	General funding for non-TI topics	*capital expenditure, PPP for cloud* / استثمار، شراكات	Q2–Q4

B. Organizational Readiness (OR)

Code ID	Name	Definition & Inclusion Cues	Exclusions	Examples (EN / AR)	Survey Link
OR1	Workforce & Skills	Training, certification, AI literacy, talent pipelines, hiring	University research grants w/out public-sector upskilling	*upskilling, reskilling, academy* / رفع المهارات، تدريب، أكاديمية	Q5–Q7
OR2	Leadership & Strategy	Executive sponsorship, governance bodies inside ministries, change mgmt	High-level national "vision" only	*steering committee, CoE, leadership mandate* / لجنة توجيه، مركز تميز	Q5–Q7
OR3	Process & Integration	Operating models, procurement/process reform, M&E for AI projects	Pure technology mentions (use TI)	*operationalization, SOPs, agile in gov, PMO* / نماذج تشغيل، إجراءات، متابعة وتقييم	Q5–Q7
OR4	Partnerships for Capability	Collaborations that transfer skills to gov (academia/industry/IFIs)	Pure research MoUs with no capacity transfer	*secondments, joint training, bootcamps* / شراكات تدريبية، إعارات	Q5–Q7

C. Governance Environment (GE)

Code ID	Name	Definition & Inclusion Cues	Exclusions	Examples (EN / AR)	Survey Link
GE1	Policy & Regulation	Acts, decrees, standards, compliance regimes; sectoral rules	Internal ministry process tweaks (use OR3)	*AI regulation, standard, conformity assessment* / تشريعات، معايير، لوائح	Q8–Q9
GE2	Ethics & Oversight	Ethics boards, risk frameworks, impact assessments, transparency	General "ethics" rhetoric with no instrument	*AIA, DPIA, bias mitigation, transparency* / أطر أخلاقيات، تقييم الأثر	Q8–Q9

Code ID	Name	Definition & Inclusion Cues	Exclusions	Examples (EN / AR)	Survey Link
GE3	Data Governance & Privacy	PDPLs, cross-border data rules, retention/consent, security baselines	Data infrastructure plumbing (use TI3)	*PDPL, data sharing policy, privacy by design* / حماية البيانات، حوكمة البيانات	Q8–Q9

D. Outcome Tags (AIO) - secondary tags only

Tag ID	Name	When to Tag	Examples	Survey Link
AIO1	Service Efficiency & Quality	Passage links inputs to service improvements	turnaround time, automation rates	Q10–Q12
AIO2	Citizen Satisfaction & Trust	Mentions of trust, satisfaction, grievance redress	surveys, trust indices	Q10–Q12
AIO3	Performance & ROI/KPIs	KPIs, benefits tracking	cost savings, ROI	Q10–Q12

E. Context Tags (optional, multi-select)

- **Sectors/Use-Cases:** health, education, justice, smart city, public finance, etc.
- **Localization & Inclusion:** Arabic NLP, accessibility, multilingual services.
- **Culture & Society:** norms, values, religion-sensitive guidance.

A1.5 Coding Procedure

1. Segment each NAS into paragraphs/bullets.
2. Screen for capacity-building scope (A1.3).
3. Assign one primary code (TI/OR/GE + specific sub-code).
4. Add secondary tags (AIO, Context) where applicable.
5. Record evidence: copy short quote + page/section.

6. Log synonyms you used to justify the code (EN/AR).
7. Escalate edge cases to a shared log; adjudicate weekly.
8. Quality control: double-code ≥15% per document; resolve; compute κ.

Inter-coder reliability: target Cohen's κ ≥ 0.79 (substantial).
If κ < 0.79, revise examples/synonyms and recalibrate.

A1.6 Term-Salience Scoring (TF–IDF) for Capabilities

We quantify how strongly specific AI capabilities (for example, NLP, CV, XAI, robotics, cybersecurity) are emphasized within capacity-building passages across the six NASs.

- **Vocabulary & Synonyms:** build a controlled list with EN/AR forms (for instance, *NLP / معالجة اللغة الطبيعية*).
- **Term Frequency (TF):** count each capability mention per NAS within coded passages only; de-duplicate obvious repeats within the same sentence/bullet.
- **Document Frequency (df):** number of NASs where the capability appears at least once.
- **Smoothed IDF:**

$$\boldsymbol{IDF(t) = ln\left(\frac{N+1}{df(t)+1}\right) + 1, \qquad N = 6}$$

(prevents zero IDF for ubiquitous terms and preserves comparability).

- **Average TF per doc:** $\boldsymbol{TF(t) = \frac{\sum dTFd(t)}{df(t)}}$
- **TF–IDF score:** $\boldsymbol{TFIDF(t) = TF(t) \times IDF(t)}$

Reporting: for each capability list df, total mentions, $TF, IDF, TF - IDF$.
Interpretation: higher TF–IDF = more concentrated emphasis (specialized but stressed); lower TF–IDF with high df = ubiquitous baseline capability.

A1.7 From Codes to Construct-Level Signals

To link NAS content to the survey/model:

- **TI signal (per country):** normalized count of TI1–TI5 passages (for example, per 1,000 words) + presence of TI investments.
- **OR signal:** normalized count of OR1–OR4 passages + presence of structural change items.

- **GE signal:** normalized count of GE1–GE3 passages + presence of binding instruments (weight +½ for advisory, +1 for binding).
- **AIO tags:** used for descriptive triangulation with Q10–Q12 (not part of TI/OR/GE tallies).

Normalize each signal to [0,1] within the corpus for comparability. These document-level signals do not alter the survey-based PLS-SEM; they are used for contextual validation and narrative alignment.

A1.8 Examples (coding shorthand)

- ***"Establishing a government cloud region with compliant data lakes"*** → TI2 + TI3 (primary TI3 if one only).
- ***"Launching an AI academy for civil servants with certification tracks"*** → OR1 (+Context: cross-ministry).
- ***"Mandating algorithmic impact assessments for high-risk uses"*** → GE2 (+AIO2 if linked to trust).
- ***"Approving PDPL amendments enabling cross-border data with safeguards"*** → GE3.

A1.9 Quality & Reproducibility

- **Pre-processing:** OCR only if needed; keep a synonym lexicon (EN/AR) in the repo; record version and date per NAS.
- **Audit trail:** retain coder, timestamp, document locators, code IDs, and snippet.
- **Change log:** track any updates to definitions or examples; re-run κ after substantial edits.

A1.10 Limitations

- Strategy documents differ in length/detail; normalize by word count to mitigate bias.
- Translations may miss nuance; when possible, cross-check Arabic terms in the lexicon.
- TF–IDF captures emphasis, not quality of implementation; triangulate with survey/PLS-SEM findings.