UNIVERSITY OF THE PHILIPPINES
CENTER FOR INTEGRATIVE AND DEVELOPMENT STUDIES

POLICY BRIEF 2026-25
ISSN 2619-7278 (Print)
ISSN 2619-7286 (Online)

PROGRAM ON ALTERNATIVE DEVELOPMENT

From Prompts to Specifications:

A Transitional Hybrid Governance Framework for Generative AI in the Public Sector

Mark Joseph M. Garrovillas[1]

Executive Summary

- Policy Problem (Status Quo): Public sector reliance on Conversational Prompts (Option 1) for high-stakes workflows introduces systemic risks to institutional memory. These ephemeral instructions lead to inconsistent reliability and compromised reproducibility, which complicates administrative oversight and erodes public trust.

- Root Cause (Problem Statement): A common failure mode in current AI interactions is the practice of Instructional Decay, where users treat instructions as temporary text blocks and discard them once a result is received. This retains the lossy projection of the output while losing the underlying policy logic, leaving no verifiable audit trail for high-stakes decisions.

- Policy Recommendation: This brief proposes the Hybrid Approach (Option 3) as a transitional governance framework, integrating conversational reasoning with schema-validated execution to bridge the gap toward Mandatory Specifications (Option 2). This model requires users to pair conversational prompts (the Reasoning Framework) with standardized, machine-readable specification files (the Execution Schema). This workflow functions as a governance anchor to ensure that institutional intent is captured in a persistent, version-controlled digital asset.

- System Integrity: By encoding policy constraints into specifications, agencies provide the formal documentation required for Explainable AI (XAI) validation. This hybrid model serves as a necessary bridge for all mission-critical government functions.

- Call to Action: To ensure institutional accountability, agencies must immediately mandate specification-based AI workflows and launch comprehensive upskilling programs in Specification Architecture, treating AI instructions with the same legal and procedural rigor as foundational regulatory contracts.

Problem Statement

Current reliance on unstructured prompts for high-stakes Generative AI introduces critical risks to reliability and institutional memory. This practice—termed 'Instructional Decay'—retains the output while discarding the underlying policy logic, creating a 'lossy projection' that lacks a verifiable audit trail. Without these persistent, formalized instructions, public sector AI interactions remain unsuitable for functions requiring accountability, reproducibility, and alignment with legal norms.

Background and Context

As government agencies integrate Generative AI into high-stakes decision-making, the method of instruction becomes a critical regulatory concern because an AI model's output is only as reliable as the instructions it receives. Public sector agencies

1 Junior ICT Associate, Core Facility for Bioinformatics - Philippine Genome Center, University of the Philippines - System

must move beyond casual conversational interactions toward a framework of formalized contracts, ensuring that machine execution remains strictly aligned with institutional policy intent.

Adopting this structured approach operationalizes the University of the Philippines Principles for Responsible AI, specifically fulfilling mandates for Accountability (Principle 5) and Transparency (Principle 7) (University of the Philippines 2023). By treating AI instructions as persistent, versioned digital assets—similar to how a nation's Constitution serves as a foundational specification for its laws—public sector professionals must evolve from prompt engineers into Specification Architecture practitioners to ensure that institutional values and logic are never lost through Instructional Decay (Grove 2025; Tiwari 2025).

Key Evidence and Findings

To understand the necessity of institutional standards for Generative AI, we must distinguish between the two primary ways we interact with it. This distinction centers on the contrast between discretionary conversational inputs and formalized technical documentation.

Prompt-Only Approach (The Conversation)

Prompts represent the fluid, conversational side of the interface. They function as a flexible reasoning guide that excels at creative ideation, natural dialogue, and brainstorming. However, prompts are inherently variable in reliability; running the same prompt twice often yields different results.

Prompts are frequently treated as ephemeral blocks of text. A common failure mode is instructional decay, defined here as the aforementioned procedural loss where original instructions are discarded once a result is received.

This ephemeral approach compromises system integrity, a risk best illustrated through these three analogies:

1. In Software Engineering, discarding instructions is analogous to deleting the source code while only version-controlling the binary output—a practice Fuentes (2025) rightly characterizes as creating 'lossy projections' that preclude verification of underlying logic.
2. Similarly, Architectural and Construction standards require retaining blueprints to verify structural decisions. Without this digital blueprint, auditors cannot confirm that final outputs adhere to safety constraints or design specifications.
3. Even in the Culinary Arts, reproducibility relies on the recipe. Discarding these instructions makes results impossible to replicate, rendering the process unsuitable for administrative environments where accountability is paramount.

To visualize the governance risk, consider a public health agency using conversational AI to triage financial support. Because the 'recipe' vanishes upon output, the agency loses the persistent record of the policy logic—a critical gap for institutional accountability and judicial review.

Sample Prompt (The Discarded Blueprint):

> "Review our registry of chronic patients and identify four individuals who need financial aid for their treatments. Provide a clear reason for their eligibility regarding their health status and financial status, and include a remarks section. Keep the tone professional, anonymize all identities for privacy, and ensure each explanation is under 15 words."

Unstructured AI Output (The Constructed Result):

> "Sure! I checked the registry and found four people who really need help.
>
> - PATIENT-K-105: Has Stage 4 Kidney Disease and income under 100k PHP. Priority for dialysis.
> - PATIENT-H-210: Chronic respiratory failure; needs oxygen support. High priority for low-income households.
> - PATIENT-R-315: Advanced Diabetic Neuropathy. Struggles with maintenance medication costs.
> - PATIENT-S-420: Chronic hypertension with heart risks. Needs supplemental aid for maintenance medication.
>
> Hope this helps your review process!"

While this response is "helpful," it illustrates the governance risks of a purely conversational approach:

the logic is unstructured, the "recipe" is lost, and there is no structured data for an Explainable AI (XAI) tool to verify if the AI actually adhered to official policy limits.

Specification Equivalent (The Contract)

Conversely, the Specification Equivalent approach provides a technical framework—a rigid, machine-readable rulebook ensuring tasks are executed with repeatable precision. This method replaces conversational ambiguity with a structured layout (YAML or JSON), allowing the AI to follow policy constraints deterministically.

Specifications act as an Authoritative Reference and a formalized contract. As Grove (2025) notes, this principle mirrors legal governance, where foundational documents—much like a nation's Constitution—serve as versioned national model specifications that derive all subsequent laws and actions. For government agencies, specifications are superior to prompts for three reasons:

1. Establishes a Governance Anchor: Specifications act as version-controlled references, ensuring that the logic governing a decision remains immutable and retrievable.

2. Ensures Full Auditability: By providing formal documentation, specifications enable Explainable AI (XAI) tools to validate that outcomes adhere strictly to legal and policy constraints.

3. Guarantees Deterministic Precision: Standardizing input protocols converts ambiguous conversational intent into exact, repeatable execution—a fundamental prerequisite for judicial review.

Synthesizing the frameworks of Fuentes (2025), Grove (2025), and Tiwari (2025), adopting this transition requires selecting text formats that bridge the gap between human reasoning and machine precision.

Ultimately, this transition requires public sector professionals to master the Specification Architecture of technical design. Rather than just talking to a model, staff must learn to embed regulatory constraints and institutional values into reproducible formats to ensure that the government remains accountable for every AI-assisted action.

The Structured Output	Schema Output: Sample JSON output from YAML specs for triaging financial support for patients with chronic illnesses scenario
This YAML specification acts as the rigid rulebook, mandating exactly four patients and splitting the rationale into discrete keys for granular auditing. Inline comments can be added to facilitate understanding of the specification.	

```yaml
# FILE: financial_aid_spec_v2.6.yaml
# PURPOSE: Encode policy for clinical and financial aid triage.
---
task_id: "Chronic_Illness_Financial_Aid_Triage"
version: "2.6"
source_data: "District_B_Clinical_Registry_2026.csv"
required_output_format: "JSON"

# CONSTRAINTS: The "Rules of Engagement"
# These encode policy intent and legal requirements for aid eligibility.
constraints:
  data_privacy: "Mandatory Patient ID Anonymization"
  eligibility_criteria: "Diagnosis: Chronic AND Income: < 250k PHP/yr"
  bias_prevention: "Exclude Gender/Religion from aid justification"
  logging_intent: "Full trace for Judicial Review" # Activates XAI validation

# OUTPUT SCHEMA: Defining Success
# Translates "fuzzy" intent into a structured, digital checklist.
output_schema:
  aid_recommendations:
    type: "array"
    max_items: 5 # Strictly aligned with monthly budget for 5 patients
    items:
      anon_id: "string"
      clinical_priority_score: "integer (1-10)"
      financial_aid_tier: "string (Tier_A, Tier_B, or Tier_C)"
      # Discrete keys allow XAI to test health vs. finance logic separately
      current_health_status: "string (max 15 words)"
      current_financial_status: "string (max 15 words)"
      remarks: "string (max 15 words)"
...
```

The Structured Output	Schema Output: Sample JSON output from YAML specs for triaging financial support for patients with chronic illnesses scenario

This JSON output demonstrates the predictability of the Specification Approach. By following the "skeleton" of the YAML specification, the AI provides structured, de-identified data that adheres to all legal and policy constraints.	{
	"task_metadata": {
	"task_id": "Chronic_Illness_Financial_Aid_Triage",
	"specification_version": "2.5",
	"timestamp": "2026-08-06T20:55:00Z",
	"audit_log_id": "XAI-AID-LOG-1022"
	},
	"aid_recommendations": [
	{
	"anon_id": "PATIENT-K-105",
	"clinical_priority_score": 10,
	"financial_aid_tier": "Tier_A",
	"current_health_status": "Stage 4 Kidney Disease; requires immediate dialysis.",
	"current_financial_status": "Annual income below 100k PHP threshold.",
	"remarks": "Urgent full subsidy recommended for life-saving care."
	},
	{
	"anon_id": "PATIENT-M-210",
	"clinical_priority_score": 9,
	"financial_aid_tier": "Tier_A",
	"current_health_status": "Chronic respiratory failure requiring oxygen support.",
	"current_financial_status": "Verified low-income household status.",
	"remarks": "High priority to prevent emergency clinical escalation."
	},
	{
	"anon_id": "PATIENT-R-315",
	"clinical_priority_score": 8,
	"financial_aid_tier": "Tier_B",
	"current_health_status": "Advanced Diabetic Neuropathy with complications.",
	"current_financial_status": "Significant treatment cost barriers documented.",
	"remarks": "Recommended for essential maintenance medication support."
	},
	{
	"anon_id": "PATIENT-S-420",
	"clinical_priority_score": 7,
	"financial_aid_tier": "Tier_B",
	"current_health_status": "Chronic hypertension with secondary cardiac risks.",
	"current_financial_status": "Qualifies for Tier B based on income.",
	"remarks": "Ensures medication continuity to prevent complications."
	}
	]
	}

File Format Suitability: Bridging Human Intent and Machine Execution

(Fuentes, 2025; Grove, 2025; Ramadurai, 2025; Tiwari, 2025)

The transition to a hybrid framework requires selecting text formats that bridge the gap between human reasoning and machine precision. Text-based formats are the native language of Generative AI systems, which are designed to ingest and output structured text with high fidelity. By standardizing input protocols, agencies transform ephemeral conversational intent into a verifiable digital contract.

- JSON and YAML—The Optimal Machine Interface: While humans prefer the flexibility of natural language, AI systems excel at processing structured formats like YAML and JSON. These serve as the execution schema or formal rulebook required for consistent performance. They allow for versionable, precise, machine-readable communication that avoids the ambiguous intent of conversational prose, making them ideal for integration with existing government IT automation.

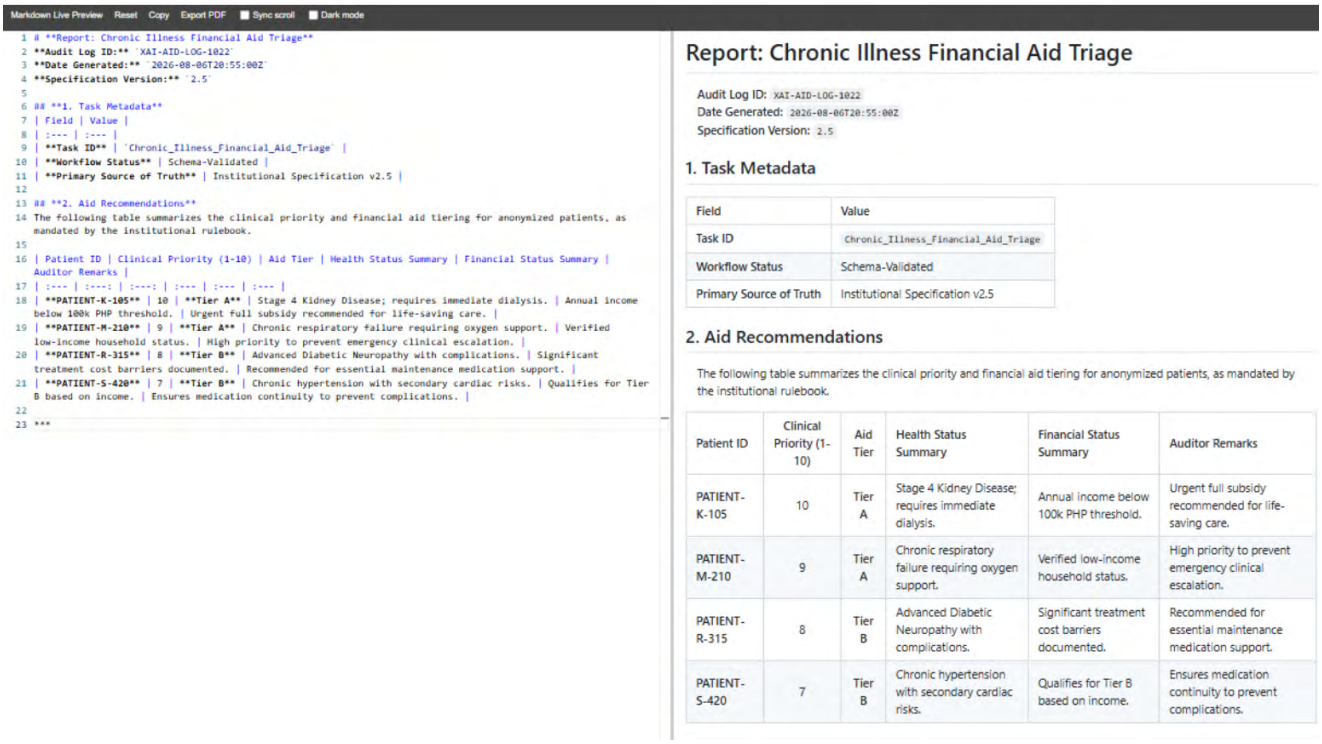
  - JSON (JavaScript Object Notation) is the industry standard for machine-to-machine communication, making it ideal for automation and integration with IT systems.

  - YAML (YAML Ain't Markup Language) is uniquely suited as a compromise format because it acts as a human-readable "executable contract" for input while maintaining the deterministic execution required for AI output.

- Markdown—The Optimal Human Interface: Markdown, on the other hand, is the superior format for authorship because it utilizes natural language, making policy specifications accessible to technical and non-technical stakeholders alike.

As a universal artifact that aligns human teams on shared goals, it establishes a foundational record that can be seamlessly rendered into formatted reports for administrative or judicial review. It allows for plain-text files that are easily versioned and change-logged.

Figure 1 illustrates Markdown outputs: A comparison of raw Markdown text versus its rendered formatted output for triaging financial support.



Strategic Advantages of Structured Formats:

- Token Optimization and Efficiency: Standardizing on these structured text formats is critical for Generative AI token optimization. By replacing lengthy, ambiguous prompts with precise specifications, agencies reduce the "inference-time compute" the AI model must spend on parsing natural language, resulting in more efficient performance at a lower cost.

- Auditability and Truth: Adopting these structured formats converts ambiguous intent into a version tracked foundational record. This creates the deterministic execution necessary for institutional accountability, ensuring that model behavior can be measured against explicitly formalizing administrative intent during judicial review.

TOON format as an Emerging Technical Standard:

Token-Oriented Object Notation (TOON) on the other hand, is an experimental and optional format designed for the AI era. While it can achieve significant token savings for high-volume automation, JSON and YAML remain the recommended stable industry standards for the current transitional phase. TOON should be treated as a next-generation optimization for high-volume tasks rather than a foundational governance requirement, as its specification remains in flux (Adhikary 2025; Schopplich 2025; Ramadurai 2025; Cappello 2025).

Legacy Protocol Constraints: Rationale for Excluding XML and High-Overhead Formats:

While traditional legacy structured formats like XML (eXtensible Markup Language) provide clear hierarchy, they are explicitly excluded from this framework due to extreme verbosity (Ramadurai 2025) and a high token tax (Cappello 2025). Benchmarks indicate that XML is the most computationally expensive common format, requiring nearly triple the tokens of next-generation standards like TOON for uniform records (Schopplich 2025). This punctuation tax creates unnecessary noise that disrupts model reasoning and increases inference-time compute costs (Grove 2025). Furthermore, XML provides a poor return on investment, yielding a retrieval accuracy of only 14.4% per 1,000 tokens (Schopplich 2025).

Side-by-Side Comparison:

Having established the fundamental risks of conversational prompting and the technical necessity of specifications, we now turn to the policy options available to mitigate these systemic threats.

Table 1. Comparative Analysis of AI Input Methods: Ephemeral Conversational Prompts vs. Persistent Institutional

Feature	Prompts (The Conversation)	Specifications (The Contract)
Reliability	Variable	Strong
Reproducibility	Low	High
Purpose	Creativity and Ideas	Automation and Workflows
AI Role	Guides reasoning (Reasoning framework)	Handles execution (Execution schema)
Auditability	Procedural loss (No record)	Authoritative Reference (Versioned)

The Evolving Skillset: From Prompting to Spec Authorship

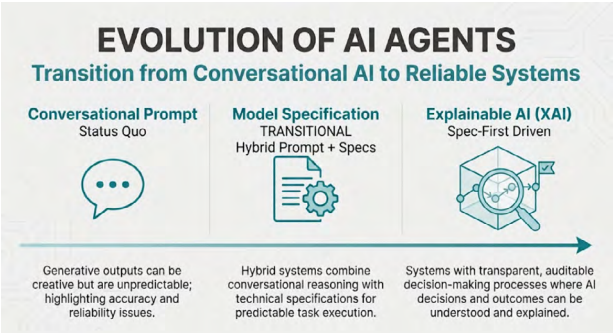
As government agencies increasingly rely on Generative AI for complex decisions, the responsibility of the public sector professional evolves from prompting to Specification Architecture. This new scarce skill involves the precise encoding of procedural norms into a reproducible and auditable format.

Writing robust specifications—treating them as foundational records—ensures that the AI stays on track, transforming helpful but ambiguous intent into deterministic execution (Fuentes 2025).

Bridging the Gap: The Compromise Hybrid Workflow

Transitioning from conversational interaction to structured documentation is a significant hurdle, as Specification Architecture—the precise engineering of machine-readable contracts—demands a specialized, high-level competency. To mitigate this barrier, the Hybrid Approach leverages the AI's conversational reasoning to bridge the critical gap between conceptual intent and technical execution.

Figure 2 illustrates Workflow Evolution: The transition from creative conversational logic to schema-validated execution. Adapted from Tiwari (2025) and generated with Google Gemini.



In this compromise workflow, conversational prompts are used for ideation and brainstorming to help the user translate policy goals into a formal specification. The AI acts as the Reasoning Framework to generate the Execution Schema. This ensures that while prompts may be used for iteration, only the version-controlled specification is used for production output. Managing the specification like a source code in a repository ensures it remains the Authoritative Reference, establishing the governance anchor necessary for Explainable AI (XAI) tools to validate that the results align with original policy goals.

The Hybrid Workflow Scenario illustrates the process of using a conversational prompt to generate a formal specification. Under the Hybrid Approach, a novice can use a simple conversational prompt to have the AI build a formal Execution Schema that ensures the task is repeatable and fair to generate the formalized documentation required for the Chronic Illness Financial Support Triage scenario.

Highly Simplified Spec-Authoring Prompt:

> "Help me write a formal spec file for an AI task called 'Chronic_Illness_Financial_Aid_Triage'.
>
> Goal: Use 'District_B_Clinical Registry_2026.csv' to find 4 patients who have a 'Chronic' illness and earn less than 250,000 PHP.
>
> Rules:
>
> - Privacy: Hide all names; use IDs only.

- Fairness: Don't use gender or religion to make decisions.
- Audit: Record every step the AI takes for audit review later.

Output: Create a list with an ID, a 1-10 priority score, an aid tier (A, B, or C), and short (under 15 words) summaries for health, money, and final remarks.

File Formats: output a YAML spec file that generates JSON output."

Policy Options

1. Option 1: Status Quo (Discretionary Prompting). This approach relies on ad-hoc conversational input, which prioritizes user creativity but introduces unacceptable variability and poor reproducibility. By treating instructions as ephemeral, this method precludes the creation of an Authoritative Reference essential for institutional oversight and judicial review.

2. Option 2: Mandatory Specifications. This framework mandates a spec-first protocol for all Generative AI tasks to maximize structural integrity and traceability. While it provides the highest level of reliability and auditability, it may restrict the agile, iterative workflows that are often necessary for initial brainstorming and creative ideation.

3. Option 3: Hybrid Governance (Recommended). This framework integrates discretionary prompting for ideation with mandatory, schema-validated specifications for production. By establishing conversational prompts as a Reasoning Framework initially and generating structured specifications as the Execution Schema transitionally, this approach ensures institutional intent is captured in an executable contract. It formalizes XAI logging, providing a sustainable transition where professionals evolve into Specification Architecture designers, effectively mitigating the risks of procedural loss.

Recommended Action: The Hybrid Transition Pathway

The primary recommendation is the immediate adoption of Option 3 (Hybrid Approach). This formalizes the natural workflow where users initiate tasks via Option 1 (Conversational Prompting) before refining them into Option 2 (Mandatory Specifications) for production. While discretionary prompting remains effective for ideation, the transition to structured, machine-readable specifications is imperative for institutional alignment. By mandating this shift, agencies preserve the documentation necessary for rigorous system accountability and external audit (Tiwari 2025; Batista 2025; Fuentes 2025).

All specifications—whether generated via Option 3 or authored from scratch—must be managed as version-controlled source code. This establishes a Primary Source of Truth from which all AI-assisted production flows, effectively eliminating the risk of 'Instructional Decay' and ensuring the persistence of policy intent.

Additional Applications for the Hybrid Approach

While high-stakes government functions necessitate a shift toward Mandatory Specifications (Option 2), the Hybrid Approach remains essential for specific professional applications:

- Low-Stakes and Creative Tasks: Agencies can potentially utilize Option 3 indefinitely for brainstorming, creative ideation, or tasks where human feel and flexibility are prioritized over strict, schema-validated reproducibility. In these contexts, conversational prompts function effectively as a reasoning framework to explore possibilities without the need for a rigid execution contract.

- Initial Reasoning and Heuristic Exploration: The hybrid model will likely remain a permanent feature of the initial heuristic exploration phase, where users explore ideas and test logic before formalizing those instructions into a final version-controlled contract for production execution. This allows staff to utilize AI's natural language strengths to distill complex stories into the execution schemas required for official records.

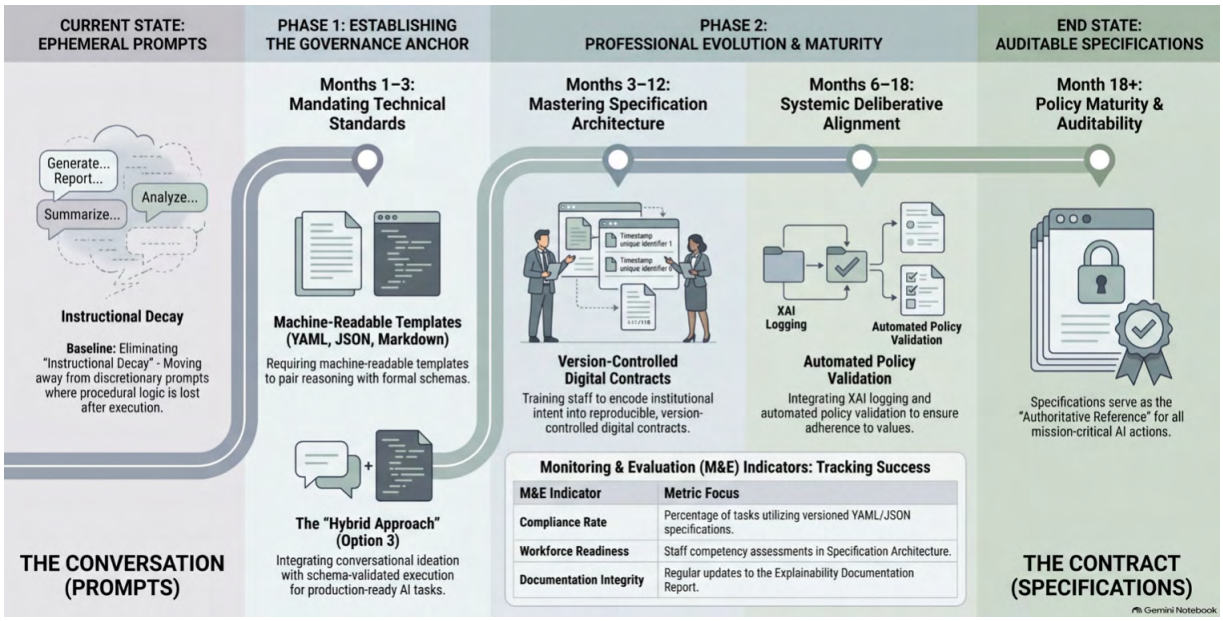
Implementation Plan

Operationalizing this hybrid governance framework requires the lead regulatory body to issue mandatory technical standards that treat AI specifications with the same rigor as traditional source code. The following roadmap outlines the critical, phased actions necessary to transition agency workflows from ephemeral prompting to persistent, auditable, specification-based execution.

Table 2. Policy Implementation Roadmap: Operationalizing the Hybrid Transition (Months 1-18+)

Critical Action	Responsible/ Accountable Unit	Timeframe	Budget
Phase 1: Hybrid Foundations (Option 3)			
Issue Technical Standards: Establish mandatory templates for structured machine readable specification files (YAML, JSON, Markdown) and provide guidelines for pairing them with conversational prompts.	Lead Regulatory Body (e.g., ICT Office or DICT)	Immediate (Months 1–3)	Agency Administrative & Operating Budget
Phase 2: Repository & Procurement Integration			
Deploy Standardized Digital Interfaces: Build and deploy standardized digital interfaces that ensure AI specifications are automatically version-controlled and treated as persistent source code. These YAML/JSON specifications are designed for high-interoperability with existing government legacy IT systems, such as databases and workflow management tools (Tiwari 2025; Batista 2025; Fuentes 2025).	Technical Resources/ IT Department	Short-term (Months 3–6)	Capital Outlay/ ICT Infrastructure Fund
Align Procurement Policies: Mandate that all procured Generative AI systems must validate and log input data from these structured specifications to create a verifiable audit trail (Grove 2025; Batista 2025; Fuentes 2025).	Procurement Unit/ Regulatory Body	Short-term (Months 3–6)	General Appropriations/ Procurement Budget
Phase 3: Workforce & Systemic Alignment			
Launch Comprehensive Training Modules: Launch comprehensive training modules to help policymakers and engineers evolve into technical designers. As noted in the frameworks synthesized by Fuentes (2025) and Tiwari (2025), this focuses on the Specification Architecture required for formalizing administrative intent and institutional values into reproducible formats.	HR Department/ Training Partners (e.g., DSPPP)	Continuous (Months 6–18)	Agency HR Development Fund/Partnership Grants
Implement Automated Training Procedures: Implement automated training procedures where human-authored specifications are used to update an institutional AI system's internal logic, internalizing policy as sort of muscle memory.	Data Science Team/Training Partners	Short-term (Months 6–18)	Research & Development Grants
Phase 4: Policy Maturity (Option 2)			
Mandate Schema-Validated Specifications: Mandate schema-validated, versioned specifications for all mission-critical AI tasks. This phase prioritizes systemic reliability and professional spec-authoring capacity, ensuring high-stakes decision-making remains auditable and reproducible.	Agency Executive / Head of Office	Long-Term (Month 18+)	General Appropriations

Figure 3 illustrates Governance Maturity: A phased 18-month strategic roadmap toward institutional policy alignment. Based from Tiwari (2025) and generated as infographic within NotebookLM.



Operationalizing Accountability

By treating the specification as the Authoritative Reference, the government establishes a governance anchor against which all AI behavior can be measured for compliance and legal adherence. Synthesizing the research of Fuentes (2025), Grove (2025), and Tiwari (2025), this plan ensures that the institutional focus shifts from writing traditional code to designing precise, communicable specifications that guarantee Generative AI outputs remain predictable and aligned with the public interest.

Monitoring and Evaluation Approach

To ensure institutional compliance and systemic alignment throughout the 18-month transition, the monitoring and evaluation (M&E) framework employs a dual-track strategy focusing on process integrity and outcome validation.

- Process-Based Tracking (Compliance & Governance): The Planning and Monitoring Unit will oversee institutional adherence to technical standards, ensuring all high-stakes specifications are version-controlled and managed as immutable source code (Fuentes 2025). Key metrics include:
  - Compliance Rate: Percentage of high-stakes AI tasks featuring versioned YAML/JSON specifications.
  - Workforce Readiness: 100% of mission-critical staff complete master-level competency assessments in Specification Architecture by Month 12.
  - Documentation Integrity: The regular update of the Explainability Documentation Report, a living artifact that tracks model behavior, data provenance, and fairness mitigation strategies to prevent concept drift over time.
- Outcome-Based Verification (Systemic Alignment): The Data Science Team will utilize Explainable AI (XAI) tools, such as Deliberative Alignment, to stress-test model outputs against defined policy intent. By establishing the specification as the Authoritative Reference, this framework creates a persistent governance anchor that prevents the erosion of policy intent during automated execution.

  This rigorous testing process functions analogously to judicial review, where evaluation results establish precedents—structured input/output pairs—that disambiguate and reinforce the original policy specification. According to Grove (2025), this ensures continuous administrative accountability, transforming Generative AI workflows into verifiable, auditable systems that maximize institutional reproducibility.

By formalizing the transition from ephemeral prompting to persistent, auditable specifications, agencies can secure the integrity of the public trust in an AI-driven administrative landscape.

Technical Appendix

Token-Oriented Object Notation (TOON) serves as a compact, human-readable encoding of the JSON data model specifically optimized for Large Language Model (LLM) prompts. (Adhikary 2025, Cappello 2025, Ramadurai 2025, Schopplich 2025).

- Structural Guardrails: TOON utilizes schema-aware headers, such as explicit array lengths ([N]) and field lists ({fields}), which give models a clear skeleton to follow. This reduces ambiguity and improves parsing reliability compared to unstructured prose.

- The Tabular Advantage: For uniform object arrays (e.g., person registries or logs), TOON collapses data into a table-like format that declares fields only once. This eliminates the redundant keys found in JSON, which are a primary source of token tax, an unnecessary overhead in token consumption caused by the verbose grammar and repetitive structure of some traditional data formats.

- Lossless Representation: TOON is entirely lossless relative to JSON, meaning any government data structure can be converted from JSON to TOON for LLM processing and back again without data degradation.

- Operational Reliability: By providing models with explicit lengths and field counts, TOON acts as a structural validator. This helps mitigate hallucinations and omissions when the model is asked to reason about high-stakes datasets.

Deliberative Alignment (OpenAI 2024) is a safety framework (developed by OpenAI) designed to ensure that an AI model's behavior remains strictly consistent with a human-authored rulebook (or specification). Rather than relying on the AI to "guess" the right thing to do during a conversation, this method proactively trains the system to follow encoded institutional values. To make this accessible for policymakers, the process can be viewed as a three-step digital training loop:

- Testing with Hard Scenarios: The system is presented with "challenging prompts"—difficult or controversial questions specifically designed to see if the AI will break the rules or show bias.

- Automated Policy Validation: A high-capacity evaluator model performs an automated audit of the output. It assesses the congruence between the stimulus, the AI response, and the governing policy specifications, assigning a quantitative alignment score based on established regulatory benchmarks.

- Internalized Policy Alignment: The validation metrics are utilized for iterative model optimization. Through this reinforcement process, the AI system internalizes the human-authored specifications, ensuring that policy adherence becomes a primary component of the model's core reasoning architecture.

This framework functions similarly to judicial review in a legal system. Just as a court uses the Constitution to decide if a new law is valid, Deliberative Alignment uses the specification to decide if an AI's action is valid. This process creates a library of "precedents" (successful examples), making the entire system testable and reproducible for government auditors.

Methodological Note

This policy brief employs a hybrid methodology, synthesizing original human-authored technical specifications with AI-assisted drafting and iterative editorial refinement. It serves as an independent scholarly contribution and does not reflect the official positions of the University of the Philippines System, the Philippine Genome Center, or the Center for Integrative and Development Studies (UP CIDS). All findings and recommendations are the author's own; readers are encouraged to exercise independent critical judgment regarding their application.

References

Adhikary, T. (2025). *What is TOON? How token-oriented object notation could change how AI sees data*. freeCodeCamp. https://www.freecodecamp.org/news/what-is-toon-token-oriented-object-notation/.

Batista, P. (2025, July 15). *Prompt engineering is dead: Spec-writing is the new superpower*. Medium. https://medium.com/@paulocsb/prompt-

engineering-is-dead-spec-writing-is-the-new-superpower-ee532894121e.

Cappello, A. F. (2025, November 18). *What the TOON format is (Token-oriented object notation)*. OpenAPI. https://openapi.com/blog/what-the-toon-format-is-token-oriented-object-notation.

Fuentes, A. (2025, July 15). *The new code: Everything is a spec. Medium*. https://falexm.medium.com/the-new-code-everything-is-a-spec-565be5b8a4b3.

Grove, S. (2025, July 12). *The new code* [Video]. YouTube. https://youtu.be/8rABwKRsec4.

OpenAI. (2024). *Introducing OpenAI's model specification*. OpenAI. https://openai.com/index/deliberative-alignment/.

Ramadurai, S. (2025, November 8). *TOON vs JSON: A modern data format showdown*. DEV Community. https://dev.to/sreeni5018/toon-vs-json-a-modern-data-format-showdown-2ooc.

Schopplich, J. (2025). *Token-oriented object notation (TOON): Compact, human-readable serialization of JSON data for LLM prompts*. GitHub. https://github.com/toon-format/toon.

Tiwari, A. (2025, September 10). *Prompts vs. specifications: How we'll talk to AI in the future. The Tech Nexus*. https://wisdomdeployed.substack.com/p/prompts-vs-specifications-how-well.

University of the Philippines. (2023, August 24). *Principles for responsible and trustworthy artificial intelligence*. University of the Philippines Board of Regents.

THE UP CIDS POLICY BRIEF SERIES

The UP CIDS Policy Brief Series features short reports, analyses, and commentaries on issues of national significance and aims to provide research-based inputs for public policy.

Policy briefs contain findings on issues that are aligned with the core agenda of the research programs under the University of the Philippines Center for Integrative and Development Studies (UP CIDS).

Center For Integrative And Development Studies

Established in 1985 by University of the Philippines (UP) President Edgardo J. Angara, the UP Center for Integrative and Development Studies (UP CIDS) is the policy research unit of the University that connects disciplines and scholars across the several units of the UP System. It is mandated to encourage collaborative and rigorous research addressing issues of national significance by supporting scholars and securing funding, enabling them to produce outputs and recommendations for public policy.

The UP CIDS currently has twelve research programs that are clustered under the areas of education and capacity building, development, and social, political, and cultural studies. It publishes policy briefs, monographs, webinar/conference/forum proceedings, and the Philippine Journal for Public Policy, all of which can be downloaded free from the UP CIDS website.

The Program

The Program on Data Science for Public Policy (DSPPP) aims to build the data science knowledge and capacities of academics, researchers, and policymakers, as well as, decision-makers from various sectoral stakeholders, and apply this learned skill to public policy and governance. DSPPP strives to engage a community of researchers within the university and encourage the pursuit of interdisciplinary problem-oriented research using high-level quantitative analysis. It seeks to convene multidisciplinary teams of social scientists, humanists, and scientists to research issues in the public sector.

Get your policy papers published.
Download open-access articles.

The Philippine Journal of Public Policy: Interdisciplinary Development Perspectives (PJPP), the annual peer-reviewed journal of the UP Center for Integrative and Development Studies (UP CIDS), welcomes submissions in the form of full-length policy-oriented manuscripts, book reviews, essays, and commentaries. The PJPP provides a multidisciplinary forum for examining contemporary social, cultural, economic, and political issues in the Philippines and elsewhere. Submissions are welcome year-round.

For more information, visit cids.up.edu.ph. All issues/articles of the PJPP can be downloaded for free.

Get news and the latest publications.

Join our mailing list to get our publications delivered straight to your inbox! Also, you'll receive news of upcoming webinars and other updates.

bit.ly/signup_cids

We need your feedback.

Have our publications been useful? Tell us what you think.

bit.ly/dearcids

UNIVERSITY OF THE PHILIPPINES
CENTER FOR INTEGRATIVE AND DEVELOPMENT STUDIES

Lower Ground Floor, Ang Bahay ng Alumni, Magsaysay Avenue
University of the Philippines Diliman, Quezon City 1101

cids.up.edu.ph

Telephone (02) 8981-8500 loc. 4266 to 4268
(02) 8426-0955

Email cids@up.edu.ph
cidspublications@up.edu.ph

Website cids.up.edu.ph