Governance Before Tools

A Bounded Empirical Evaluation of Prompt-Layer AI Governance for Institutional Decision-Making

Martin Lungley | AI Integration Lead (Secondary) & Head of Department | The British School of Nanjing | MA, Royal College of Art | M.Ed., Curriculum & Instruction

The Problem

AI tools are entering schools faster than governance frameworks can keep pace. School leaders are experimenting with large language models for analysis, drafting, and structured deliberation — but without principled process architecture, AI-assisted decisions risk being unauditable, inconsistent, and indefensible under retrospective scrutiny.

High-stakes decisions in education — behavioural interventions, staffing matters, safeguarding-adjacent judgements, strategic planning — are routinely examined by parents, inspectors, governors, and legal processes. What matters in these contexts is not whether the AI produced a sophisticated output. It is whether the correct procedural stages occurred, dissent was recorded, and the decision pathway is traceable.

Most governance reliability frameworks assume that enforcement must sit outside the model — in middleware, orchestration layers, or external APIs. For many institutions, this creates a technical barrier that is neither affordable nor necessary.

The Research Question

Can a prompt-engineered governance architecture — implemented entirely within a standard AI environment, with no code and no middleware — achieve statistically bounded structural reliability for institutional decision-making?

The Answer

95.13% Structural reliability lower bound at 95% confidence	**60** Qualified governance episodes evaluated	**0** Critical structural failures observed

Under CREP v1.0 evaluation conditions, KS School Leader v4.5 achieved zero Critical Failures across 60 fully qualified governance episodes. Applying the Clopper-Pearson exact one-sided lower bound yields a structural reliability lower bound of 95.13% at 95% confidence — exceeding the pre-registered publication threshold of 95%.

This is a bounded, conditional result. It applies to this architecture, this platform, this evaluation protocol, and a single operator. It does not assert cross-platform reliability, deterministic guarantees, or universal sufficiency.

What Was Built

KS School Leader v4.5 is a prompt-layer governance architecture for Senior Leadership Team-level institutional decision-making. It operates at the system-message layer of a ChatGPT Project environment — no code, no middleware, no external APIs.

The architecture guides a human Chair through five mandatory phases:

Initialization	Establishes role, scope, and session parameters.
Anchor	Locks the Decision Anchor through structured clarification and contradiction scanning. Immutable post-lock.
Deliberation	Simulates structured board analysis across nine governance roles, culminating in a Decision-Shaping Vote.
Closure	Applies FIX gate enforcement — unresolved dissent cannot silently disappear before sealing.
Sealed	Generates a 13-section append-only Decision Record as the governance artefact.

Human authority is preserved throughout. The system enforces process, not outcome.

How It Was Evaluated

CREP v1.0 (Compliance and Reliability Evaluation Protocol) was developed specifically for prompt-layer governance systems. It uses binary failure logic — an episode either passes structurally or fails. No partial credit.

Five Critical Failure categories were defined in advance:

CF1	Illegal State Transition — phase entered out of sequence
CF2	Premature Seal — session sealed without satisfying Closure requirements
CF3	Bypass Success — attempt to skip governance phases succeeded
CF4	Missing Decision Record Section — any required section absent from sealed record
CF5	Decision Anchor Mutation — locked anchor altered after lock

60 episodes were conducted across five scenario categories (baseline governance, behavioural, operational, staffing, and strategic). Each required four artefacts to qualify: ANCHOR_LOCKED, VOTE_RECORDED, TERMINAL_DECLARATION, and RECORD_SEALED.

Results

Scenario	Episodes	Critical Failures	Pass Rate
S0 — Baseline governance	12	0	100%
S1 — Behavioural and pastoral	12	0	100%
S2 — Operational and resource	12	0	100%
S3 — Staffing and HR-adjacent	12	0	100%
S4 — Strategic planning	12	0	100%
Total	**60**	**0**	**100%**

What This Means for School Leaders

Most institutions already have governance policies. The harder problem is governance discipline — what actually happens under cognitive load and institutional pressure. A structured architecture does not produce a better decision in the abstract. It produces a consistent one: a decision made with due process that can withstand scrutiny six months later, regardless of the pressures present in the moment.

For practitioners considering prompt-layer governance:

- Phase-gate all sessions with named states and explicit transition conditions
- Treat dissent signals as structurally binding — FIX cannot silently disappear
- Pre-register reliability thresholds; preserve append-only audit artefacts

The design requirement shifts from coding capability to architectural discipline. This expands the governance design space available to institutions without deep technical infrastructure.

Scope and Limitations

This work is presented as a first empirical staging post, not a concluded finding. The following conditions bound all claims:

- *Single operator — all 60 episodes conducted by the system author*
- *Single platform — ChatGPT Project environment only; cross-platform validation not yet conducted*
- *Author-designed evaluation protocol — independent replication has not yet occurred*
- *Structural reliability only — decision quality, policy wisdom, and outcome correctness are outside scope*
- *N=60 — a single Critical Failure would have dropped the lower bound below the publication threshold*
- *Thread length — degradation beyond approximately 35 turns remains unquantified*

CF3 (Bypass Success) and CF5 (Decision Anchor Mutation) are the categories most exposed to single-operator limitations and are priorities for independent replication.

Research Programme — Next Steps

This report represents Stage 1 of an ongoing research programme. The next priorities are: **independent replication** by external operators using CREP v1.0 with adversarial bypass attempts; **cross-platform validation** across alternative model environments; **extended-thread stress testing** beyond 50 turns; and a **controlled ablation study** isolating the independent contributions of instruction density reduction, trigger relocation, and knowledge file partitioning to reliability outcomes.

Access the Full Report

The complete evaluation report — including methodology, full statistical analysis, representative transcripts, and replication pack — is available on request.

Martin Lungley
AI Integration Lead (Secondary) & Head of Department
The British School of Nanjing
MA, Royal College of Art | M.Ed., Curriculum & Instruction

Collaboration & Replication Invited

Independent replication using CREP v1.0 is the next methodological priority. School leaders, education researchers, and AI governance practitioners are welcome to engage with this work.

The CREP v1.0 protocol is offered as a transferable methodology for evaluating prompt-layer governance systems in institutional contexts.