

Article

Cognitive Supervisory Control with LLM Reasoning Agent for Fault-Tolerant Process Systems: A Digital Twin Perspective

Alexios Papacharalampopoulos , Olga Maria Karagianni and Panagiotis Stavropoulos * 

Laboratory for Manufacturing Systems & Automation (LMS), Mechanical Engineering & Aeronautics Department, University of Patras, Panepistimioupoli—Rio, 26504 Patras, Greece

* Correspondence: pstavr@lms.mech.upatras.gr; Tel.: +30-2610-91-01-60

Abstract

Fault-tolerant control systems deployed in manufacturing and process industries must maintain output regulation under actuator degradation and parametric drift while producing operator-readable intervention records. This paper proposes a three-layer cognitive supervisory control framework: a continuously synchronized nominal plant model serves as a lightweight digital twin, driving residual-based anomaly detection that responds to model-mismatch faults before tracking error accumulates; an event-triggered supervisory layer applies composite anomaly scoring, risk estimation, and dwell-time-constrained mode switching over a bank of pre-verified stable controllers; and an optional bounded LLM reasoning agent forming Layer 3 of the cognitive architecture produces real-time natural-language audit records at each supervisory event. The framework is functionally non-intrusive—the closed-loop response is numerically identical to standalone LQI in three of four benchmark scenarios and on an unstable plant variant, with zero supervisory interventions. Under abrupt 75% actuator loss, a single targeted intervention achieves lower overshoot and better tracking than all baselines, including MPC whose feasibility-degradation under reduced actuator authority is confirmed empirically. On near-undamped plants representative of compressor and pipeline dynamics, integral absolute error improves significantly relative to standalone LQI; on a distributed thermal process representative of laser welding and induction heating, IAE improves substantially under abrupt heater power loss. Parametric sensitivity analysis across fault severity, detection threshold, onset timing, measurement noise, score weights, and dwell-time confirms robustness of the detection architecture. The LLM layer's contribution is regulatory rather than performative: its value is the causally connected audit record generated at the moment of each decision, not IAE improvement—a form of real-time explainability that after-the-fact attribution methods such as SHAP and LIME cannot provide by construction.



Academic Editors: Sheng Du, Li Jin, Pan Yu and Hao Liu

Received: 29 May 2026

Revised: 11 July 2026

Accepted: 12 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: cognitive supervisory control; fault-tolerant control; large language model; digital twin; switched systems; LQI; explainable AI; process systems; anomaly detection

1. Introduction

1.1. The Problem

Process control systems operating under real-world conditions, especially in manufacturing, must maintain output regulation despite component degradation, parametric drift, and unforeseen disturbances. The scientific challenge addressed in this paper is the following: how can a supervisory control layer be added to an existing controller that (a) detects faults before tracking error accumulates, (b) intervenes in a non-invasive to real-time way,

(c) produces an operator-readable record of every decision, and (d) adds zero overhead during healthy operation? No existing framework addresses all four simultaneously—classical FTC addresses (a) and partially (b) but not (c) or (d); explainable AI addresses (c) but not (a), (b), or (d); monitoring architectures address (c) and (d) but not (a) or (b). Despite sustained research effort in fault-tolerant control, three gaps remain persistently unresolved across the literature: real-time adaptability under uncertainty, formal stability guarantees under mode switching, and operator-readable intervention logic [1–3]. Classical controllers such as the LQR (Linear Quadratic Regulator) and model predictive control deliver strong nominal performance but are fundamentally limited under faults: standard LQR designs can lose stability guarantees under actuator degradation [4], while MPC (model predictive control) feasibility regions shrink and constraint satisfaction can no longer be ensured once actuator effectiveness drops below a critical threshold [5,6]. Supervisory and hierarchical architectures address this by coordinating pre-designed controllers through logic-based mode switching, with stability under switching established via average dwell-time conditions [7,8]; such schemes have been applied to aircraft and microgrid fault tolerance [9,10]. However, existing frameworks rely on pre-defined switching logic and explicit fault models, limiting adaptability to gradual or unanticipated degradation [2,3]. Supervisory FTC schemes with dwell-time-constrained switching have been proposed that jointly handle fault detection and controller reconfiguration [11,12], but these rely on pre-defined fault models and fixed switching logic, and produce no human-readable intervention records.

In summary, classical controllers lack fault adaptability; supervisory switching frameworks require pre-defined fault models and do not produce audit records; adaptive FTC sacrifices formal stability guarantees during adaptation; and no existing approach simultaneously addresses all three gaps of real-time adaptability, formal stability under switching, and operator-readable intervention logic.

1.2. Cognitive Control and Supervision Based on an LLM (Large Language Model)

The engineering community has pursued the combination of control and reasoning under the framework of cognitive control—defined as a self-aware, goal-directed paradigm that integrates perception, learning, and reasoning within a closed loop, enabling a system to select among actions based on context and monitor outcomes [13,14].

Current, LLM-based supervisory architectures face four deployment-critical limitations. Latency: LLM inference operates on second-scale timescales, incompatible with millisecond-scale closed-loop control demands, and cloud-hosted inference adds non-deterministic delay unsuitable for latency-sensitive automation [15,16]. Hallucination: hallucination appears intrinsic to current autoregressive LLMs and is unlikely to be eliminated by scaling alone [17]. Consistency: outputs vary across prompt phrasings and model versions, creating reliability barriers in safety-critical deployment [18]. Formal guarantees: no LLM-based supervisory architecture has demonstrated closed-loop stability certificates. The present work addresses each directly: latency is resolved by placing the LLM above the control loop with a non-blocking call; hallucination and consistency are contained by a bounded action set and zero-temperature inference; and stability is guaranteed by the pre-verified controller bank independently of LLM availability.

In practice, recent work has placed LLM agents above industrial control loops in supervisory and diagnostic roles. Vyas and Mercangöz [15] demonstrate agentic LLM control using digital twin validation and FSM-based fault recovery planning, achieving high action accuracy in laboratory temperature control experiments. In contrast, this work uses the continuously evaluated nominal-model residual as the online fault signal driving controller switching, capable of responding to model-mismatch faults before large tracking error accumulates. Additionally, LLMs have also been used to translate natural-language specifi-

cations into formal artefacts for symbolic control synthesis, where a downstream solver provides the stability certificate while the LLM handles specification parsing [19]. Separately, diagnostic and decision-support architectures have been proposed for batch process supervision, combining deterministic rule-based agents with LLM-driven conversational interfaces for preventive maintenance planning over industrial cleaning processes [20]; in these frameworks, the LLM has no authority to switch controllers or command actuators, and no stability conditions are required, since the problem addressed is process monitoring rather than closed-loop fault recovery.

Most recently, Gill et al. [21] proposed an LLM-agent architecture for fault handling in process plants using digital twins for corrective-action validation before physical deployment—one of the closest existing architectures to the present work. The key distinctions are three: their framework uses the digital twin as a simulation environment for action screening rather than as a continuously evaluated residual signal, without considering formal switched-controller stability analysis; and intervention records are generated only when the LLM agent is queried, not automatically at every supervisory event.

These recent studies illustrate the growing interest in integrating LLMs with digital twins for supervisory fault handling and operator support, while leaving open the challenge of combining explainable supervisory reasoning with formally verified fault-tolerant control.

Recent studies and reviews consistently position LLMs as supervisory or decision-support layers integrated with digital twins, validation mechanisms, or structured verification pipelines, rather than as unconstrained low-level controllers, with reliability, latency, and industrial-data governance remaining the primary barriers to autonomous deployment [21–23].

Recent literature published between 2023 and 2026 highlights three complementary developments that further contextualize the current work. Digital twin research has progressed from virtual representations of physical assets toward AI-assisted frameworks supporting real-time monitoring, fault diagnosis, and predictive maintenance, although model fidelity, synchronization, interoperability, and cybersecurity remain significant barriers to industrial deployment [24–26]. Explainable AI for industrial automation has similarly evolved from general trust-and-transparency principles toward domain-specific fault-diagnosis and process-monitoring applications, while recent reviews consistently identify the lack of standardized evaluation methodologies and deployment practices as major unresolved challenges [27,28]. In parallel, fault-tolerant control research has gradually shifted from classical passive, active, and hybrid architectures toward intelligent, adaptive, and digital twin-enabled supervisory frameworks, with rapid fault estimation and validated real-time deployment identified as the principal remaining challenges [29,30]. These developments collectively motivate the integration proposed in this work: a digital-twin-derived residual as the online fault indicator, an LLM-assisted audit record providing a practical mechanism for traceable, operator-oriented explanations, and a switched fault-tolerant control architecture that retains formal dwell-time stability guarantees.

1.3. The Auditability Gap and Added Value

Beyond the control-theoretic gaps above, classical FTC (Fault-Tolerant Control) approaches do not produce human-readable intervention records—a gap increasingly pressing under emerging regulatory frameworks [31]. The explainable AI literature has established that interpretability and traceable reasoning are prerequisites for deploying AI in safety-critical industrial environments [32,33]. Existing approaches address interpretability through after-the-fact attribution methods such as SHAP and LIME, applied after a decision has already been made [32,33]. The distinction from the present work is categorical rather

than incremental: SHAP and LIME attribute importance to inputs after the fact, requiring additional instrumentation and producing explanations that are not causally connected to the specific decision made; the LLM audit record here is generated as part of the decision process itself, at the moment it occurs, and requires no post hoc computation. This places the proposed approach in the output-level explanation category identified in [32]—the least explored and most practically relevant for operator-facing deployment.

Existing FTC logging architectures record inputs, outputs, and timestamps to support post hoc audit [34,35], but these capture what happened rather than why a specific controller mode was selected. Rule-based supervisory systems can expose which rule fired and under what conditions, but such explanations are generated independently of formally constrained controller selection and do not constitute causal intervention records in the sense required here.

While industrial alarm systems, expert systems, and rule-based supervisory architectures have long produced contemporaneous explanations for operational decisions, such explanations are typically advisory, diagnostic, or generated independently of formally constrained controller selection. The novelty here is not natural-language explanation *per se*, but its integration into a fault-tolerant supervisory control architecture in which each audit record is generated contemporaneously with a bounded controller-selection decision, while closed-loop stability is preserved through pre-verified controller modes and dwell-time-constrained switching independently of LLM availability. To the authors' knowledge, no prior cognitive supervisory FTC framework has been reported that jointly combines online residual-driven detection, bounded LLM-assisted mode selection over pre-verified stable closed-loop controllers with dwell-time-constrained switching, automatic real-time natural-language audit logging, and functionally non-intrusive behavior that remains numerically identical to a standalone LQI (linear quadratic integral) during healthy operation. The motivation for combining these three elements is that each addresses a gap the other two cannot fill. The digital twin provides a real-time fault signal derived from model-plant divergence—responding to parametric faults before tracking error accumulates, which neither a standalone controller nor an LLM can do. The supervisory layer provides the stability guarantee and the intervention mechanism—the LLM alone cannot be trusted to command actuators directly in safety-critical environments, and the digital twin alone provides no control action. The LLM provides the decision rationale—neither the digital twin nor the supervisory layer produces human-readable audit records by construction. The three elements are therefore complementary by design, not by convenience. The specific contributions realizing this are:

1. A cognitive supervisory architecture that is functionally non-intrusive in fault-free conditions—the closed-loop response is numerically identical to standalone LQI in three of four benchmark scenarios with zero supervisory interventions, confirmed also on an unstable open-loop plant variant.
2. An event-triggered detection scheme based on composite anomaly scoring and risk estimation. The detection threshold is set as a fixed multiple of the nominal settled score; in the benchmark scenarios evaluated, the nominal and fault-case scores are separated by more than an order of magnitude, making the detection boundary practically interpretable and adjustable by practitioners without requiring closed-loop fault simulations.
3. An empirical evaluation across plant variants with varying open-loop damping ratios, showing that cognitive supervisory benefit increases as damping decreases. Improvement ranges from 0% on well-damped and unstable plants to 20.7% on near-undamped plants representative of compressor and pipeline dynamics. This gives practitioners a qualitative indicator: plants with lightly damped open-loop dynamics

are the primary candidates for cognitive supervisory augmentation, and open-loop frequency response measurements can guide deployment decisions without requiring closed-loop fault simulations.

4. A safe Layer 3 LLM integration scheme, completing the cognitive supervisory architecture, with bounded actions, hardcoded fallback to the nominal controller on inference failure, and real-time natural-language audit logging—providing intervention records without after-the-fact attribution, addressing a transparency dimension that emerging regulatory frameworks such as the EU AI Act are designed to promote [31–33].
5. A switched LQI architecture in which each selectable mode corresponds to a pre-verified stable closed-loop matrix and a dwell-time constraint satisfies average dwell-time stability conditions [7], providing the formal connection between LLM action selection and closed-loop stability that existing frameworks lack.

Each contribution is validated empirically in the results: Contribution 1 in Sections 4.1.2 and 4.1.5; Contribution 2 in Section 4.1.6; Contribution 3 in Section 4.1.5; Contribution 4 in Section 4.3; Contribution 5 in Section 2.2 and the Lyapunov approach.

2. Method

The proposed framework comprises three strictly separated layers as shown in Figure 1. Strict separation is enforced through three constraints on information flow and decision authority. First, Layer 2 has read access to Layer 1 signals (measured states, control effort, and tracking error) but write access only to the mode-selection output—it cannot modify controller gains, reference signals, or any Layer 1 parameter directly. Second, Layer 3 has read access to the Layer 2 detection state (scores, risk, and detection time) but write access only to the bounded action selection—it cannot bypass Layer 2 action limits, modify thresholds, or issue direct actuator commands. Third, failure propagation is strictly downward: an LLM inference failure causes fallback to the Layer 2 surrogate decision; a Layer 2 failure causes fallback to the nominal controller. No upward failure propagation is possible by construction. The framework is a switched control system [7,8]: one or more controllers are pre-designed offline, and the supervisory layer selects between them at runtime based on measured process signals. What distinguishes this from standard supervisory switching is the decision mechanism—a joint anomaly-rate condition, a risk estimator, and an optional LLM—replacing the fixed scalar thresholds of conventional supervisory switching. The theoretical foundation of the switched architecture rests on three elements: pre-verification of each controller mode via CARE (Section 2.1), event-triggered mode selection with AND detection logic that rejects isolated anomalies (Section 2.2, Step 5), and average dwell-time-constrained switching with a computed minimum bound $\tau_D^* = 11.03$ s (Section 2.2, Step 5). Together these ensure that inner-loop stability is maintained in every reachable mode and that the framework reduces exactly to a standalone LQI when no fault is detected.

2.1. Layer 1—Inner Control Loop

The inner loop executes the currently active controller, implementing the control law $u(t) = -K x_a(t)$ where K is the gain matrix of whichever mode has been selected by the supervisory layer. At initialization and during fault-free operation, this is the nominal LQI controller, whose gain matrix is computed by solving CARE for designer-chosen cost matrices [36]. Every controller in the bank is verified to produce a stable closed-loop matrix by construction of CARE [36]. The inner loop runs at full simulation rate and does not adapt online; all adaptation is handled by Layer 2 above.

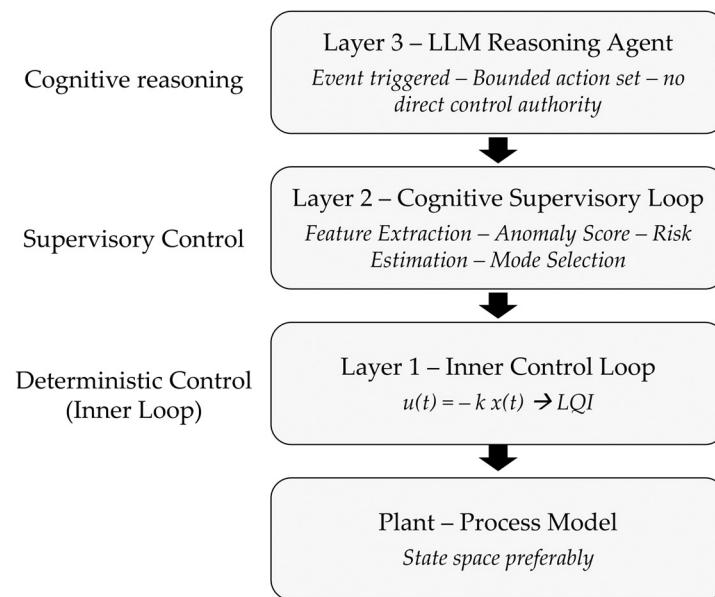


Figure 1. Multi-Layer Cognitive Supervisory Control Architecture. Strict layer separation ensures inner-loop stability is unaffected by supervisory or LLM decisions.

2.2. Layer 2—Supervisory Loop

The supervisory layer operates in an event-driven manner and never modifies inner-loop gains directly. Figure 2 shows the processing pipeline: six sequential steps convert raw measurements into a mode-switching decision.

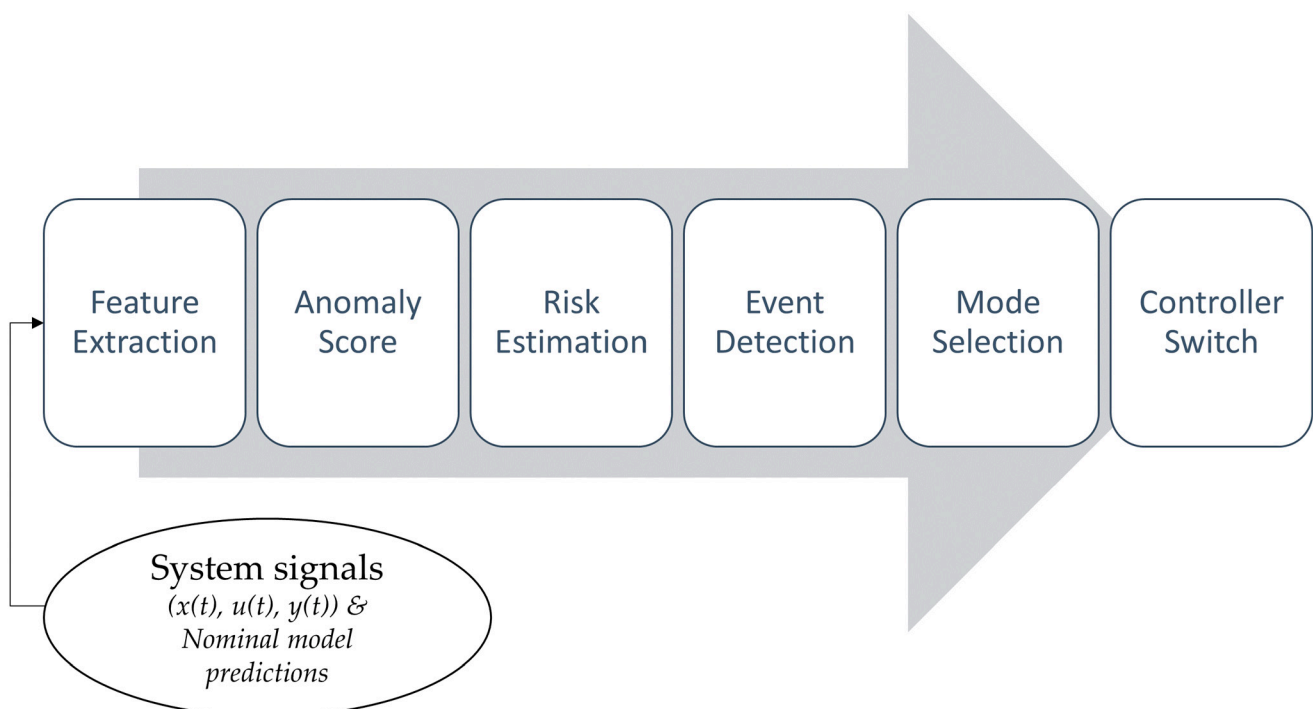


Figure 2. Cognitive supervisory loop signal processing pipeline. Six sequential steps convert raw measurements into a mode-switching decision.

Step 1—Feature extraction. At each timestep, the layer computes three signals from plant measurements: the absolute tracking error $|e(t)|$, the one-step model residual $\|x(k) - \hat{x}(k)\|_2$ where $\hat{x}(k)$ is the nominal model's prediction of the current state from the previous state and control input, and the squared prior control effort $u^2(t-1)$. The

one-step prediction $\hat{x}(k)$ constitutes a lightweight real-time digital twin of the plant: a continuously synchronized nominal model whose divergence from the measured state provides the primary fault signal, capable of responding to model-mismatch faults before a large tracking error accumulates [37].

Step 2—Anomaly score. A composite scalar anomaly score combines the three features into a single number (Equation (1)):

$$\text{score}(t) = w_1 |e(t)| + w_2 \cdot \text{residual}(t) + w_3 \cdot u^2(t-1) \quad (1)$$

The residual is weighted more heavily than the tracking error because the model residual responds to parametric faults before tracking error builds to detectable levels [37]. This earlier response is the principal advantage of residual-based detection over output-error-only detection. The control effort term captures situations where the controller is working unusually hard—a precursor of actuator stress—without yet producing large tracking errors. During the initial step-response transient, the score is suppressed to avoid false detections driven by normal startup dynamics rather than faults. The score rate is defined as the one-step finite difference (Equation (2)):

$$\text{score_rate}(t) = \text{score}(t) - \text{score}(t-1) \quad (2)$$

Step 3—Threshold selection. The detection threshold is set as a fixed multiple of the nominal settled score, providing a clear safety margin above normal operation while remaining well below the minimum score observed under the fault conditions to be detected. A rate threshold is applied in parallel to require that the score is actively rising, not merely elevated. Practitioners can adjust the multiplier depending on the measurement noise level in their specific plant: a noisier environment requires a larger multiplier. For plants whose nominal behavior is unknown prior to deployment, the threshold multiplier can be set conservatively high ($\kappa \geq 10$) and refined as nominal operating data accumulates—the 15-unit safe window identified in Sweep B provides practical guidance. For plants with slowly time-varying nominal behavior, a fixed threshold based on initial settled scores will drift over time and eventually produce false positives. The thermal case study addresses this directly: instead of a fixed nominal settled score, an adaptive baseline tracks the exponential moving average of the score, so the detection threshold updates continuously with the evolving nominal operating point. This approach transfers directly to any plant where nominal behavior changes slowly relative to the fault detection timescale.

Step 4—Risk estimation. A risk estimator combines instantaneous score, score rate, tracking error, model residual, and control effort into a single number between 0 and 1. This is a logistic function of a linear combination of these features, where the coefficients are chosen so that the output stays below 0.1 during normal settled operation and rises sharply when multiple fault indicators align simultaneously. For the benchmark plant, the calibrated form is $\text{risk}(t) = \sigma(-2.2 + 3.0 \cdot |e| + 4.0 \cdot \text{residual} + 0.015 \cdot u^2 + 0.12 \cdot \text{score_rate})$, where σ denotes the logistic function; coefficients were fixed by requiring $\text{risk} < 0.1$ during settled nominal operation and $\text{risk} > 0.6$ within 1 s of the abrupt actuator fault. The risk value serves two purposes: it determines whether an event triggers at all (together with the score threshold), and it determines which mode to select if a trigger occurs. Thinking of it as a continuous fault severity indicator distinguishes it from binary threshold detection—it allows graduated responses proportional to the estimated fault severity. The calibration procedure requires only data from normal plant operation. The coefficients are chosen so that the risk output stays below 0.1 during healthy settled conditions and rises above 0.6 within 1 s of a fault—either a simulated fault or a historically observed one. These two conditions are sufficient to determine the coefficients and can be satisfied using any

standard optimizer. Sweep E confirms that the exact coefficient values are not critical provided the residual term carries the largest weight. To apply the framework to a different plant, the same two conditions are re-applied to nominal data from that plant—no fault experiments are required. The dwell-time parameter is insensitive to its exact value in single-event scenarios (Sweep F) and should be set $\geq \tau^*_D$ for plants where repeated fault-recovery cycles are expected.

Step 5—Event detection and mode selection. The controller bank contains two pre-computed modes: a nominal LQI controller designed for the healthy plant, and a degraded LQI controller pre-computed offline for the fully degraded operating point. Both are available for immediate activation without any online computation. An event triggers when the anomaly score exceeds the detection threshold AND simultaneously either the score rate exceeds the rate threshold or the risk exceeds a high-risk level. The AND condition requires simultaneous evidence of fault presence and fault escalation, rejecting isolated anomalies. Upon triggering, the risk value determines the action: low risk retains the nominal controller; medium risk switches to the degraded controller; high risk combined with high actuator effort selects an energy-saving mode. This energy-saving mode applies the degraded controller gains with an additional control effort cap, reducing actuator stress when the controller is operating near its authority limit; it is selected over the degraded controller when high actuator effort suggests that unrestricted degraded gains would drive the actuator further into saturation rather than relieving it. A dwell-time counter is activated to prevent immediate mode reversion. Lyapunov analysis of the two switching modes confirms that all four reachable closed-loop matrices are Hurwitz (maximum $\text{Re}(\lambda)$ of -0.144 , -0.196 , -0.313 , and -0.370 , respectively), establishing individual mode stability. The theoretical average dwell-time bound following Hespanha and Morse [7] is $\tau_D = 11.03$ s, computed from the Lyapunov mismatch ratio $\mu = 23.9$ and slowest decay rate $\lambda = 0.144$. The operating regime addressed here—sparse process faults with at most one supervisory intervention per fault event—satisfies this condition trivially: the system switches once and remains in the selected mode until autonomous recovery. For deployments with repeated fault-recovery cycles, τ_{dwell} should be set $\geq \tau^*_D$; the `request_human_review` action provides an additional governance pathway for intermediate reconfiguration without autonomous switching. Each selectable mode maps to a pre-verified stable closed-loop matrix, ensuring inner-loop stability is maintained in every reachable mode.

When no supervisory event triggers, Layer 2 performs read-only computations on measured signals and issues no mode-switch command. The active controller remains the nominal LQI throughout, and the closed-loop dynamics reduce to $\dot{x}(t) = (A - BK_{\text{nom}})x_a(t)$, which is identical to standalone LQI by construction. Non-intrusiveness is therefore a structural property of the architecture, not an empirical observation: it holds whenever the joint AND detection condition is not satisfied, regardless of plant or fault type.

The complete per time-step execution sequence is summarized as Algorithm A1 in Appendix B.

2.3. Layer 3—LLM Reasoning Agent

The LLM layer is activated once per fault event, after a short delay following initial detection that allows the risk signal to stabilize. It receives a structured summary of the current detection state: detection time, instantaneous and recent-averaged anomaly score and rate, model residual, control effort, risk value, current mode, and time since detection. The LLM is instructed to return only a valid structured response selecting from a bounded set of actions—keep nominal, switch degraded, energy-save mode, or request human review—together with a confidence level and a plain-language rationale. The full system prompt is reproduced in Appendix A.

The LLM output is strictly bounded: if the returned action falls outside the allowed set, the supervisory layer’s own risk-based decision is retained. If LLM inference fails or times out, the fallback is the nominal controller, which is always stable. This ensures that an LLM failure propagates no further than one supervisory timestep and that inner-loop stability is never contingent on LLM availability. The natural-language rationale produced at each intervention is logged automatically as part of the decision record. Table 1 directly compares the Cognitive (no LLM) and Cognitive + LLM configurations: both achieve identical IAE and intervention count, with the audit record as the sole differentiator—confirming that the LLM layer’s contribution is governance and explainability rather than performance.

Table 1. Configuration comparison across evaluated methods. IAE values correspond to the abrupt actuator fault scenario ($\alpha = 0.25$, $t = 10$ s). Stability refers to formal closed-loop guarantees under the stated fault conditions; the MPC entry is marked with an asterisk because its feasibility guarantee degrades once actuator effectiveness drops below a critical threshold, as demonstrated in Section 4.1.3.

Configuration	IAE	Detection	Interventions	Audit Record	Stability	Performance Gain vs. LQI
LQI	2.366	no	none	no	yes	0%
MPC	7.089	no	none	no	yes *	−200%
CFG	3.214	no	none	no	yes	−36%
Cognitive (no LLM)	2.313	yes	1	no	yes	2.3%
Cognitive + LLM	2.313	yes	1	yes	yes	2.3%

* MPC stability guarantee degrades under reduced actuator authority.

Table 1 summarizes the three configurations evaluated in this paper—LQI alone, Cognitive without the LLM layer, and Cognitive with the LLM layer—alongside the MPC and CFG baselines, characterizing each by the properties that distinguish them: fault detection capability, intervention authority, audit record generation, and closed-loop stability guarantee. This comparison makes explicit what each layer of the cognitive architecture contributes and what remains unchanged relative to the baselines.

3. Case Studies

3.1. Benchmark Plant

3.1.1. Open-Loop Dynamics

The benchmark plant is a fourth-order coupled oscillatory system representing two coupled subsystems, each with a position and velocity state. The system is described by the standard linear state-space representation (Equation (3)):

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}(\theta) \mathbf{x}(t) + \mathbf{B}(\alpha) \mathbf{u}(t), \\ \mathbf{y}(t) &= \mathbf{C} \mathbf{x}(t), \\ \mathbf{D} &= \mathbf{0},\end{aligned}\tag{3}$$

where $\mathbf{x}(t) = [q_1, \dot{q}_1, q_2, \dot{q}_2]^T \in \mathbb{R}^4$ represents two coupled state pairs (position and velocity of each subsystem), $\mathbf{u}(t) \in \mathbb{R}$ is the scalar control input, and $\mathbf{y}(t) = q_1 \in \mathbb{R}$ is the measured output of the first subsystem. The system parameters are $m_1 = 1$, $m_2 = 0.8$, $k_1 = 2$, $k_2 = 1.5$, $c_1 = 0.4$, with coupling damping $c_2 = 0.35\theta$ scaled by the fault parameter θ . The full nominal matrices ($\theta = 1$, $\alpha = 1$) are given by Equation (4):

$$\begin{aligned}
 A &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ -3.5 & -0.75 & 1.5 & 0.35 \\ 0 & 0 & 0 & 1 \\ 1.875 & 0.4375 & -1.875 & -0.4375 \end{bmatrix}, \\
 B^T &= [0 \ 1 \ 0 \ 0], \\
 C &= [1 \ 0 \ 0 \ 0], \\
 D &= 0
 \end{aligned} \tag{4}$$

The open-loop eigenvalues are $\lambda_{1,2} = -0.507 \pm 2.072j$ and $\lambda_{3,4} = -0.087 \pm 0.904j$ —two pairs of stable but lightly damped complex poles, characteristic of oscillatory process dynamics. Under full damping degradation ($\theta = 0.15$), the eigenvalues shift to $-0.198 \pm 2.118j$ and $-0.061 \pm 0.908j$, becoming significantly more lightly damped and explaining the increased sensitivity to actuator faults. Under actuator degradation ($\alpha = 0.25$), the B matrix scales to $B^T = [0, 0.25, 0, 0]$, reducing input authority by 75%. This eigenvalue migration toward the imaginary axis under both fault types explains the increased oscillatory sensitivity observed in the benchmark scenarios below, and motivates the use of a pre-computed degraded controller matched to the fully degraded operating point.

This system class directly captures coupled oscillatory dynamics encountered in process engineering: pressure and flow oscillations in pipeline networks subject to surge [38,39], liquid level coupling in interconnected vessel systems, torsional vibration in compressors and pump trains [38], and coupled drive-train dynamics in manufacturing systems. The control objective—output regulation of q_1 under parametric faults and actuator degradation—maps directly to setpoint tracking problems in process control under equipment degradation. The LQI inner-loop controller is designed with cost matrices $Q_x = \text{diag}(40, 4, 10, 2)$, $Q_i = 200$, $R = 0.8$. The degraded controller is designed for the fully degraded plant ($\theta = 0.15$). Detection is gated for $t \leq 8$ s to suppress the initial step-response transient.

The benchmark plant is intentionally simplified relative to full industrial process models. As a fourth-order linear system it captures the essential coupled oscillatory dynamics of compressor and pipeline applications but omits nonlinearities, multi-input multi-output interactions, time delays, and higher-order modal content present in real equipment. The thermal case study similarly uses a one-dimensional finite-difference model of a steel sheet, which captures the essential heat diffusion and boundary condition dynamics but omits three-dimensional effects, material property variation with temperature, and surface geometry. These simplifications are appropriate for a framework paper establishing proof-of-concept: the cognitive supervisory architecture is agnostic to plant order and operates on any state-space model that can produce a one-step residual. The limitations of the benchmark models are acknowledged as a motivation for hardware-in-the-loop validation on higher-fidelity models as future work.

The parameters and the workflow are visualized in Figure 3.

3.1.2. Fault Scenario Rationale

The selection of fault types and magnitudes is grounded in established industrial evidence, as detailed below. The two fault types and their magnitudes are grounded in established industrial evidence. Actuator degradation is represented by scaling the input matrix by $\alpha = 0.25$. This follows the standard multiplicative loss-of-effectiveness model widely used in the fault-tolerant control literature, in which a scalar effectiveness factor multiplying the control input represents partial actuator failure [3]. The chosen severity—75% loss, $\alpha = 0.25$ —falls within the severe-fault benchmark range routinely tested across fault-tolerant control studies, where effectiveness levels between 20% and

60% of nominal are commonly evaluated [3,4]. Damping degradation is represented by scaling the coupling damping coefficient via θ , reducing it to 15% of its nominal value. This parameterization is motivated by the dynamics of compressor and pipeline systems, where lightly damped coupled oscillatory modes are characteristic [38,39]. The dynamic consequence of damping reduction in coupled oscillators—eigenvalue migration toward the imaginary axis and increased sensitivity to disturbances—is consistent with findings in coupled nonlinear oscillator models [40–42]. The abrupt and gradual fault profiles represent two physically distinct degradation mechanisms: sudden failures such as valve seizure or power supply loss, and progressive wear-driven drift. Both are standard fault classes in the fault-tolerant control literature [1,2].

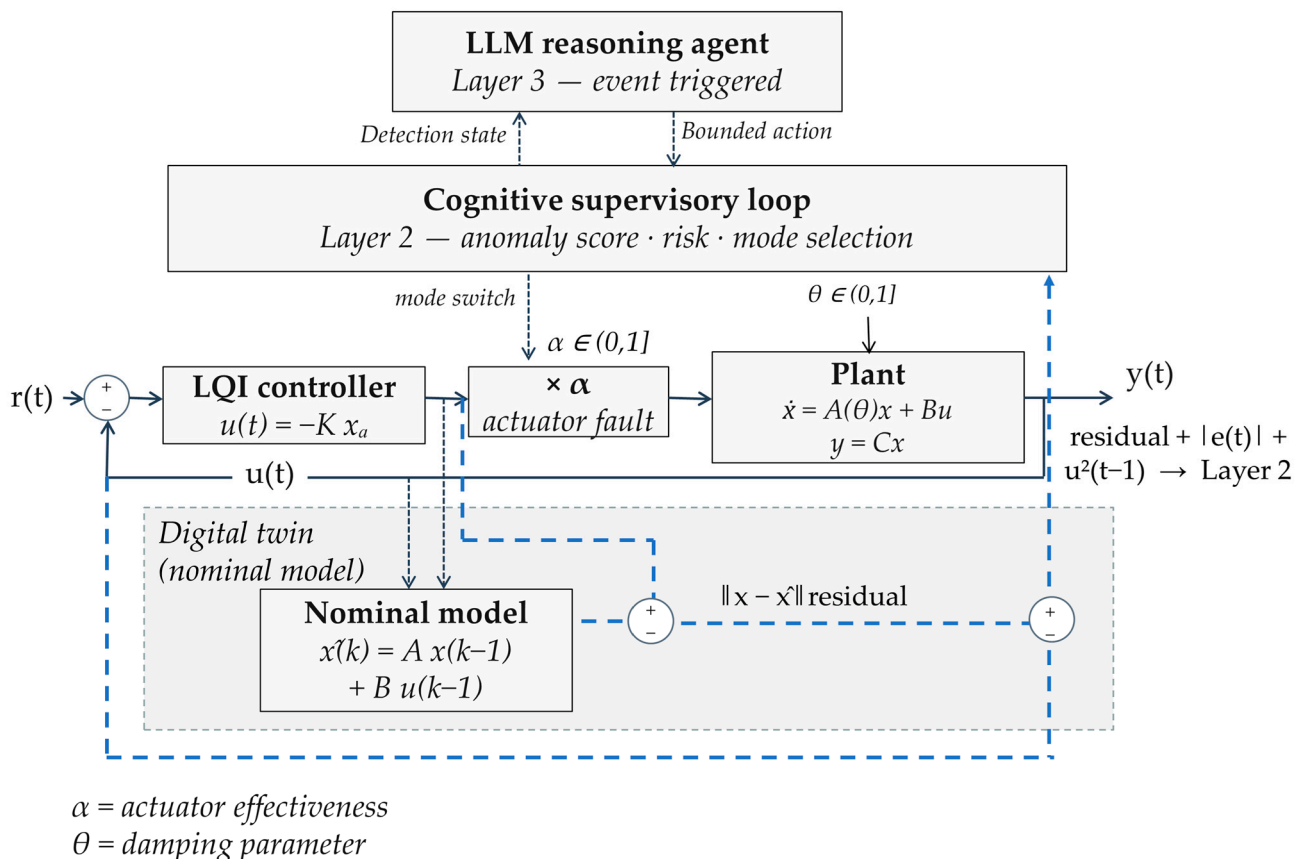


Figure 3. Closed-loop block diagram with fault parameters α and θ . The actuator effectiveness α scales the input channel B; the damping parameter θ enters the state matrix A. The nominal model runs in parallel to produce the residual fault signal fed to Layer 2.

3.1.3. Evaluation Scenarios

Four scenarios are evaluated over $T = 25$ s at $dt = 0.01$ s with a unit step reference $r(t) = 1$ for $t \geq 2$ s:

1. Nominal: $\theta = 1$, $\alpha = 1$ throughout. No fault injected.
2. Gradual damping drift: θ transitions linearly from 1 to 0.15 between $t = 8$ s and $t = 18$ s.
3. Abrupt structural change: θ drops from 1 to 0.15 at $t = 10$ s.
4. Abrupt actuator degradation: α drops from 1 to 0.25 at $t = 10$ s, representing 75% effectiveness loss.

3.1.4. Baselines

Three baselines are compared against the proposed framework. The first is the LQI alone—the inner loop without any supervisory layer—representing the current state of

practice. The second is a score-interpolated gain-scheduling (SIGS) baseline that continuously interpolates between nominal and degraded LQI gains in proportion to the anomaly score; this represents a smoothly adaptive scheme with no discrete switching logic. The third is a conservative fixed-gain (CFG) baseline applying $0.5\times$ nominal LQI gains throughout—representative of deliberately detuned controllers that trade tracking performance for robustness to reduced actuator authority [4]. Unlike MPC, which cannot guarantee optimization feasibility once actuator effectiveness drops below a critical threshold [5,6], CFG provides a deterministic, always-feasible fallback at a permanent tracking cost.

A fourth baseline, MPC (model predictive control), is included to characterize the behavior of an optimization-based controller under actuator degradation. The MPC is implemented as a receding-horizon linear quadratic regulator on the same augmented five-state system used by LQI, with prediction horizon $N = 30$ steps (0.3 s), the same state and input cost matrices as the LQI design, and symmetric input constraints $u \in [-8, 8]$. The QP is solved at each timestep using a dense quadratic program (quadprog). A paired MPC+Cognitive variant applies the same cognitive supervisory layer to the MPC inner loop, switching between a nominal and a degraded MPC design under the same detection and dwell-time logic. MPC is included to demonstrate the feasibility-degradation effect noted in Section 1.1: under full actuator authority, its receding-horizon optimization produces the lowest energy consumption across all fault-free scenarios, but under 75% actuator loss, the QP solution degrades severely, as demonstrated in Section 4.3. The residual-based fault signal used here follows the performance residual formulation of [37]; that work addresses detection only and does not propose a switched control response, so it does not constitute a control baseline.

Adaptive MPC and reconfigurable MPC approaches specifically designed for actuator degradation [5,6] were not included as explicit baselines. Such approaches address actuator faults by updating the MPC prediction model or constraint set online; however, the standard MPC already evaluated here demonstrates that reduced actuator authority shrinks the feasible region and degrades QP solution quality—a fundamental limitation that online model updates partially mitigate but do not eliminate without additional conservatism. Comparison with adaptive MPC variants is identified as future work.

3.2. Plant Variants

To characterize the boundaries of the cognitive supervisory benefit, three alternative plant variants are evaluated under the same abrupt actuator fault ($\alpha = 0.25$ at $t = 10$ s). These variants share the same detection and switching architecture; only the plant parameters differ. MPC is not evaluated on plant variants; its behavior under the abrupt actuator fault is fully characterized in the main benchmark (Section 4.3), where it produces the worst tracking performance of all methods. Results across all plant variants are reported in Table 3 in Section 4.1.5.

Plant A—Unstable open-loop dynamics ($k_1 = -0.5$). A destabilizing spring constant introduces one open-loop eigenvalue with a positive real part. The LQI gain matrix is redesigned with a heavier state penalty to stabilize the closed loop. Under $\alpha = 0.25$, the cognitive framework correctly remains dormant—zero interventions—producing performance identical to standalone LQI. This confirms the non-intrusiveness property on a structurally harder plant: the fault is not severe enough relative to the stabilizing gain margin to require supervisory action.

Plant B—Stiff fast dynamics ($k_1 = 16$, $k_2 = 12$). Spring constants increased eight times raise the natural frequency to approximately 6 rad/s. The stiff dynamics generate residuals during the normal step-response transient that cross the detection threshold before the

fault occurs. This constitutes a false positive. The score weights, calibrated for the nominal 2 rad/s dynamics, do not automatically generalize to high-frequency plants. This is an inherent limitation of fixed-weight composite scores and motivates adaptive threshold calibration as future work.

Plant C—Near-undamped dynamics ($c_1 = 0.02$). Damping reduced by a factor of 20 yields damping ratios of 0.035 and 0.175 for the two mode pairs, representative of lightly damped oscillatory process equipment including pipeline pressure surge and compressor inter-stage dynamics [38,39]. The cognitive framework detects the fault, switches to the degraded controller, and achieves a 20.7% IAE (integral absolute error) reduction. This is the condition under which the cognitive mode switch provides maximum benefit: reduced actuator authority eliminates the controller’s ability to provide active damping, and the degraded controller carries the gain structure to compensate.

3.3. Thermal Process

3.3.1. Plant Description

The third case study applies the cognitive supervisory framework to a spatially distributed thermal process representative of laser welding, induction heating, and surface treatment operations. The plant is a 1 mm steel sheet modeled by the one-dimensional heat equation (Equation (5)) discretized using an explicit finite-difference scheme with $N_x = 41$ nodes through the thickness:

$$\rho C_p \frac{\partial T}{\partial t} = k \frac{\partial^2 T}{\partial x^2}, \quad x \in [0, 1] \quad L = 1 \text{ mm} \quad (5)$$

Boundary conditions are a Neumann boundary at $x = 0$ —a controlled heat flux u in W/m^2 representing the laser or induction heater—and a Robin convection boundary at $x = L$ (Equation (6)) representing heat loss to ambient:

$$-k \left. \frac{\partial T}{\partial x} \right|_{x=L} = h(T(L) - T_{amb}), \quad (6)$$

Nominal steel parameters are thermal diffusivity $\alpha = 1.2 \times 10^{-5} \text{ m}^2/\text{s}$, volumetric heat capacity $\rho C_p = 3.8 \times 10^6 \text{ J}/(\text{m}^3 \cdot \text{K})$, thermal conductivity $k = 45.6 \text{ W}/(\text{m} \cdot \text{K})$, convection coefficient $h = 100 \text{ W}/(\text{m}^2 \cdot \text{K})$, and ambient temperature $T_{amb} = 22 \text{ }^\circ\text{C}$. The measurement point is at mid-thickness $x = 0.5 \text{ mm}$. The control objective is setpoint tracking $T(0.5 \text{ mm}) = 200 \text{ }^\circ\text{C}$ under a unit step reference applied at $t = 1 \text{ s}$, with heat flux constrained to $u \in [0, 5 \times 10^7 \text{ W}/\text{m}^2]$. The inner-loop controller is a PI (proportional–integral) with $K_p = 1 \times 10^5 \text{ W}/(\text{m}^2 \cdot ^\circ\text{C})$ and $K_i = 5 \times 10^3 \text{ W}/(\text{m}^2 \cdot ^\circ\text{C} \cdot \text{s})$, discretized at 1 ms. The nominal finite-difference model—computed online at each control step using the nominal steel parameters—serves as the process digital twin: its residual against the measured temperature profile drives the adaptive anomaly score, providing fault detection specific to thermal property changes rather than generic output deviation.

3.3.2. Fault Scenario Rationale

The three fault types are selected to represent distinct physical degradation mechanisms relevant to laser, induction, and resistance heating operations. Each is grounded in industrial evidence.

Abrupt heater power loss (effectiveness $1 \rightarrow 0.25$ at $t = 5 \text{ s}$) represents a 75% reduction in heat input. In high-power laser systems, thermal lensing and catastrophic optical damage cause sudden and severe power collapse [43,44]. In laser-based manufacturing processes, reduced laser power is a detectable and documented fault mode associated

with optics contamination, misalignment, and thermal degradation [45]. In induction heating, supply voltage faults and module failures produce comparable sudden reductions in delivered power [46,47]. The 75% magnitude mirrors the actuator fault severity used in the mechanical benchmark, enabling direct comparison of cognitive supervisory benefit across plant types.

Gradual surface fouling (convection coefficient h increasing linearly to five times its nominal value between $t = 3$ s and $t = 8$ s) represents progressive thermal resistance build-up at the back face of the workpiece. Oxide scale accumulation on steel surfaces is a well-established degradation mechanism in high-temperature processing: iron oxide scales have thermal conductivities of approximately $1.6\text{--}3\text{ W/(m}\cdot\text{K)}$ [48,49]—more than one order of magnitude below the base steel value of $45.6\text{ W/(m}\cdot\text{K)}$ used here—making even thin scale layers a substantial boundary resistance and increasing the effective convection coefficient seen by the control system. More broadly, fouling consistently degrades heat transfer coefficients and challenges temperature control, with reported efficiency losses up to 30–40% in process heat exchangers [50,51]. Critically, under closed-loop temperature control, fouling can grow while heat duty remains nominally stable, making output-based detection insufficient [52]; residual-based detection of the type used here is therefore specifically motivated for this fault type [52,53].

Phase change and microstructure alteration (thermal diffusivity α dropping 40% at $t = 5$ s) represent a severe microstructural fault scenario. Steel phase transformations such as austenite-to-martensite produce measurable reductions in thermal diffusivity, with reported values ranging from approximately 20% for single transformations [54] to larger reductions under combined phase and microstructural changes [55]. The 40% reduction used here is intended as a stress test of the cognitive detection and classification capability.

3.3.3. Evaluation Scenarios

Four scenarios are evaluated over $T = 15$ s. The same three baselines as the mechanical benchmark are used: LQI (nominal PI only), SIGS (score-interpolated PI gain scheduling), and CFG ($0.5 \times$ nominal PI gains). The cognitive framework uses an adaptive baseline anomaly score—the excess of residual and tracking error above their respective exponential moving averages—to avoid false positives from the initial heating transient. The detection gate opens at $t = 4$ s. A fault-type heuristic distinguishes power-loss from phase-change signatures at detection time: if the model residual excess exceeds the tracking error excess by a factor of three and the absolute error is below $5\text{ }^{\circ}\text{C}$, the fault is classified as phase-like and routed to a dedicated phase-change controller ($K_p = 1 \times 10^5\text{ W/(m}^2\cdot^{\circ}\text{C)}$, $K_i = 2 \times 10^3\text{ W/(m}^2\cdot^{\circ}\text{C}\cdot\text{s)}$). Otherwise the fault is classified as authority-loss and routed to the degraded high-gain controller ($K_p = 4 \times 10^5\text{ W/(m}^2\cdot^{\circ}\text{C)}$, $K_i = 2 \times 10^4\text{ W/(m}^2\cdot^{\circ}\text{C}\cdot\text{s)}$).

4. Results

The results are organized to validate each contribution listed in Section 1.3 in sequence.

4.1. Mechanical Benchmark

4.1.1. Benchmark Performance Table

Table 2 reports complete performance results across all four mechanical benchmark scenarios and all methods. The metrics reported are integral absolute error (IAE), integral squared error (ISE), control energy, output overshoot, settling time, and detection time where applicable. Cognitive refers to the proposed framework. SIGS is the score-interpolated gain-scheduling baseline. CFG is the conservative fixed-gain baseline.

Table 2. Complete benchmark results—all scenarios and methods.

Scenario	Method	IAE	ISE *	Energy	Overshoot	Settle (s)	Detect (s)
Nominal	LQI	1.218	0.744	86.8	0.043	8.89	–
	SIGS	1.227	0.745	86.7	0.043	9.31	–
	CFG	1.284	0.841	86	0.027	8.37	–
	MPC	2.179	0.971	78.8	0.012	15.41	–
	Cognitive	1.218	0.744	86.8	0.043	8.89	–
	MPC+Cogniti	2.176	0.967	79.3	0.024	15.28	8.01
Gradual damping	LQI	1.234	0.744	86.8	0.043	8.92	–
	SIGS	1.242	0.745	86.7	0.043	9.32	–
	CFG	1.314	0.841	86	0.027	8.37	–
	MPC	2.216	0.972	78.8	0.016	20.25	–
	Cognitive	1.234	0.744	86.8	0.043	8.92	–
	MPC+Cogniti	2.222	0.968	79.4	0.025	20.38	8.01
Abrupt damping	LQI	1.254	0.745	86.8	0.043	8.89	–
	SIGS	1.26	0.745	86.7	0.043	9.31	–
	CFG	1.344	0.841	86	0.027	8.37	–
	MPC	2.238	0.973	78.8	0.02	20.41	–
	Cognitive	1.254	0.745	86.8	0.043	8.89	–
	MPC+Cogniti	2.251	0.97	79.4	0.028	20.48	8.01
Abrupt actuator	LQI	2.366	0.883	936.6	0.129	–	–
	SIGS	2.332	0.887	929.9	0.099	–	–
	CFG	3.214	1.22	892.1	0.18	–	–
	MPC	7.089	3.378	437.3	0	–	–
	Cognitive	2.313	0.883	928.6	0.094	–	10.66
	MPC+Cogniti	6.987	3.352	442.9	0	–	8.01

* Integral squared error.

4.1.2. Non-Intrusive Behavior in Three Scenarios

In the nominal, gradual damping, and abrupt damping scenarios, the cognitive framework produces responses numerically identical to a standalone LQI, with zero supervisory interventions. This non-intrusiveness is a direct consequence of the joint AND detection condition and the transient gating period. Evidently, as seen through the IAE (and in Figure 4), the proposed framework matches the LQI exactly in scenarios 1–3 and achieves the lowest IAE in scenario 4. MPC+Cognitive is not non-intrusive in fault-free scenarios: it triggers a supervisory intervention at $t = 8.01$ s in all three undisturbed scenarios, as the longer MPC settling transient (≈ 15 – 20 s) generates residuals that satisfy the detection condition before the plant has reached steady state. This indicates that the detection threshold and gating period require recalibration when the MPC inner loop replaces LQI.

In the gradual drift scenario, the anomaly score rises moderately during the drift but never simultaneously satisfies both the score magnitude and score rate conditions. In the abrupt damping scenario, the step change produces a brief residual spike, but the LQI integrator compensates rapidly enough that the score rate drops before both conditions are simultaneously satisfied.

4.1.3. Abrupt Actuator Degradation

In scenario 4 (Figure 5), the 75% reduction in input authority forces the LQI loop to push control effort toward its saturation limit. The composite score spikes and the score rate rises sharply, triggering the supervisory event at $t = 10.66$ s— 0.66 s after fault onset.

The medium-risk assessment selects switch_degraded. The LLM layer is called at $t = 11.06$ s and, having identified high actuator stress from the near-saturation control effort, overrides the surrogate decision by selecting energy_save_mode.

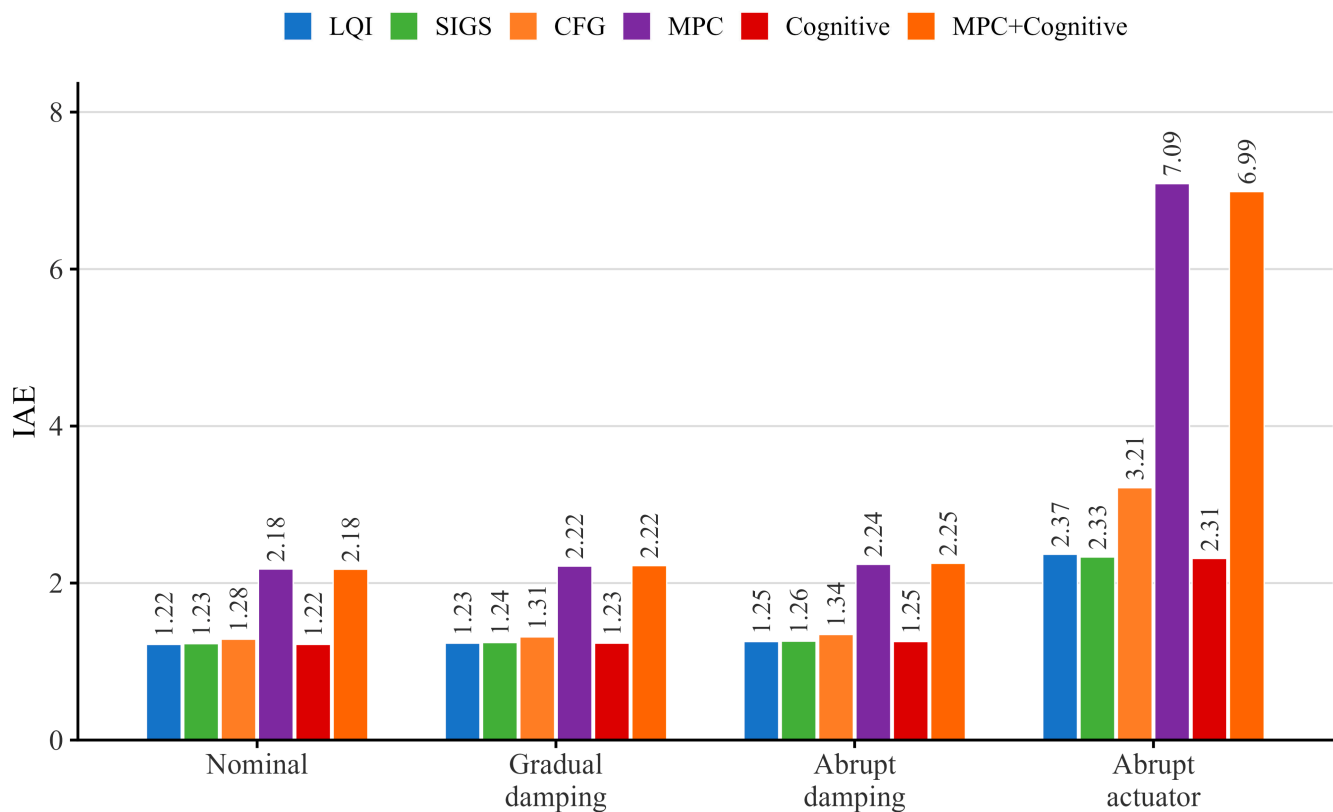


Figure 4. IAE comparison across all scenarios and methods. The cognitive framework (red) is numerically identical to LQI in scenarios 1–3 and achieves the lowest IAE in the actuator fault scenario.

The cognitive controller achieves overshoot of 0.094 and IAE of 2.313—the best among all methods. The CFG baseline achieves lower energy (892.1) but at a substantially degraded tracking cost (IAE = 3.214, overshoot = 0.180) that would be unacceptable in process applications with output constraints. The SIGS baseline achieves competitive IAE (2.332) through continuous gain interpolation but lacks discrete safety logic and an auditable intervention record. MPC performs worst of all methods under this fault, yielding IAE = 7.089—three times that of standalone LQI—with zero overshoot because the QP solver reduces control effort drastically once reduced actuator authority shrinks the feasible region. This confirms empirically the feasibility-degradation effect described in Section 1.1: MPC’s constraint-satisfaction guarantee becomes its liability when actuator effectiveness drops below a critical level.

An additional comparison was conducted against an adaptive fault-tolerant baseline that continuously estimates how much the actuator has degraded and adjusts controller gains accordingly, using a recursive least squares (RLS) estimator [56,57]. The estimator correctly identifies the 75% actuator degradation within approximately 1 s of fault onset. Despite this, the adaptive baseline achieves IAE = 2.364—almost identical to a standalone LQI (2.366) and worse than the cognitive framework (2.313). The reason is that continuously blending between nominal and degraded gains never fully reaches the controller specifically designed for this fault condition, whereas the cognitive framework switches decisively to it. Additionally, the adaptation process itself does not carry formal stability guarantees during the transient, unlike the cognitive framework where every selectable controller mode is pre-verified stable. A variant combining the RLS estimator with an LLM audit call

(AFTC+LLM) was also evaluated: the LLM correctly identifies the degradation event and confirms the adaptive response, but produces no IAE change ($IAE = 2.364$), consistent with the finding that the LLM contribution is governance rather than performance regardless of the underlying control architecture.

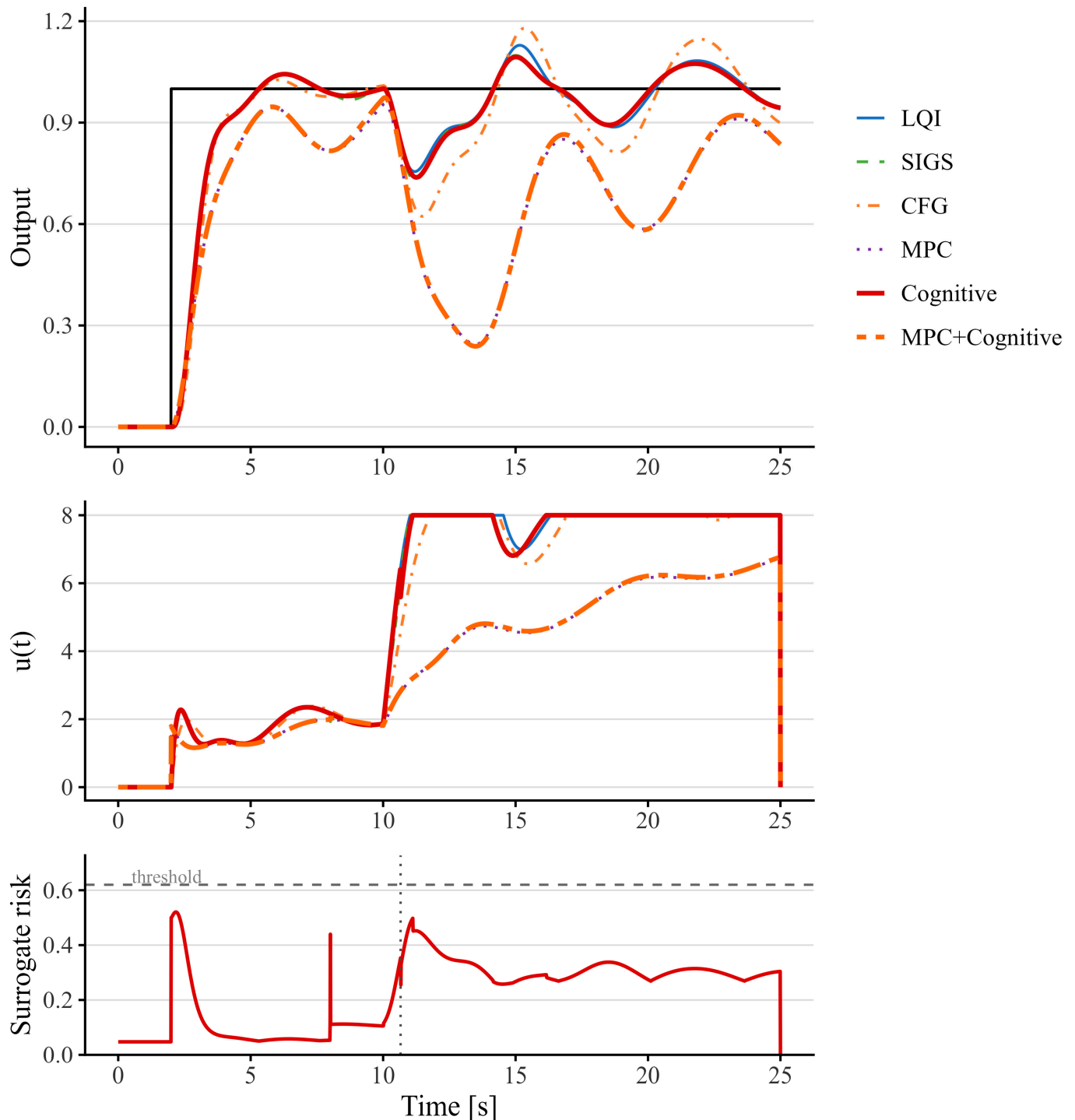


Figure 5. Time-domain response under abrupt actuator degradation (fault at $t = 10$ s, $\alpha = 0.25$). **(Top):** system output for all methods; the cognitive framework (red) shows reduced overshoot relative to LQI following supervisory intervention. **(Middle):** control effort; the cognitive framework activates energy-save mode, capping actuator output below the LQI saturation level. **(Bottom):** surrogate risk signal; the dashed horizontal line marks the detection threshold ($\kappa \cdot \text{score}_{\text{nom}}$) and the dotted vertical line marks the detection instant at $t = 10.66$ s.

4.1.4. Gradual Drift Response

Figure 6 shows the gradual damping drift scenario. All methods track the reference identically. The risk signal rises moderately during the drift but remains consistently below the detection threshold, confirming correct supervisory restraint under gradual, manageable degradation.

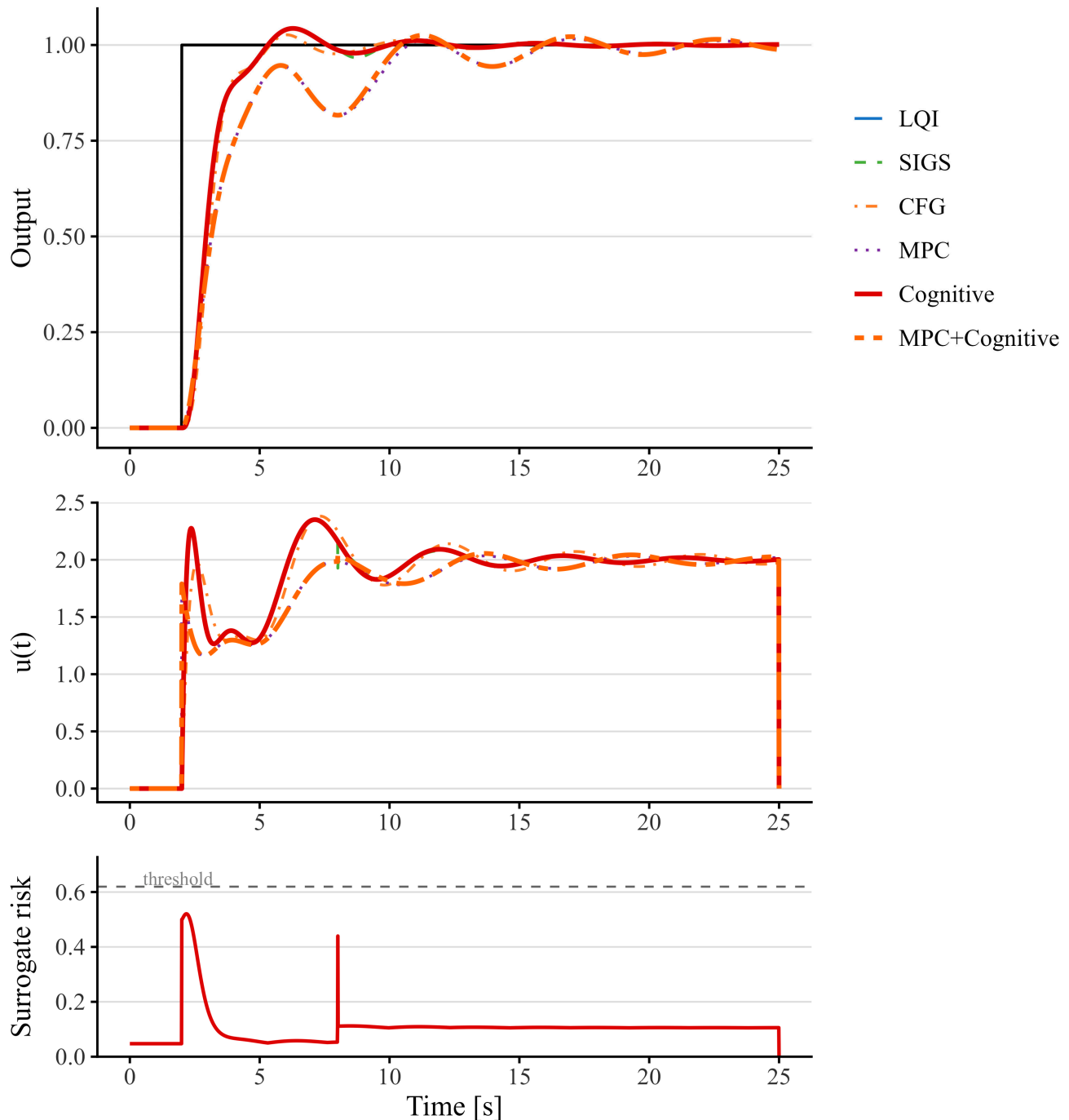


Figure 6. Gradual damping drift scenario. All methods produce identical output responses. The RQ2 surrogate risk (bottom) stays below threshold—no intervention is triggered.

4.1.5. Plant Variant Results

Table 3 summarizes results across all plant variants. On the unstable Plant A, the cognitive framework remains dormant and produces IAE = 1.492, identical to a standalone LQI, confirming non-intrusiveness on a structurally harder plant. On the stiff Plant B, a

false positive triggers at $t = 8.01$ s before the fault occurs at $t = 10$ s—an inherent limitation of fixed-weight scores on high-frequency plants. On the near-undamped Plant C, the framework detects the fault at $t = 10.60$ s and achieves $IAE = 2.363$ versus LQI's 2.980, a 20.7% reduction.

Table 3. Plant variant results—abrupt actuator fault ($\alpha = 0.25$, $t = 10$ s).

Plant	Description	LQI IAE	Cognitive IAE	Improvement	Detection
Nominal	$c_1 = 0.4$, $k_1 = 2$	2.366	2.313	2.3%	$t = 10.66$ s
Plant A	Unstable OL, $k_1 = -0.5$	1.492	1.492	0% (dormant)	–
Plant B	Stiff, $k_1 = 16$	21.752	21.752	0% (false trigger)	$t = 8.01$ s
Plant C	Near-undamped, $c_1 = 0.02$	2.980	2.363	20.7%	$t = 10.60$ s

The results across plant variants show that cognitive supervisory benefit is a function of open-loop damping. Results suggest plants with damping ratios below approximately 0.05 may benefit most. Practitioners can estimate potential benefit from open-loop frequency response measurements prior to deployment.

4.1.6. Parametric Sensitivity—Fault Severity, Detection Threshold, and Onset Timing

Figures 7–10 summarize the parametric sensitivity analysis across three sweep dimensions. Figure 7 (fault severity, $\alpha \in [0.05, 1.0]$) confirms that the cognitive benefit is concentrated at $\alpha \leq 0.25$, with a peak ΔIAE of 0.053 under the paper's benchmark scenario. Figures 8 and 9 show that varying the detection threshold multiplier ($\kappa \in [2, 20]$) identifies a robust operating window of $\kappa \in [6, 20]$, providing reliable fault detection without false positives. The selected value, $\kappa = 8$, lies conservatively within this range. Figure 10 (fault onset, $t_{fault} \in [4, 22]$ s) shows that the cognitive framework matches or outperforms standalone LQI in seven of the 10 tested fault-onset timings. The three exceptions correspond to faults at $t = 8$ s, located at the detection-gate boundary, and at $t \geq 20$ s, where the remaining simulation horizon is insufficient for the benefits of controller switching to accumulate.

Three additional sweeps characterize the sensitivity of the remaining design parameters. Sweep D (measurement noise $\sigma \in \{0, 0.005, 0.01, 0.02, 0.05, 0.10\}$ on $y(t)$) was extended to a Monte Carlo analysis of $N = 50$ independent noise realizations per level. Zero false positives were observed across all 300 nominal-scenario runs (50 seeds $\times$ six noise levels). The cognitive advantage ΔIAE remains positive and statistically significant at all noise levels up to $\sigma = 0.05$, with 95% confidence intervals strictly above zero; at $\sigma = 0.10$, $\Delta IAE = 0.003 [-0.013, +0.018]$, statistically indistinguishable from zero but not reversed. Detection time remains stable between 10.60 s and 10.79 s across all noise levels (mean $\pm$ 95% CI), confirming that noise affects detection timing by less than 0.2 s. Sweep E (five score weight configurations) confirms that the paper's weight design is the only configuration that detects the fault without producing false positives in the nominal run; residual-only and error-only weights both fail to detect the fault entirely, while equal and effort-heavy weights detect early but trigger false positives in the nominal scenario. This validates the joint use of all three features and the residual-dominant weighting. Sweep F (dwell-time $\tau_{dwell} \in \{0.2, \dots, 6.0\}$ s) produces identical IAE across all tested values, confirming that the 2.0 s design choice is a conservative default rather than a tuned parameter—the single-intervention structure of the fault scenario is insensitive to dwell duration.

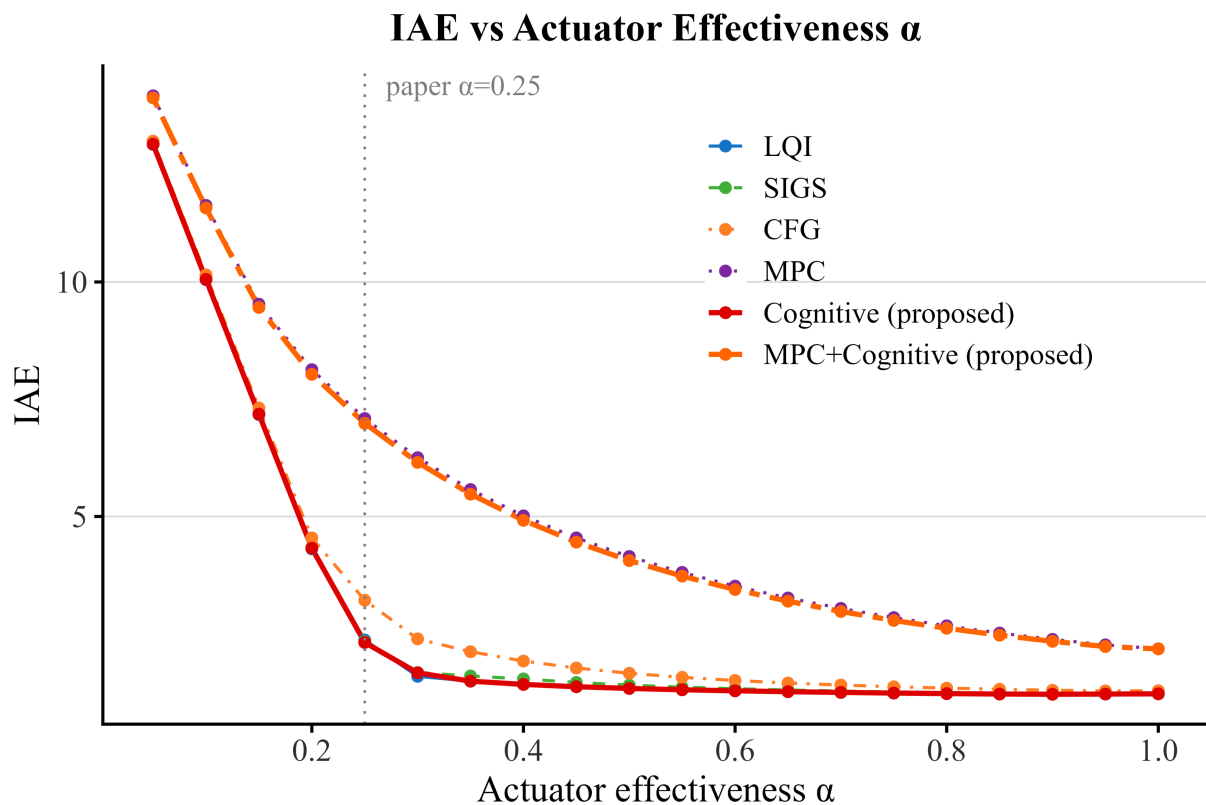


Figure 7. Parametric sensitivity of closed-loop performance to actuator effectiveness. IAE versus actuator effectiveness α for all benchmark controllers. The cognitive supervisory framework provides its greatest benefit under severe actuator degradation ($\alpha \leq 0.25$), while converging to nominal LQI performance as actuator effectiveness approaches unity.

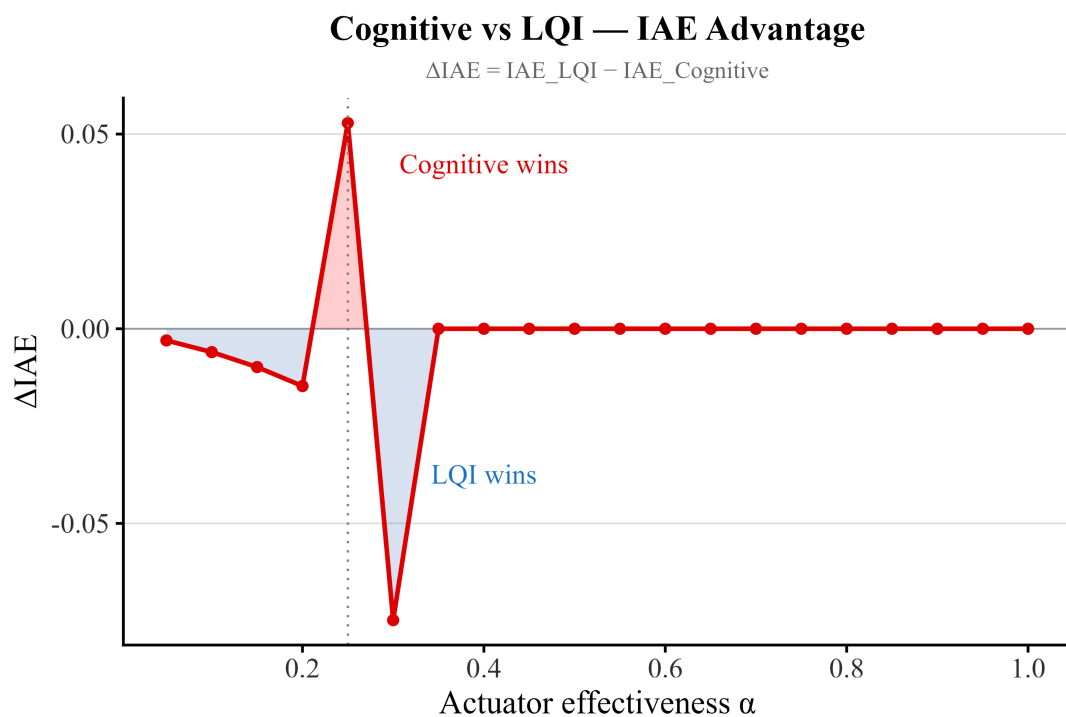


Figure 8. Performance advantage of the cognitive supervisory framework over a standalone LQI. $\Delta IAE = IAE_{LQI} - IAE_{Cognitive}$ as a function of actuator effectiveness α . Positive values indicate improved tracking performance of the proposed cognitive supervisory architecture relative to a standalone LQI.

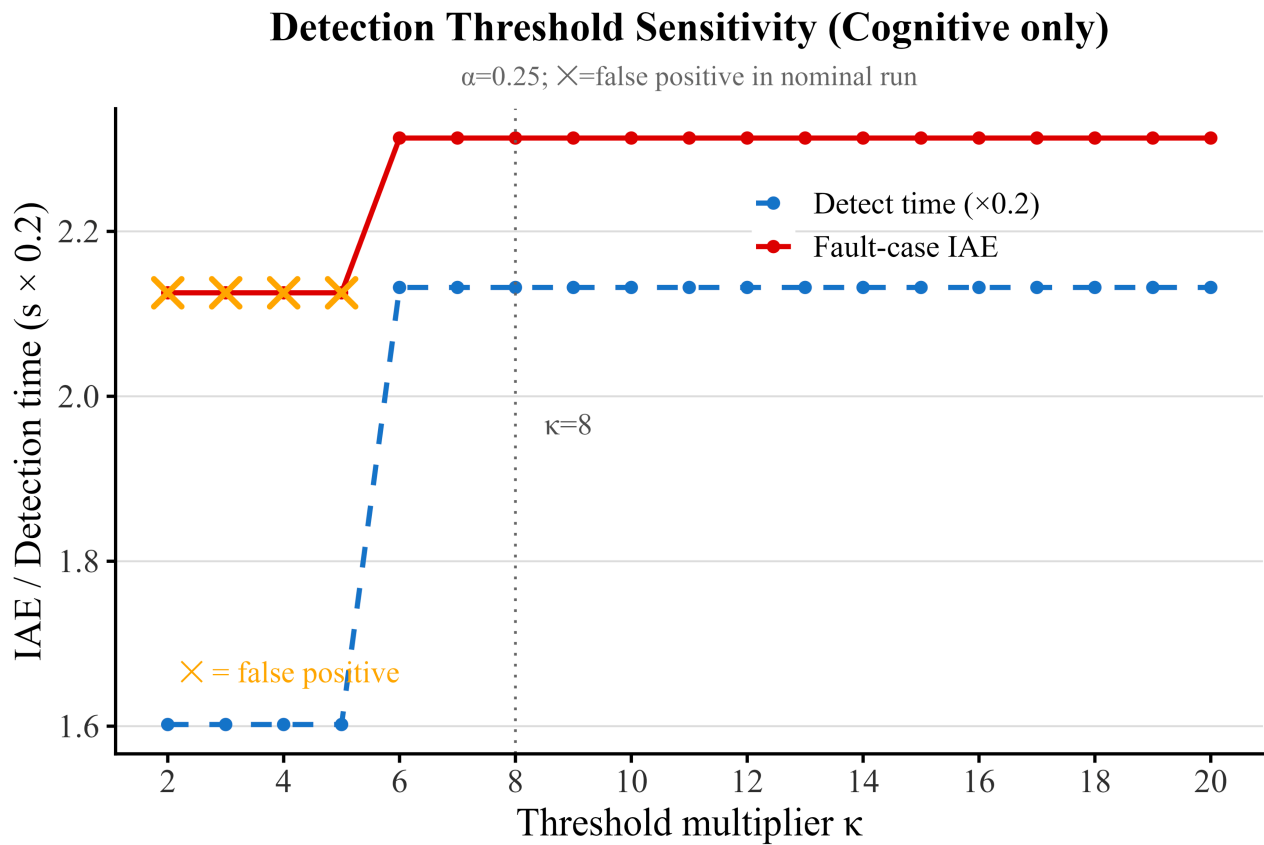


Figure 9. Detection threshold sensitivity analysis. Influence of the detection threshold multiplier κ on fault-case IAE and detection time. The operating range $\kappa \in [6, 20]$ provides stable detection without false positives; $\times$ denotes false-positive detections during the nominal (fault-free) simulation.

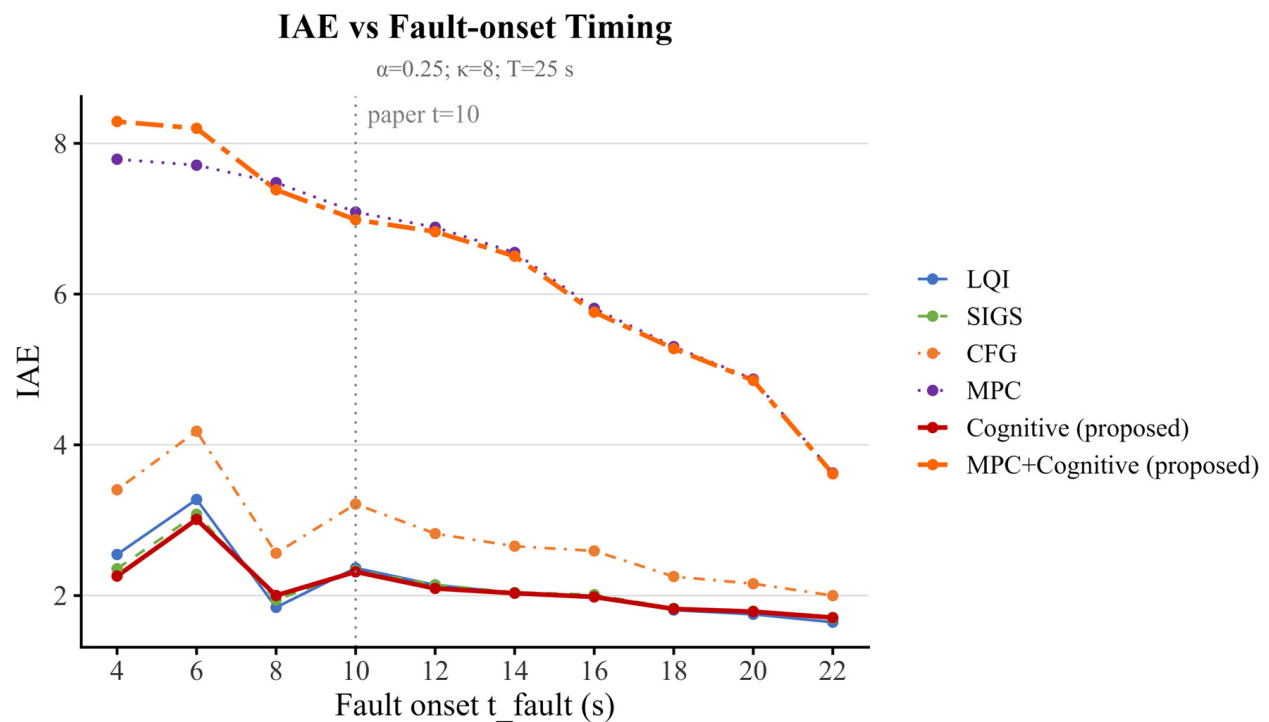


Figure 10. Sensitivity to fault-onset timing. IAE as a function of actuator fault occurrence time. The proposed cognitive supervisory framework matches or improves upon a standalone LQI for most tested fault-onset times while maintaining consistent performance across the operating range.

4.2. Thermal Process

The following results correspond to the spatially distributed thermal plant described in Section 3.3, evaluated across four fault scenarios over $T = 15$ s. Table 4 and Figure 11 summarize IAE results across all four thermal scenarios and all methods.

Table 4. Thermal process benchmark results—IAE ($^{\circ}\text{C}\cdot\text{s}$).

Scenario	LQI	SIGS	CFG	Cognitive	Improvement vs. LQI
Nominal	16.72	16.72	33.01	16.72	0.0%
Power loss	20.70	20.70	40.60	16.89	+18.4%
Gradual fouling	22.21	22.21	43.78	22.21	0.0%
Phase change	16.72	16.73	33.01	18.72	−11.9%

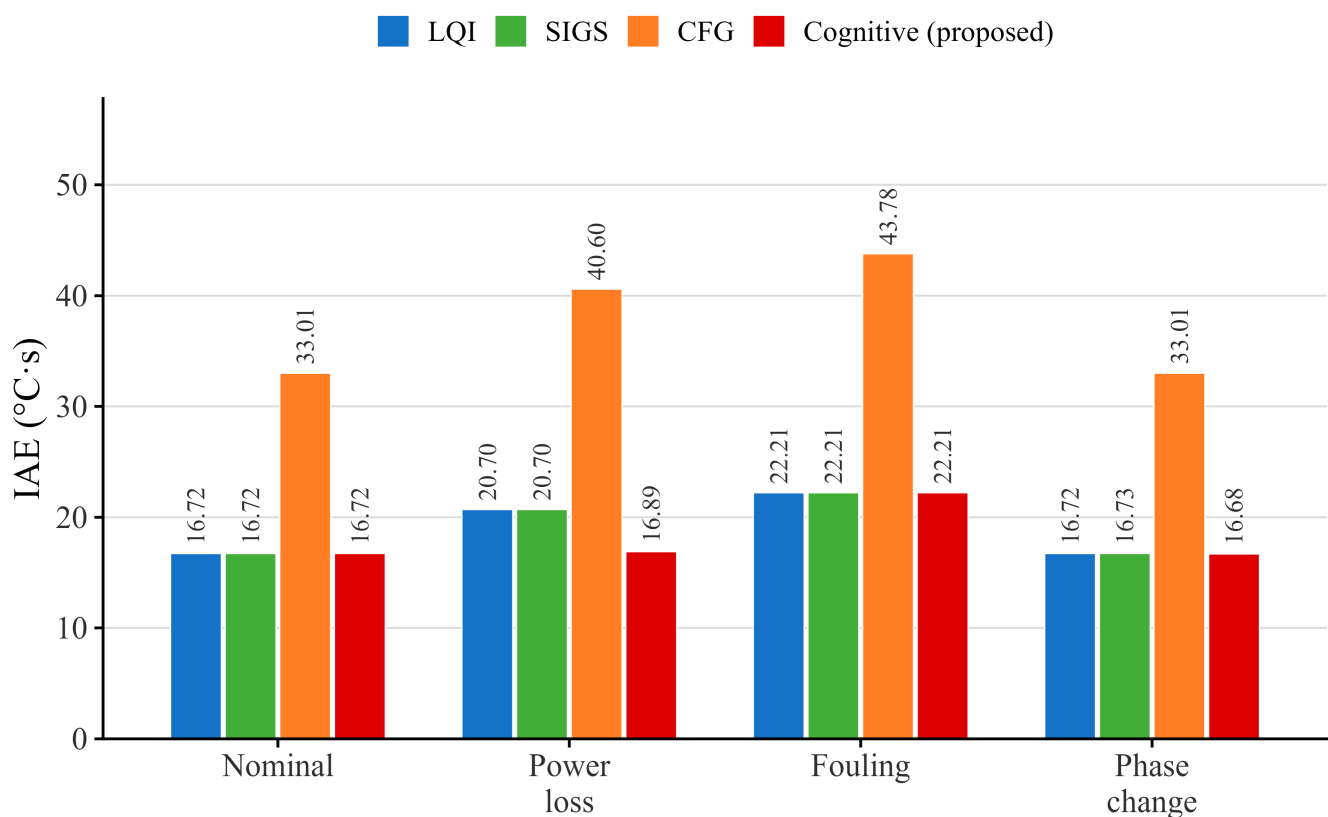


Figure 11. IAE comparison across all four thermal scenarios and methods. The cognitive framework (red) matches standalone PI in nominal and fouling conditions (non-intrusive) and achieves the lowest IAE under abrupt power loss (+18.4% improvement). CFG is consistently worst due to permanently detuned gains.

Nominal. The cognitive framework produces an IAE identical to a standalone PI with zero supervisory interventions, confirming non-intrusiveness on a distributed thermal plant. The risk signal rises during the initial step-response transient but remains below the adaptive detection threshold throughout.

Abrupt power loss. The cognitive framework detects the fault at $t = 5.00$ s—zero detection lag—switches to the high-gain degraded PI, and achieves an IAE = 16.89 versus the LQI's 20.70, an 18.4% improvement. Detection is instantaneous because the 75% power reduction creates an immediate model residual spike that exceeds the adaptive baseline at the first eligible control step after $t = 4$ s. The LLM is called at $t = 5.40$ s and confirms

switch_degraded with medium confidence, reasoning that the power loss fault is confirmed and high-gain authority compensation is appropriate. Figure 12 visualizes the response.

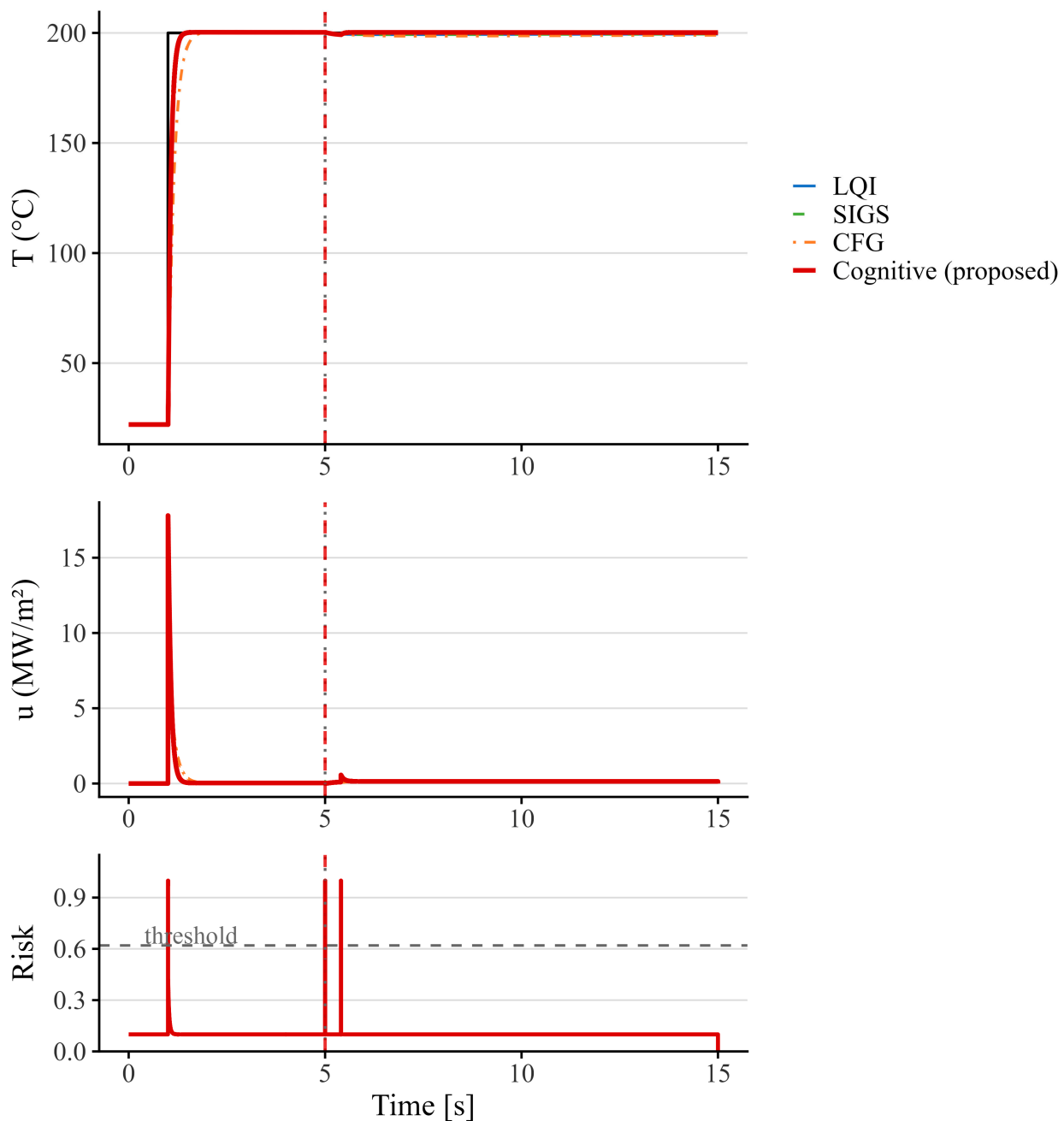


Figure 12. Time-domain response under abrupt heater power loss (fault at $t = 5$ s, effectiveness $1 \rightarrow 0.25$, 75% power reduction). **(Top):** temperature at measurement point; cognitive (red) detects at $t = 5.00$ s and recovers to setpoint faster than LQI. **(Middle):** heat flux u (MW/m^2); the cognitive controller switches to high-gain degraded PI at fault onset, providing additional authority. **(Bottom):** adaptive surrogate risk signal; the risk spike at $t = 5$ s crosses the detection threshold instantaneously, triggering the supervisory event with zero lag. Claude Sonnet is called at $t = 5.40$ s and confirms switch_degraded with medium confidence.

Gradual fouling. The risk signal crosses the detection threshold at the onset of fouling, correctly identifying the anomaly. However, IAE improvement is zero: fouling increases heat loss through the Robin boundary condition, not through the heater input channel, so neither the authority-compensating nor the phase-change degraded controller addresses the fault mechanism. The framework detects correctly and switches, but the pre-designed

controller bank does not contain a controller matched to increased boundary convection. This is a controller bank coverage gap, not a detection failure.

Phase change. Detection fires at $t = 5.00$ s. The fault-type heuristic correctly classifies the fault as phase-like (residual-dominant) and routes to the phase-change controller. However, the IAE increases to 18.72 versus the LQI's 16.72. When thermal diffusivity drops 40%, heat diffuses more slowly; the reduced integrator gain of the phase-change controller prevents overshoot but accumulates additional re-regulation error.

4.3. LLM-Assisted Supervision

With the LLM layer enabled, one call is made per fault event, 0.4 s after initial detection. On the nominal benchmark plant, the call fires at $t = 11.06$ s and selects `energy_save_mode` with medium confidence, reasoning that near-saturation actuator effort justifies modulation rather than unrestricted degraded gain application. The effect on IAE is marginal (2.312 versus 2.313). On Plant C, the call fires at $t = 11.00$ s and selects `switch_degraded` with high confidence; on the thermal power-loss scenario, the call fires at $t = 5.40$ s and selects `switch_degraded` with medium confidence, consistent with the authority-compensating degraded controller activated at fault onset. Table 5 records the LLM calls across all case studies.

Table 5. LLM audit log—supervisory events with LLM layer enabled.

Plant	t (s)	Action	Confidence	Reason (Truncated)
Nominal	11.06	<code>energy_save_mode</code>	medium	Near-saturation prior effort with confirmed degradation...
Plant C	11.00	<code>switch_degraded</code>	high	Near-undamped plant with reduced actuator authority...
Thermal	5.40	<code>switch_degraded</code>	medium	Power loss fault confirmed; high-gain authority compensation appropriate...

The primary contribution of the LLM layer is not the marginal energy saving but the audit record itself—a natural-language rationale generated as part of the decision process, causally connected to the action taken, and readable without control engineering expertise [58]. This addresses the output-level explanation category [32] and the operator-facing interpretability gap identified in process monitoring [32,33]. Unlike the conversational query responses produced in batch process monitoring architectures [20], these audit records are generated at the moment of intervention, not retrieved after the fact in response to an operator query.

The consistency of LLM outputs is ensured by three architectural constraints. The action set is bounded to four options—any response outside this set is discarded and the surrogate decision is retained. The system prompt specifies structured JSON output with explicit decision rules, reducing sensitivity to prompt phrasing variations. Temperature is set to zero, making outputs deterministic for a given input state. Across the three supervisory events reported in Table 5, the LLM selected actions consistent with the surrogate risk assessment in all cases, with the single override (`energy_save_mode` over `switch_degraded`) reflecting additional reasoning about near-saturation actuator effort that the surrogate logic does not capture. Sensitivity to model version is mitigated by the bounded action set and deterministic fallback: a future model version producing an out-of-set response automatically reverts to the surrogate decision without affecting closed-loop stability.

Non-causal explanations are structurally prevented because the LLM receives only current and recent-averaged detection signals—it has no access to future states and cannot

reason beyond the bounded detection state summary; any rationale it produces is therefore causally grounded in the signals present at the moment of the call. When evidence is genuinely conflicting or uncertain, the *request_human_review* action escalates the decision to an operator rather than forcing a potentially unreliable classification. The broader question of LLM reliability, hallucination, and consistency across model versions is an active research area; the work herein does not address this at the model level but instead contains its impact at the architecture level through bounded actions, deterministic fallback, and human escalation.

The *request_human_review* action is triggered when the LLM encounters conflicting or uncertain evidence—specifically when risk falls in the intermediate range (0.30–0.55) and the residual signal dominates without a corresponding tracking error rise, indicating a parametric change whose severity cannot be confidently classified. In all benchmark scenarios, detection signals are either jointly below threshold or jointly unambiguous, placing every event outside this uncertainty zone, which explains the absence of human review calls in the reported results. During a human escalation interval, the surrogate decision remains active and the dwell-time counter continues, preventing mode reversion until nominal conditions are restored or an operator issues an explicit override. The assumed organizational context is a supervised process environment where an operator is reachable within the dwell-time window; for fully unattended operation, the surrogate decision is retained until autonomous recovery.

5. Discussion

The cognitive supervisory framework affects system performance in three measurable ways: it eliminates performance degradation during healthy operation; it improves tracking under severe faults when the fault mechanism matches the available controller bank; and it produces a causally connected audit record at each intervention that no baseline method generates.

The practical significance of the performance gains should be evaluated in context. The 2.3% IAE improvement on the nominal benchmark is modest in absolute terms, but it is achieved with zero overhead during healthy operation. The larger improvements on Plant C and the thermal power-loss scenario are practically significant: in process applications where output constraints are tight, reducing overshoot and recovery time can determine whether a batch remains within specification. More importantly, the added complexity of the cognitive layer is not justified by IAE improvement alone, but by the combination of three properties no baseline provides simultaneously: guaranteed non-intrusiveness, targeted fault response, and automatic audit generation.

The results across both case studies converge on a single architectural principle: separating detection from control. Because the supervisory layer monitors residuals continuously but intervenes only when jointly necessary, it remains inactive during healthy operation and responds decisively when conditions warrant. This separation—non-intrusiveness by design, not by accident—distinguishes the proposed framework from adaptive schemes that continuously modify gains regardless of fault status.

The detection architecture is robust because it does not rely on a single indicator. The AND condition requires simultaneous evidence of fault presence and fault escalation, while the composite score combines tracking error, model residual, and control effort. This explains the absence of false positives across the Monte Carlo noise analysis and the stable detection timing across most fault-onset cases. At the same time, the Plant B result shows the boundary of this robustness: high-frequency plants can generate transient residuals large enough to approach the detection threshold, requiring recalibration of score weights or adaptive thresholding.

The effectiveness of the proposed framework ultimately depends on the availability of controllers matched to the underlying fault mechanism. The largest performance improvements were obtained for abrupt actuator or heater power-loss faults, where dedicated degraded controllers directly compensated for the loss of input authority. By contrast, gradual fouling produced essentially no improvement because the fault altered the boundary heat-transfer coefficient rather than the heater input channel addressed by the available controller bank. Likewise, the phase-change scenario produced only marginal benefit because the controller designed to reduce overshoot under slower thermal dynamics accumulated additional regulation error after the transient. These observations demonstrate that supervisory intelligence cannot compensate for an incomplete controller bank; rather, the digital twin identifies when the most appropriate pre-designed controller should be activated.

Furthermore, the LLM layer's contribution is qualitative rather than performative. Its value lies in producing an operator-readable record that is contemporaneous with the supervisory decision and causally connected to the selected controller mode. This differs from post hoc attribution methods such as SHAP and LIME, which explain a decision only after it has already been made. It also differs from historian or alarm systems, which record what happened but do not necessarily explain why a formally constrained controller-selection decision was taken. The LLM-assisted audit layer therefore supports traceability and regulatory interpretability without affecting closed-loop stability or real-time controller execution.

Compared with recent LLM-enabled supervisory architectures, the proposed framework intentionally confines the LLM to an advisory and audit role rather than allowing direct interaction with the control loop. This design choice reduces autonomous decision-making flexibility but preserves deterministic controller execution and enables formal stability guarantees through pre-verified controller modes and dwell-time-constrained switching. Likewise, unlike adaptive fault-tolerant controllers that continuously update controller parameters online, the proposed switched-controller architecture relies on a finite bank of offline-designed controllers. Although this limits adaptation to previously designed operating conditions, it simplifies verification, improves predictability, and facilitates certification for safety-critical industrial environments. These trade-offs suggest that the proposed architecture is particularly suited to applications where reliability, explainability, and regulatory compliance are prioritized over unrestricted online adaptation.

Limitations, Deployment Considerations, and Future Work

The proposed framework is a proof-of-concept cognitive supervisory architecture, and several limitations should therefore be recognized before industrial deployment can be claimed.

First, the current implementation has been evaluated using two simulation-based case studies representing distinct classes of process dynamics. Although these benchmarks demonstrate the proposed architectural properties—non-intrusiveness, residual-driven fault detection, switched fault-tolerant control, and automatic audit generation—validation on higher-fidelity digital twins, hardware-in-the-loop platforms, and industrial process data remains necessary before practical deployment can be claimed, consistent with recent reviews identifying model fidelity and deployment validation as continuing challenges for AI-enabled digital twins [53].

Second, scalability has not yet been evaluated. The present implementation considers a small controller bank, a single supervisory layer, and a single-fault assumption. Simultaneous actuator, sensor, and process faults would require richer diagnostic logic to distinguish overlapping fault signatures, while the number of required degraded controllers increases with fault type and severity. Future work should therefore investigate hierarchical supervi-

sory architectures and systematic controller-bank synthesis methods, including automated redesign and data-driven approaches, while preserving the formal dwell-time stability guarantees established here.

Third, the computational overhead of the supervisory layer is negligible for the evaluated case studies. Layer 2 processing requires approximately $0.01\ \mu\text{s}$ per control step, while digital twin synchronization contributes approximately $0.002\ \mu\text{s}$ per cycle. Scalability of this computational margin to substantially larger controller banks has not yet been evaluated. The LLM introduces separate deployment considerations, including inference latency, external service availability, model updates, and computational cost. In the proposed architecture, however, the LLM remains outside the real-time control loop: controller switching is determined entirely by deterministic supervisory logic, whereas the LLM is invoked only after fault detection to generate operator-oriented audit records. Consequently, temporary LLM unavailability cannot compromise closed-loop stability or fault accommodation. The observed $0.4\ \text{s}$ audit-generation delay is acceptable for the relatively slow process dynamics considered here but would require optimization for high-bandwidth mechanical systems.

Cybersecurity also represents an important consideration for LLM-assisted supervisory systems. Potential threats include adversarial prompt injection, unauthorized access to audit records, compromised communication channels, and dependence on external inference services. These risks motivate authenticated communication, network isolation, local model deployment where appropriate, and deterministic fallback strategies that preserve safe controller operation independently of LLM availability, consistent with recent recommendations for trustworthy AI-enabled digital twins [54].

Finally, the proposed framework assumes that suitable degraded controllers have been designed and verified offline for the anticipated fault scenarios. As demonstrated by the plant-variant and thermal case studies discussed above, supervisory intervention provides limited benefit when the controller bank does not contain a controller appropriately matched to the active fault mechanism. Likewise, the Plant B results indicate that the composite score requires recalibration when process dynamics change sufficiently for residual behavior to approach the detection threshold. These limitations define the next stage of development from proof-of-concept toward industrial deployment, including adaptive controller-bank synthesis, online digital-twin adaptation, multi-fault supervision, and experimental validation on realistic industrial systems.

6. Conclusions

This paper presented a cognitive supervisory control architecture that combines a lightweight digital twin, residual-based fault detection, switched fault-tolerant control, and an optional LLM-assisted audit layer within a strictly layered architecture. The proposed framework addresses four complementary objectives: preserving nominal closed-loop behavior during healthy operation, detecting faults before substantial tracking degradation develops, selecting among pre-verified stable controllers under formal dwell-time constraints, and automatically generating operator-oriented natural-language audit records at each supervisory intervention.

Evaluation across two representative process classes—a coupled oscillatory system representative of compressor and pipeline dynamics and a distributed thermal process representative of laser welding and induction heating—demonstrated that the proposed architecture remains functionally non-intrusive during healthy operation while improving fault response when the controller bank contains a controller matched to the active fault mechanism. The largest improvements were obtained for abrupt actuator and heater power-loss faults, whereas little or no benefit was observed for faults whose dynamics were not represented within the available controller bank. These results indicate that the primary

engineering constraint is controller-bank coverage rather than fault detection capability, suggesting that practitioners should prioritize characterizing anticipated fault modes and verifying controller-bank coverage before deployment, rather than assuming detection alone guarantees improved fault response.

The results indicate that the LLM contributes primarily to governance rather than control performance. By operating outside the real-time control loop, the LLM generates contemporaneous, traceable audit records without influencing deterministic controller selection or compromising formal closed-loop stability, avoiding the need for post hoc explainability methods. This separation allows explainability to be integrated into supervisory fault-tolerant control while preserving the formal stability guarantees provided by the switched control architecture.

The present work should be regarded as a proof-of-concept supervisory architecture rather than a deployment-ready industrial solution; the specific limitations and deployment considerations bounding this claim are discussed in the Limitations, Deployment Considerations, and Future Work section. Future work will focus on adaptive controller-bank synthesis, scalable supervisory architectures, and experimental validation on realistic industrial process systems.

Author Contributions: Conceptualization, A.P.; investigation, A.P. and O.M.K.; resources, P.S.; writing—original draft preparation, A.P.; writing—review and editing, P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been supported by EC project SkillAbility, under contract number 101177783.

Data Availability Statement: Data is contained within the article.

Acknowledgments: During the preparation of this manuscript/study, the authors used Claude Sonnet (claude-sonnet-4-6, Anthropic) for the purposes of language editing and polishing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript: CARE Continuous Algebraic Riccati Equation CFG Conservative Fixed-Gain DT Digital Twin FSM Finite State Machine FTC Fault-Tolerant Control IAE Integral Absolute Error ISE Integral Squared Error LLM Large Language Model LQI Linear Quadratic Integral LQR Linear Quadratic Regulator MPC Model Predictive Control PI Proportional–Integral SIGS Score-Interpolated Gain-Scheduling XAI Explainable Artificial Intelligence.

Appendix A

The following system prompt is sent to the LLM at each supervisory event call:

Query A1. “You are a supervisory control reasoning module. Output ONLY valid JSON. Allowed actions: keep_nominal, switch_degraded, energy_save_mode, request_human_review. Decision rules: (1) weak evidence → keep_nominal; (2) degradation likely → switch_degraded; (3) high effort and likely degradation → energy_save_mode; (4) conflicting or uncertain evidence → request_human_review. Do not output a direct control signal. Return exactly this schema: {“action”: “...”, “confidence”: “low | medium | high”, “reason”: “short explanation”}”

Appendix B

Algorithm A1: Cognitive Supervisory Layer (Layer 2)—per-timestep execution

Input: $x(k)$, $u(k-1)$, $r(k)$, t

Output: $u(k)$, $\text{mode}(k)$

1. Compute residual: $\text{res} \leftarrow \|x(k) - [A \cdot x(k-1) + B \cdot u(k-1)]\|_2$
 2. Compute features: $e \leftarrow r(k) - C \cdot x(k)$

$$\text{score}(k) \leftarrow w_1 |e| + w_2 \cdot \text{res} + w_3 \cdot u^2(k-1) \text{ [if } t > t_{\text{gate}}, \text{ else } 0]$$

$$\text{score_rate}(k) \leftarrow \text{score}(k) - \text{score}(k-1)$$
 3. Compute risk: $\text{risk}(k) \leftarrow \sigma(-2.2 + 3.0 |e| + 4.0 \cdot \text{res} + 0.015 \cdot u^2 + 0.12 \cdot \text{score_rate})$
 4. Detect event: if $\text{score}(k) > \kappa \cdot \text{score_nom}$

$$\text{AND } (\text{score_rate}(k) > \tau_{\text{rate}} \text{ OR } \text{risk}(k) > \rho_{\text{high}})$$

$$\text{AND } \text{dwell_counter} = 0:$$

$$\text{trigger} \leftarrow \text{TRUE}$$
 5. Select mode: if trigger:

$$\text{if } \text{risk} < 0.35: \text{mode} \leftarrow \text{nominal}$$

$$\text{elif } u_{\text{prev}} > u_{\text{sat}}: \text{mode} \leftarrow \text{energy_save}$$

$$\text{else: mode} \leftarrow \text{degraded}$$

$$\text{dwell_counter} \leftarrow \tau_{\text{dwell}} / dt$$
 else:

$$\text{mode} \leftarrow \text{previous mode}$$

$$\text{dwell_counter} \leftarrow \max(0, \text{dwell_counter} - 1)$$
 6. [Optional] LLM: if trigger AND $(t - t_{\text{detect}}) \geq 0.4$:

$$\text{mode} \leftarrow \text{LLM_call}(\text{detection_state}) \text{ [bounded, with fallback]}$$
 7. Apply controller: $u(k) \leftarrow -K_{\text{mode}} \cdot [x(k); \xi(k)]$
 8. Update integrator: $\xi(k+1) \leftarrow \xi(k) + dt \cdot e$
-

References

1. Abbaspour, A.; Mokhtari, S.; Sargolzaei, A.; Yen, K. A Survey on Active Fault-Tolerant Control Systems. *Electronics* **2020**, *9*, 1513. [\[CrossRef\]](#)
2. Amin, A.A.; Hasan, K.M. A review of fault tolerant control systems: Advancements and applications. *Measurement* **2019**, *143*, 58–68. [\[CrossRef\]](#)
3. Zhang, Y.; Jiang, J. Bibliographical review on reconfigurable fault-tolerant control systems. In Proceedings of the 5th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes, Washington, DC, USA, 9–11 June 1997; pp. 265–276.
4. Owczarkowski, A.; Horla, D. Robust LQR and LQI control with actuator failure of a 2DOF unmanned bicycle robot stabilized by an inertial wheel. *Int. J. Appl. Math. Comput. Sci.* **2016**, *26*, 325–334. [\[CrossRef\]](#)
5. Xue, D.; El-Farra, N. Forecast-triggered model predictive control of constrained nonlinear processes with control actuator faults. *Mathematics* **2018**, *6*, 104. [\[CrossRef\]](#)
6. Hu, X.; Wang, L. Improved robust constrained model predictive control design for industrial processes under partial actuator faults. *IEEE Access* **2019**, *7*, 34095–34103. [\[CrossRef\]](#)
7. Hespanha, J.; Morse, A. Stability of switched systems with average dwell-time. In *Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, USA; IEEE: Piscataway, NJ, USA, 1999*; pp. 2655–2660.
8. Wang, R.; Hou, L.; Zong, G.; Fei, S.; Yang, D. Stability and stabilization of continuous-time switched systems: A multiple discontinuous convex Lyapunov function approach. *Int. J. Robust. Nonlinear Control* **2019**, *29*, 1499–1514.
9. Ouyang, H.; Lin, Y. Supervisory adaptive fault-tolerant control against actuator failures with application to an aircraft. *Int. J. Robust. Nonlinear Control* **2018**, *28*, 536–551.
10. Hosseinzadeh, M.; Salmasi, F. Fault-tolerant supervisory controller for a hybrid AC/DC micro-grid. *IEEE Trans. Smart Grid* **2018**, *9*, 2809–2823. [\[CrossRef\]](#)
11. Efimov, D.; Cieslak, J.; Henry, D. Supervisory fault-tolerant control with mutual performance optimization. *Int. J. Adapt. Control Signal Process.* **2013**, *27*, 251–279.
12. Cieslak, J.; Efimov, D.; Henry, D. Transient Management of a Supervisory Fault-Tolerant Control Scheme Based on Dwell-Time Conditions. *Int. J. Adapt. Control Signal Process.* **2015**, *29*, 123–142. [\[CrossRef\]](#)

13. Liu, Z.; Yin, X. A Review of Cognitive Control: Advancement, Definition, Framework, and Prospect. *Actuators* **2025**, *14*, 32. [\[CrossRef\]](#)
14. Haykin, S.; Fatemi, M.; Setoodeh, P.; Xue, Y. Cognitive Control. *Proc. IEEE* **2012**, *100*, 3156–3169. [\[CrossRef\]](#)
15. Vyas, J.; Mercangöz, M. Autonomous industrial control using an agentic framework with large language models. *arXiv* **2024**, arXiv:2411.05904.
16. Vyas, J.; Mercangöz, M. Autonomous Control Leveraging LLMs: An Agentic Framework for Next-Generation Industrial Automation. *arXiv* **2025**, arXiv:2507.07115.
17. Xu, Z.; Jain, S.; Kankanhalli, M. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv* **2024**, arXiv:2401.11817. [\[CrossRef\]](#)
18. Hiriyanna, S.; Zhao, W. Multi-Layered Framework for LLM Hallucination Mitigation in High-Stakes Applications: A Tutorial. *Computers* **2025**, *14*, 332. [\[CrossRef\]](#)
19. Bayat, A.; Abate, A.; Ozay, N.; Jungers, R.M. LLM-Enhanced Symbolic Control for Safety-Critical Applications. *arXiv* **2025**, arXiv:2505.11077.
20. González-Potes, A.; Martínez-Castro, D.; Paredes, C.M.; Ochoa-Brust, A.; Mena, L.J.; Martínez-Pelaez, R.; Felix, V.G.; Felix-Cuadras, R.A. Hybrid AI and LLM-Enabled Agent-Based Real-Time Decision Support Architecture for Industrial Batch Processes: A Clean-in-Place Case Study. *AI* **2026**, *7*, 51. [\[CrossRef\]](#)
21. Gill, M.S.; Vyas, J.; Markaj, A.; Gehlhoff, F.; Mercangöz, M. Leveraging LLM Agents and Digital Twins for Fault Handling in Process Plants. In *Proceedings of the 2025 IEEE 30th International Conference on Emerging Technologies and Factory Automation (ETFA)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–8. [\[CrossRef\]](#)
22. Song, L.; Zhang, C.; Zhao, L.; Bian, J. Pre-Trained Large Language Models for Industrial Control. *arXiv* **2023**, arXiv:2308.03028.
23. Rasheed, A.; Ravik, O.; San, O. Large Language Models for Control. *arXiv* **2025**, arXiv:2511.00337.
24. Bofill, J.; Abisado, M.B.; Villaverde, J.; Sampedro, G.A.R. Exploring Digital Twin-Based Fault Monitoring: Challenges and Opportunities. *Sensors* **2023**, *23*, 7087. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Nakti, I.; Mansouri, M.; Al-Hmouz, R.; Khedher, A. Artificial Intelligence Techniques With Digital Twin for Fault Diagnosis in Interconnected Systems: A Review. *IEEE Access* **2025**, *13*, 91860–91874. [\[CrossRef\]](#)
26. Sarker, I.H.; Janicke, H.; Mohsin, A.; Gill, A.; Maglaras, L.A. Explainable AI for Cybersecurity Automation, Intelligence and Trustworthiness in Digital Twin: Methods, Taxonomy, Challenges and Prospects. *ICT Express* **2024**, *10*, 935–958. [\[CrossRef\]](#)
27. Madathil, A.P.; Luo, X.; Liu, Q.; Walker, C.; Madarkar, R.; Qin, Y. A Review of Explainable Artificial Intelligence in Smart Manufacturing. *Int. J. Prod. Res.* **2025**, *63*, 8654–8697. [\[CrossRef\]](#)
28. Di Bonito, L.P.; Campanile, L.; Di Natale, F.; Mastroianni, M.; Iacono, M. eXplainable Artificial Intelligence in Process Engineering: Promises, Facts, and Current Limitations. *Appl. Syst. Innov.* **2024**, *7*, 121. [\[CrossRef\]](#)
29. Quamar, M.M.; Nasir, A. Review on Fault Diagnosis and Fault-Tolerant Control Scheme for Robotic Manipulators: Recent Advances in AI, Machine Learning, and Digital Twin. *arXiv* **2024**, arXiv:2402.02980.
30. Safa, N. A Survey of Decentralized Fault-Tolerant Control Systems. *Acad. J. Electr. Comput. Eng.* **2025**, *1*, 51–60. [\[CrossRef\]](#)
31. European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, 1–144.
32. Mumuni, A.; Mumuni, F. Explainable artificial intelligence (XAI): From inherent explainability to large language models. *arXiv* **2025**, arXiv:2501.09967.
33. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Di Der, J.; Diaz-Rodriguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [\[CrossRef\]](#)
34. Ahmad, A.; Saad, M.; Mohaisen, A. Secure and Transparent Audit Logs with BlockAudit. *J. Netw. Comput. Appl.* **2019**, *145*, 102406. [\[CrossRef\]](#)
35. Kulothungan, V. Using Blockchain Ledgers to Record AI Decisions in IoT. *IoT* **2025**, *6*, 37. [\[CrossRef\]](#)
36. Anderson, B.D.O.; Moore, J.B. *Optimal Control: Linear Quadratic Methods*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1990.
37. Liu, R.; Li, L.; Yang, Y. Performance residual based fault detection for feedback control systems. *IEEE Trans. Circuits Syst. II* **2021**, *68*, 3291–3295. [\[CrossRef\]](#)
38. Helvoirt, J.; De Jager, A. Dynamic model including piping acoustics of a centrifugal compression system. *J. Sound. Vib.* **2007**, *302*, 361–378. [\[CrossRef\]](#)
39. Shahriyari, M.; Firouzabadi, A.; Khaleghi, H.; Esmailifar, S. A new model for compressor surge and stall control. *Sci. Rep.* **2024**, *14*, 5347. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Hernández-Cedillo, J.; Gallegos, A.; Urenda-Cazares, E.; Barba-Franco, J.J. Modulation effects on the dynamics of a ring of three Duffing oscillators coupled unidirectionally with time-dependent damping. *Nonlinear Dyn.* **2024**, *113*, 313–327. [\[CrossRef\]](#)
41. Chen, R.; Lv, J.; Tian, J.; Ai, Y.; Zhang, F.; Yao, Y. Modeling and Nonlinear Dynamic Characteristics Analysis of Fault Bearing Time-Varying Stiffness-Flexible Rotor Coupling System. *Mathematics* **2024**, *12*, 3591. [\[CrossRef\]](#)

42. Cheng, H.; Zhang, Y.; Lu, W.; Yang, Z. Mechanical characteristics and nonlinear dynamic response analysis of rotor-bearing-coupling system. *Appl. Math. Model.* **2021**, *93*, 708–727. [\[CrossRef\]](#)
43. Wu, S.; Liu, J.; Zhang, J.; Chen, L.; Zhang, J.; Huang, W.; Meng, Q.; Ling, L.; Zhang, R.; Wang, Z.; et al. Transient thermal response of quasi-continuous-wave laser diodes and its impact on saturation power. *Opt. Laser Technol.* **2026**, *193*, 114269. [\[CrossRef\]](#)
44. Song, Y.; Lv, Z.; Bai, J.; Niu, S.; Wu, Z.; Qin, L.; Chen, Y.; Liang, L.; Lei, Y.; Jia, P.; et al. Processes of the Reliability and Degradation Mechanism of High-Power Semiconductor Lasers. *Crystals* **2022**, *12*, 765. [\[CrossRef\]](#)
45. Kenda, M.; Klobčar, D.; Bračun, D. Condition based maintenance of the two-beam laser welding in high volume manufacturing of piezoelectric pressure sensor. *J. Manuf. Syst.* **2021**, *59*, 117–126. [\[CrossRef\]](#)
46. Lucia, O.; Maussion, P.; Dede, E.J.; Burdío, J.M. Induction heating technology and its applications: Past developments, current technology, and future challenges. *IEEE Trans. Ind. Electron.* **2013**, *61*, 2509–2520. [\[CrossRef\]](#)
47. Xiang, X.; Luo, A.; Li, Y. Intelligent Control Method of Power Supply for Tundish Electromagnetic Induction Heating System. *J. Mod. Power Syst. Clean. Energy* **2020**, *8*, 1188–1195. [\[CrossRef\]](#)
48. Li, M.; Endo, R.; Akoshima, M.; Tanei, H.; Okada, H.; Susa, M. Thermal Conductivity of Oxide Scale Thermally Grown on Iron Substrate Corrected by Temperature-dependent Interfacial Thermal Resistance in Laser Flash Measurement. *ISIJ Int.* **2019**, *59*, 398–403. [\[CrossRef\]](#)
49. Wendelstorf, R.; Spitzer, K.; Wendelstorf, J. Effect of oxide layers on spray water cooling heat transfer at high surface temperatures. *Int. J. Heat Mass Transf.* **2008**, *51*, 4892–4901. [\[CrossRef\]](#)
50. Berce, J.; Zupančič, M.; Može, M.; Golobič, I. A Review of Crystallization Fouling in Heat Exchangers. *Processes* **2021**, *9*, 1356. [\[CrossRef\]](#)
51. Ben-Mansour, R.; El-Ferik, S.; Al-Naser, M.; Qureshi, B.A.; Eltoym, M.A.M.; Abuelyamen, A.; Al-Sunni, F.; Mnsour, R.B. Experimental/Numerical Investigation and Prediction of Fouling in Multiphase Flow Heat Exchangers: A Review. *Energies* **2023**, *16*, 2812. [\[CrossRef\]](#)
52. Patil, P.; Srinivasan, B.; Srinivasan, R. Monitoring Fouling in Heat Exchangers Under Temperature Control Based on Excess Thermal and Hydraulic Loads. *Chem. Eng. Res. Des.* **2022**, *181*, 41–54. [\[CrossRef\]](#)
53. De Carvalho, C.; Carvalho, E.; Ravagnani, M. Tuning Strategies for Overcoming Fouling Effects in Proportional Integral Derivative Controlled Heat Exchangers. *Ind. Eng. Chem. Res.* **2018**, *57*, 10518–10527. [\[CrossRef\]](#)
54. Matteis, P.; Campagnoli, E.; Firrao, D.; Ruscica, G. Thermal diffusivity measurements of metastable austenite during continuous cooling. *Int. J. Therm. Sci.* **2008**, *47*, 695–708. [\[CrossRef\]](#)
55. Kuepferle, J.; Wilzer, J.; Weber, S.; Theisen, W. Thermo-physical properties of heat-treatable steels in the temperature range relevant for hot-stamping applications. *J. Mater. Sci.* **2015**, *50*, 2594–2604. [\[CrossRef\]](#)
56. Liang, W.; Chen, Z.; Yao, B. High-accuracy adaptive robust fault-tolerant control for quadrotor with actuator uncertainties and aerodynamic drag compensation. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 8123–8136. [\[CrossRef\]](#)
57. Liu, X.; Chen, M.; Sheng, L.; Zhou, D. Adaptive fault-tolerant control for nonlinear high-order fully-actuated systems. *Neurocomputing* **2022**, *495*, 75–85. [\[CrossRef\]](#)
58. Papacharalampopoulos, A.; Foteinopoulos, P.; Karagianni, O.M.; Stavropoulos, P. A Human–AI Collaborative Framework for Additive Manufacturing Modeling and Decision-Making. *Processes* **2025**, *13*, 3877. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.