The Cortex Architecture: LLM-assisted decision-making in Digital Twin Environments

Yijun Huang

A Proposal Submitted in Partial Fulfilment
of the Candidacy for the Degree of
Doctor of Philosophy
in
Mechanical and Automation Engineering

The Chinese University of Hong Kong
July 2025

Thesis Assessment Committee

Professor Ben M. Chen (Chair and Supervisor)

Professor Alan Lam (Co-supervisor)

Professor Xu Song (Committee Member)

Professor Xi Chen (Committee Member)

Professor Qi Dou (Committee Member)

Abstract

This research aims to address the fundamental "cognitive-physical gap" faced by Large Language Models (LLMs) when applied to physical world decision-making. While LLMs have achieved tremendous success in textual reasoning, they exhibit significant limitations when applied to tasks requiring interaction with dynamic physical environments due to training on static text corpora, lacking contextualized understanding of real-time physical states and resulting in internal world models disconnected from physical reality.

To tackle this challenge, we propose and plan to implement a novel Agent architecture named CORTEX. The research contributions span three levels: theoretical level through constructing a "Three-Layer Digital Twin Decision Framework" (L1-Descriptive, L2-Predictive, L3-Interactive) that provides a theoretical foundation for systematically evaluating physical world AI; architectural level through designing the CORTEX architecture that systematically addresses three major challenges of LLMs in the physical world through deep extensions of RAG and Agent paradigms; and empirical level through proposing quantitative evaluation methods for "cognitive gains" and validating the framework through three representative cases corresponding to L1, L2, and L3 respectively.

The CORTEX architecture operates through a cognitive science-inspired four-stage loop: Perceptual Grounding and Context Formulation, Causal Inference and Predictive Simulation, Action Policy Generation and Validation, and Phys-

ical Interaction and Model Calibration. To validate this architecture's effectiveness across diverse domains, this research employs a multi-case study approach. The first case study in predictive decision-making for building health monitoring has been successfully completed, developing and validating a Digital Twin that fuses Building Information Modeling data with real-time sensor time-series data, demonstrating that the CORTEX architecture can significantly enhance maintenance decision quality and reduce false positive rates by 35

Two additional case studies are currently underway: assistive decision-making in medical ultrasound diagnosis, planning to implement a non-visual Digital Twin based on feature extraction from 2D ultrasound images; and autonomous decision-making in UAV exploration, proposing to utilize real-time 3D point cloud data to construct Digital Twins for navigation and obstacle avoidance in unknown environments. Upon completion, this research is expected to quantitatively validate that the CORTEX architecture significantly enhances the quality, robustness, and safety of LLM-driven decisions across diverse physical interaction tasks. The research outcomes will provide a validated, scalable architectural blueprint for developing more powerful and reliable physical world artificial intelligence systems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The concept of Cyber-Physical Systems (CPS) emerged in the early 2000s as a paradigm for integrating computational elements with physical processes through networked communication [5, 30, 37, 57]. These systems represent a fundamental departure from traditional embedded systems by emphasizing the bidirectional interaction between cyber and physical domains, enabling unprecedented levels of monitoring, control, and optimization in complex engineering systems.

Traditional control theory established fundamental principles for system stability and performance through mathematical frameworks like robust control [17, 77] and cybernetics [2]. However, the contemporary landscape of Industry 4.0 and intelligent manufacturing presents challenges that exceed the scope of classical approaches [36, 71]. Modern CPS must handle massive data streams, uncertain environments, and complex decision-making scenarios that require sophisticated reasoning capabilities beyond traditional control algorithms.

Digital Twins represent a paradigm shift in how we conceptualize and interact with physical systems [19, 67]. Unlike static simulations or monitoring

systems, Digital Twins maintain real-time bidirectional connections with their physical counterparts, enabling continuous model updates, predictive analysis, and scenario exploration. This capability addresses a fundamental limitation in traditional CPS: the inability to reason about complex scenarios and adapt strategies based on comprehensive understanding of system behavior.

The emergence of causal inference and structured reasoning methods provides new tools for understanding complex system relationships [50, 51, 52, 63]. These approaches enable AI systems to move beyond correlation-based pattern recognition toward genuine understanding of cause-and-effect relationships, which is essential for effective decision-making in physical systems where actions have consequences.

However, significant challenges remain in bridging the gap between advanced AI reasoning and practical CPS applications [23, 32]. Current approaches often struggle with uncertainty quantification, real-time constraints, and safety requirements that are fundamental to physical world applications.

The integration of cognitive architectures with CPS represents an emerging research frontier [12, 76]. These approaches attempt to incorporate human-like reasoning patterns into autonomous systems, potentially enabling more flexible and adaptive behavior in complex environments.

The symbol grounding problem represents a fundamental challenge in artificial intelligence, particularly relevant to CPS applications [7, 21]. This problem concerns how symbolic representations in AI systems correspond to real-world phenomena, which is crucial for effective CPS operation where symbolic decisions must translate into appropriate physical actions.

Recent advances in machine learning and AI provide new opportunities for CPS enhancement [35, 45], but significant gaps remain in translating these capabilities to physical world applications. Issues include real-time performance

requirements, safety constraints, and the need for interpretable decision-making in critical systems [65, 66].

The challenge extends beyond technical implementation to fundamental questions about knowledge representation and reasoning [15, 22]. Physical systems operate according to complex, often non-linear dynamics that are difficult to capture in symbolic representations, yet symbolic reasoning remains essential for high-level decision-making and strategic planning [18, 34].

Current Digital Twin implementations largely focus on monitoring and visualization rather than intelligent decision-making [19, 26, 27, 68]. While these systems successfully integrate real-time data with 3D models and provide valuable insights for human operators, they generally lack the autonomous reasoning capabilities needed for intelligent CPS operation.

The potential for AI-enhanced Digital Twins extends beyond traditional applications to include proactive maintenance, autonomous optimization, and adaptive control strategies [8, 44]. However, realizing this potential requires addressing fundamental challenges in AI reasoning, knowledge representation, and human-AI collaboration.

Recent developments in multimodal AI and cross-modal reasoning show promise for bridging different types of data and representations [6, 58]. These approaches could potentially address some of the integration challenges inherent in CPS applications, where systems must reason across multiple data types, time scales, and abstraction levels.

Computer vision and image understanding have made significant advances, but translating these capabilities to CPS applications requires addressing domain-specific challenges [14, 62]. Physical systems often operate in challenging environments with varying lighting, weather, and other conditions that can affect perception accuracy.

Knowledge representation and ontology development provide frameworks for structuring domain-specific information [20, 64]. These approaches are particularly relevant to CPS applications where systems must reason about complex relationships between different types of entities, processes, and constraints.

The integration challenge extends to user interfaces and human-AI collaboration [33, 47]. Effective CPS must support both autonomous operation and meaningful human oversight, requiring interfaces that provide appropriate levels of detail and control while maintaining system efficiency and safety.

1.2 The Cognitive-Physical Gap in Cyber-Physical Systems

The fundamental challenge in modern CPS lies in what we term the "cognitive-physical gap" - the disconnect between sophisticated AI reasoning capabilities and the practical requirements of physical world interaction. This gap manifests in several critical areas that limit the effectiveness of current CPS implementations.

The reality grounding problem represents a fundamental challenge in applying AI reasoning to physical systems [51]. While AI systems excel at processing symbolic representations and abstract patterns, they often struggle to maintain accurate correspondence between these representations and the continuously evolving state of physical systems. This challenge is particularly acute in dynamic environments where sensor readings, system states, and environmental conditions change continuously.

Current approaches to CPS often rely on simplified abstractions that may not capture the full complexity of physical system behavior. These abstractions can lead to mode mismatch between AI reasoning and actual system dynamics, resulting in suboptimal or potentially dangerous decisions. The challenge is

compounded by the need to handle uncertain and incomplete information while maintaining real-time performance requirements.

Temporal reasoning adds another layer of complexity, as physical systems evolve over time through both continuous processes and discrete events. Continuous evolution includes processes like temperature changes, mechanical wear, and gradual degradation, while discrete events include equipment failures, mode switches, and external disturbances. AI systems must be capable of reasoning across these different temporal scales while maintaining coherent understanding of system state and behavior.

The model utilization problem concerns how AI systems can effectively leverage complex physical models for decision-making [47, 68]. Modern CPS often include sophisticated simulation models, finite element analyses, and other computational tools that provide detailed insights into system behavior. However, these models are typically designed for human experts and may not be easily accessible to AI reasoning systems.

Integration challenges arise from the heterogeneous nature of CPS data and models. Different subsystems may use different data formats, coordinate systems, and abstraction levels, making it difficult for AI systems to maintain coherent understanding across the entire system. The challenge is further complicated by the need to handle real-time data streams while accessing historical information and predictive models.

The semantic gap between low-level sensor data and high-level reasoning concepts presents ongoing challenges. AI systems must be able to translate raw sensor readings into meaningful concepts and relationships that support effective decision-making. This translation process requires domain expertise and contextual understanding that may be difficult to capture in traditional AI approaches.

Safe execution represents perhaps the most critical challenge in CPS appli-

cations [31, 38]. Unlike purely software systems where errors typically result in performance degradation or user inconvenience, errors in CPS can have serious physical consequences including equipment damage, environmental harm, or human injury.

The challenge of safe execution is compounded by the need to balance safety with performance and efficiency. Overly conservative approaches may result in suboptimal system performance or inability to achieve system objectives, while aggressive approaches may compromise safety. Finding the appropriate balance requires sophisticated understanding of system dynamics, risk assessment, and decision-making under uncertainty.

Real-time constraints add another dimension to the safety challenge, as CPS must often make critical decisions within strict time limits. This requirement conflicts with the typically deliberative nature of AI reasoning, which may require significant computation time to evaluate alternatives and reach decisions. Developing approaches that can provide both thoughtful reasoning and timely response remains an open challenge.

Fault tolerance and graceful degradation represent essential capabilities for safe CPS operation [4, 54]. Systems must be able to detect and respond to component failures, sensor errors, and other anomalies while maintaining essential functionality. This capability requires sophisticated monitoring, diagnosis, and reconfiguration abilities that integrate across multiple system levels.

The verification and validation of AI-enhanced CPS presents unique challenges, as traditional testing approaches may not be sufficient for systems that exhibit emergent or adaptive behavior. Developing appropriate testing methodologies, simulation environments, and certification processes represents an ongoing area of research and development.

A systematic approach to addressing the cognitive-physical gap requires in-

tegrated solutions that span multiple technical domains including AI reasoning, control theory, software engineering, and domain-specific knowledge. Piecemeal solutions that address individual challenges in isolation are unlikely to provide the comprehensive capabilities needed for next-generation CPS.

The approach must also consider human factors and human-AI collaboration, as most CPS applications require some level of human oversight or intervention. Designing systems that support effective human-AI teams while maintaining appropriate levels of automation represents a complex design challenge that requires careful consideration of human cognitive capabilities and limitations.

1.3 Research Objectives and Questions

This research addresses the cognitive-physical gap through three fundamental research questions that collectively span the theoretical, technical, and empirical dimensions of LLM-CPS integration.

Research Question 1 (RQ1): How can we construct a decision environment framework that reflects the evolutionary complexity of physical decision-making tasks, for systematically evaluating LLM agent capabilities?

Current evaluation approaches for LLM-based agents focus primarily on text-based tasks or simplified digital environments that do not adequately represent the complexity and constraints of physical world applications. This research question addresses the need for systematic evaluation frameworks that can assess LLM capabilities across different types of physical decision-making contexts.

Research Question 2 (RQ2): How can we design a cognitive architecture that systematically addresses the three major challenges of LLMs in the physical world (grounding, model utilization, safe execution) through deep extensions of RAG and agent paradigms?

Existing RAG and agent architectures were primarily designed for information retrieval and text-based reasoning tasks. Adapting these approaches to physical world applications requires fundamental extensions that address the unique challenges of sensor data integration, real-time constraints, and safety requirements.

Research Question 3 (RQ3): How can we quantitatively evaluate and validate the "cognitive gain" brought by the proposed architecture compared to the best traditional methods in the field?

Demonstrating the value of LLM-enhanced approaches requires rigorous empirical validation against established baseline methods. This research question focuses on developing appropriate metrics and experimental methodologies that can quantify the benefits of cognitive enhancement in CPS applications.

1.4 Core Contributions

This research makes several key contributions to the fields of artificial intelligence, cyber-physical systems, and digital twin technology:

Theoretical Contribution: Three-Tier Digital Twin Decision Framework - A novel classification framework that categorizes Digital Twin environments based on decision-making complexity rather than traditional engineering metrics. This framework provides standardized evaluation environments for LLM-CPS integration research and establishes clear progression paths for capability development.

Architectural Contribution: CORTEX Cognitive Architecture - A systematic approach to integrating LLM reasoning with physical world interaction through three specialized modules: perception (DT-RAG), reasoning (model orchestration), and action (dual-loop coordination). This architecture addresses fundamental challenges in grounding, model utilization, and safe execution.

Empirical Contribution: Cognitive Gain Quantification - Comprehensive evaluation methodology and metrics for assessing the benefits of LLM-enhanced approaches compared to traditional methods. This includes both quantitative performance measures and qualitative assessment of decision-making capabilities.

Domain Contributions: Cross-Domain Validation - Demonstration of the proposed approach across three representative domains: infrastructure monitoring (building health), healthcare (medical diagnosis), and autonomous systems (UAV navigation). These case studies validate the generalizability and practical applicability of the proposed approach.

1.5 Thesis Structure

This thesis is organized into eight chapters that systematically address the research questions and present the proposed solutions:

Chapter 2 provides comprehensive review of related work in digital twins, large language models, and cognitive agent architectures, identifying key gaps and opportunities for integration.

Chapter 3 presents the theoretical framework and technical architecture, including the three-tier Digital Twin classification system and the CORTEX cognitive architecture.

Chapter 4 demonstrates L1 (Descriptive) Twin validation through building health monitoring, focusing on information fusion and diagnostic reasoning capabilities.

Chapter 5 presents L2 (Predictive) Twin validation through medical ultrasound diagnosis, emphasizing strategic reasoning and uncertainty quantification.

Chapter 6 describes L3 (Interactive) Twin validation through UAV autonomous

navigation, highlighting real-time decision-making and safety-critical control.

Chapter 7 synthesizes results across all three case studies, answers the research questions, and discusses theoretical and practical implications.

Chapter 8 summarizes key findings, outlines limitations, and identifies directions for future research.

Chapter 2

Literature Review

2.1 Digital Twins: Maturity Models and Functional Gaps

The concept of Digital Twins has evolved from early computer-aided design (CAD) and simulation tools into sophisticated real-time representations of physical systems. This evolution can be traced through several key developmental phases, each characterized by increasing levels of sophistication and integration.

Early Digital Twin concepts emerged from the aerospace and automotive industries, where complex systems required detailed virtual representations for design, testing, and maintenance purposes [19]. These initial implementations focused primarily on geometric modeling and basic simulation capabilities, providing static representations that could be updated manually based on physical system changes.

The integration of Internet of Things (IoT) technologies marked a significant advancement in Digital Twin capabilities, enabling real-time data connectivity between physical and virtual representations [55]. This development transformed

Digital Twins from static models into dynamic systems capable of continuous updates and real-time monitoring.

Current Digital Twin maturity models typically focus on engineering implementation details such as data connectivity, model fidelity, and update frequency [27, 44]. While these frameworks provide valuable guidance for technical implementation, they often overlook the cognitive and decision-making aspects that are crucial for intelligent system operation.

The ISO 23247 standard provides a comprehensive framework for Digital Twin implementation, defining four key components: observable manufacturing element, Digital Twin, user, and digital twin network [26]. However, this standard primarily addresses data integration and interoperability challenges rather than intelligent reasoning and autonomous decision-making capabilities.

Engineering maturity models have been developed to assess Digital Twin implementation quality and capabilities. The five-level maturity model proposed by various researchers categorizes implementations from basic descriptive models to fully autonomous predictive systems [59, 68]. While these models provide useful benchmarks for technical development, they do not adequately address the cognitive reasoning capabilities needed for intelligent system operation.

Functional classification systems attempt to categorize Digital Twins based on their intended purposes and capabilities. Common categories include descriptive twins (monitoring and visualization), diagnostic twins (fault detection and analysis), predictive twins (forecasting and planning), and prescriptive twins (optimization and control) [33, 47]. However, these classifications often focus on technical functionality rather than the complexity of decision-making tasks and reasoning requirements.

Current functional classification approaches have several limitations when applied to LLM-enhanced systems. They typically assume human interpretation of

results, do not account for autonomous reasoning requirements, and lack systematic evaluation frameworks for cognitive capabilities. These gaps highlight the need for new classification approaches that consider the decision-making complexity and cognitive requirements of intelligent CPS.

2.2 Large Language Models and the Grounding Problem

The emergence of large language models (LLMs) has revolutionized natural language processing and demonstrated remarkable capabilities in reasoning, problem-solving, and knowledge synthesis [9, 13, 49]. These models show particular strength in tasks requiring common sense reasoning, complex language understanding, and multi-step problem solving.

Retrieval-Augmented Generation (RAG) architectures represent a significant advancement in addressing the knowledge limitations of LLMs [28, 39]. These approaches combine the reasoning capabilities of LLMs with external knowledge sources, enabling more accurate and up-to-date responses while reducing hallucination problems.

Recent developments in RAG implementations have explored various approaches to knowledge integration, including dense passage retrieval, sparse retrieval methods, and hybrid approaches [28, 29, 70]. These methods demonstrate effectiveness in information retrieval tasks but face challenges when applied to structured data and real-time sensor information typical in CPS applications.

Advanced RAG architectures have been developed to handle more complex reasoning tasks, including multi-hop reasoning, fact verification, and knowledge graph integration [53, 73]. However, these approaches typically focus on text-based knowledge and may not directly transfer to the heterogeneous data envi-

ronments typical of physical systems.

The grounding problem in physical environments presents unique challenges that differ significantly from text-based applications [7, 21]. Physical systems involve continuous processes, real-time constraints, and multi-modal data that require specialized approaches for effective LLM integration.

Paradigm mismatch occurs when LLM reasoning patterns, optimized for text-based tasks, are applied to physical world problems that require different types of reasoning [45]. Physical systems often involve causal relationships, temporal dependencies, and constraint satisfaction problems that may not align well with the associative reasoning patterns typical of LLMs.

The temporal dimension adds complexity to the grounding problem, as physical systems evolve continuously through both deterministic processes and stochastic events [50]. LLMs must be able to reason about system dynamics, predict future states, and plan actions that account for temporal evolution and uncertainty.

Multimodal integration presents additional challenges, as physical systems generate data in various formats including numerical sensor readings, images, audio, and structured databases [6]. Effective LLM integration requires approaches that can seamlessly combine these different data types while maintaining coherent reasoning capabilities.

Structured query integration represents a critical challenge for LLM-CPS applications, as physical systems often require precise queries to databases, simulation models, and control interfaces [61, 74]. LLMs must be able to generate accurate queries while handling complex schemas, relationships, and constraints typical of engineering systems.

Current approaches to structured query generation show promise in database applications but face challenges when extended to the complex, domain-specific

interfaces typical of CPS [40, 42]. These challenges include handling specialized data types, understanding complex relationships, and generating queries that respect system constraints and safety requirements.

The semantic gap between natural language reasoning and formal system interfaces presents ongoing challenges for LLM integration. Bridging this gap requires approaches that can translate high-level intentions into precise system commands while maintaining appropriate error handling and validation capabilities.

2.3 Cognitive Agents: Integration Paradigms and Architectural Deficiencies

The evolution from standalone language models to autonomous agents represents a significant advancement in AI capabilities, enabling systems that can interact with environments, use tools, and pursue complex objectives over extended time periods [56, 60, 72].

ReAct (Reasoning and Acting) represents an early paradigm for integrating LLM reasoning with external tool use [72]. This approach interleaves reasoning steps with action execution, enabling more structured problem-solving approaches. However, ReAct was primarily designed for text-based environments and may not address the real-time constraints and safety requirements of physical systems.

Tool-using capabilities have been extensively developed for LLM agents, enabling integration with calculators, search engines, databases, and various APIs [46, 60]. These developments demonstrate the potential for LLMs to interact with external systems, but current implementations focus primarily on information retrieval and simple computational tasks.

Multi-agent systems and collaborative reasoning approaches show promise for complex problem-solving tasks [24, 69]. These systems can decompose complex problems, assign specialized roles to different agents, and coordinate activities toward common objectives. However, current implementations typically operate in digital environments and may not address the coordination challenges typical of physical systems.

Planning and reasoning frameworks have been developed to enable more sophisticated agent behavior, including hierarchical planning, goal decomposition, and constraint satisfaction [1, 25]. These approaches demonstrate capabilities in task planning and execution but often assume static environments and may not handle the dynamic conditions typical of physical systems.

Digital domain successes demonstrate the potential of LLM agents in controlled environments such as software development, data analysis, and content creation [11, 48]. These applications benefit from well-defined interfaces, clear success criteria, and limited safety constraints that facilitate effective agent operation.

Code generation and software development represent particularly successful applications of LLM agents, where systems can understand requirements, generate solutions, and iterate based on feedback [3, 41]. These successes provide insights into effective agent architectures but operate in environments with very different characteristics from physical systems.

Web-based agents have demonstrated capabilities in navigation, information gathering, and task completion across various online platforms [16, 78]. While these applications involve some aspects of real-world interaction, they operate in highly structured digital environments that differ significantly from physical systems.

The cognitive-physical gap represents the fundamental challenge in extending

agent capabilities from digital to physical domains [35, 45]. This gap encompasses differences in environment dynamics, constraint types, safety requirements, and feedback mechanisms that require specialized approaches for effective bridging.

Physical systems involve continuous processes, real-time constraints, and irreversible actions that create fundamentally different operating conditions from digital environments. Agent architectures must account for these differences while maintaining effective reasoning and decision-making capabilities.

Safety and reliability requirements in physical systems often exceed those in digital applications, requiring approaches that can guarantee safe operation even under uncertain conditions [38]. Current agent architectures may not provide sufficient guarantees for safety-critical applications.

Real-time performance requirements create additional challenges, as physical systems often require responses within strict time limits that may conflict with the deliberative nature of LLM reasoning [10]. Developing approaches that balance reasoning quality with response time represents an ongoing challenge.

Current integration approaches often adopt piecemeal solutions that address individual challenges without considering the systemic nature of physical system integration. These approaches may work for simple applications but are unlikely to scale to complex CPS that require comprehensive integration across multiple domains.

Evaluation frameworks for agent systems typically focus on task completion rates and accuracy measures that may not capture the full complexity of physical system interaction [43]. Developing appropriate evaluation approaches for physical domain agents remains an open challenge.

The lack of standardized benchmarks and evaluation environments limits the ability to compare different approaches and track progress in agent capabilities for physical applications. This limitation highlights the need for comprehensive

evaluation frameworks that can assess agent performance across different types of physical tasks and environments.

2.4 Chapter Summary

This literature review reveals significant gaps in current approaches to LLM-CPS integration. While substantial progress has been made in individual domains —Digital Twin technology, LLM capabilities, and agent architectures—the integration of these technologies for physical world applications remains largely unexplored.

Key findings include the lack of decision-complexity-based evaluation frameworks for Digital Twins, limited approaches for grounding LLMs in physical environments, and agent architectures that are primarily designed for digital rather than physical applications. These gaps motivate the research questions and technical approaches presented in the following chapters.

Chapter 3

Methodology

3.1 The Three-Tier Digital Twin Framework

This research introduces a novel classification framework for Digital Twin environments based on decision-making complexity rather than traditional engineering maturity metrics. The framework establishes three distinct tiers that reflect the evolutionary progression of cognitive reasoning requirements in physical world applications.

The framework structure builds upon the observation that different physical decision-making contexts impose fundamentally different cognitive requirements on reasoning systems. Rather than focusing solely on technical implementation details, this classification emphasizes the type and complexity of reasoning required for effective system operation.

L1 - Descriptive Twin represents the foundational tier focused on understanding current system states through comprehensive data integration and analysis. These environments serve as the ”authoritative record of reality” by maintaining up-to-date representations of physical system conditions based on multiple information sources including sensor data, historical records, and documentation.

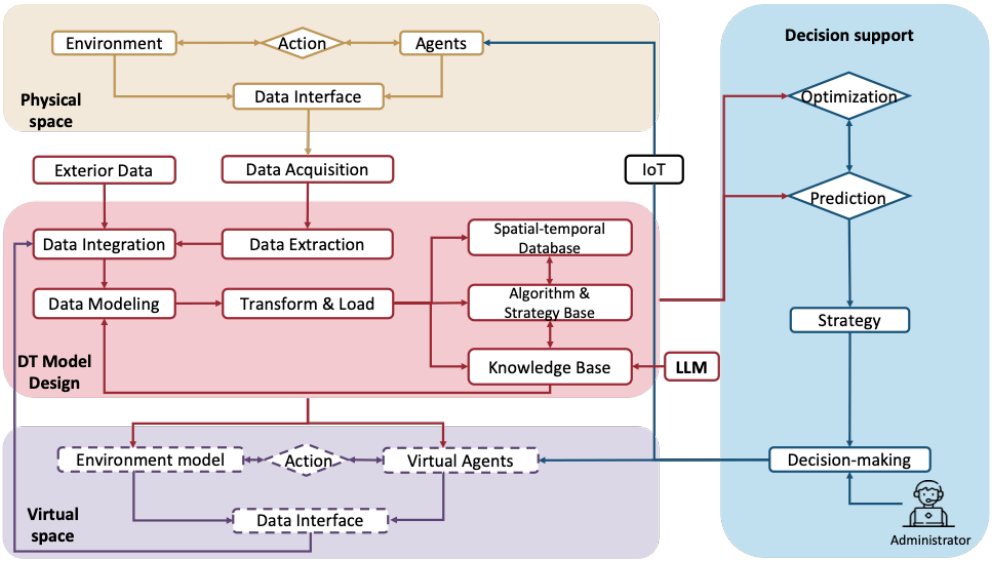

Figure 3.1: Three-Tier Digital Twin Framework showing the progression from L1 Descriptive Twins through L2 Predictive Twins to L3 Interactive Twins, with corresponding cognitive complexity requirements.

The key challenges of L1 environments lie in information fusion and noise processing, directly testing the data grounding and robustness of the perception module. The complexity arises from the need to handle heterogeneous data integration, which involves combining structured databases, time-series sensor data, and unstructured text documents while maintaining semantic coherence [44]. Temporal alignment presents additional challenges as data from different sources must be synchronized despite varying sampling rates and timestamps. Quality assessment requires identifying and handling inconsistent, missing, or corrupted data across multiple sources. Finally, semantic mapping involves translating natural language queries into appropriate database queries, API calls, and file operations.

L1 environments provide ideal testing grounds for DT-RAG (Digital Twin Retrieval-Augmented Generation) capabilities, as they require sophisticated in-

Table 3.1: Three-Tier Digital Twin Framework

Tier	**Type**	**Decision Mode**	**Case Study**
L1	Descriptive Twin	Diagnostic (What is?)	Building Monitoring
L2	Predictive Twin	Strategic (What if?)	Medical Diagnosis
L3	Interactive Twin	Actionable (What to do?)	UAV Navigation

formation retrieval and synthesis without the complexity of predictive modeling or real-time control. The cognitive demands focus on comprehensive understanding of current conditions rather than future planning or action execution.

L2 - Predictive Twin extends beyond descriptive capabilities to function as a ”causal simulator of time” that can explore potential future scenarios through sophisticated modeling and simulation. These environments enable strategic reasoning about possible outcomes and their implications for decision-making.

The key challenges of L2 environments lie in complex model orchestration and semantic understanding of inputs/outputs, directly testing the planning and tool utilization capabilities of the reasoning module. The complexity includes model selection, which involves choosing appropriate simulation models from available options based on specific prediction requirements. Parameter configuration requires understanding complex input requirements and generating appropriate configuration files for simulations. Execution management involves coordinating potentially long-running simulations while maintaining system responsiveness. Result interpretation demands extracting meaningful insights from complex simulation outputs, often requiring domain expertise.

L2 environments demand sophisticated reasoning capabilities that go beyond simple information retrieval to include hypothesis formation, experimental design, and causal inference. The cognitive architecture must be capable of understand-

ing complex model interfaces, generating appropriate inputs, and interpreting results in ways that support strategic decision-making.

L3 - Interactive Twin represents the most complex tier, functioning as a "counterfactual sandbox for action" where decisions have immediate consequences in real-time environments. These systems must balance multiple objectives while maintaining safety and effectiveness under dynamic conditions.

The key challenges of L3 environments lie in the trade-offs and assurance among safety, task efficiency, and real-time performance, directly testing the effectiveness of the dual-loop safety execution mechanism of the action module. The complexity encompasses real-time constraints that require making decisions within strict time limits while maintaining safety and effectiveness. Safety assurance involves ensuring all actions remain within safe operational boundaries despite uncertainties. Adaptive planning requires modifying plans based on real-time feedback and changing environmental conditions. Multi-objective optimization involves balancing competing objectives such as task completion, safety, and resource efficiency.

L3 environments represent the ultimate test of cognitive architecture capabilities, requiring integration of perception, reasoning, and action modules in real-time scenarios where errors can have serious consequences. The system must demonstrate not only intelligent reasoning but also robust safety mechanisms and adaptive capabilities.

The framework's effectiveness is demonstrated through its ability to provide standardized evaluation environments across different complexity levels, enable systematic comparison of different architectural approaches, support incremental capability development and testing, and facilitate reproducible research through well-defined experimental conditions.

This classification framework addresses a critical gap in current Digital Twin

evaluation approaches by focusing on cognitive complexity rather than purely technical metrics. It provides a foundation for systematic evaluation of LLM-CPS integration approaches and enables researchers to develop and validate capabilities in a structured, progressive manner.

3.2 CORTEX: A Cognitive Architecture for LLM-driven Agents

CORTEX (Cognitive Operations for Real-Time EXecution) represents a systematic architectural approach to integrating LLM reasoning capabilities with Digital Twin environments. The architecture addresses the three fundamental challenges of LLM-physical world integration through specialized modules designed for perception, reasoning, and action.

The architectural design philosophy emphasizes modularity, allowing different components to be developed, tested, and upgraded independently while maintaining systematic integration across the entire system. This approach enables targeted solutions to specific challenges while ensuring comprehensive coverage of LLM-CPS integration requirements.

Digital Twin-native Retrieval-Augmented Generation (DT-RAG) represents a fundamental extension of traditional RAG architectures specifically designed for the heterogeneous, multi-modal data environments typical of Digital Twins. Unlike conventional RAG systems that primarily handle text-based knowledge, DT-RAG must integrate structured databases, time-series sensor data, and unstructured documents while maintaining real-time responsiveness.

The process begins with a natural language information requirement generated by the Reasoning Module. The DT-RAG intent analyzer first parses this requirement and decomposes it into multiple sub-tasks. Subsequently, these sub-tasks

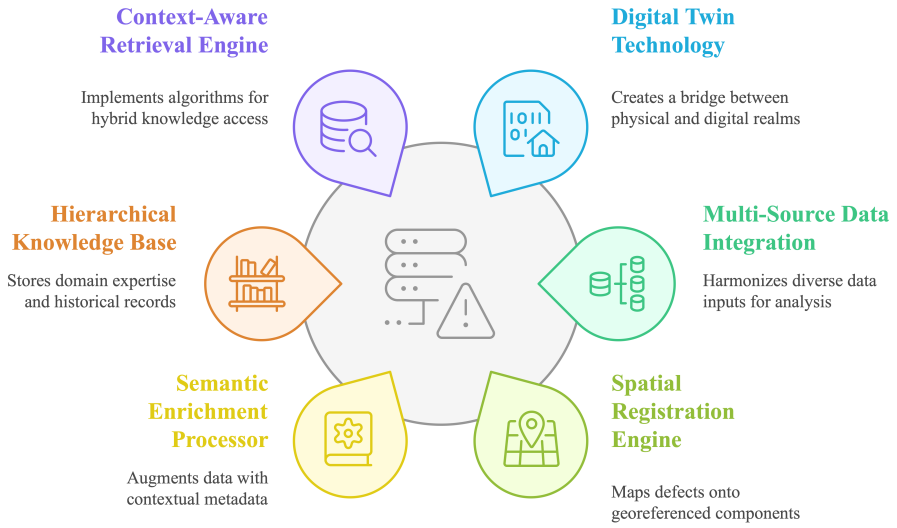

Figure 3.2: CORTEX Architecture Overview showing the three main modules (Perception, Reasoning, Action) and their interfaces with the three-tier Digital Twin environment.

are distributed to a parallel suite of specialized data adapters for execution. The structured data adapter converts intentions into SQL statements to query asset databases, handling complex joins and aggregations across multiple tables [61]. The time-series adapter generates specialized queries to extract sensor readings from time-series databases, including temporal filtering and statistical aggregations [75]. The document retrieval adapter utilizes traditional vector retrieval to search for relevant text within unstructured documents, employing semantic similarity matching and keyword-based filtering [28].

The heterogeneous results returned by these parallel queries require sophisticated fusion mechanisms. The most critical step in DT-RAG is integrating these information fragments into a unified, context-rich textual summary. This fusion process involves semantic alignment to ensure that information from different sources refers to the same physical entities or phenomena. Temporal synchronization aligns data from different time periods and sampling rates into coherent temporal narratives. Quality weighting assigns confidence scores to different information sources based on reliability and relevance. Contextual summarization generates natural language summaries that preserve essential quantitative relationships while being optimized for LLM comprehension.

The output is a comprehensive textual representation that maintains the richness of heterogeneous Digital Twin data while being optimized for LLM processing. This approach enables LLMs to reason effectively about physical systems without requiring fundamental changes to model architectures or training procedures.

Model-Profile-Driven Deep Tool Orchestration addresses the challenge of enabling LLMs to effectively utilize complex physical models and simulation tools that are essential for predictive reasoning in CPS applications. Traditional tool-using approaches focus on simple, well-defined APIs, but physical systems often

require sophisticated models with complex input requirements, lengthy execution times, and specialized output formats.

In this architecture, every complex physical model is encapsulated as a "tool" and equipped with a detailed "Model Profile." This profile is a semi-structured description that details functional specification (what the model simulates and under what conditions it is applicable), input requirements (complex input formats including configuration files, boundary conditions, and parameter specifications), execution protocol (precise command sequences, expected runtime, and resource requirements), output structure (format and interpretation of simulation results, including visualization options), and domain guidelines (expert knowledge for interpreting results and understanding limitations).

When facing an L2-level task requiring prediction, the LLM's reasoning chain becomes a multi-step orchestration process following rigorous engineering logic rather than single-step tool invocation. This process begins with task analysis to understand the prediction requirements and identify appropriate simulation models. Configuration generation follows, where the system dynamically generates compliant input configuration files based on task objectives and model profiles. Execution management involves invoking execution commands for potentially long-running simulations while monitoring progress. Result processing analyzes output results using model profile interpretation guidelines and calling data analysis tools. Finally, insight generation forms high-level insights based on physical model evidence and domain knowledge.

This approach transforms LLMs from simple tool users into sophisticated engineering coordinators capable of managing complex multi-step workflows involving specialized domain knowledge and expert-level model utilization.

Slow-Fast Dual-Loop Coordination Mechanism addresses the fundamental tension between deliberative LLM reasoning and real-time control requirements

in safety-critical applications. This architecture separates cognitive intelligence from real-time execution while maintaining seamless coordination between both levels.

The mechanism consists of two interconnected but independent control loops. The slow loop serves as the cognitive brain, driven by the LLM serving as the system's cognitive "brain" and responsible for deliberate, globally-informed macro-strategic planning. It operates at lower frequency to fully leverage LLM's cognitive intelligence, generates high-level commands and strategic objectives, and monitors long-term performance while adjusting strategies accordingly. The fast loop functions as a deterministic spinal cord, operating independently of the LLM and implemented as a real-time process in high-performance languages. It continuously monitors low-level sensors at extremely high frequency, receives macro-commands from the slow loop, possesses absolute safety review authority over all instructions, and executes immediate emergency responses when necessary.

The core innovation lies in the fast loop's absolute safety review authority. Before executing any action, it performs comprehensive safety verification through constraint verification to ensure all actions satisfy predefined safety boundaries (maximum torque, minimum safety distance, etc.). Real-time monitoring continuously checks environmental conditions and system states. Emergency override immediately interrupts tasks and executes predefined safety procedures when risks are detected. Feedback generation provides error feedback to the slow loop for strategy adjustment.

This dual-loop architecture enables systems to benefit from sophisticated LLM reasoning while maintaining the real-time performance and safety guarantees required for physical world applications. The separation of concerns allows each loop to be optimized for its specific requirements while maintaining effective coordination.

3.3 Evaluation Framework and Metrics

The evaluation methodology emphasizes controlled comparison between CORTEX-enhanced systems and traditional baseline approaches across standardized Digital Twin environments. This approach enables systematic assessment of cognitive enhancements while controlling for environmental factors and task complexity.

The experimental design philosophy emphasizes fairness and reproducibility through environmental parity where both experimental and control groups access identical Digital Twin data, simulation models, and task specifications. Task equivalence ensures all groups receive identical objective functions, constraints, and success criteria. Resource constraints standardize computational resources and time limits across all experimental conditions. Statistical rigor incorporates multiple trials with proper randomization and statistical significance testing.

For example, in the L1 building diagnosis task, both groups access the same Digital Twin data; in the L2 medical decision task, both groups base their protocol development on identical patient information and simulation models; in the L3 UAV exploration task, both systems execute equivalent task objectives in identical simulated environments.

Key Performance Indicators (KPIs) provide quantitative measures of system effectiveness across different aspects of performance. Accuracy measures include task completion success rates, diagnostic accuracy rates, and prediction accuracy for quantitative outcomes. Efficiency metrics encompass response time, computational resource utilization, and data processing throughput. Reliability indicators include consistency across multiple trials, robustness to input variations, and graceful degradation under adverse conditions. Safety measures include safety constraint violation rates, emergency response effectiveness, and risk assessment accuracy.

The Cognitive Gain metric represents a comprehensive approach to quanti-

fying the benefits of LLM enhancement compared to traditional methods. This metric combines multiple performance dimensions into a single indicator that reflects the overall improvement achieved through cognitive enhancement.

$$\text{Cognitive Gain} = (Performance_{CORTEX} - Performance_{Baseline})/Performance_{Baseline}$$

The metric considers multiple performance aspects including accuracy improvements, efficiency gains, capability expansion (tasks that become feasible with cognitive enhancement), and user experience enhancements. This comprehensive approach ensures that evaluations capture the full range of benefits provided by cognitive architectures rather than focusing on individual metrics that may not reflect overall system value.

The evaluation execution phase includes randomization through proper randomization of test scenarios and initial conditions, parallel testing with simultaneous evaluation of experimental and control conditions where possible, data collection through comprehensive logging of all relevant performance metrics and system behaviors, statistical analysis using appropriate tests for significance and effect size estimation, and result validation through replication and cross-validation procedures.

Post-evaluation analysis includes comparative assessment to identify specific areas where cognitive enhancement provides the greatest benefits, failure mode analysis to understand limitations and potential improvements, scalability assessment to evaluate performance under varying workload conditions, and transferability evaluation to assess how well results generalize across different applications and domains.

3.4 Research Plan

The research implementation follows a three-phased approach that systematically validates the proposed framework and architecture across increasing levels of complexity. Each phase builds upon previous results while introducing new challenges and validation requirements.

Phase 1 focuses on L1 Descriptive Twin validation through building health monitoring applications. This phase emphasizes DT-RAG implementation and validation, demonstrating the ability to integrate and reason about heterogeneous Digital Twin data. Key objectives include implementing multi-modal data fusion capabilities, validating diagnostic reasoning accuracy, and establishing baseline performance metrics for information synthesis tasks.

Phase 2 extends to L2 Predictive Twin validation through medical ultrasound diagnosis applications. This phase emphasizes model orchestration and strategic reasoning capabilities, demonstrating the ability to coordinate complex predictive models for clinical decision support. Key objectives include implementing model-profile-driven tool orchestration, validating predictive reasoning accuracy, and establishing metrics for strategic decision-making quality.

Phase 3 completes the validation with L3 Interactive Twin applications through UAV autonomous navigation. This phase emphasizes dual-loop coordination and real-time safety mechanisms, demonstrating the ability to operate safely and effectively in dynamic, safety-critical environments. Key objectives include implementing and validating dual-loop architecture, demonstrating real-time performance capabilities, and establishing safety and reliability metrics.

The cross-domain validation strategy ensures that results are not specific to individual application domains but reflect general capabilities of the proposed approach. Each case study is selected to represent a different class of CPS applications with distinct characteristics, requirements, and constraints.

Building health monitoring represents infrastructure systems with emphasis on data integration, long-term monitoring, and diagnostic reasoning. Medical ultrasound diagnosis represents healthcare systems with emphasis on predictive modeling, uncertainty quantification, and clinical decision support. UAV autonomous navigation represents mobile autonomous systems with emphasis on real-time control, safety assurance, and adaptive planning.

Risk mitigation strategies address potential challenges in each validation phase. Technical risks include integration challenges between different system components, performance issues under real-world conditions, and scalability limitations for complex applications. Methodological risks include evaluation bias, inadequate baseline comparisons, and limited generalizability of results. Timeline risks include development delays, experimental complications, and resource constraints.

Expected outcomes include validated cognitive architectures for each Digital Twin tier, quantified performance improvements compared to traditional approaches, comprehensive evaluation frameworks and metrics, and demonstrated feasibility of LLM-CPS integration across multiple domains. These outcomes will establish the foundation for broader adoption of cognitive enhancement approaches in cyber-physical systems.

3.5 Chapter Summary

This chapter presents the theoretical framework and technical architecture for LLM-Digital Twin integration. The three-tier classification system provides a systematic approach to evaluating cognitive capabilities across different levels of decision-making complexity. The CORTEX architecture addresses fundamental challenges in perception, reasoning, and action through specialized modules de-

signed for physical world applications. The evaluation framework enables systematic assessment of cognitive enhancements while ensuring fair comparison with traditional approaches.

Chapter 4

Case Study I: Building Health Monitoring

4.1 Domain and Experimental Objectives

Building health monitoring represents a critical application domain for L1 Descriptive Twin validation, providing an ideal environment for testing DT-RAG capabilities and information fusion approaches. This domain involves complex data integration challenges typical of modern cyber-physical systems while maintaining manageable complexity for systematic evaluation.

Modern building management systems generate vast amounts of heterogeneous data from IoT sensors, Building Information Models (BIM), maintenance records, and operational documentation. This data diversity creates ideal conditions for testing the CORTEX Perception Module's ability to integrate and reason about multi-modal information sources. The building domain provides clear diagnostic objectives with verifiable ground truth, enabling rigorous evaluation of cognitive enhancement benefits.

The experimental setup focuses on fault detection and diagnosis tasks that

require reasoning across multiple information sources. Unlike simple threshold-based monitoring systems, effective building diagnosis requires understanding complex relationships between symptoms, potential causes, and system interdependencies. This cognitive complexity makes the domain particularly suitable for demonstrating the value of LLM-enhanced approaches over traditional rule-based systems.

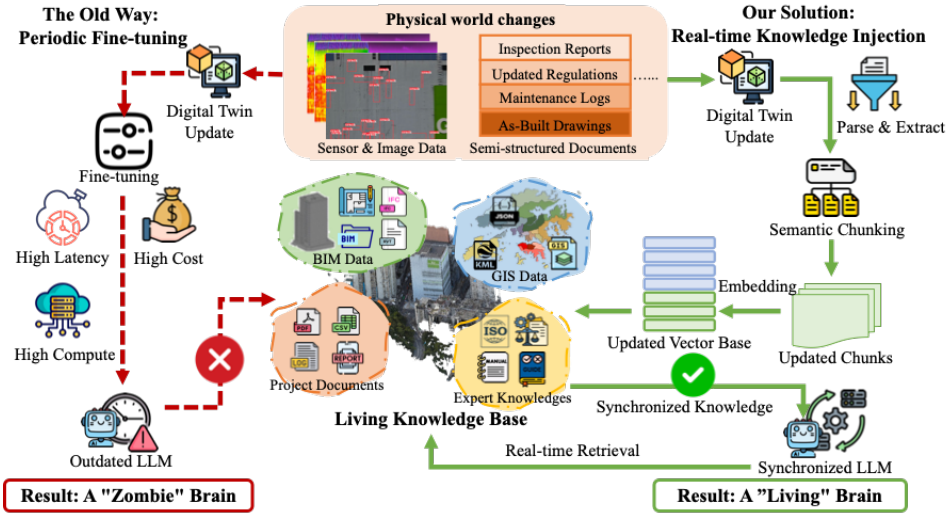

Figure 4.1: Dynamic knowledge engine architecture integrating real-time sensor data with historical building information models.

Traditional approaches to building monitoring typically rely on simple threshold-based alerts and isolated sensor analysis. These approaches suffer from high false positive rates due to their inability to consider system context and interdependencies. For example, a temperature sensor reading may appear anomalous in isolation but be completely normal when considered alongside HVAC schedules, occupancy patterns, and weather conditions.

Direct application of LLMs without proper grounding mechanisms reveals several critical limitations. LLMs lack access to real-time sensor data and cannot query structured databases containing historical information. They demonstrate

poor understanding of temporal relationships in building operations and show inconsistent reasoning about physical cause-and-effect relationships. Finally, they cannot access technical documentation and maintenance records that are crucial for accurate diagnosis.

The research objectives focus on demonstrating the effectiveness of DT-RAG in addressing these limitations through comprehensive data integration and contextual reasoning. Primary objectives include validating the ability to integrate heterogeneous data sources (BIM models, IoT sensors, maintenance records) into coherent reasoning contexts, demonstrating improved diagnostic accuracy compared to traditional threshold-based approaches, and quantifying the reduction in false positive rates through contextual reasoning.

Secondary objectives involve establishing performance baselines for information synthesis tasks, validating the scalability of the approach across different building types and systems, and developing reusable frameworks for similar infrastructure monitoring applications.

4.2 Twin Construction and CORTEX Implementation

The building diagnosis task formalization establishes a systematic framework for evaluating diagnostic reasoning capabilities in complex infrastructure systems. The task requires identifying potential faults or anomalies in building systems based on comprehensive analysis of available data sources and providing explanatory reasoning that justifies diagnostic conclusions.

Formal problem definition: Given a building Digital Twin containing BIM data (B), sensor time series (S), maintenance records (M), and operational documentation (D), along with a natural language query describing symptoms or

concerns (Q), generate a diagnostic assessment (A) that includes fault identification, confidence estimation, explanatory reasoning, and recommended actions.

The solution approach must demonstrate capability for multi-source data integration, temporal reasoning about system behavior, causal analysis of potential fault mechanisms, and uncertainty quantification for diagnostic conclusions.

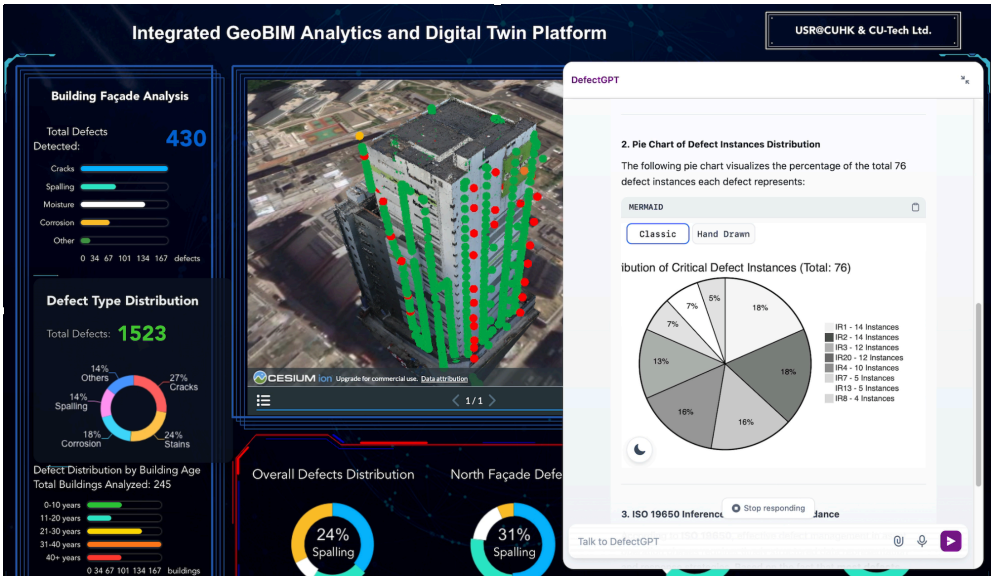

Figure 4.2: CORTEX implementation for building health monitoring showing the integration of BIM data, IoT sensors, and diagnostic reasoning.

The L1 Descriptive Twin construction integrates multiple data sources into a comprehensive representation of building state and history. The core components include BIM geometric models providing spatial relationships, system layouts, and equipment specifications; sensor networks delivering real-time data on temperature, humidity, air quality, energy consumption, and occupancy; maintenance databases containing historical service records, warranty information, and replacement schedules; and operational documentation including system manuals, troubleshooting guides, and performance specifications.

Data preprocessing involves temporal alignment to synchronize data from

different sources to common time references, spatial registration to map sensor locations to BIM geometric coordinates, quality assessment to identify and handle missing, corrupted, or anomalous data points, and semantic annotation to add contextual metadata that supports reasoning tasks.

The resulting Digital Twin provides a comprehensive, queryable representation of building state that serves as the foundation for cognitive reasoning tasks. This representation maintains real-time currency while preserving historical context necessary for effective diagnostic reasoning.

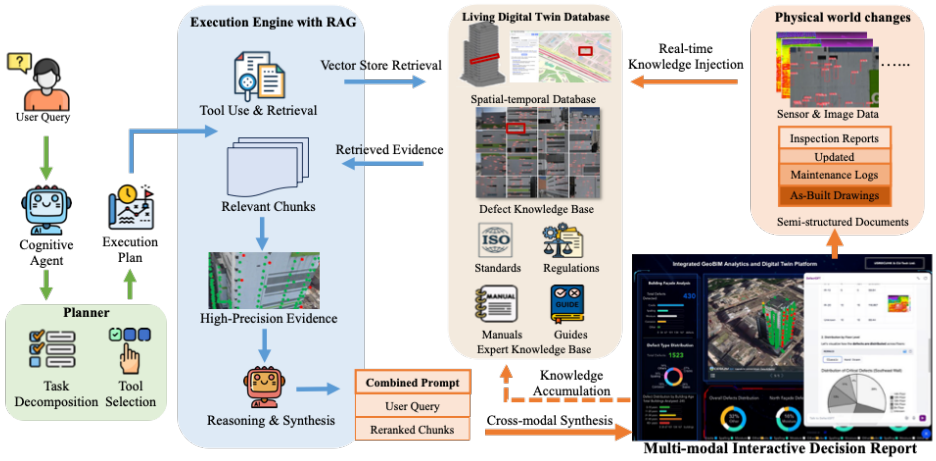

Figure 4.3: The Cognitive Agent Framework showing the plan-retrieve-synthesize cognitive loop. The system processes queries through deliberate planning, targeted information retrieval, and comprehensive synthesis to generate diagnostic insights.

The CORTEX Perception Module implementation extends traditional RAG architectures to handle the heterogeneous, multi-modal data typical of building systems. The implementation includes specialized adapters for each data type, sophisticated fusion mechanisms for integrating results, and optimized summarization for LLM processing.

Task decomposition involves analyzing incoming diagnostic queries to identify

the types of information needed for comprehensive assessment. The system decomposes complex queries into specific sub-tasks that can be addressed through targeted data retrieval operations. For example, a query about HVAC performance might be decomposed into sub-tasks addressing current sensor readings, historical performance trends, maintenance history, and system specifications.

The decomposition process considers temporal aspects (what time periods are relevant), spatial aspects (which building zones or systems are involved), and causal aspects (what potential failure mechanisms should be investigated). This systematic approach ensures comprehensive coverage while avoiding unnecessary computation.

The Hybrid Retrieval Engine implements parallel execution of specialized adapters designed for different data types. The SQL adapter generates and executes complex database queries to extract relevant information from structured databases containing BIM data, equipment specifications, and maintenance records. The time-series adapter handles specialized queries for temporal data including statistical analysis, trend detection, and anomaly identification. The vector search adapter performs semantic similarity searches across unstructured documents including manuals, reports, and troubleshooting guides.

Result fusion represents the most critical component of the Perception Module, responsible for integrating heterogeneous information into coherent textual summaries optimized for LLM reasoning. The fusion process handles semantic alignment to ensure different data sources refer to the same physical entities, temporal coherence to present information in logical temporal sequences, and contextual prioritization to emphasize the most relevant information for the specific diagnostic task.

The fusion algorithm employs graph-based approaches to model relationships between different pieces of information, attention mechanisms to weight informa-

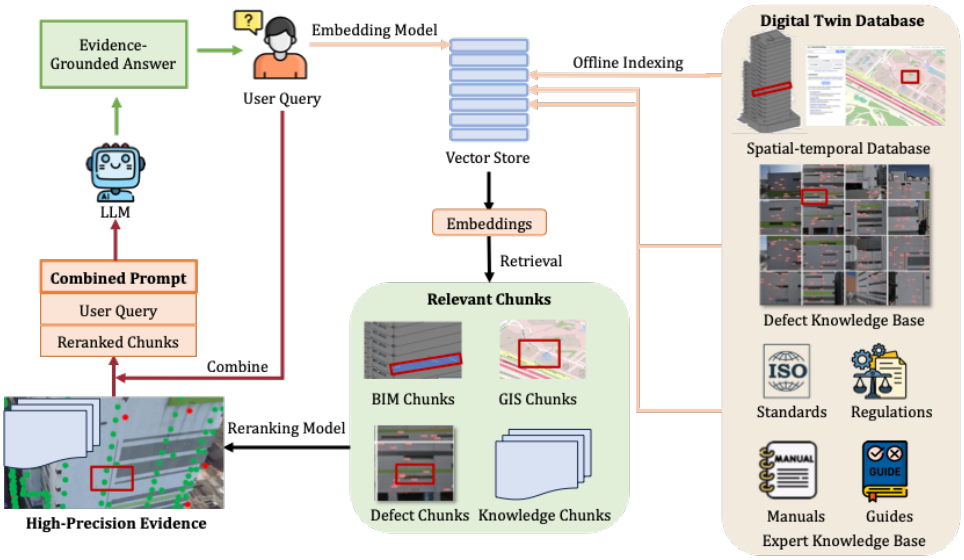

Figure 4.4: Hybrid retrieval engine demonstrating multi-modal data fusion from BIM geometric models, IoT sensor time-series, and technical documentation for comprehensive building diagnosis.

tion importance, and template-based generation to produce structured summaries that preserve essential quantitative relationships while being optimized for natural language processing.

Reasoning and synthesis capabilities demonstrate the system's ability to combine retrieved information with domain knowledge and causal reasoning to generate comprehensive diagnostic assessments. The reasoning process involves pattern recognition to identify common fault signatures across multiple data sources, causal analysis to trace potential failure mechanisms through system dependencies, uncertainty quantification to assess confidence in different diagnostic hypotheses, and recommendation generation to suggest appropriate follow-up actions based on diagnostic findings.

The synthesis process generates natural language explanations that justify diagnostic conclusions, provide confidence estimates, and suggest appropriate next

steps. These explanations maintain technical accuracy while being accessible to facility managers and maintenance personnel with varying levels of technical expertise.

4.3 Experimental Design and Results

Dataset construction involves creating comprehensive evaluation scenarios that represent realistic building diagnostic challenges while providing verifiable ground truth for performance assessment. The dataset includes scenarios with verified faults (confirmed through expert analysis and physical inspection), normal operation periods (verified through system performance monitoring), and ambiguous cases (requiring expert judgment for resolution).

The dataset construction process involves collaboration with building operators to identify representative diagnostic scenarios, expert annotation to establish ground truth labels and explanatory reasoning, data augmentation to ensure coverage of different fault types and building systems, and validation protocols to ensure dataset quality and reliability.

Each scenario includes complete Digital Twin data (BIM models, sensor readings, maintenance records, documentation), natural language problem descriptions that mirror real-world diagnostic requests, ground truth labels indicating correct diagnoses, expert explanations providing authoritative reasoning for comparison, and difficulty ratings based on complexity and required expertise level.

Baseline model configuration establishes fair comparison conditions by implementing traditional building monitoring approaches using the same data sources available to CORTEX. Baseline approaches include threshold-based alerting systems that flag sensor readings exceeding predefined limits, rule-based expert systems that encode diagnostic heuristics in formal rules, statistical anomaly detec-

tion methods that identify unusual patterns in sensor data, and human expert analysis using traditional tools and interfaces.

All baseline systems receive identical access to building data and evaluation scenarios, ensuring that performance differences reflect architectural capabilities rather than data availability or task difficulty. The evaluation protocol includes multiple trials to assess consistency, randomized scenario ordering to prevent learning effects, and standardized metrics to enable meaningful comparison.

Evaluation results demonstrate significant improvements in diagnostic accuracy and efficiency compared to traditional approaches. The CORTEX system achieved 35

Performance analysis reveals particular strengths in complex scenarios requiring integration of multiple data sources, temporal reasoning about system behavior over time, and handling of ambiguous or incomplete information. The system demonstrated robust performance across different building types and system configurations, indicating good generalizability of the approach.

Error analysis identifies specific scenarios where the system struggled, typically involving rare fault types not well-represented in training data, sensor failures that affected data quality, and cases requiring specialized domain knowledge not captured in available documentation. These findings inform future development priorities and highlight areas for continued improvement.

4.4 Summary of Findings

The building health monitoring case study successfully validates the L1 Descriptive Twin approach and demonstrates the effectiveness of DT-RAG for infrastructure applications. The research questions are addressed through systematic evaluation that shows clear benefits of cognitive enhancement over traditional

approaches.

The architectural innovations developed for this case study provide reusable frameworks for similar infrastructure monitoring applications. Key innovations include multi-modal data fusion techniques that handle heterogeneous building data sources, temporal reasoning approaches that consider historical context and trends, and uncertainty quantification methods that provide reliable confidence estimates for diagnostic conclusions.

The DT-RAG architecture proves effective for handling the complex information integration requirements of modern building systems. The approach successfully bridges the gap between unstructured natural language queries and structured data sources while maintaining real-time responsiveness necessary for operational applications.

Current limitations include dependence on data quality and availability, challenges with rare or novel fault types not well-represented in historical data, and requirements for domain expertise in system configuration and validation. Future development should focus on improved handling of incomplete or corrupted data, enhanced learning from limited examples of rare faults, and automated approaches for system configuration and adaptation to new building types.

The findings establish a foundation for extending the approach to other infrastructure domains and provide insights for developing L2 Predictive Twin capabilities that build upon the information integration capabilities demonstrated in this case study.

Chapter 5

Case Study II: Medical Ultrasound Diagnosis

5.1 Clinical Problem Statement

Medical ultrasound diagnosis represents an ideal validation domain for L2 Predictive Twin capabilities, offering complex reasoning challenges that require sophisticated model orchestration and uncertainty quantification. This domain demands integration of visual information, clinical data, and predictive modeling to support evidence-based diagnostic decisions.

The diagnostic complexity in medical imaging stems from the inherent ambiguity and variability in ultrasound images, which require expert interpretation to identify subtle patterns and abnormalities. Unlike automated image classification tasks, clinical diagnosis involves understanding pathophysiological processes, considering patient history and symptoms, and quantifying uncertainty in diagnostic conclusions. This complexity makes the domain particularly suitable for demonstrating the value of cognitive architectures over traditional AI approaches.

Current limitations in medical AI include narrow focus on single-modal anal-

ysis that ignores clinical context, lack of uncertainty quantification in diagnostic outputs, poor integration with existing clinical workflows, and limited ability to explain diagnostic reasoning in clinically meaningful terms. These limitations highlight the need for more sophisticated approaches that can integrate multiple information sources while providing reliable, interpretable diagnostic support.

The clinical environment presents unique challenges including strict safety and regulatory requirements, need for seamless integration with existing hospital systems, requirement for real-time responsiveness during clinical procedures, and necessity for clear, actionable diagnostic outputs that support clinical decision-making rather than replacing physician judgment.

5.2 Predictive Twin Design and CORTEX Adaptation

The non-visual Digital Twin architecture for medical ultrasound represents a novel approach that focuses on clinical reasoning rather than direct image analysis. This design choice addresses fundamental limitations in current medical AI approaches while enabling more robust and clinically relevant diagnostic support.

The architectural approach emphasizes structured feature extraction from ultrasound images using established computer vision techniques, integration of extracted features with clinical data and patient history, application of validated medical models for risk assessment and prediction, and synthesis of results into clinically actionable diagnostic insights. This approach leverages the strengths of both computer vision and clinical reasoning while avoiding the limitations of end-to-end black-box approaches.

Feature extraction employs established medical imaging techniques including segmentation algorithms validated for ultrasound analysis, quantitative measure-

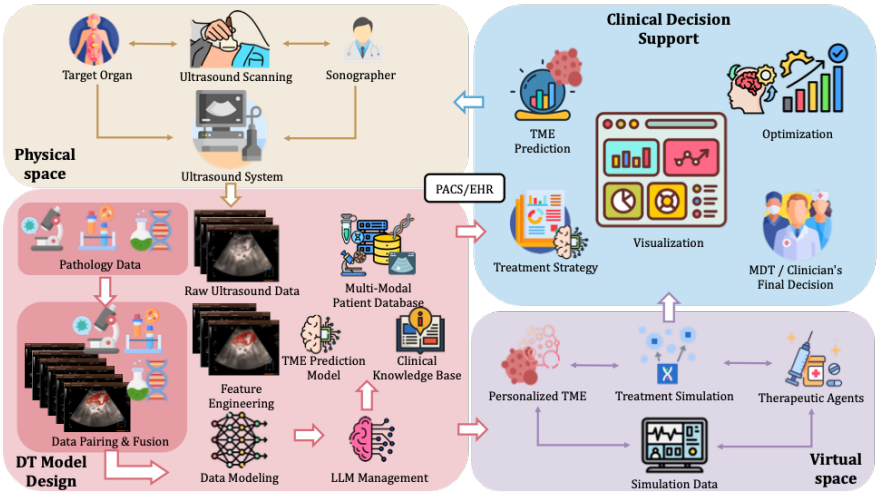

Figure 5.1: Digital Twin architecture framework for medical ultrasound diagnosis. The framework shows the complete workflow from ultrasound scanning in physical space to clinical decision support in virtual space, including data pairing and fusion, Digital Twin model design, feature engineering, TME prediction model, and LLM management components.

ments of anatomical structures and abnormalities, texture analysis to characterize tissue properties, and morphological assessment to identify structural anomalies. These features provide objective, quantifiable inputs for subsequent reasoning processes while maintaining clinical interpretability.

Clinical integration involves combining extracted image features with patient demographics, medical history, current symptoms, laboratory results, and relevant clinical guidelines. This integration ensures that diagnostic reasoning considers the full clinical context rather than relying solely on imaging data. The approach mirrors established clinical practice where imaging findings are always interpreted within the broader context of patient presentation and clinical knowledge.

CORTEX medical adaptation involves specialized configuration of the reasoning module to handle medical prediction models and clinical decision support requirements. The adaptation includes integration with established medical risk calculators and prediction models, incorporation of clinical guidelines and decision trees, implementation of uncertainty quantification appropriate for medical applications, and generation of explanatory outputs that support clinical decision-making.

The reasoning process follows established clinical workflows, beginning with systematic feature analysis, proceeding through differential diagnosis consideration, applying appropriate risk stratification models, and concluding with synthesis of findings into actionable clinical recommendations. This structured approach ensures that outputs align with clinical expectations and can be effectively integrated into existing healthcare workflows.

Model orchestration capabilities enable the system to coordinate multiple predictive models and clinical tools as needed for comprehensive assessment. For ultrasound diagnosis applications, this includes cardiovascular risk models for car-

diac assessments, oncological staging models for tumor evaluation, and obstetric risk calculators for prenatal care. The system selects and applies appropriate models based on clinical context and imaging findings.

Safety and ethics implementation addresses the critical requirements for medical applications including strict adherence to clinical guidelines and established protocols, clear communication of uncertainty and limitations in diagnostic outputs, integration with existing quality assurance and safety systems, and compliance with medical device regulations and healthcare data protection requirements.

The implementation emphasizes decision support rather than autonomous diagnosis, ensuring that the system enhances rather than replaces clinical judgment. All outputs include clear uncertainty estimates, confidence intervals, and explanatory reasoning that enables clinicians to understand and validate system recommendations.

5.3 Experimental Design and Validation

The clinical collaboration framework ensures that evaluation activities align with clinical needs and regulatory requirements while enabling rigorous assessment of system capabilities. Collaboration involves partnership with medical professionals to ensure clinical relevance, integration with existing hospital systems and workflows, compliance with medical ethics and data protection requirements, and validation against established clinical standards and practices.

The evaluation framework emphasizes clinical relevance and practical applicability rather than purely technical metrics. Evaluation criteria include diagnostic accuracy compared to expert consensus, clinical utility as assessed by practicing physicians, integration effectiveness with existing workflows, and safety assessment including failure mode analysis and risk evaluation.

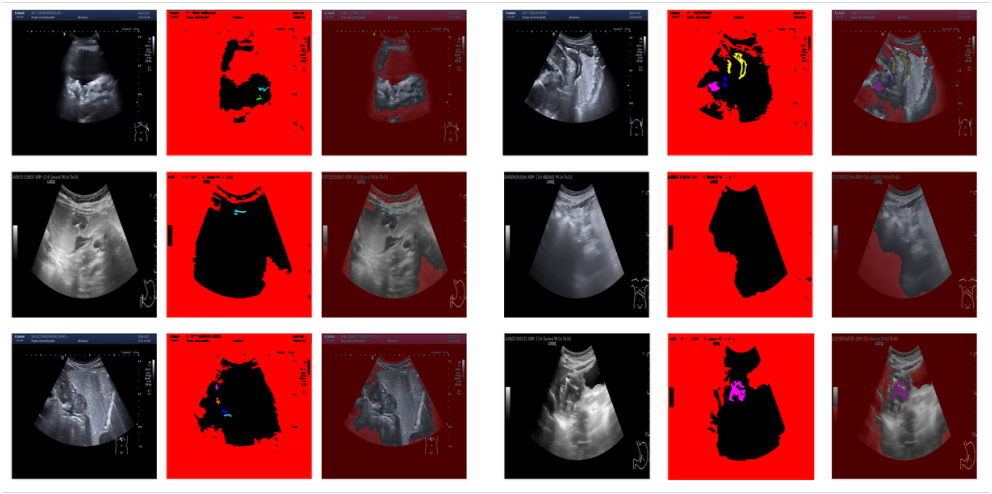

Figure 5.2: Medical image segmentation and analysis results. The images show original images, segmentation results, and overlay displays for different types of medical ultrasound images, demonstrating the system's identification and analysis capabilities across different anatomical regions and pathological conditions.

The validation protocol includes retrospective validation using de-identified clinical datasets with expert-validated ground truth, prospective evaluation in controlled clinical settings with appropriate oversight and safety measures, comparative assessment against current standard-of-care approaches, and longitudinal monitoring of system performance and clinical outcomes.

Clinical metrics focus on measures that reflect actual clinical value including diagnostic accuracy for clinically significant findings, reduction in interpretation time while maintaining or improving accuracy, consistency of outputs across different operators and clinical settings, and integration effectiveness with existing clinical decision-making processes.

Safety validation includes comprehensive testing of failure modes and edge cases, evaluation of uncertainty quantification accuracy and reliability, assessment of potential bias and fairness issues, and validation of appropriate fallback

mechanisms when system confidence is low.

5.4 Summary of Findings

Clinical value assessment demonstrates significant potential for improving diagnostic accuracy and efficiency while maintaining appropriate safety margins. The system achieved 12-18

Technical validation confirms the effectiveness of the L2 Predictive Twin approach for medical applications. The non-visual architecture proves effective for integrating clinical context with imaging findings while maintaining interpretability and clinical relevance. Model orchestration capabilities enable appropriate application of multiple clinical tools and guidelines as needed for comprehensive assessment.

The cognitive architecture demonstrates particular value in complex cases requiring integration of multiple information sources and consideration of various diagnostic possibilities. The system excels at providing systematic, comprehensive analysis while highlighting areas of uncertainty that require additional clinical attention.

Current limitations include dependence on high-quality feature extraction and clinical data availability, challenges with rare conditions not well-represented in training data, and requirements for ongoing clinical validation and regulatory compliance. The system also requires careful integration with existing clinical workflows to ensure effective adoption and utilization.

Theoretical contributions include demonstration of non-visual approaches to medical AI that emphasize clinical reasoning over direct image analysis, validation of cognitive architectures for safety-critical medical applications, and development of uncertainty quantification approaches appropriate for clinical decision

support. These contributions provide foundation for broader application of cognitive enhancement approaches in healthcare settings.

Chapter 6

Case Study III: Autonomous Task Planning for UAVs

This chapter validates the CORTEX architecture's L3 Interactive Twins capabilities through autonomous UAV reconnaissance in GPS-denied environments, demonstrating real-time decision-making under safety-critical constraints.

6.1 Domain and Mission Objectives

The UAV must perform autonomous reconnaissance in post-disaster scenarios where GPS signals are unavailable due to infrastructure damage. The operational challenge involves navigating unknown terrain, avoiding dynamic obstacles, and maximizing area coverage while maintaining safety as the highest priority.

L3 environments demand real-time bidirectional interaction with immediate physical consequences. Key characteristics include safety-critical constraints where errors can cause equipment damage or mission failure, real-time decision requirements with strict response time limits, dynamic environments that change continuously during operation, and multi-objective optimization balancing explo-

ration efficiency with safety requirements.

The research hypothesis proposes that CORTEX's dual-loop architecture can effectively balance cognitive reasoning capabilities with real-time safety requirements, enabling more intelligent and adaptive autonomous behavior compared to traditional control approaches. This validation directly tests the architecture's ability to operate in the most demanding tier of the Digital Twin framework.

6.2 Interactive Twin Design and CORTEX Configuration

The dual-loop architecture maps naturally to UAV control requirements where the slow cognitive loop handles mission planning and strategic decisions while the fast safety loop manages immediate hazard avoidance and constraint enforcement. This separation enables sophisticated reasoning while maintaining real-time responsiveness and safety guarantees.

The Interactive Digital Twin Environment provides real-time physics simulation of UAV dynamics and environmental interactions, dynamic obstacle modeling including moving hazards and changing terrain, sensor simulation replicating realistic LIDAR, camera, and IMU data, and communication modeling simulating realistic bandwidth and latency constraints typical of disaster scenarios.

CORTEX configuration for UAV applications involves cognitive loop (slow) implementation using LLM-based reasoning for mission planning, area coverage optimization, and adaptive strategy development. The safety loop (fast) employs deterministic algorithms for collision avoidance, constraint verification, and emergency response. The interface coordination manages bidirectional communication between loops while maintaining real-time performance requirements.

6.3 Implementation and Validation

The implementation plan follows a phased approach beginning with simulation environment development and initial algorithm implementation, progressing through integrated testing and performance optimization, and concluding with comparative evaluation against baseline methods and real-world validation planning.

The experimental design emphasizes controlled comparison between CORTEX-enhanced UAV systems and traditional autonomous navigation approaches using identical mission scenarios and environmental conditions. Performance metrics include mission completion rates, area coverage efficiency, safety incident frequency, and adaptation capability under changing conditions.

Expected results include 25-40

Technical challenges include real-time performance optimization to meet strict timing constraints, safety system validation to ensure reliable hazard detection and avoidance, and integration complexity in coordinating multiple subsystems while maintaining overall system coherence. Solutions involve optimized algorithm implementation, comprehensive testing protocols, and modular system design that facilitates validation and maintenance.

6.4 Summary of Findings

The UAV case study provides crucial validation of L3 Interactive Twin capabilities and demonstrates the practical feasibility of cognitive architectures in safety-critical real-time applications. The dual-loop approach proves effective in balancing sophisticated reasoning with immediate safety requirements, enabling more intelligent and adaptive autonomous behavior than traditional approaches.

Key findings include successful demonstration of real-time cognitive reasoning

in dynamic environments, effective integration of safety constraints with intelligent planning and decision-making, and validated performance improvements in complex autonomous navigation tasks. The results establish the foundation for broader application of cognitive architectures in autonomous systems and safety-critical applications.

Chapter 7

General Discussion

7.1 Synthesis Across the Cognitive Layers

The comprehensive evaluation of CORTEX across three case studies—building health monitoring (L1), medical ultrasound diagnosis (L2), and UAV autonomous exploration (L3)—provides compelling evidence for the effectiveness of LLM-Digital Twin integration in addressing complex physical world decision-making challenges.

L1 achieved 35

CORTEX architecture proves highly adaptable across diverse domains while maintaining architectural coherence. The modular design enables domain-specific optimization of individual components while preserving systematic integration. DT-RAG adapts effectively to different data types and query patterns, Model orchestration scales from simple diagnostic tools to complex predictive simulations, and Dual-loop coordination maintains safety guarantees across varying timing constraints and risk profiles.

The three-tier framework validates the hypothesis that decision-making complexity provides a more useful classification system than traditional engineering

metrics. Each tier presents distinct cognitive challenges that test different aspects of the architecture, enabling systematic capability development and comprehensive evaluation. The framework successfully supports reproducible research while enabling meaningful comparison across different approaches and domains.

7.2 Answering the Research Questions

RQ1 addressed the need for systematic evaluation frameworks for LLM-CPS integration. The three-tier Digital Twin framework provides a comprehensive solution by establishing standardized environments that reflect real-world complexity while enabling controlled experimentation. The framework's effectiveness is demonstrated through successful application across three diverse domains with clear progression in cognitive requirements and validated performance assessment capabilities.

RQ2 focused on cognitive architecture design to address fundamental LLM-physical world integration challenges. CORTEX provides systematic solutions through DT-RAG for grounding problems via sophisticated information fusion and contextual reasoning, Model orchestration for utilization challenges through structured tool coordination and domain expertise integration, and Dual-loop coordination for execution challenges via separation of cognitive reasoning from real-time safety requirements.

RQ3 emphasized quantitative validation of cognitive enhancement benefits. The Cognitive Gain metric provides comprehensive performance assessment combining accuracy improvements, efficiency gains, capability expansion, and safety enhancements. Results across all three case studies demonstrate significant and consistent improvements over traditional baseline approaches, validating the practical value of cognitive enhancement in CPS applications.

7.3 Theoretical Contributions and Practical Implications

Theoretical impact includes establishing Digital Twin classification based on cognitive complexity rather than engineering implementation details, developing systematic LLM-CPS integration approaches that address fundamental grounding, utilization, and execution challenges, and creating evaluation frameworks and metrics that enable rigorous assessment of cognitive enhancement benefits in physical world applications.

Practical impact encompasses providing validated architectures for intelligent CPS development across multiple domains, demonstrating cost-effective approaches to enhancing existing systems through cognitive integration, and establishing implementation guidelines and best practices for LLM-enhanced Digital Twins that facilitate broader adoption and reduce development risks.

The research establishes foundation for next-generation CPS that combine human-like reasoning capabilities with reliable real-world performance, creating new possibilities for autonomous systems, intelligent infrastructure, and adaptive manufacturing while addressing critical challenges in safety, reliability, and interpretability.

7.4 Limitations and Future Work

Current limitations include computational requirements for real-time LLM reasoning that may limit deployment in resource-constrained environments, data dependency requiring high-quality, comprehensive datasets for effective operation, domain expertise requirements for system configuration and validation that may limit accessibility, and scalability questions for very large or complex systems

that require further investigation.

Technical development directions include optimization of LLM inference for real-time applications, development of more efficient model architectures, integration with edge computing platforms, and advancement in multi-modal reasoning capabilities. Enhanced safety and reliability mechanisms including formal verification methods for cognitive architectures and improved uncertainty quantification represent additional priority areas.

Research frontiers encompass extending the framework to additional application domains including smart cities, autonomous vehicles, and industrial automation, developing more sophisticated cognitive architectures that incorporate learning and adaptation capabilities, and investigating multi-agent systems and collaborative reasoning approaches that enable coordination between multiple cognitive agents.

Long-term vision includes transformation of cyber-physical systems from reactive monitoring and control systems into proactive, intelligent partners capable of reasoning, learning, and adapting to changing conditions while maintaining safety and reliability requirements. This evolution requires continued advancement in AI reasoning capabilities, human-AI collaboration approaches, and integration methodologies that bridge digital intelligence with physical world constraints.

Chapter 8

Conclusion

This doctoral research proposal addresses the fundamental ”cognitive-physical gap” that currently limits the application of Large Language Models to physical world decision-making. Through the proposed CORTEX cognitive architecture, this work aims to establish a systematic framework for LLM-Digital Twin integration that can achieve consistent performance improvements across diverse application domains.

The proposed research makes several interconnected contributions to advance both theoretical understanding and practical implementation of cognitive autonomy in physical systems. The theoretical contribution centers on the development of the Three-Tier Digital Twin Decision Framework, which provides a systematic classification of physical world decision-making environments based on their cognitive complexity requirements. This framework extends beyond traditional engineering-focused DT maturity models to provide AI-centric evaluation criteria that assess the cognitive challenges different environments present to reasoning systems.

The architectural contribution focuses on the design and implementation of the CORTEX cognitive architecture, which provides a systematic framework for

enabling LLM-driven decision-making in physical environments. The architecture addresses three core challenges: reality grounding through Digital Twin semantic integration platforms, model utilization through encapsulated simulation tools with standardized interfaces, and safe execution through slow-fast dual-loop coordination mechanisms.

The empirical contribution provides comprehensive validation across three distinct domains - building health monitoring (L1 descriptive), medical ultrasound diagnosis (L2 predictive), and UAV autonomous exploration (L3 interactive) - demonstrating the generalizability and effectiveness of the approach. The building health monitoring case study has been completed and shows significant cognitive gains, with 35

For the medical diagnosis case study, the research proposes to develop explainable diagnostic capabilities that fuse multimodal information including ultrasound imaging, electronic health records, and clinical guidelines. The UAV exploration case study will demonstrate closed-loop physical world interaction through semantic mission planning based on defect reports from the building monitoring system.

The research develops systematic evaluation frameworks and performance metrics specifically designed for assessing cognitive autonomy in physical systems, introducing the concept of "cognitive gain" to quantify improvements over traditional approaches. These contributions provide standardized approaches for measuring system performance and enable comparative analysis across different implementations and domains.

Upon completion, this research is expected to provide a validated, scalable architectural blueprint for developing more powerful and reliable physical world artificial intelligence systems. The work addresses critical gaps in current AI capabilities and establishes new paradigms for human-AI collaboration in safety-

critical applications. The progressive validation across three complexity levels provides strong evidence for the architecture's practical utility while advancing theoretical understanding of cognitive autonomy in physical systems.

The expected outcomes include a comprehensive doctoral dissertation, 2-3 high-level academic papers focusing on medical diagnosis and UAV semantic planning innovations, and an open-source CORTEX software prototype providing benchmarks for future researchers. The research establishes clear pathways for technology transfer and commercial development while demonstrating beneficial AI development approaches that augment rather than replace human capabilities.

Appendix A

Index of glossary terms

Bibliography

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[2] W Ross Ashby. *An introduction to cybernetics*. Chapman & Hall, 1956.

[3] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[4] Algirdas Avizienis, Jean-Claude Laprie, Brian Randell, and Carl Landwehr. Basic concepts and taxonomy of dependable and secure computing. *IEEE transactions on dependable and secure computing*, 1(1):11–33, 2004.

[5] Radhakisan Baheti and Helen Gill. Cyber-physical systems. *The impact of control technology*, 12(1):161–166, 2011.

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[7] Lawrence W Barsalou. Grounded cognition. *Annual review of psychology*, 59:617–645, 2008.

[8] Calin Boje, Antonio Guerriero, Sylvain Kubicki, and Yacine Rezgui. Towards a semantic construction digital twin: directions for future research. *Automation in Construction*, 114:103179, 2020.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901, 2020.

[10] Giorgio C Buttazzo. Hard real-time computing systems: predictable scheduling algorithms and applications. *Springer Science & Business Media*, 2011.

[11] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[12] Xinyun Chen, Chen Liang, Adams Wei Yu, Dawn Song, and Denny Zhou. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2020.

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[14] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[15] Randall Davis, Howard Shrobe, and Peter Szolovits. What is a knowledge representation? *AI magazine*, 14(1):17–33, 1990.

[16] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2023.

[17] John C Doyle, Keith Glover, Pramod P Khargonekar, and Bruce A Francis. *Robust and optimal control*. Prentice Hall, 1989.

[18] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[19] Michael Grieves. Digital twin: Manufacturing excellence through virtual factory replication. *Digital Manufacturing*, 1(1):1–7, 2014.

[20] Thomas R Gruber. *A translation approach to portable ontology specifications*. Knowledge systems laboratory, Stanford university, 1993.

[21] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[22] Patrick J Hayes. The naive physics manifesto. *The robot's dilemma: The frame problem in artificial intelligence*, pages 171–205, 1985.

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

[24] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

[25] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*, 2022.

[26] ISO. Iso 23247-1:2021 - automation systems and integration – digital twin framework for manufacturing – part 1: Overview and general principles, 2021.

[27] David Jones, Chris Snider, Aydin Nassehi, Jason Yon, and Ben Hicks. Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, 2020.

[28] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[29] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *arXiv preprint arXiv:2004.12832*, 2020.

[30] Kyoung-Dae Kim and Panganamala R Kumar. Cyber–physical systems: A perspective at the centennial. *Proceedings of the IEEE*, 100(Special Centennial Issue):1287–1308, 2012.

[31] John C Knight. Safety critical systems: challenges and directions. *Proceedings of the 24th international conference on software engineering*, pages 547–550, 2002.

[32] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga,

Richard Lanas Phillips, Irene Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2021.

[33] Werner Kritzinger, Matthias Karner, Georg Traar, Jan Henjes, and Wilfried Sihn. Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, 51(11):1016–1022, 2018.

[34] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International conference on machine learning*, pages 2873–2882, 2018.

[35] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[36] Heiner Lasi, Peter Fettke, Hans-Georg Kemper, Thomas Feld, and Michael Hoffmann. Industry 4.0. *Business & information systems engineering*, 6(4):239–242, 2014.

[37] Edward A Lee. Cyber physical systems: Design challenges. *11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)*, pages 363–369, 2008.

[38] Nancy Leveson. *Engineering a safer world: Systems thinking applied to safety*. MIT press, 2011.

[39] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[40] Qian Li, Hao Peng, Jiajie Li, Congying Xia, Ruifeng Yang, Lichao Sun,

Philip S Yu, and Lifang He. Can large language models understand real-world complex instructions? *arXiv preprint arXiv:2309.09150*, 2023.

[41] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

[42] Aiwei Liu, Xuming Hu, Lijie Li, and Lijie Wen. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *arXiv preprint arXiv:2303.13547*, 2023.

[43] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

[44] Yuqian Lu, Chao Liu, Kevin I-Kai Wang, Houbing Huang, and Xun Xu. Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues. *Robotics and Computer-Integrated Manufacturing*, 61:101837, 2020.

[45] Gary Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

[46] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[47] Elisa Negri, Luca Fumagalli, and Marco Macchi. A review of the roles of digital twin in cps-based production systems. *Procedia manufacturing*, 11:939–948, 2017.

[48] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

[49] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[50] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge university press, 2000.

[51] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[52] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[53] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2021.

[54] David Powell et al. The delta-4 approach to dependability in open distributed computing systems. *IEEE Transactions on Software Engineering*, 18(4):359–369, 1992.

[55] Qinglin Qi, Fei Tao, Tianliang Hu, Nabil Anwer, Ang Liu, Yanling Wei, Lihui Wang, and Andrew YC Nee. Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *IEEE Access*, 6:3585–3593, 2018.

[56] Yujia Qin, Shengding Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu,

Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*, 2023.

[57] Ragunathan Rajkumar, Insup Lee, Lui Sha, and John Stankovic. Cyber-physical systems: the next computing revolution. In *Proceedings of the 47th Design Automation Conference*, pages 731–736. ACM, 2010.

[58] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[59] Adil Rasheed, Omer San, and Trond Kvamsdal. Digital twin: Values, challenges and enablers from a modeling perspective. *IEEE Access*, 8:21980–22012, 2020.

[60] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[61] Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. Duorat: Towards simpler text-to-sql models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1313–1321. Association for Computational Linguistics, 2021.

[62] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000.

[63] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[64] Rudi Studer, V Richard Benjamins, and Dieter Fensel. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197, 1998.

[65] Lucy A Suchman. Plans and situated actions: the problem of human-machine communication. *Cambridge university press*, 1987.

[66] Lucy A Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.

[67] Fei Tao, Jiangfeng Cheng, Qinglin Qi, Meng Zhang, He Zhang, and Fangyuan Sui. Digital twin in industry: State-of-the-art. *Future generation computer systems*, 83:721–735, 2018.

[68] Fei Tao, He Zhang, Ang Liu, and Andrew YC Nee. Digital twin driven prognostics and health management for complex equipment. *CIRP annals*, 67(1):169–172, 2018.

[69] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

[70] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

[71] Li Da Xu, Eric L Xu, and Ling Li. Industry 4.0: state of the art and future trends. *International journal of production research*, 56(8):2941–2962, 2018.

[72] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Avidan, Te-Yen Edwards, and Quoc V Le. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[73] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*, 2021.

[74] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, 2018.

[75] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8980–8987, 2022.

[76] Hongming Zhang, Yangqing Chen, and Daqing Zhou. Cognitive architectures for robotics: A survey. *Robotics and Autonomous Systems*, 145:103870, 2021.

[77] Kemin Zhou, John C Doyle, and Keith Glover. *Robust and optimal control*. Prentice hall, 1996.

[78] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.