LARGE LANGUAGE MODELS AS COGNITIVE COMPANIONS TO REINFORCEMENT LEARNING FOR TRUSTWORTHY POWER SYSTEM OPERATIONS

by

Shrenik Amol Jadhav

Master of Science in Data Science
Department of Computer and Information Science
University of Michigan–Dearborn

Thesis Committee:

Van-Hai Bui, Ph.D., Assistant Professor (Thesis Advisor)
Srijita Das, Ph.D., Assistant Professor
Guilherme Hollweg, Ph.D., Assistant Professor

A thesis presented in partial fulfillment of the requirements for the degree of Master of Science in Data Science, administered through the College of Engineering and Computer Science, University of Michigan–Dearborn, Winter 2026.

Shrenik Jadhav

shrenikj@umich.edu

ORCID ID: 0009-0003-6906-7465

© Shrenik Jadhav 2026

Dedication

*To my father, **Amol**, for your steady strength and for teaching me the value of hard work.*

*To my mother, **Shailaja**, for your endless love and for being my greatest cheerleader.*

*To my grandparents, **Dattatrya and Prabhavati**, whose prayers and blessings have always protected and guided me.*

To my friends, who shared the long nights and kept me grounded through the challenges of this journey.

And to my professors at the University of Michigan-Dearborn, whose mentorship has shaped both my research and my character.

List of Figures

1.1 The LLM as cognitive companion paradigm 8
1.2 Thesis roadmap 12

2.1 Thematic landscape of the reviewed literature 14

3.1 FairMarket-RL training architecture 38
3.2 FairMarket-RL inference architecture 39
3.3 Algorithm comparison: IPPO vs. baselines 39
3.4 Fairness curriculum comparison 40
3.5 FairMarket-RL learning dynamics 40
3.6 Fairness-efficiency Pareto frontier 41

4.1 Pilot community training dynamics 49
4.2 Scaled community results 50
4.3 Real-world community results 50
4.4 Sensitivity analysis 52

5.1 XRL-LLM framework overview 56
5.2 PPO training dynamics 62
5.3 Voltage magnitude profile before and after RL control 62
5.4 Radar chart comparing three explanation methods 63
5.5 Forest plot of effect sizes 64
5.6 Counterfactual faithfulness across load levels 65
5.7 LLM backbone comparison 67

List of Tables

1.1 Summary of thesis contributions 10

2.1 Mapping of research gaps to thesis contributions 28

3.1 Case study environment and training parameters 34
3.2 FairMarket-RL final evaluation results 35

4.1 Comparison of Chapters 3 and 4 53

5.1 Mapping of discrete actions to physical control operations 57
5.2 PPO training hyperparameters 58
5.3 Computational cost breakdown 61
5.4 RL agent voltage violation reduction 61
5.5 G-Eval scores across quality dimensions 63
5.6 Statistical significance analysis 64

Contents

List of Figures **I**

List of Tables **I**

Abstract **VIII**

1 Introduction **1**

1.1 Reinforcement Learning for Power System Operations 3

1.2 The Trust Gap in RL for Power Systems 4

1.2.1 Fairness in Decentralized Energy Markets 4

1.2.2 Scalability Under Realistic Operating Conditions 5

1.2.3 Explainability of RL Control Decisions 6

1.3 Large Language Models as Cognitive Companions to RL 6

1.4 Research Questions and Contributions 8

1.5 Thesis Organization 10

2 Literature Review **13**

2.1 Reinforcement Learning for Distribution System Voltage Control 14

2.2 Multi-Agent Reinforcement Learning for Peer-to-Peer Energy Trading 16

2.3 Fairness in Multi-Agent Systems and Energy Markets 18

2.3.1 Fairness Definitions and Metrics 18

2.3.2 Algorithmic Approaches to Fair MARL 18

2.3.3 Energy Equity and Fairness in P2P Markets 19

2.3.4 Value Alignment and Constitutional AI 19

2.4 Explainable AI for Sequential Decision Making 20

2.4.1 Feature Attribution Methods 20

2.4.2 Explainability for Reinforcement Learning 20

2.4.3 Counterfactual Explanations . . . 21
2.4.4 Evaluation of Explanation Quality . . . 22
2.5 Large Language Models as Decision-Making Companions . . . 22
2.5.1 LLMs for Reward Design and Shaping . . . 22
2.5.2 LLMs for Task Planning and Decision Support . . . 23
2.5.3 LLMs for Explanation Generation . . . 23
2.5.4 LLM Applications in Power and Energy Systems . . . 24
2.6 Trustworthy AI for Critical Infrastructure . . . 24
2.6.1 Regulatory Frameworks . . . 24
2.6.2 Trust in Automation . . . 25
2.6.3 The Deployment Gap for RL in Power Systems . . . 26
2.7 Synthesis: Research Gaps and Thesis Positioning . . . 26

3 FairMarket-RL: LLM-Driven Fairness Shaping in Multi-Agent Energy Trading 29
3.1 Introduction . . . 30
3.2 System Architecture . . . 31
3.3 Mathematical Modeling . . . 31
3.3.1 State Representation . . . 32
3.3.2 Action Spaces and Transitions . . . 32
3.3.3 Raw Reward Functions . . . 32
3.3.4 LLM-Based Fairness Shaping . . . 33
3.4 Learning Methodology and Experimental Results . . . 33
3.4.1 Algorithm Selection: Independent Proximal Policy Optimization . . . 33
3.4.2 Case Study Environment . . . 34
3.4.3 Fairness Integration via Curriculum Learning . . . 34
3.4.4 Learning Dynamics . . . 34
3.4.5 Key Performance Indicators . . . 35
3.4.6 Ablation Studies on Learning Dynamics . . . 35
3.4.7 Fairness-Efficiency Trade-off Analysis . . . 36
3.4.8 Robustness Under Adversarial and Noisy Conditions . . . 36
3.5 Potential Applications in Power and Energy Systems . . . 36
3.6 Summary . . . 37

4 Scalable Fairness Shaping with LLM-Guided Multi-Agent Reinforcement Learning for Peer-to-Peer Electricity Markets 42
4.1 Introduction . . . 43
4.2 Market Environment and Agent Modeling . . . 44
4.2.1 Market Structure . . . 44
4.2.2 Environment Dynamics . . . 44
4.2.3 Battery Energy Storage . . . 45
4.3 Multi-Agent Game Formulation . . . 45
4.3.1 State and Observation Spaces . . . 46
4.3.2 Action Spaces . . . 46
4.3.3 Transition and Clearing . . . 46
4.4 LLM-Based Fairness Shaping . . . 46
4.4.1 Fairness Metrics . . . 46
4.4.2 Shaped Reward Function . . . 47
4.4.3 Raw Reward Functions . . . 48
4.5 Training and Policy Architecture . . . 48
4.6 Experimental Validation . . . 48
4.6.1 Case 1: 30-Day Pilot Community . . . 49
4.6.2 Case 2: 90-Day Scaled Community . . . 49
4.6.3 Case 3: 90-Day Real-World Community . . . 49
4.6.4 Computational Scalability . . . 51
4.7 Sensitivity Analysis . . . 51
4.8 Comparison with Chapter 3 . . . 51
4.9 Summary . . . 51

5 XRL-LLM: Explainable Reinforcement Learning Framework for Voltage Control 54
5.1 Introduction . . . 55
5.2 Methodology . . . 56
5.2.1 Power System Environment . . . 57
5.2.2 Reinforcement Learning Agent . . . 58
5.2.3 SHAP Attribution Pipeline . . . 59
5.2.4 LLM Explanation Generation . . . 59
5.2.5 Counterfactual Verification . . . 60
5.2.6 Evaluation Framework . . . 60

5.2.7 Experimental Protocol . . . 60
5.3 Results and Discussion . . . 61
5.3.1 RL Agent Training and Voltage Control Performance . . . 61
5.3.2 Explanation Quality Assessment . . . 63
5.3.3 Human Validation . . . 64
5.3.4 Statistical Significance Analysis . . . 64
5.3.5 Counterfactual Verification . . . 65
5.3.6 Failure Case Analysis . . . 66
5.4 Ablation Studies . . . 66
5.4.1 SHAP Feature Budget . . . 66
5.4.2 LLM Backbone Comparison . . . 66
5.4.3 Prompt Component Decomposition . . . 66
5.4.4 Prompt Structure Ablation . . . 67
5.4.5 Temperature Sensitivity . . . 67
5.4.6 Comparison with Alternative Explainable RL Methods . . . 67
5.4.7 Robustness Under Network Perturbations . . . 68
5.5 Summary . . . 68

6 Synthesis, Conclusions, and Future Work 70
6.1 Answering the Research Questions . . . 70
6.2 Cross-Cutting Insights . . . 72
6.2.1 Prompt Engineering as a First-Class Design Concern . . . 72
6.2.2 The LLM Companion Paradigm Generalizes Across Interaction Modes 72
6.2.3 Fairness and Explainability Are Complementary Trust Dimensions . . 73
6.2.4 Computational Overhead Patterns . . . 73
6.3 Thesis-Level Limitations . . . 74
6.4 Future Research Directions . . . 75
6.4.1 Toward an Integrated LLM-RL Companion System . . . 75
6.4.2 Scaling to Utility-Grade Networks . . . 75
6.4.3 Dynamic Prompt Engineering . . . 76
6.4.4 Open-Source LLM Deployment . . . 76
6.4.5 Human Factors Studies . . . 76
6.4.6 Regulatory Alignment . . . 76
6.5 Concluding Remarks . . . 77

Acknowledgments **78**

References **79**

Abstract

Reinforcement learning (RL) agents have demonstrated strong performance in power distribution system tasks, including voltage regulation and peer-to-peer energy trading, yet their deployment in operational settings remains limited by three interconnected trust barriers: the inability to ensure fair outcomes in decentralized markets, the fragility of fairness mechanisms under realistic operating conditions, and the opacity of learned control policies to human operators. This thesis investigates a unifying approach to these barriers by positioning large language models (LLMs) as cognitive companions to RL agents, providing fairness evaluation, scalable reward shaping, and natural language explanation capabilities that the RL paradigm does not inherently offer.

Three studies are presented, each exploring a distinct mode of LLM-RL interaction within the power systems domain. The first study introduces FairMarket-RL, a framework in which an instruction-tuned LLM serves as a real-time fairness critic within a multi-agent RL training loop for peer-to-peer energy trading. The LLM evaluates each trading episode and returns normalized fairness scores that modulate agent rewards through a curriculum-based scheduling mechanism. In a controlled microgrid setting, Independent PPO agents achieve Fairness-To-Buyer and Fairness-Between-Sellers scores above 0.85, fulfill over 90% of buyer demand, and maintain balanced seller profit margins, substantially outperforming a baseline without LLM guidance where fairness stalls at 0.35–0.40.

The second study extends this framework to realistic market conditions, reformulating the environment as a continuous double auction with partial observability, discrete price-quantity actions, realistic residential load and photovoltaic profiles, and battery energy storage. The LLM critic is expanded to three fairness metrics with staggered activation schedules. Across a progression of experiments from a 5-agent pilot community to a 13-agent neighborhood using a European residential dataset, the framework shifts energy exchanges toward local peer-to-peer trades, lowers consumer costs by 27–30% relative to grid-only procurement, and sustains strong fairness across participants. The same LLM-guided reward design and policy architecture transfer from the small pilot to larger deployments without modification, demonstrating

scalable fairness shaping.

The third study broadens the LLM companion role from behavior shaping to decision explanation. XRL-LLM combines KernelSHAP feature attribution with domain-context-engineered LLM prompting to generate natural language explanations of PPO-based voltage control decisions on the IEEE 33-bus distribution network. The framework achieves a G-Eval explanation quality score of 4.13/5, representing a 33.7% improvement over template-based generation and 67.9% over raw SHAP output, with statistically significant gains in accuracy (Cohen's $d = 1.34$), actionability ($d = 4.07$), and completeness ($d = 2.61$), all at $p < 0.001$. A physics-grounded counterfactual verification procedure confirms causal faithfulness of 0.81 under critical loading conditions. Five ablation studies demonstrate that the framework is backbone-agnostic, that structured domain context produces synergistic quality gains, and that explanation reliability increases with operational severity.

Cross-cutting analysis reveals that prompt engineering functions as a first-class systems design concern across all three interaction modes, that the LLM companion paradigm generalizes through a consistent extract-format-process-integrate architectural pattern, and that fairness and explainability address complementary dimensions of operator trust. Together, the three studies establish that LLMs can serve as versatile cognitive companions to RL agents in power systems, bridging the gap between autonomous decision-making capability and the trustworthiness requirements of fairness, scalability, and explainability that operational deployment and emerging regulations demand.

Keywords: deep reinforcement learning, distribution systems, explainable artificial intelligence, fairness, large language models, multi-agent systems, peer-to-peer energy trading, SHAP, trustworthy AI, voltage control

Chapter 1

Introduction

The electric power grid is undergoing its most significant structural transformation since electrification began over a century ago. Driven by climate targets, falling hardware costs, and supportive policy, distributed energy resources (DERs) such as rooftop photovoltaic (PV) systems, battery energy storage, and electric vehicles are proliferating across distribution networks at an unprecedented rate. The U.S. Energy Information Administration projects that distributed solar capacity alone will more than double between 2023 and 2050 [1], fundamentally altering the physics and economics of electricity delivery. Where distribution systems once carried power in a single direction from centralized substations to passive consumers, they must now accommodate bidirectional flows, volatile generation profiles, and millions of active participants who simultaneously consume and produce energy. This evolution demands a new generation of intelligent control and market mechanisms that can operate autonomously, adapt to uncertainty, and earn the trust of human operators who remain ultimately responsible for system safety and equity.

Reinforcement learning (RL) has emerged as one of the most promising computational paradigms for meeting this challenge. Unlike traditional optimization methods that require accurate mathematical models of the system, RL agents learn effective control policies directly from interaction with their environment, making them well suited to the stochastic, high dimensional, and partially observable settings that characterize modern distribution systems [2, 3]. In recent years, RL has demonstrated impressive performance on two critical distribution system tasks: voltage regulation, where agents coordinate capacitor banks and transformer tap changers to maintain bus voltages within statutory limits [4, 5], and peer-to-peer (P2P) energy trading, where multiple autonomous agents negotiate prices and quantities in local electricity markets [6, 7]. These successes have positioned RL as a viable pathway

toward autonomous grid operation.

Yet despite strong laboratory performance, the deployment of RL in operational power systems remains limited. The gap between demonstrated capability and real world adoption is not primarily technical; it is a gap of *trust*. Grid operators, market regulators, and end users are reluctant to hand decision making authority to agents whose internal reasoning is opaque, whose outcomes may be unfair, and whose behavior may not generalize reliably beyond controlled test conditions. This trust deficit manifests along three distinct but interconnected dimensions. First, in P2P energy markets, RL agents trained to maximize individual or aggregate profit can produce *inequitable* outcomes, where dominant sellers inflate prices or certain buyers absorb disproportionate costs, without any mechanism to detect or correct such imbalances [8, 9]. Second, practical market environments involve *partial observability*, discrete action spaces, and heterogeneous participant populations that challenge the scalability of fairness interventions proven effective in idealized settings [10]. Third, even when RL agents make correct decisions, their inability to *explain* those decisions in terms a human operator can understand creates a fundamental barrier to oversight and accountability [11, 12]. Taken together, these three dimensions of fairness, scalability, and explainability form the core of the trust problem that currently limits RL adoption in power systems.

Addressing these trust barriers requires augmenting RL agents with capabilities that RL alone does not naturally provide: normative judgment about what constitutes a fair outcome, the ability to maintain that judgment under increasingly complex and realistic conditions, and the capacity to articulate the reasoning behind control decisions in human understandable language. This thesis argues that large language models (LLMs) offer a uniquely versatile means of providing all three capabilities. Trained on vast corpora of human text, LLMs encode broad knowledge of ethical norms, economic reasoning, and technical domain concepts [13]. When integrated with RL systems through carefully designed interfaces, they can serve as real time fairness critics that evaluate trading outcomes against equity principles, as scalable reward shapers that adapt to growing market complexity, and as post hoc explainers that translate opaque numerical attributions into actionable natural language guidance. In each of these roles, the LLM functions not as a replacement for the RL agent but as a *cognitive companion* that supplies the judgment, context, and communication abilities the agent lacks.

The remainder of this chapter develops these ideas in detail. Section 1.1 reviews the state of RL for power system operations, establishing the technical foundation for the thesis. Section 1.2 examines each dimension of the trust gap, identifying the specific research challenges that motivate the three studies comprising this thesis. Section 1.3 introduces the LLM as companion paradigm and positions the thesis within the emerging literature on LLM augmented

decision making. Section 1.4 states the research questions and summarizes the contributions of each study. Finally, Section 1.5 presents the organization of the thesis.

1.1 Reinforcement Learning for Power System Operations

Reinforcement learning formalizes the sequential decision making problem in which an agent interacts with an environment over discrete time steps, observing states, selecting actions, and receiving scalar rewards that encode the desirability of outcomes [14]. The agent's objective is to learn a policy, a mapping from states to actions, that maximizes the expected cumulative reward over time. When the state and action spaces are large or continuous, deep neural networks parameterize the policy and value functions, giving rise to deep reinforcement learning (DRL) [15]. Among DRL algorithms, Proximal Policy Optimization (PPO) [16] has become a widely adopted baseline due to its stable training dynamics and strong empirical performance across diverse environments.

The application of RL to power systems has accelerated rapidly since Duan et al. [2] introduced Grid Mind, one of the earliest DRL frameworks for autonomous voltage control. Their work demonstrated that deep Q networks and deep deterministic policy gradient agents could learn effective voltage regulation policies on large scale distribution systems without explicit network models. Subsequent research expanded the scope of DRL based voltage control along several fronts. Cao et al. [3] proposed a multi-agent DRL framework that coordinates PV inverters for voltage regulation on the IEEE 33-bus system. Yang et al. [17] developed a two timescale scheme combining physics based optimization for slow capacitor bank switching with DRL for fast inverter dispatch. Wang et al. [5] formulated active voltage control as a decentralized partially observable Markov decision process and published the MAPDN benchmark at NeurIPS, establishing an open source evaluation platform. Fan et al. [4] introduced PowerGym, a standardized OpenAI Gym compatible RL environment for Volt-VAR control. These frameworks typically use pandapower [18] as the underlying AC power flow simulation engine and evaluate performance on IEEE benchmark networks, particularly the 33-bus radial distribution test system introduced by Baran and Wu [19].

In parallel, RL has been applied to decentralized energy market design. P2P energy trading allows prosumers, participants who both produce and consume electricity, to transact directly with one another without a central market operator [7]. Multi-agent reinforcement learning (MARL) provides a natural framework for this setting, with each prosumer modeled as an independent agent learning a bidding or allocation policy. Chen and Liu [6] demonstrated a networked MARL platform for P2P trading with voltage control, though their system required

ad hoc price caps to maintain socially acceptable outcomes. Independent PPO (IPPO), where each agent trains its own PPO policy using only local observations, has emerged as a surprisingly effective approach in cooperative multi-agent settings [20, 21], offering decentralization and scalability advantages over centralized training methods such as MAPPO [21] or value decomposition approaches such as QMIX [22].

Despite these advances, two fundamental limitations persist across the RL for power systems literature. First, none of the voltage control frameworks cited above provides any form of explanation for the agent's decisions, rendering learned policies opaque to the operators who must ultimately trust and supervise them. Second, MARL based trading frameworks optimize for efficiency or profit without incorporating fairness objectives, or rely on static, hand crafted fairness constraints that degrade under distributional shift [9]. These limitations are not minor engineering gaps; they represent structural barriers to real world deployment that motivate the research presented in this thesis.

1.2 The Trust Gap in RL for Power Systems

The reluctance to deploy RL agents in operational power systems can be traced to three interconnected trust barriers, each corresponding to a distinct dimension of trustworthy autonomous operation. This section examines each barrier and the state of existing approaches to address it.

1.2.1 Fairness in Decentralized Energy Markets

Decentralized P2P trading offers significant benefits in terms of local energy utilization, grid congestion relief, and prosumer autonomy. However, without explicit fairness mechanisms, RL agents trained to maximize aggregate welfare or individual profit can produce outcomes where dominant sellers accumulate disproportionate revenue while smaller participants or budget constrained buyers bear excessive costs [8]. This is not merely an academic concern; energy equity is increasingly a regulatory priority, and market designs that systematically disadvantage certain participants face both ethical objections and practical barriers to adoption.

Traditional approaches to fairness in MARL include potential based reward shaping [9], proportional fair scoring, and profit variance penalties. Siddique et al. [10] optimized a lexicographic maximin welfare objective, and Zimmer et al. introduced an auxiliary equity network, both achieving roughly 20% fairness improvements but at the cost of bespoke metric design

and fragile hyperparameter tuning. A fundamental challenge with all rule based fairness interventions is their brittleness: the reward shaping terms encode a fixed definition of fairness that may not generalize across market configurations, participant populations, or operating conditions. When load profiles shift, new participants enter the market, or generation patterns change, static fairness constraints often degrade or produce unintended side effects [23].

What is needed is a fairness mechanism that can interpret trading outcomes in context, apply nuanced judgment about equity, and adapt its evaluation as market conditions evolve. This motivates the use of LLMs as dynamic fairness critics, an approach explored in Chapters 3 and 4 of this thesis.

1.2.2 Scalability Under Realistic Operating Conditions

Even when a fairness shaping mechanism proves effective in a controlled experimental setting, the transition to realistic deployment conditions introduces significant challenges. Practical P2P energy markets involve partial observability, where each agent sees only its own meter data and a subset of market information. Action spaces are discrete and combinatorial, with agents selecting price quantity pairs from finite grids rather than continuous values. Participant populations are heterogeneous, comprising households with varying PV capacities, consumption patterns, and budget constraints. Trading horizons extend across multiple time slots within a day, requiring policies that reason about temporal dependencies. Load and generation profiles exhibit stochastic variation driven by weather, occupancy, and seasonal patterns.

These realistic conditions stress test the assumptions underlying fairness interventions validated on small, fully observable, continuous action problems. A fairness critic that works well for two sellers and one buyer under full observability may fail to generalize when the market scales to dozens of prosumers with partial information. The scalability question is therefore not simply whether an approach can handle more agents, but whether the fairness shaping mechanism, the learning algorithm, and the market structure jointly remain effective under the compounding complexity of real world conditions.

Chapter 4 of this thesis directly addresses this challenge by extending the FairMarket-RL framework to a continuous double auction with partial observability, discrete price quantity actions, realistic residential load and PV profiles, three expanded fairness metrics, and a progression of experiments from a small pilot community to a larger simulated neighborhood and a mixed asset real world dataset.

1.2.3 Explainability of RL Control Decisions

The third trust barrier concerns the opacity of learned RL policies. A well trained PPO agent for voltage control may achieve a 90% reduction in voltage violations, but if a grid operator cannot understand *why* the agent chose to activate a specific capacitor bank or adjust a particular tap changer, the operator cannot verify the decision, anticipate its consequences, or intervene appropriately when conditions deviate from training scenarios. This opacity is especially problematic in safety critical infrastructure, where misguided autonomous actions can cascade into equipment damage or service interruptions.

Explainable artificial intelligence (XAI) offers a path toward addressing this challenge. Post hoc methods such as SHAP (SHapley Additive exPlanations) [24] assign each input feature a Shapley value quantifying its marginal contribution to the model's output, providing a principled decomposition of any individual decision. In the power systems domain, Zhang et al. [25] applied DeepSHAP to DRL agents for emergency control, demonstrating that Shapley value attributions can reveal which grid state variables most influence an agent's actions. However, their approach, like all prior work in this space, presents SHAP values only as raw numerical vectors. A numerical attribution telling an operator that "Bus 17 voltage has a SHAP value of 0.342" conveys importance but not *meaning*: it does not explain what physical condition at Bus 17 triggered the decision, how the chosen action will address it, or what the operator should monitor afterward.

Bridging the gap between numerical attribution and human understandable explanation requires a reasoning system that can connect feature importance scores to domain specific physical knowledge. This motivates the integration of LLMs into the explainability pipeline, an approach developed in Chapter 5 of this thesis.

1.3 Large Language Models as Cognitive Companions to RL

Large language models have demonstrated remarkable capabilities in natural language understanding, reasoning, and generation across diverse domains [13]. Trained on massive text corpora through self-supervised learning and fine-tuned via reinforcement learning from human feedback (RLHF), models such as GPT-4 encode broad knowledge spanning technical, ethical, and economic reasoning. These capabilities create opportunities for LLMs to serve roles that complement RL agents in decision making systems.

The integration of LLMs with RL is an emerging research direction that has taken several

forms. In task performance and decision support, LLMs have been used to generate reward functions [26], propose action plans [27], and guide exploration in complex environments. In the energy domain specifically, Cheng et al. [28] introduced GAIA, the first LLM tailored for power dispatch operations, while Majumder et al. [29] examined LLM capabilities across electric energy sector tasks. In explainability, Slack et al. [30] developed TalkToModel, using LLMs to explain ML predictions through natural language conversations, and Krishna et al. [31] demonstrated at NeurIPS 2023 that translating SHAP attributions into natural language rationales improves downstream task performance by 10–25% over chain of thought prompting.

This thesis positions LLMs in a specific and novel role: as *cognitive companions* to RL agents in power system applications. The companion metaphor is deliberate. The LLM does not replace the RL agent's decision making capability; rather, it augments the agent with complementary capacities that the RL paradigm does not inherently provide. Three distinct modes of LLM-RL interaction are explored across the three studies comprising this thesis:

1. **LLM as Fairness Critic (Chapter 3).** The LLM reads a structured ledger of trading outcomes after each market episode and returns normalized fairness scores that are integrated into agent rewards through scheduled coefficients. In this mode, the LLM shapes *what* the RL agents learn to value, injecting normative judgment about equity directly into the training loop.

2. **LLM as Scalable Reward Shaper (Chapter 4).** The same fairness critic architecture is extended to operate under partial observability and discrete actions within a continuous double auction, demonstrating that the LLM guided approach transfers from small pilot settings to larger, more realistic market environments without architectural changes. In this mode, the LLM enables the fairness mechanism to *scale* with operational complexity.

3. **LLM as Post Hoc Explainer (Chapter 5).** The LLM receives numerical SHAP attributions from a trained voltage control agent, along with structured domain context encoding network topology, device physics, and voltage standards, and generates natural language explanations that grid operators can understand and act upon. In this mode, the LLM bridges the gap between *what* the agent decided and *why* that decision makes physical sense.

Figure 1.1 illustrates these three interaction modes as a unified conceptual framework. The RL agent interacts with the power system environment (left), while the LLM companion

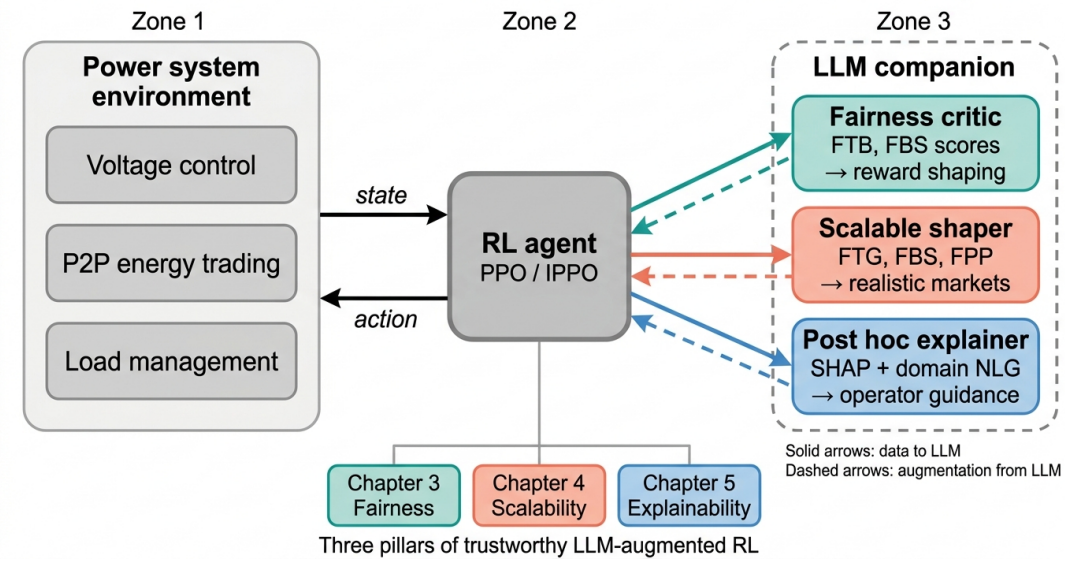

Figure 1.1: The LLM as cognitive companion paradigm for trustworthy RL in power systems. The RL agent handles sequential decision making in the grid environment, while the LLM companion augments three dimensions of trust: fairness (shaping what agents value), scalability (maintaining fairness under complexity), and explainability (articulating why decisions were made). Each mode corresponds to one study in this thesis.

provides fairness shaping, scalability, and explainability services (right), forming a complete trustworthiness augmentation layer.

The unifying insight across all three modes is that LLMs contribute capabilities that are difficult to engineer through conventional algorithmic means. Fairness evaluation requires contextual judgment that adapts to novel market configurations. Scalable reward shaping requires a representation flexible enough to process diverse outcome formats without hand-coded feature engineering. Explanation generation requires both domain knowledge and natural language fluency. In each case, the LLM's ability to process structured information and produce semantically meaningful outputs provides a capability that complements, rather than duplicates, the RL agent's strength in learning optimal sequential decisions from experience.

1.4 Research Questions and Contributions

Building on the motivations developed in the preceding sections, this thesis addresses the following overarching research question:

> *How can large language models serve as cognitive companions to reinforcement learning agents to address the critical trustworthiness barriers of fairness, scalability,*

and explainability in smart grid operations?

This overarching question is decomposed into three specific research questions, each addressed by one study:

RQ1: Can an LLM serve as a real time fairness critic to shape equitable outcomes in multi-agent energy trading? Chapter 3 introduces FairMarket-RL, a framework in which an instruction tuned LLM evaluates each trading episode and returns two normalized fairness scores, Fairness-To-Buyer (FTB) and Fairness-Between-Sellers (FBS), that are integrated into IPPO agent rewards via scheduled λ coefficients. Evaluated in a simulated P2P microgrid with two sellers and one buyer, the framework achieves FTB and FBS scores above 0.85, fulfills over 90% of buyer demand, and maintains balanced seller profit margins of 24–26%, outperforming baselines without LLM guidance where fairness stalls at 0.35–0.40. This study establishes the proof of concept that natural language fairness assessments can replace brittle rule based shaping in RL training loops for energy markets.

RQ2: Does the LLM guided fairness mechanism scale to realistic market conditions with partial observability, discrete actions, and heterogeneous participants? Chapter 4 extends FairMarket-RL to a continuous double auction with partial observability, discrete price quantity actions, and realistic residential load and PV profiles. The LLM now returns three fairness scores, Fairness-to-Grid (FTG), Fairness-Between-Sellers (FBS), and Fairness-of-Pricing (FPP), that are integrated via ramped coefficients with tunable scaling. Across a progression of experiments from a small pilot community to a larger simulated neighborhood and a mixed asset real world dataset, the framework shifts energy exchanges toward local P2P trades, lowers consumer costs relative to grid-only procurement, and sustains strong fairness across participants. This study demonstrates that the LLM guided reward design and policy class can be trained on a small pilot and transferred, without architectural changes, to larger communities and longer horizons.

RQ3: Can LLMs generate physically grounded, actionable explanations of RL voltage control decisions that are superior to template based and numerical baselines? Chapter 5 presents XRL-LLM, a framework that combines KernelSHAP feature attribution with domain context engineered LLM prompting to generate operator facing natural language explanations of PPO based voltage control decisions on the IEEE 33-bus network. Evaluated via G-Eval across 30 scenarios, the framework achieves an explanation quality score of 4.13/5, a 33.7% improvement over template based generation and 67.9% over raw SHAP output, with

Table 1.1: Summary of the three studies comprising this thesis, the trust dimension each addresses, and the principal contribution.

Study	Trust Dimension	LLM Role	Principal Contribution
FairMarket-RL [32] (Ch. 3)	Fairness	Real time fairness critic returning FTB and FBS scores	Proof of concept that LLM guided reward shaping produces equitable multi-agent trading
Scalable FairMarket-RL [33] (Ch. 4)	Scalability	Slot level fairness shaper with three metrics (FTG, FBS, FPP)	Scalable transfer from pilot to community scale under realistic conditions
XRL-LLM [34] (Ch. 5)	Explainability	Post hoc explainer translating SHAP attributions to natural language	Physics grounded, statistically validated NLG explanations for voltage control

statistically significant gains in accuracy ($d = 1.34$), actionability ($d = 4.07$), and completeness ($d = 2.61$), all at $p < 0.001$. A physics grounded counterfactual verification procedure confirms a causal faithfulness of 0.81 under critical loading. Five ablation studies demonstrate that the framework is backbone agnostic (even an 8B parameter open source model outperforms templates), that structured domain context engineering produces synergistic quality gains exceeding any single knowledge component, and that counterfactual faithfulness increases with load severity, indicating that explanations are most reliable where trustworthy guidance matters most.

Table 1.1 summarizes the three studies, their publication venues, and their principal contributions.

1.5 Thesis Organization

This thesis follows a manuscript based structure in which each core chapter presents a self-contained study that has been published or submitted for peer review. The chapters are organized to reflect the intellectual progression from establishing the LLM-RL fairness paradigm, through scaling it to realistic conditions, to broadening the LLM's role from behavior shaping to decision explanation.

Chapter 2 provides a thematic literature review that synthesizes the research landscape across the key areas intersecting this thesis: reinforcement learning for distribution system operations, fairness in multi-agent systems, explainable AI for sequential decision making,

and the emerging integration of LLMs with decision making agents. This review goes beyond the individual literature surveys in each paper to identify cross cutting themes and position the thesis within the broader trajectory of trustworthy AI for critical infrastructure.

Chapter 3 presents the first study, "FairMarket-RL: LLM-Driven Fairness Shaping in Multi-Agent Energy Trading for Smart Distribution Grids." This chapter introduces the core idea of using an LLM as a real time fairness critic within the MARL training loop, validates it in a controlled two seller, one buyer P2P microgrid, and demonstrates resilience against adversarial behaviors and noisy LLM outputs.

Chapter 4 presents the second study, "Scalable Fairness Shaping with LLM-Guided Multi-Agent Reinforcement Learning for Peer-to-Peer Electricity Markets." This chapter extends the FairMarket-RL framework to handle partial observability, discrete price quantity actions, realistic residential profiles, and multi-slot trading horizons, demonstrating scalability from small pilots to community scale deployments.

Chapter 5 presents the third study, "XRL-LLM: Explainable Reinforcement Learning Framework for Voltage Control." This chapter introduces a complementary mode of LLM-RL interaction, post hoc explanation generation, applied to a different power system task (voltage regulation on the IEEE 33-bus network). The framework combines game theoretic feature attribution with domain context engineered prompting, evaluated through multi-dimensional automated scoring, physics grounded counterfactual verification, and human validation.

Chapter 6 synthesizes the findings across all three studies, identifies cross cutting insights that emerge only from viewing the work collectively, discusses thesis level limitations, and outlines directions for future research toward an integrated LLM-RL system that simultaneously shapes fairness, enables scalability, and generates explanations for trustworthy autonomous grid operation.

Figure 1.2 presents a visual roadmap of the thesis structure and the logical connections between chapters.

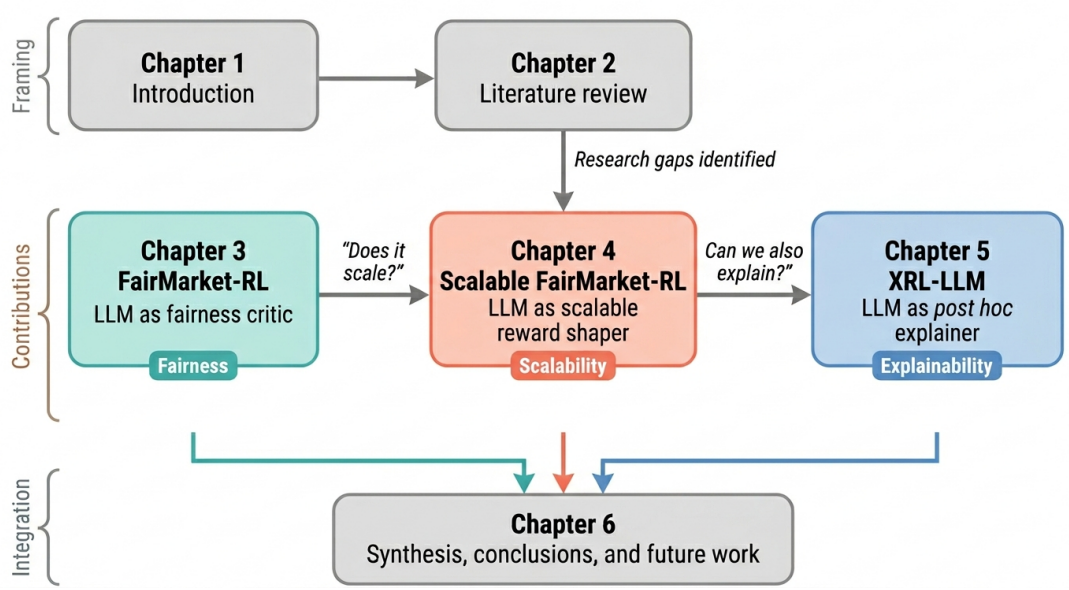

Figure 1.2: Thesis roadmap illustrating the logical progression across chapters. Chapters 3 and 4 address the fairness and scalability pillars through successive extensions of the FairMarket-RL framework. Chapter 5 broadens the paradigm to explainability, completing the trustworthiness triad. Chapter 6 synthesizes cross cutting insights.

Chapter 2

Literature Review

This chapter provides a thematic synthesis of the research landscape that underpins the three studies comprising this thesis. Rather than reviewing each paper's related work in isolation, the chapter is organized around six cross-cutting themes that collectively frame the contribution space: reinforcement learning for distribution system voltage control, multi-agent reinforcement learning for peer-to-peer energy trading, fairness in multi-agent systems and energy markets, explainable AI for sequential decision making, the integration of large language models with decision-making agents, and trustworthy AI for critical infrastructure. The chapter concludes with a synthesis of the identified research gaps and a mapping of how each gap motivates a specific study in this thesis.

The scope of this review covers the period from 2017 to 2026, with particular emphasis on work published since 2020, reflecting the rapid maturation of both deep RL for power systems and LLM-based reasoning. References were selected based on citation impact, venue prestige, and direct relevance to the thesis contributions. Survey papers are cited throughout to provide entry points for readers seeking deeper coverage of any individual theme.

Figure 2.1 provides a visual overview of these six thematic streams arranged by application domain (top row) and methodological foundation (bottom row). The intersections between regions indicate where prior work has established partial connections, while the dashed contribution zones highlight the specific intersections that this thesis occupies. Two observations are worth noting before proceeding. First, the two thesis contribution zones sit at triple intersections that no prior work has addressed: Chapter 5 combines RL voltage control, explainable AI, and LLM-RL integration, while Chapters 3 and 4 combine P2P trading, fairness in MARL, and LLM-RL integration. Second, both contribution zones connect upward to an overarching regulatory compliance contribution, reflecting the fact that the EU AI Act and NIST AI RMF

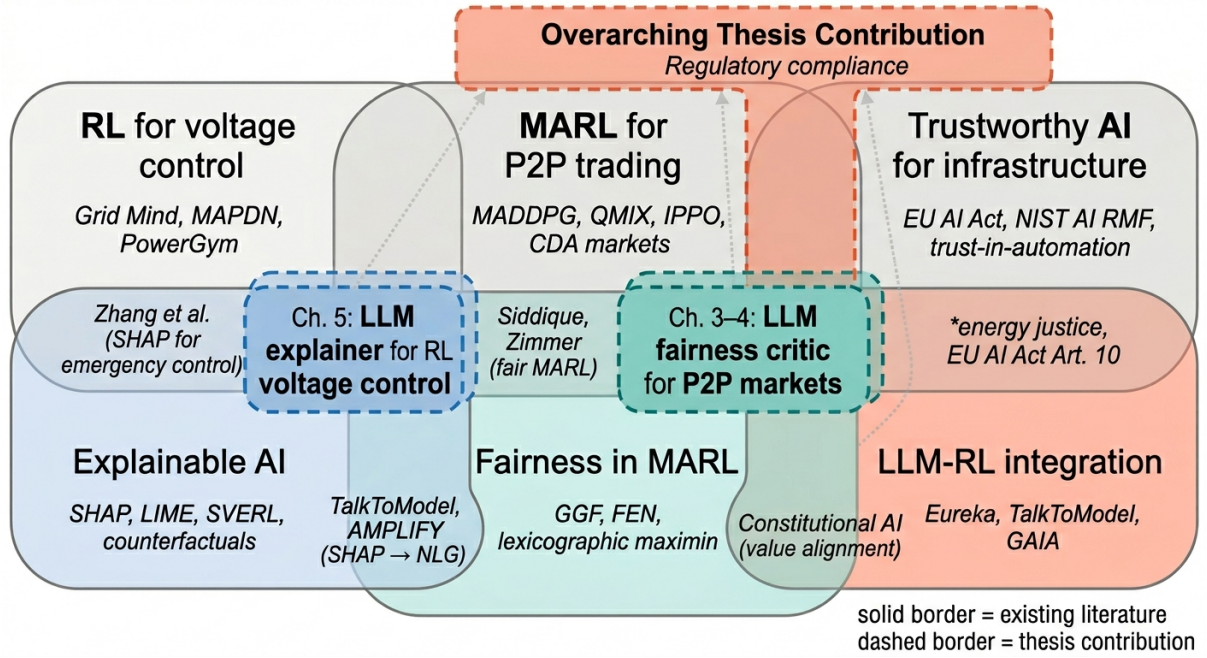

Figure 2.1: Thematic landscape of the literature reviewed in this chapter. Six research streams are arranged by application domain (top row) and methodological foundation (bottom row). Italic annotations at intersections identify the closest prior work. Dashed borders mark the thesis contribution zones at previously unoccupied triple intersections. Dotted arrows indicate that both contribution zones collectively address the overarching regulatory compliance requirements identified in the Trustworthy AI literature.

now mandate precisely the trustworthiness properties (fairness, transparency, explainability) that these studies provide.

2.1 Reinforcement Learning for Distribution System Voltage Control

The application of deep reinforcement learning to voltage regulation in distribution networks has progressed from initial proof-of-concept demonstrations to standardized multi-agent benchmarks within a span of roughly five years. This section traces that trajectory, highlighting the key methodological advances, simulation platforms, and persistent limitations that motivate the explainability work in Chapter 5.

The foundational contribution in this area is the Grid Mind framework introduced by Duan et al. [2], which demonstrated that DQN and DDPG agents could learn autonomous voltage control policies directly from simulated SCADA and PMU data on the IEEE 14-bus sys-

tem without requiring explicit power system models. This work established the viability of model-free RL for grid operations and catalyzed a wave of subsequent research. Cao et al. [3] extended the paradigm to multi-agent settings, proposing a MADRL framework that coordinates PV inverter reactive power outputs for voltage regulation on the IEEE 33-bus network. Their work demonstrated that decentralized agents could achieve comparable performance to centralized optimization when provided with appropriate reward structures.

Yang et al. [17] introduced a practically motivated two-timescale architecture that couples physics-based optimization for slow-timescale capacitor bank switching with DRL for fast-timescale smart inverter dispatch, tested on both IEEE test feeders and a real-world 47-bus distribution system. This hybrid approach acknowledged that not all control decisions benefit equally from learned policies, a design insight relevant to the hybrid LLM-RL architectures explored in this thesis.

Two benchmark contributions have been particularly influential in standardizing evaluation. Wang et al. [5] formulated active voltage control as a decentralized partially observable Markov decision process (Dec-POMDP) and published MAPDN at NeurIPS 2021, benchmarking ten MARL algorithms including IAC, IDDPG, IPPO, MAPPO, MADDPG, and COMA across bus-33, bus-141, and bus-322 networks. MAPDN established an open-source evaluation platform that bridges the power systems and MARL research communities. Fan et al. [4] introduced PowerGym, an OpenAI Gym-compatible RL environment built on OpenDSS that provides standardized interfaces for Volt-VAR control on IEEE 13, 34, 123, and 8500-node systems, finding that SAC achieved substantially better sample efficiency than PPO on larger networks.

The underlying simulation infrastructure for most of this research is pandapower [18], an open-source Python library that provides a Newton-Raphson AC power flow solver along with convenient modeling abstractions for loads, generators, transformers, and control devices. The IEEE 33-bus radial distribution test system, originally introduced by Baran and Wu [19] for network reconfiguration studies, has become the most widely used benchmark topology for RL-based voltage control, appearing in the majority of the works cited in this section.

More recent work has addressed safety constraints and architectural innovations. Qu et al. [35] formalized voltage control as a constrained Markov game with double safety estimation at IJCAI 2024. Wang et al. [36] incorporated transformer-based masked self-attention to capture grid topology in multi-agent actor-critic architectures at KDD 2022. Comprehensive surveys by Li et al. [37] in the Proceedings of the IEEE and by Glover et al. [38] provide extensive coverage of the field, with the latter highlighting reproducibility challenges and the persistent gap between simulated and operational deployment.

Complementary to RL-based control, graph neural network approaches have emerged for accelerating the underlying power flow computations and predicting grid failure modes. Jadhav et al. [39] proposed a topology-aware gated graph neural network surrogate for AC power flow estimation that generalizes across IEEE benchmark networks under line contingencies and up to 40% load variation, with both supervised and physics-informed self-supervised training variants that enforce Kirchhoff-law consistency. Sevak, Jadhav, and Bui [40] extended the physics-informed GNN paradigm to cascading failure prediction, introducing PI-GN-JODE, a hybrid continuous–discrete model that fuses an edge-conditioned GNN encoder with Neural ODEs for power redistribution and jump processes for relay trips, achieving PR-AUC of 0.991 on IEEE 24-bus and 118-bus systems at near-real-time speed. These physics-informed surrogate models are complementary to the RL control policies explored in this thesis: they can serve as fast forward models within RL training loops or as physics-consistent verification tools for the counterfactual explanations developed in Chapter 5.

Despite impressive simulation results, a consistent limitation across this entire body of work is the absence of explainability mechanisms. None of the frameworks cited above provides any form of explanation for the agent's control decisions. Grid operators who must supervise RL-based control systems cannot verify why an agent chose to activate a specific capacitor bank or adjust a particular tap changer, creating a fundamental barrier to operational trust. This gap directly motivates the XRL-LLM framework presented in Chapter 5.

2.2 Multi-Agent Reinforcement Learning for Peer-to-Peer Energy Trading

Peer-to-peer energy trading allows prosumers to exchange electricity directly without a central market operator, offering benefits in local energy utilization, grid congestion relief, and prosumer autonomy [7]. MARL provides a natural computational framework for this setting, with each prosumer modeled as an independent agent learning a bidding or allocation policy. This section reviews the MARL foundations, their application to energy markets, and the market design mechanisms that inform Chapters 3 and 4.

The algorithmic foundations of modern MARL rest on three landmark contributions. Lowe et al. [41] introduced MADDPG at NeurIPS 2017, pioneering the centralized-training-decentralized-execution (CTDE) paradigm where each agent's critic has access to global state during training but acts using only local observations at deployment. Rashid et al. [22] proposed QMIX at ICML 2018, introducing monotonic value function factorization that decomposes a joint

action-value function into per-agent utilities while preserving the monotonicity needed for tractable maximization. Yu et al. [21] demonstrated the surprising effectiveness of multi-agent PPO in cooperative settings at NeurIPS 2022, showing that Independent PPO (IPPO), where each agent trains its own PPO policy using only local observations, can match or exceed the performance of more complex centralized approaches on standard MARL benchmarks. This finding has significant practical implications because IPPO's decentralized architecture scales naturally to large agent populations without the communication overhead of centralized critics.

The most prolific research group at the intersection of MARL and energy trading has been Qiu and colleagues at Imperial College London. Qiu et al. [42] proposed a double-auction MADDPG framework at IJCAI 2021 that models P2P trading with each agent's critic augmented by public auction information rather than full state access, reducing the dimensionality of the centralized training problem. Their mean-field MARL extension [43] addressed scalability to 100 prosumers by approximating the joint effect of all other agents through mean-field theory, published in IEEE Transactions on Power Systems. Chen et al. [44] extended multi-agent energy trading to multi-energy microgrids using TD3 with cross-network interaction modeling in IEEE Transactions on Smart Grid. More recently, Feng and Liu [45] proposed consensus-based actor-critic MARL that integrates physical network constraints directly into the learning process.

Market design for P2P trading encompasses several mechanisms, each with different properties regarding efficiency, fairness, and computational tractability. Continuous double auctions (CDAs), where buyers and sellers submit bids and offers that are matched in real time, have emerged as the dominant mechanism in MARL-based P2P research due to their natural compatibility with decentralized agent architectures [46]. Alternative mechanisms include bilateral contracts, pool-based markets, and blockchain-mediated platforms [7]. The scalable FairMarket-RL framework in Chapter 4 adopts the CDA mechanism specifically because it supports partial observability (each agent sees only its own bids and public clearing prices) and discrete price-quantity actions, reflecting realistic market conditions.

Recent surveys by Keren et al. [47] and Shah et al. [48] provide comprehensive overviews of MARL for energy networks, identifying scalability, partial observability, and the integration of physical network constraints as the three primary open challenges. Notably absent from these surveys is any discussion of fairness mechanisms, a gap that the FairMarket-RL line of work directly addresses.

2.3 Fairness in Multi-Agent Systems and Energy Markets

Fairness in multi-agent reinforcement learning has received growing attention as MARL systems are deployed in settings where the distribution of outcomes across participants matters as much as aggregate performance. This section reviews the theoretical foundations, algorithmic approaches, and energy-specific fairness considerations that motivate Chapters 3 and 4.

2.3.1 Fairness Definitions and Metrics

The quantitative study of fairness in resource allocation originated with Jain, Chiu, and Hawe [49], who proposed the Jain fairness index as a bounded measure of allocation equality. In the MARL context, fairness can be defined along several axes. Distributional fairness concerns the equality of outcomes (rewards, profits, or costs) across agents. Procedural fairness concerns whether the decision-making process treats all agents equitably. Proportional fairness, formalized through Nash bargaining theory, seeks allocations that maximize the product of utilities. The Generalized Gini Social Welfare Function (GGF) provides a flexible framework that encompasses utilitarian, egalitarian, and intermediate fairness criteria through a single parameterized welfare function [10].

The most comprehensive survey of fairness in RL is provided by Reuel and Ma [50], published at AAAI/ACM AIES 2024, which distinguishes between fairness of the policy (equal treatment of similar agents), fairness of outcomes (equal distribution of rewards), and fairness during learning (equal convergence rates).

2.3.2 Algorithmic Approaches to Fair MARL

The foundational algorithmic work was established by two closely related papers. Siddique, Weng, and Zimmer [10] introduced learning fair policies via GGF optimization in multi-objective deep RL at ICML 2020, adapting both DQN and policy gradient methods to optimize lexicographic maximin and proportional fairness objectives. Zimmer, Glanois, Siddique, and Weng [51] extended this to decentralized cooperative MARL at ICML 2021, proposing a dual-subnetwork architecture where one network estimates individual utility and a second estimates social welfare, compatible with multiple fairness criteria.

Complementary approaches address different aspects of the fairness challenge. Jiang and Lu [52] proposed the Fair-Efficient Network (FEN) at NeurIPS 2019, using hierarchical RL with a meta-controller that switches between fair and efficient sub-policies. Fan et al. [53] studied Nash Social Welfare optimization for fair multi-objective RL at AAMAS 2023, prov-

ing that even approximate optimization of expected Nash Social Welfare is computationally intractable and providing a convergent Welfare Q-Learning algorithm. Ju, Ghosh, and Shroff [54] achieved fairness in multi-agent MDPs through online convex optimization within confidence bounds at ICLR 2024.

A fundamental limitation shared by all of these approaches is their reliance on fixed, handcrafted fairness metrics and reward shaping terms. The fairness objectives must be specified mathematically before training begins and remain static throughout. When market conditions shift, participant populations change, or new equity concerns arise, the reward structures must be manually redesigned and the agents retrained. This brittleness motivates the use of LLMs as dynamic fairness critics that can interpret outcomes in context and adapt their evaluation as conditions evolve, the approach taken in Chapters 3 and 4.

2.3.3 Energy Equity and Fairness in P2P Markets

The energy justice literature provides the normative foundation for fairness in electricity markets. Jenkins et al. [55] established the conceptual framework of distributional, procedural, and recognition justice in energy systems. Chen and Liu [56] provide a comprehensive systematic review covering 160 studies on energy justice definitions, measures, and practices across policy, planning, and operations. In the P2P trading context specifically, Zhou, Wu, and Long [8] demonstrated through multiagent simulation that unconstrained trading can produce highly inequitable outcomes where dominant sellers accumulate disproportionate revenue. Irusta et al. [57] proposed distributionally fair P2P optimization using the Wasserstein metric on the IEEE 33-bus network.

The gap between the MARL fairness literature and the energy equity literature is striking. The MARL community has developed increasingly sophisticated algorithms for optimizing formal fairness objectives, while the energy policy community has articulated increasingly nuanced conceptions of justice that resist reduction to a single mathematical formula. No prior work has attempted to bridge this gap using natural language reasoning about fairness, a bridge that the LLM fairness critic in FairMarket-RL provides.

2.3.4 Value Alignment and Constitutional AI

The idea of using AI systems to evaluate the normative properties of other AI systems has precedent in the value alignment literature. Bai et al. [58] introduced Constitutional AI at Anthropic, demonstrating that RL from AI Feedback (RLAIF) guided by human-written constitutional principles can produce models that are both helpful and harmless. Huang et al. [59]

extended this with democratically sourced principles in Collective Constitutional AI at ACM FAccT 2024. The FairMarket-RL framework draws conceptual inspiration from this paradigm, positioning the LLM not as a constitutional self-critic of its own outputs but as an external fairness critic of multi-agent trading outcomes, evaluated against equity principles encoded through structured prompts rather than constitutional documents.

2.4 Explainable AI for Sequential Decision Making

The opacity of deep RL policies creates a fundamental barrier to their deployment in safety-critical systems where human operators must understand, verify, and potentially override autonomous decisions. This section reviews the explainability methods that inform Chapter 5, spanning post-hoc attribution techniques, counterfactual explanations, RL-specific challenges, and evaluation methodologies.

2.4.1 Feature Attribution Methods

The two dominant frameworks for post-hoc feature attribution are SHAP and LIME. Lundberg and Lee [24] introduced SHAP at NeurIPS 2017, unifying six prior explanation methods under Shapley value theory. SHAP assigns each input feature a value quantifying its marginal contribution to the model's output, with theoretical guarantees of local accuracy, missingness, and consistency. KernelSHAP provides a model-agnostic estimation procedure that approximates Shapley values through weighted regression over randomly sampled feature coalitions. Ribeiro, Singh, and Guestrin [60] introduced LIME at KDD 2016, which explains individual predictions by fitting a locally faithful interpretable model (typically a sparse linear regression) in the neighborhood of the instance being explained.

Both methods were originally designed for supervised learning and face additional challenges when applied to RL policies, where actions affect future states and the value of a decision depends on the entire trajectory, not just immediate outcomes.

2.4.2 Explainability for Reinforcement Learning

The application of feature attribution to RL agents introduces theoretical challenges beyond those encountered in supervised learning. Beechey, Smith, and Şimşek [12] provided the first rigorous treatment of Shapley values for RL at ICML 2023, proposing the SVERL framework. SVERL distinguishes between three explanation targets: explaining agent behavior (why a

particular action was selected), explaining agent performance (which features contribute to high or low returns), and explaining value estimation (which features drive the critic's state-value predictions). This taxonomy reveals that prior ad-hoc applications of SHAP to RL often conflated these distinct explanation targets, potentially producing misleading attributions. An extended journal version appeared in 2025 with additional theoretical analysis.

Earlier work on RL explainability took visual approaches. Greydanus et al. [61] used perturbation-based saliency maps to identify which pixels influence Atari agent decisions. Mott et al. [62] proposed attention-augmented agents that create interpretable information bottlenecks, allowing the attention weights themselves to serve as explanations.

In the power systems domain specifically, Zhang, Xu, and Zhang [25] pioneered the application of DeepSHAP to DRL-based emergency control, published in IEEE Transactions on Computational Social Systems. They demonstrated that Shapley value attributions can reveal which grid state variables most influence an RL agent's emergency actions. However, their approach presents SHAP values only as raw numerical vectors without translating them into human-understandable language, does not validate attributions against physical simulation, and targets transmission-level emergency control rather than distribution-level voltage regulation. These three limitations define the specific contribution of Chapter 5.

Comprehensive surveys of explainable RL include Milani et al. [63] in ACM Computing Surveys, which organizes methods by the RL setting (model-based vs. model-free, single-agent vs. multi-agent), and Qing et al. [64], which provides a continuously updated taxonomy of XRL concepts, algorithms, and challenges.

2.4.3 Counterfactual Explanations

Counterfactual explanations answer the question "what minimal change to the input would alter the model's output?" and were formalized for algorithmic decision-making by Wachter, Mittelstadt, and Russell [65]. Mothilal, Sharma, and Tan [66] extended this with DiCE at ACM FAT* 2020, generating diverse counterfactuals using determinantal point processes. Karimi et al. [67] provided a comprehensive survey of algorithmic recourse methods.

A critical limitation of standard counterfactual approaches in physical systems is that they perturb input features without verifying whether the resulting scenario is physically realizable. In power systems, arbitrary perturbations of bus voltages may violate Kirchhoff's laws and AC power flow constraints, producing counterfactual scenarios that are physically impossible. Chapter 5 addresses this limitation through a physics-grounded counterfactual verification procedure that re-solves the full nonlinear AC power flow equations under coun-

terfactual conditions, ensuring that only physically consistent scenarios are used to evaluate explanation faithfulness.

2.4.4 Evaluation of Explanation Quality

Evaluating the quality of generated natural language explanations has historically relied on n-gram overlap metrics such as BLEU and ROUGE, which are poorly suited for open-ended explanation assessment where multiple valid phrasings exist. Liu et al. [68] introduced G-Eval at EMNLP 2023, a framework that uses GPT-4 with chain-of-thought prompting to evaluate NLG outputs across dimensions such as coherence, consistency, fluency, and relevance. G-Eval achieves approximately 0.514 Spearman correlation with human judgments in summarization tasks, substantially outperforming prior automated metrics. Zheng et al. [69] systematically validated the LLM-as-a-judge paradigm at NeurIPS 2023, demonstrating that GPT-4 achieves over 80% agreement with human evaluators.

For readability assessment, the Flesch Reading Ease formula [70] and the Flesch-Kincaid Grade Level [71] remain the most widely used metrics in technical communication. Chapter 5 combines G-Eval scoring, Flesch-Kincaid readability analysis, and rigorous statistical testing (Mann-Whitney U, Cohen's d) to provide the most comprehensive evaluation of XAI-generated explanations for power systems to date.

2.5 Large Language Models as Decision-Making Companions

The integration of large language models with reinforcement learning is an emerging research direction that has crystallized rapidly since 2022 into several distinct interaction paradigms. This section reviews the key paradigms, with emphasis on the three modes of LLM-RL interaction explored in this thesis: reward shaping, decision support, and explanation generation.

2.5.1 LLMs for Reward Design and Shaping

One of the most productive LLM-RL integration patterns uses LLMs to design or shape reward functions.As noted in recent surveys of the field [72], this LLM-as-Reward paradigm is essential for bridging the gap between high-level human intent and numerical optimization.Yu et al. [26] demonstrated Language to Rewards at CoRL 2023, using LLMs to define reward parameters for MuJoCo controllers and achieving 90% success across 17 robotic tasks. Kwon,

Xie, Bullard, and Sadigh [73] explored LLMs as proxy reward functions at ICLR 2023, evaluating agent behavior against textual descriptions of desired outcomes. Ma et al. [74] introduced Eureka at ICLR 2024, performing evolutionary optimization over LLM-generated reward code that outperformed expert human-engineered rewards on 83% of 29 tasks, including the first simulated pen-spinning with a Shadow Hand.

These approaches use LLMs to generate static reward code before training begins. FairMarket-RL (Chapters 3 and 4) takes a distinct approach: the LLM evaluates trading outcomes during training and returns dynamic fairness scores that modulate agent rewards in real time. This online evaluation pattern is closer to Constitutional AI [58] than to static reward code generation, because the LLM must interpret each specific outcome in context rather than producing a fixed reward function.

2.5.2 LLMs for Task Planning and Decision Support

Huang, Abbeel, Pathak, and Mordatch [27] demonstrated that LLMs can decompose high-level task descriptions into executable action sequences at ICML 2022. Ahn et al. [75] grounded LLM-generated plans in robotic affordances via value functions at CoRL 2022, achieving 84% plan success on 101 real-world tasks. The foundational work on instruction-following through RLHF is Ouyang et al. [13], whose InstructGPT demonstrated the three-step pipeline of supervised fine-tuning, reward model training, and PPO optimization that underpins modern instruction-tuned LLMs.

2.5.3 LLMs for Explanation Generation

The use of LLMs to translate numerical model outputs into natural language explanations has been explored primarily in the supervised learning context. Slack, Krishna, Lakkaraju, and Singh [30] developed TalkToModel, published in Nature Machine Intelligence, which enables interactive dialogue-based ML explanation with 73% of healthcare workers preferring it over existing explanation systems. Krishna, Ma, Slack, Ghandeharioun, Singh, and Lakkaraju [31] demonstrated at NeurIPS 2023 that post-hoc SHAP attributions translated into natural language rationales improve LLM task performance by 10–25% over chain-of-thought prompting alone [76]. Kroeger et al. [77] systematically evaluated multiple prompting strategies for converting feature attributions into textual narratives.

Chapter 5 extends this line of work in two important ways. First, it applies the SHAP-to-NLG pipeline to RL policies rather than supervised models, navigating the additional challenges identified by Beechey et al. [12]. Second, it incorporates structured domain context

(network topology, device physics, voltage standards) into the LLM prompt, enabling physically grounded explanations rather than generic statistical descriptions.

2.5.4 LLM Applications in Power and Energy Systems

Domain-specific LLM applications in the energy sector are emerging rapidly. Cheng et al. [28] introduced GAIA, the first LLM fine-tuned specifically for power dispatch operations, evaluated on the ElecBench benchmark. Majumder et al. [29] systematically explored LLM capabilities and limitations across electric energy tasks including optimal power flow and EV scheduling. Recent review papers in IEEE and Applied Energy [78] comprehensively survey LLM integration into energy systems, identifying grid operations, market design, and consumer engagement as the three primary application areas.

The broader LLM-RL integration landscape is mapped by several surveys. Cao et al. [79] proposed the first comprehensive taxonomy in IEEE Transactions on Neural Networks and Learning Systems, categorizing LLM roles as information processor, reward designer, decision-maker, and generator. Pternea et al. [80] defined three integration classes: RL for LLM (using RL to improve LLM alignment), LLM for RL (using LLMs to enhance RL components), and RL+LLM (bidirectional synergy). Sun, Huang, and Pompili [81] surveyed LLM-based MARL, discussing LLMs as reward designers and communication enablers in multi-agent systems.

This thesis contributes to the "LLM for RL" class of the taxonomy, demonstrating three previously unexplored interaction modes in the power systems domain: fairness critic, scalable reward shaper, and post-hoc explainer.

2.6 Trustworthy AI for Critical Infrastructure

The deployment of AI systems in safety-critical infrastructure requires trustworthiness properties that extend beyond predictive accuracy to include fairness, explainability, robustness, and human oversight. This section reviews the regulatory landscape, trust-in-automation theory, and the specific deployment gap for RL in power systems.

2.6.1 Regulatory Frameworks

The EU AI Act (Regulation 2024/1689), which entered into force on August 1, 2024, represents the most consequential regulatory development for AI in energy systems. Annex III, Section 2 explicitly classifies AI systems used for management and operation of critical infrastructure,

including electricity supply, as high-risk. High-risk system obligations encompassing risk management, transparency, human oversight, and accuracy become fully applicable on August 2, 2026, with penalties reaching up to 35 million euros or 7% of global turnover [82]. This regulatory timeline creates direct and urgent motivation for the trustworthiness properties that this thesis addresses.

The NIST AI Risk Management Framework (AI RMF 1.0) [83] provides a complementary voluntary framework organized around four functions: Govern, Map, Measure, and Manage. It promotes seven characteristics of trustworthy AI: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. The U.S. Department of Homeland Security released a Roles and Responsibilities Framework for AI in Critical Infrastructure in November 2024 [84], defining obligations across cloud providers, AI developers, infrastructure owners, and civil society. The U.S. Department of Energy's "AI for Energy" report [85] identifies grid planning, operations and reliability, and resilience as priority areas for AI deployment, explicitly calling for trustworthy AI techniques.

IEEE standards provide process-level guidance. IEEE 7000-2021 establishes Value-Based Engineering methodology for translating stakeholder values into system requirements during design [86]. The IEEE Ethically Aligned Design initiative [87] contains over 100 recommendations for ensuring that autonomous and intelligent systems prioritize human well-being.

2.6.2 Trust in Automation

The human factors literature provides essential theoretical grounding for understanding why operators do or do not trust automated systems. Lee and See [88] developed the most widely cited model of trust in automation, identifying three cognitive bases: performance-based (does the system work reliably?), process-based (do I understand how it works?), and purpose-based (does the system's intent align with mine?). Their framework predicts that explainability directly strengthens process-based trust, the weakest link for opaque RL agents. Parasuraman and Riley [89] defined four pathologies of human-automation interaction: use (appropriate reliance), misuse (over-reliance), disuse (under-reliance), and abuse (automation of tasks that should not be automated). Unexplainable RL agents are particularly vulnerable to disuse, where operators bypass the automation because they cannot predict or understand its behavior. The updated relational approach by Chiou and Lee [90] reframes trust through situation, semiotics, strategy, and sequence dimensions, providing a more dynamic model appropriate for iterative human-AI interaction.

2.6.3 The Deployment Gap for RL in Power Systems

The gap between RL performance in simulation and its deployment in operational power systems is concretely documented in several studies. The Learning to Run a Power Network (L2RPN) challenge, organized by Marot, Donnot, Dulac-Arnold et al. [91] at NeurIPS 2020, showed that RL agents could achieve super-human performance in grid management simulation while revealing the chasm between simulated and operational environments. Their follow-up work, "Learning to Run a Power Network with Trust" [92], directly addresses human-AI trust-building for grid operations, arguing that operators will remain in charge of critical decisions and proposing trust mechanisms for RL agents in congestion management.

The RL2Grid benchmark [93], released in 2025, further documented that RL methods struggle with complex dynamics, aleatoric uncertainty, and hard physical constraints in realistic grid settings. A recent graph RL survey [94] identified the simulation-to-reality gap as the primary barrier to RL deployment in power systems. Beyond distribution networks, ML-based optimization has been applied to other energy-critical infrastructure where the same trust and deployment challenges arise. Jadhav and Liu [95] developed a physics-guided gradient-boosting framework for diagnosing cooling inefficiencies at the Frontier exascale supercomputer, identifying approximately 85 MWh/year of excess cooling energy through a monotonicity-constrained surrogate with sub-1% PUE prediction error. Jadhav and Liu [96] extended this with a Modelica-based digital twin of Frontier's liquid-cooling loop, validated to ASHRAE Guideline 14 standards, and demonstrated through layered optimization that co-optimizing flow rates and supply temperatures under actuator ramp-rate constraints yields 27.8% energy savings. These works illustrate that physics-informed ML and digital-twin approaches can bridge the gap between theoretical optima and deployable control strategies in critical infrastructure, a challenge that parallels the simulation-to-deployment gap for RL in power systems. These findings reinforce the thesis argument that technical performance alone is insufficient for deployment; trustworthiness properties including fairness and explainability are necessary preconditions for operator acceptance.

2.7 Synthesis: Research Gaps and Thesis Positioning

The six literature streams reviewed in this chapter converge on three critical, interconnected research gaps that define the contribution space of this thesis.

Gap 1: Fairness in MARL energy markets lacks scalable, adaptive mechanisms. The MARL fairness literature has produced increasingly sophisticated algorithms for optimizing formal fairness objectives [10, 51, 53], but these rely on static, hand-crafted reward terms that degrade under distributional shift. The energy equity literature articulates nuanced conceptions of justice [55, 56] that resist reduction to fixed mathematical formulas. No prior work has bridged this gap using natural language reasoning about fairness. Chapters 3 and 4 address this gap by introducing an LLM-based fairness critic that interprets trading outcomes in context and provides dynamic, adaptable fairness evaluation.

Gap 2: RL policies for power systems lack operator-facing explainability. The XAI literature provides principled attribution methods (SHAP [24], LIME [60]) and theoretical frameworks for their application to RL (SVERL [12]). The NLG literature demonstrates that LLMs can translate feature attributions into natural language [30, 31]. Power systems XAI work has applied SHAP to RL emergency control but stopped at numerical vectors [25]. No existing framework combines SHAP-based RL attribution with LLM-powered natural language explanation generation for power system voltage control. Chapter 5 fills this gap with a complete pipeline from attribution to operator-facing guidance, validated through physics-grounded counterfactual verification and multi-dimensional quality assessment.

Gap 3: The regulatory landscape now mandates trustworthiness properties that current RL deployments cannot satisfy. The EU AI Act classifies AI in energy infrastructure as high-risk, requiring transparency, explainability, and human oversight by August 2026. The NIST AI RMF promotes fairness and explainability as core trustworthiness characteristics. The trust-in-automation literature [88, 89] predicts that unexplainable agents will be disused by operators regardless of their technical performance. Current RL frameworks for power systems address none of these requirements. The three studies in this thesis collectively demonstrate that LLMs can serve as cognitive companions that provide the fairness evaluation, scalable adaptation, and natural language explanation capabilities needed to move RL-based power system control from the laboratory toward trustworthy operational deployment.

Table 2.1 summarizes the mapping between the identified gaps, the prior work that partially addresses each gap, the specific limitation that remains, and the thesis chapter that provides the contribution.

Table 2.1: Mapping of identified research gaps to the prior work, remaining limitations, and thesis contributions that address each gap.

Gap	Closest Prior Work	Remaining Limitation	Thesis Ch.
Adaptive fairness in MARL energy markets	Siddique et al. (ICML 2020), Zimmer et al. (ICML 2021)	Static metrics; no language-based contextual evaluation	Ch. 3, 4
Operator-facing RL explainability	Zhang et al. (IEEE TCSS 2022), Beechey et al. (ICML 2023)	Numerical SHAP vectors only; no NLG or physical verification	Ch. 5
Regulatory compliance for RL in energy	EU AI Act (2024), NIST AI RMF (2023)	No RL framework addresses mandated transparency and fairness	Ch. 3–5

Chapter 3

FairMarket-RL: LLM-Driven Fairness Shaping in Multi-Agent Energy Trading

---

Publication status. This chapter is based on the following peer-reviewed publication:

> S. Jadhav, B. Sevak, S. Das, A. Hussain, W. Su, and V.-H. Bui, "FairMarket-RL: LLM-Driven Fairness Shaping in Multi-Agent Energy Trading for Smart Distribution Grids," *The Electricity Journal*, under review, 2025.
> *Preprint:* `https://arxiv.org/abs/2506.22708`
> *Submission status:* `https://track.authorhub.elsevier.com/?uuid=80446a85-e6fd-4ea7-a39c-e35599219e45`

Research question addressed. *RQ1: Can an LLM serve as a real-time fairness critic to shape equitable outcomes in multi-agent energy trading?*

Author contributions. Shrenik Jadhav: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing (original draft and review). Birva Sevak: Methodology, Software, Investigation, Visualization. Srijita Das: Data curation, Investigation, Validation. Akhtar Hussain: Methodology, Validation, Resources. Wencong Su: Conceptualization, Supervision, Funding acquisition. Van-Hai Bui: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition (Corresponding author).

Connection to thesis. This chapter establishes the foundational proof of concept for the LLM-as-companion paradigm introduced in Chapter 1. It demonstrates that an instruction-tuned LLM can function as a real-time fairness critic within a MARL training loop, addressing

the first pillar of the trustworthiness framework (fairness). The fairness shaping mechanism developed here is subsequently scaled to realistic market conditions in Chapter 4.

3.1 Introduction

Peer-to-peer energy trading on distribution feeders allows prosumers to transact locally without a central market operator, enhancing autonomy and resilience. However, this decentralization also creates equity risks when dominant sellers inflate prices or certain buyers consistently absorb excessive costs [8, 7]. Recent multi-agent reinforcement learning (MARL) prototypes can clear local energy bids while respecting grid constraints, yet pronounced payoff disparities persist; Chen and Liu's networked-MARL testbed, for example, required ad hoc price caps to remain socially acceptable [6]. Classical fairness interventions such as potential-based shaping, proportional-fair scoring, and profit-variance penalties mitigate extreme outcomes only under static conditions and degrade as load, generation, or topology shift [9].

More principled solutions recast equity as an explicit learning objective. Siddique et al. [10] optimized a lexicographic maximin welfare, and Zimmer et al. [51] introduced an auxiliary equity network, both yielding roughly 20% fairness improvements at the cost of bespoke metrics and fragile hyperparameter tuning. However, these approaches encode a fixed definition of fairness that cannot adapt to changing market conditions. The emergence of instruction-tuned large language models offers a more adaptive path: modern LLMs absorb broad ethical and economic priors through human feedback training [13], enabling them to reason about concepts like justice and equity in free-form language. Such models can ingest structured ledgers of prices, unmet demand, and profits, then return nuanced fairness assessments without hand-crafted features or domain-specific tuning.

This chapter presents FairMarket-RL, a framework that embeds a pretrained LLM as a real-time fairness critic in a decentralized MARL framework. After each trading episode, the LLM observes the collective outcome and outputs two fairness metrics, Fairness-To-Buyer (FTB) and Fairness-Between-Sellers (FBS), which modulate agent rewards through scheduled λ-coefficients. A linear λ-ramping curriculum gradually introduces fairness objectives without destabilizing convergence. The resulting framework achieves both equitable and efficient outcomes in a simulated P2P microgrid market, demonstrating that natural-language fairness assessments can be injected directly into the training loop of non-textual environments.

3.2 System Architecture

FairMarket-RL is a decentralized multi-agent framework in which N_S seller agents and N_B buyer agents interact through a shared market environment. Each episode unfolds in a two-stage trading interaction: first, a seller offer stage where sellers $(S_1, \ldots, S_{N_S})$ simultaneously post price-quantity offers based on their local states; second, a buyer allocation stage where buyers $(B_1, \ldots, B_{N_B})$ simultaneously allocate demand across the posted offers subject to budget and demand limits. After these two stages, the environment compiles an outcome ledger containing offers, allocations, residual shortfall, and realized payoffs. Figure 3.1 illustrates the training loop with the LLM fairness path.

Seller and buyer policies are independent decentralized networks that output offers and allocations from local observations. A prompt skeleton serializes the episode's outcome ledger into a deterministic JSON with fixed field order, units, and keys, and queries the LLM fairness critic. The LLM returns FTB (a per-buyer vector) and FBS (a single market-level scalar). Outputs pass through FTB/FBS decoders that perform type and range checks and normalization; episodes with invalid or unparsable responses are discarded to preserve a clean training signal. A reward shaping module merges raw economic returns with the FTB/FBS signals, scaled by a λ-scheduler that gradually increases the fairness weights, so agents first learn profitability and then internalize fairness. Trajectories with shaped returns are written to an on-policy rollout buffer, and an IPPO trainer updates each agent's policy independently.

At inference time, the same seller-buyer interaction proceeds with the LLM and shaping path disabled: learned policies act directly, keeping operation simple and low-latency. Figure 3.2 illustrates the inference pipeline.

3.3 Mathematical Modeling

The P2P trading environment is formalized as a finite-horizon, turn-based multi-agent game:

$$G = \left\langle S,\ \{A^{S_i}\}_{i=1}^{N_S},\ \{A^{B_j}\}_{j=1}^{N_B},\ T,\ R \right\rangle \tag{3.1}$$

where S is the state space, A^{S_i} and A^{B_j} are the action sets for seller S_i and buyer B_j, T is the deterministic transition function, and R is the reward vector. Each episode consists of two stages: first sellers act, then buyers act.

3.3.1 State Representation

The global state at time t is:

$$s_t = \left[I^1, \ldots, I^{N_S},\ D^1, \ldots, D^{N_B},\ p^1, q^1, \ldots, p^{N_S}, q^{N_S},\ \sigma_t\right] \tag{3.2}$$

where I^i is the remaining inventory of seller S_i, D^j is the unmet demand of buyer B_j, the tuples (p^i, q^i) are the active price-quantity offers of seller S_i, and $\sigma_t \in \{\text{seller_stage}, \text{buyer_stage}\}$ indicates whose turn it is.

3.3.2 Action Spaces and Transitions

At $t = 0$, each seller S_i simultaneously chooses an action $(p_i, q_i) \in \{1, \ldots, 10\} \times [0, I^i]$, posting a price from a discrete grid and a quantity not exceeding its inventory. Once all sellers have posted, the stage switches to buyer allocation. Each buyer B_j simultaneously chooses a demand allocation vector $\mathbf{b}^{(j)} = (b_1^{(j)}, \ldots, b_{N_S}^{(j)})$, where $b_i^{(j)}$ is the quantity B_j purchases from seller S_i. These allocations satisfy:

$$\sum_{i=1}^{N_S} p_i\, b_i^{(j)} \leq \text{Budget}_j, \qquad \sum_{i=1}^{N_S} b_i^{(j)} \leq D_j, \qquad 0 \leq b_i^{(j)} \leq q_i \ \ \forall i \tag{3.3}$$

After all buyers have acted, inventories and demands are updated deterministically:

$$I^i_{\text{new}} = I^i - \sum_{j=1}^{N_B} b_i^{(j)}, \qquad D^j_{\text{new}} = D^j - \sum_{i=1}^{N_S} b_i^{(j)} \tag{3.4}$$

3.3.3 Raw Reward Functions

Before fairness shaping, each agent receives an economic reward based on trade outcomes:

$$r^{\text{raw}}_{S_i} = (p_i - c) \sum_{j=1}^{N_B} b_i^{(j)} - \alpha\, D_{\text{unsat}} - \beta\, I^i_{\text{unsold}} \tag{3.5}$$

where c is the constant production cost per unit, $\alpha\, D_{\text{unsat}}$ penalizes unmet total demand, and $\beta\, I^i_{\text{unsold}}$ penalizes leftover inventory.

$$r^{\text{raw}}_{B_j} = -\sum_{i=1}^{N_S} p_i\, b_i^{(j)} - \alpha\, D_{\text{unsat}} \tag{3.6}$$

The buyer's raw reward is the negative of its spending, with the same shortfall penalty to align all agents toward meeting overall demand.

3.3.4 LLM-Based Fairness Shaping

Two fairness scores are computed by the LLM: a Fairness-To-Buyer vector $\mathrm{FTB} = (\mathrm{FTB}_1, \ldots, \mathrm{FTB}_{N_B})$ and a Fairness-Between-Sellers scalar FBS. Both lie in $[0, 1]$, where 1 indicates the ideal fair outcome. The LLM takes the final episode ledger and outputs these normalized scores, which augment the agents' rewards:

$$R_{S_i} = r_{S_i}^{\mathrm{raw}} + \lambda_{\mathrm{buy}}\, w_B \left(\frac{1}{N_B} \sum_{j=1}^{N_B} \mathrm{FTB}_j \right) + \lambda_{\mathrm{peer}}\, w_P \left(\frac{\sum_j b_i^{(j)}}{\sum_{k,l} b_k^{(l)}} \right) \mathrm{FBS} \tag{3.7}$$

$$R_{B_j} = r_{B_j}^{\mathrm{raw}} + \lambda_{\mathrm{buy}}\, w_B\, \mathrm{FTB}_j \tag{3.8}$$

where w_B and w_P are fixed scaling coefficients, and $\lambda_{\mathrm{buy}}(t)$, $\lambda_{\mathrm{peer}}(t)$ are scheduled weights that ramp linearly from 0 to 1 over training. By design, all fairness bonuses vanish initially ($\lambda = 0$), allowing agents to first learn basic market dynamics before fairness is emphasized. Any episode yielding an invalid or unparsable LLM output is discarded to maintain a clean signal.

3.4 Learning Methodology and Experimental Results

3.4.1 Algorithm Selection: Independent Proximal Policy Optimization

The framework adopts Independent Proximal Policy Optimization (IPPO) for multi-agent training. In IPPO, each agent learns its own policy via PPO with only local observations and rewards, without a centralized critic, which keeps the learning process decentralized and scalable [16, 21]. This choice is justified by recent findings that decentralized PPO agents can achieve performance on par with or superior to centralized algorithms in cooperative tasks [21].

To validate this selection, 5,000-episode experiments compared IPPO against three baselines: MAPPO (centralized critic with global state information), MADDPG (multi-agent DDPG with centralized training), and QMIX (value factorization). Figure 3.3 presents the results. IPPO attained stable, high social welfare and fairness throughout training, whereas MADDPG

Table 3.1: Case study environment and training parameters for FairMarket-RL.

Parameter	Value / Range	Comment
Seller 1 inventory I_1	8–25 units	Sampled i.i.d. each episode
Seller 2 inventory I_2	10–30 units	More stock, same cost
Buyer demand D	20–50 units	Uniform distribution
Training horizon	20,000 episodes	On-policy IPPO
Discount factor γ	0.95	All agents
$\lambda_{\text{buy}}, \lambda_{\text{peer}}$	Linear ramps $0 \rightarrow 1$	Fairness-weight curriculum

exhibited noisy learning curves and high variance, and QMIX consistently underperformed in both dimensions.

3.4.2 Case Study Environment

The framework is grounded in a controlled case study with two competing sellers and one budget-constrained buyer. Although modest, this configuration retains the core strategic tensions of price competition, inventory allocation, and buyer welfare that motivate the LLM-guided shaping mechanism. Table 3.1 summarizes the environment and training parameters.

3.4.3 Fairness Integration via Curriculum Learning

To incorporate fairness objectives without destabilizing training, a linear λ-ramping curriculum was employed. During the full 20,000-episode training run, λ was initialized at zero and increased linearly to its maximum value. This gradual schedule ensures agents first learn to optimize efficiency (social welfare) and then progressively emphasize fairness. Figure 3.4 compares linear versus abrupt introduction of the fairness weight. With linear ramping, agents achieved smooth and reliable fairness convergence, whereas abrupt application of the full fairness weight caused slower and less stable improvements.

3.4.4 Learning Dynamics

Figure 3.5 presents the learning dynamics through two complementary views. Panel (a) shows 500-episode moving-average returns for both sellers and the buyer. In the profit-only warm-up ($\lambda = 0$), sellers raise prices at the buyer's expense. When λ_{buy} ramps up (approximately episode 2,000), sellers cut prices and buyer returns rebound. A second shift occurs once λ_{peer} saturates (approximately episode 4,500), merging the two seller profit curves as the peer signal

Table 3.2: Final evaluation results over the last 2,000 training episodes.

Metric	Result	Target
Episodes with full demand met	92.1%	$\geq 90\%$
Average FTB	0.88	≥ 0.80
Average FBS	0.87	≥ 0.80
Seller margins	24–26%	20–30%
Max seller sales share	57%	$\leq 60\%$
Buyer budget violations	0	0

equalizes gains. Panel (b) shows the LLM-driven fairness signals: FTB and FBS climb early, dip during transitions, then surpass the 0.80 target and stabilize above 0.85. The alignment of FTB and FBS turning points with the λ ramps confirms that real-time LLM feedback, not static heuristics, drives the market toward equitable outcomes.

3.4.5 Key Performance Indicators

Over the final 2,000 episodes, the system achieves strong performance across all targets. Table 3.2 summarizes the results.

When the language model critic is disabled ($\lambda_{\text{buy}} = 0$, $\lambda_{\text{peer}} = 0$), performance drops sharply: fairness stalls at 0.35–0.40, demand fulfillment slips to approximately 70%, and Seller 2 out-earns Seller 1 by approximately 35%. These degradations underscore that real-time LLM feedback is essential, not merely helpful, for maintaining equity and efficiency in peer-to-peer energy trading.

3.4.6 Ablation Studies on Learning Dynamics

Shorter 5,000-episode ablation experiments were conducted to analyze early-to-mid learning dynamics under different settings. These runs used the same codebase and hyperparameters as the main 20,000-episode runs, with the linear λ curriculum scaled to the shorter horizon (ramped from 0 to maximum over 5,000 episodes instead of 20,000). The relative performance patterns emerge early: IPPO's advantages in fairness and welfare are already apparent, MAPPO's strong reward but fairness shortfall manifests, and the off-policy methods show instability or stagnation (Figure 3.3). The consistency of these trends with longer runs indicates that the ablations are representative of true learning dynamics.

3.4.7 Fairness-Efficiency Trade-off Analysis

To characterize the fairness-efficiency trade-off, agents were trained under different fixed λ settings (held constant after the curriculum phase) to obtain the Pareto relationship between social welfare and fairness outcomes. Figure 3.6 shows a clear inverse relationship: as λ increases, the fairness index improves monotonically, but social welfare declines. This behavior reflects the fundamental trade-off in multi-agent systems between equitable outcome distribution and overall efficiency. By tuning λ, regulators or system designers can navigate along this spectrum, choosing a balance between maximizing total welfare and ensuring fair access. The linear λ-ramping methodology enables smooth convergence to any desired balance without abrupt performance drops.

3.4.8 Robustness Under Adversarial and Noisy Conditions

To probe robustness beyond the idealized setting, IPPO-only ablations were conducted while injecting two stressors during training. The first stressor was corrupted fairness feedback: Gaussian perturbations on FTB and FBS in 20% of episodes plus a 5% label-flip rate, mimicking noisy or manipulated LLM outputs. The second stressor was self-serving seller tactics in 15% of episodes, consisting of temporary overpricing at 30% and hoarding of 20% of inventory. Relative to the clean baseline, IPPO remained stable with moderate degradations: mean fairness declined from 0.88 to 0.77, normalized welfare from 0.90 to 0.83, demand met from 92% to 86%, seller profit disparity widened by 6 percentage points, and price variance increased by 18%. A simple validation gate that discards unparsable fairness responses and clips fairness bonuses contained the impact of corrupted feedback, and performance recovered once stressors subsided. These results indicate that fairness-guided IPPO is resilient to reasonable levels of LLM noise and opportunistic agent behavior.

3.5 Potential Applications in Power and Energy Systems

The rise of distributed energy resources has transformed distribution networks into decentralized webs of prosumers that both consume and generate power, creating bidirectional flows that require new market rules. Without fairness-oriented mechanisms, peer-to-peer trades can result in inefficient pricing, inequitable outcomes, and added strain on distribution system operators. FairMarket-RL addresses these challenges by combining scalable multi-agent reinforcement learning with a large language model fairness critic that shapes rewards in real time, keeping autonomous buying and selling decisions balanced for all parties.

An initial and practical application is in isolated community microgrids, where each household functions as an independent buying or selling agent. The framework promotes local energy exchange, enhances system flexibility, and reduces dependency on central grid support. Because the fairness layer and learning architecture scale linearly with agent count, the framework can expand to thousands of prosumers across an interconnected smart grid, paving the way for reliable, economically efficient, and socially equitable energy trading in future distribution systems.

3.6 Summary

This chapter presented FairMarket-RL, a language-model-guided reward-shaping framework for multi-agent reinforcement learning in peer-to-peer markets. By embedding an instruction-tuned LLM as a real-time fairness critic, the method replaces brittle hand-crafted shaping rules with dense, human-interpretable feedback. In the controlled two-seller, one-buyer case study, IPPO agents fulfilled more than 90% of demand, maintained average FTB and FBS scores of 0.88 and 0.87, and prevented profit monopolization, all while outperforming a baseline without LLM guidance. The framework demonstrated resilience under noisy LLM outputs and adversarial seller behavior, and the fairness-efficiency Pareto analysis showed that the λ-ramping mechanism enables smooth navigation along the equity-welfare spectrum.

The results in this chapter establish that an LLM can function as an effective fairness critic for multi-agent energy trading under idealized conditions: full observability, continuous action spaces, and a small participant population. A natural question follows: does this approach hold under more realistic operational conditions? Specifically, can the LLM-guided fairness mechanism scale to larger markets with partial observability, discrete price-quantity actions, realistic load and PV profiles, and heterogeneous prosumer populations? Chapter 4 directly addresses this question by extending FairMarket-RL to a continuous double auction under realistic distribution system conditions.

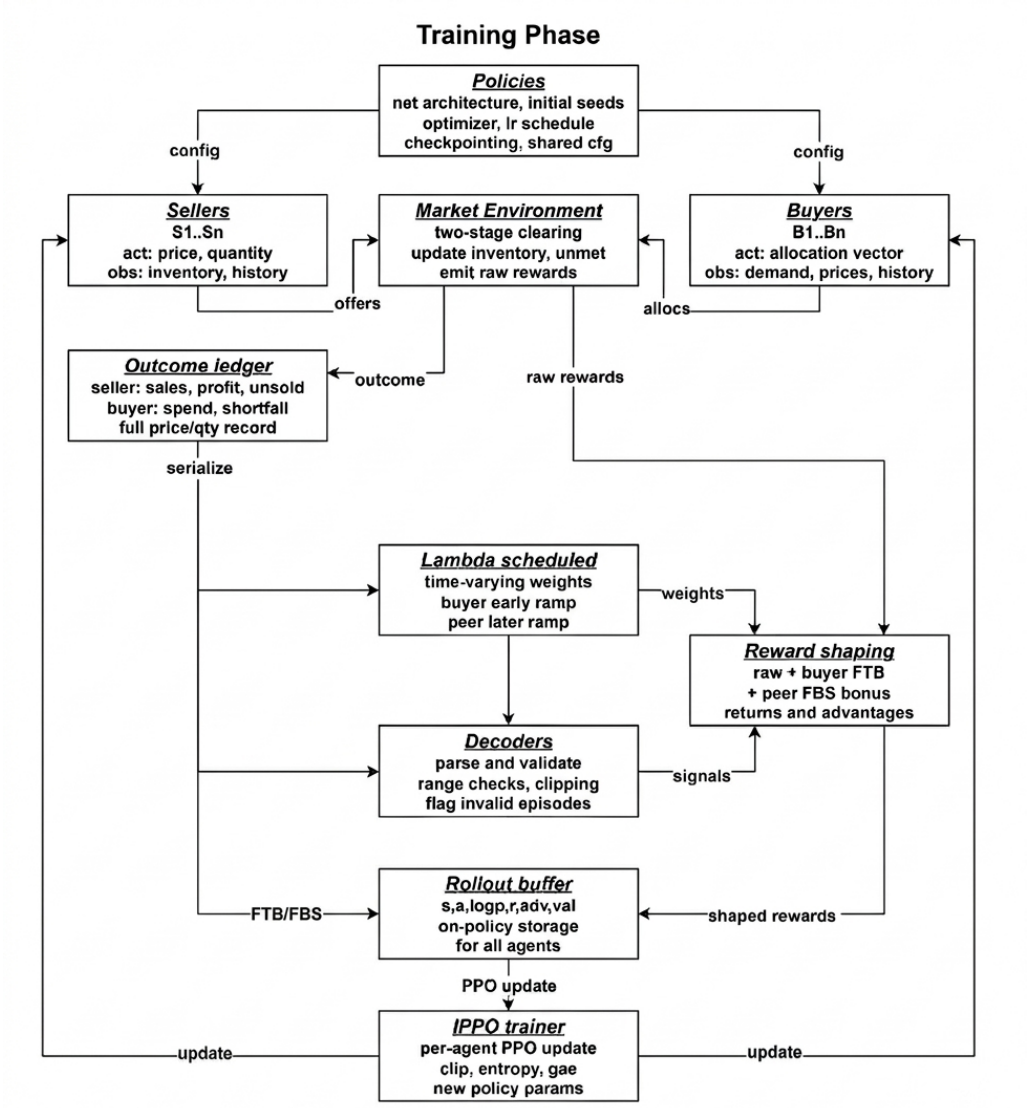

Figure 3.1: Training architecture of FairMarket-RL. Seller and buyer agents interact with the market environment. The outcome ledger is serialized and sent to the LLM fairness critic, which returns FTB and FBS scores. These are blended into agent rewards via λ-scheduled weights and stored in the rollout buffer for IPPO policy updates.

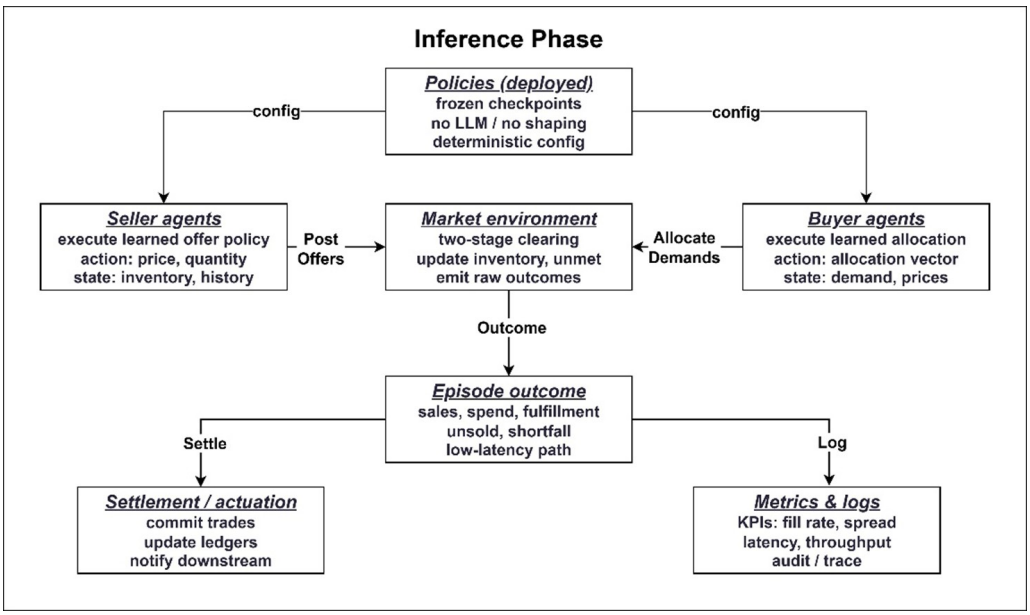

Figure 3.2: Inference architecture of FairMarket-RL. Frozen seller and buyer policies act directly in the environment to settle trades without LLM evaluation or reward shaping, yielding low-latency outcomes.



Figure 3.3: (a) Learning curves over 5,000 training episodes for IPPO and baselines showing social welfare (total reward). (b) Corresponding fairness index trajectories. IPPO achieves the best balance of welfare and fairness, while off-policy methods (MADDPG, QMIX) show instability.

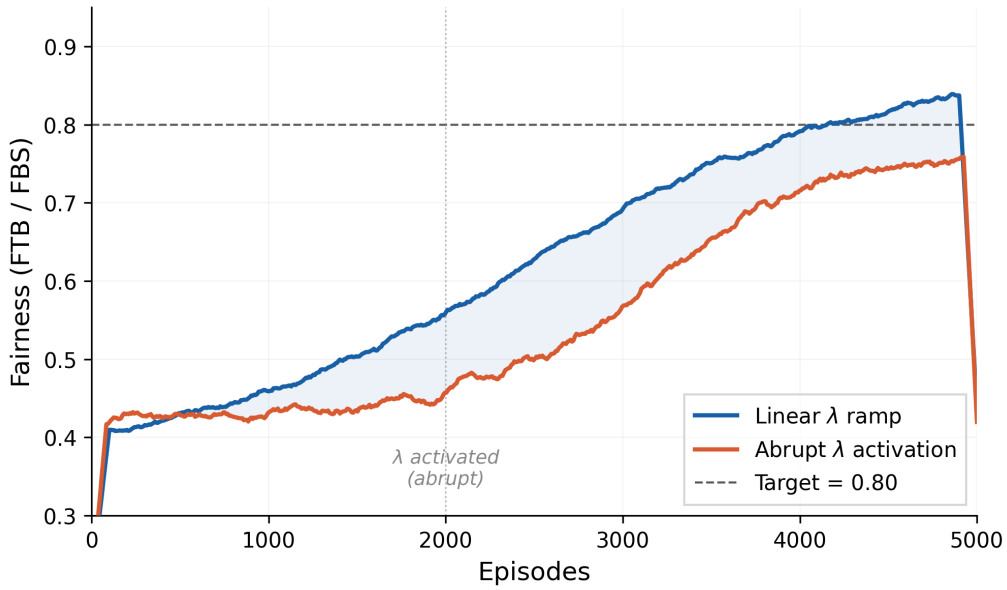

Figure 3.4: Comparison of fairness convergence behavior under a linearly increasing λ schedule versus abrupt introduction of fairness weight. Linear ramping produces smoother, more stable convergence.

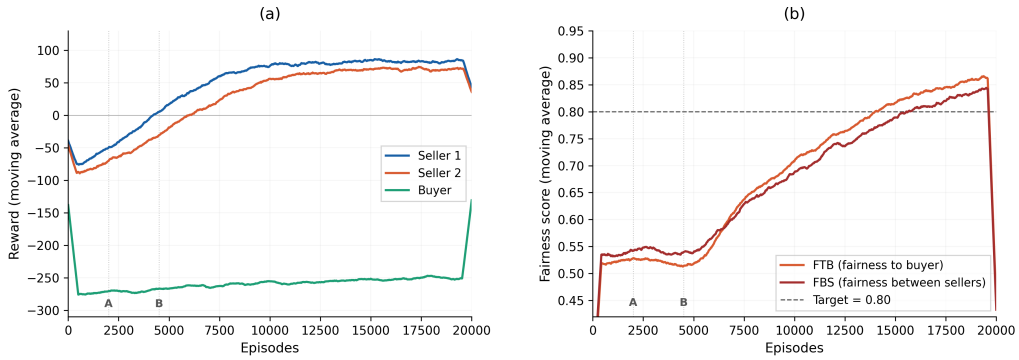

Figure 3.5: (a) Learning dynamics for the two-seller, one-buyer case study showing agent rewards over training. (b) LLM-derived fairness scores (FTB, FBS) over the same training horizon. The alignment of fairness score transitions with λ-ramp milestones confirms that LLM feedback drives equitable convergence.

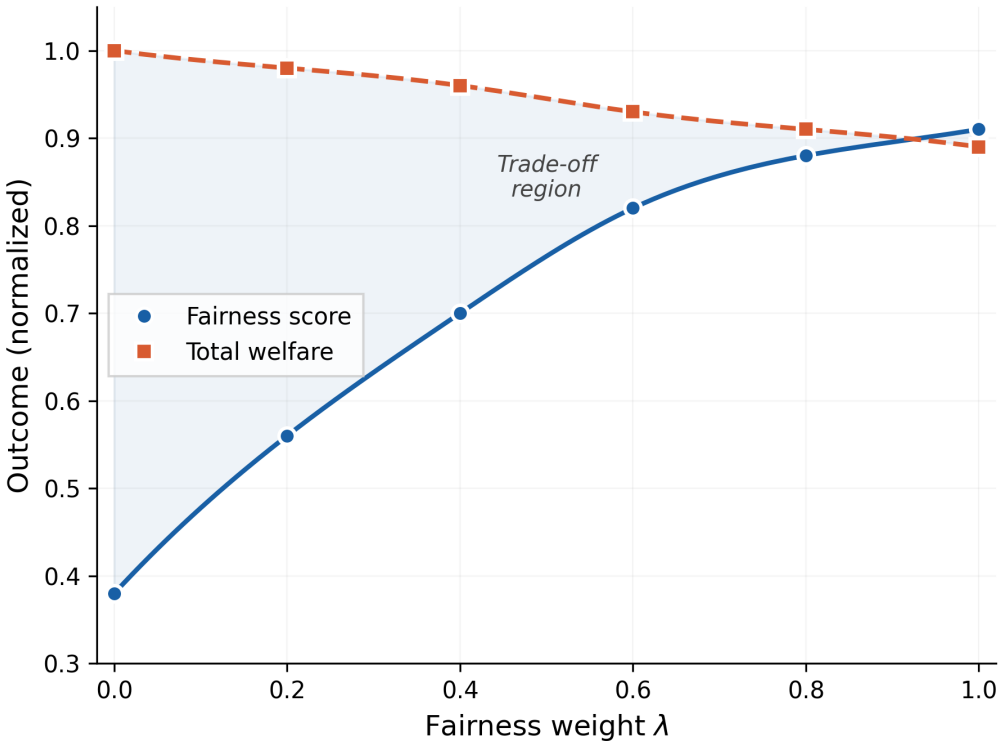

Figure 3.6: The Pareto frontier showing the fairness-efficiency trade-off. Higher values of the fairness weight λ result in higher fairness scores but lower total social welfare, reflecting the fundamental tension between equitable distribution and aggregate efficiency.

Chapter 4

Scalable Fairness Shaping with LLM-Guided Multi-Agent Reinforcement Learning for Peer-to-Peer Electricity Markets

---

Publication status. This chapter is based on the following peer-reviewed publication:

> S. Jadhav, B. Sevak, S. Das, A. Hussain, W. Su, and V.-H. Bui, "Scalable Fairness Shaping with LLM-Guided Multi-Agent Reinforcement Learning for Peer-to-Peer Electricity Markets," *Utilities Policy*, vol. 100, 102168, June 2026.
> *Available at:* `https://www.sciencedirect.com/science/article/pii/S0957178726000275`

Research question addressed. *RQ2: Does the LLM-guided fairness mechanism scale to realistic market conditions with partial observability, discrete actions, and heterogeneous participants?*

Author contributions. Shrenik Jadhav: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing (original draft and review). Birva Sevak: Methodology, Software, Investigation, Visualization. Srijita Das: Data curation, Investigation, Validation. Akhtar Hussain: Methodology, Validation, Resources. Wencong Su: Conceptualization, Supervision, Funding acquisition. Van-Hai Bui: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition (Corresponding author).

Connection to thesis. Chapter 3 established that an LLM can serve as an effective fairness critic under idealized conditions: full observability, continuous actions, and a small participant population. This chapter directly tests whether that mechanism survives the transition to realistic deployment conditions. The environment now features partial observability, discrete price-quantity actions, realistic residential load and PV profiles, battery storage, multi-slot trading horizons, and a progression from small pilot communities to larger neighborhoods. By demonstrating that the same LLM-guided reward design transfers without architectural changes, this chapter validates the scalability pillar of the trustworthiness framework. The final chapter in the contribution sequence, Chapter 5, will broaden the LLM's role from behavior shaping to decision explanation.

4.1 Introduction

Peer-to-peer energy trading is becoming central to modern distribution systems as rooftop PV installations and home energy management systems proliferate. Yet most existing market designs and reinforcement learning frameworks emphasize efficiency or private profit and offer little real-time guidance to ensure equitable outcomes under the uncertainty introduced by variable renewable generation, heterogeneous consumption patterns, and partial market information [7, 8]. The FairMarket-RL framework introduced in Chapter 3 demonstrated that an LLM fairness critic can shape equitable trading behavior in a controlled setting. However, that proof of concept operated under several simplifying assumptions: full observability (all agents see the complete market state), continuous action spaces, a single trading round per episode, and a minimal two-seller, one-buyer population. Real P2P electricity markets differ from these conditions in fundamental ways.

This chapter presents the scalable extension of FairMarket-RL to address these limitations. The framework is reformulated around a continuous double auction (CDA) mechanism in which prosumers and consumers submit discrete price-quantity bids and asks that are matched through greedy order-book clearing [46]. Each agent operates under partial observability, seeing only its own load and PV forecasts plus public market signals (weather, time of day), without access to the global order book or other agents' states. The trading horizon extends across multiple hourly slots within multi-day episodes, requiring policies that reason about temporal dependencies. Load and PV generation profiles are modeled with realistic res-

idential templates modulated by stochastic weather and log-normal noise. Some prosumers are equipped with battery energy storage systems that introduce charge/discharge decisions subject to physical constraints.

The LLM fairness critic is extended from two metrics (FTB, FBS) to three: Fairness-to-Grid (FTG), measuring the degree to which demand is met through local P2P trades rather than grid procurement; Fairness-Between-Sellers (FBS), measuring the equality of trading volume across sellers; and Fairness-of-Pricing (FPP), measuring the tightness of clearing prices around the market median. These slot-level scores are integrated into agent rewards via ramped coefficients with tunable scaling, so that fairness guidance complements rather than overwhelms economic incentives.

The framework is validated across a progression of three experimental scenarios: a 30-day pilot community with 5 participants, a 90-day scaled community with 13 participants, and a 90-day real-world community using a European residential dataset with battery-equipped prosumers. The key finding is that the same LLM-guided reward design and policy architecture transfer from the small pilot to the larger deployments without architectural changes, demonstrating the scalable fairness shaping that gives this chapter its title.

4.2 Market Environment and Agent Modeling

4.2.1 Market Structure

The P2P market consists of m prosumers (households with PV generation capability), n pure consumers (households without generation), and a single price-taking grid agent that settles residual imbalances at a fixed retail tariff. The market operates as a continuous double auction: at each hourly slot, prosumers with surplus generation post sell orders (asks) and prosumers or consumers with net demand post buy orders (bids). The CDA clears greedily by matching the highest bid to the lowest ask whenever the trade condition $p_{\text{buy}} \geq p_{\text{sell}}$ holds, continuing until no further matchable orders remain. Residual demand after CDA clearing is served by the grid agent at the retail tariff.

4.2.2 Environment Dynamics

Each learning episode spans D days with $H = 24D$ hourly decision slots. At each slot t, identified by hour-of-day h_t and day index d_t, a public daily weather flag $\omega_{d_t} \in \{1, 0\}$ (sunny or

cloudy) modulates generation via an intensity factor:

$$\kappa_{d_t} = \alpha + (1-\alpha)\cdot\omega_{d_t}, \quad \alpha \in (0,1) \tag{4.1}$$

True electricity demand and PV output are generated from scaled canonical 24-hour templates with household-specific peaks, daily weather intensity, and log-normal noise:

$$L_{a,t} = \bar{L}_a \cdot \varphi_{\text{load}}(h_t) \cdot \eta^L_{a,t} \tag{4.2}$$

$$G_{a,t} = o_a \cdot \bar{G}_a \cdot \kappa_{d_t} \cdot \varphi_{\text{pv}}(h_t) \cdot \eta^G_{a,t} \tag{4.3}$$

where $\bar{L}_a$ and $\bar{G}_a$ are household-specific peak values, $\varphi_{\text{load}}(\cdot)$ and $\varphi_{\text{pv}}(\cdot)$ are normalized diurnal profile templates, $o_a \in \{0,1\}$ indicates PV ownership, and $\eta^L_{a,t}, \eta^G_{a,t}$ are log-normal noise terms drawn from $\mathcal{N}(1, 0.05^2)$.

4.2.3 Battery Energy Storage

For prosumers equipped with battery storage ($s_a = 1$), the state of charge $B_{a,t}$ (kWh) evolves according to:

$$B_{a,t+1} = \text{clip}\left(B_{a,t} + \eta_c \cdot q^{\text{ch}}_{a,t} - \frac{q^{\text{dis}}_{a,t}}{\eta_d},\ 0,\ B^{\max}_a\right) \tag{4.4}$$

where η_c and η_d are charge and discharge efficiencies, $B^{\max}_a$ is the battery capacity, and power-rate limits are enforced:

$$0 \le q^{\text{ch}}_{a,t} \le P^{\text{ch,max}}_a \cdot \Delta t, \qquad 0 \le q^{\text{dis}}_{a,t} \le P^{\text{dis,max}}_a \cdot \Delta t \tag{4.5}$$

The storage-adjusted net demand becomes $D^{\text{net}}_{a,t} = L_{a,t} - G_{a,t} - s_a \cdot q^{\text{dis}}_{a,t} + s_a \cdot q^{\text{ch}}_{a,t}$, which determines whether the prosumer posts a buy or sell order.

4.3 Multi-Agent Game Formulation

The market is formalized as a partially observable stochastic game:

$$G = \left\langle S,\ \{A_{s_i}\}_{i=1}^{m},\ \{A_{b_j}\}_{j=m+1}^{m+n},\ T,\ R \right\rangle \tag{4.6}$$

4.3.1 State and Observation Spaces

The global state S at slot t includes true loads $L_{a,t}$, PV outputs $G_{a,t}$, the weather flag ω_{d_t}, time indices h_t and d_t, the current order book, and battery states of charge for storage-equipped prosumers. Critically, the decision problem is partially observable: each household observes only its own forecasts and public signals, not the global order book or other agents' states. The local observation for prosumer i is:

$$s_i^t = \left[L'_{i,t} - G'_{i,t},\ L'_{i,t},\ G'_{i,t},\ s_i \cdot B_{i,t},\ h_t,\ \kappa_{d_t},\ d_t\right] \tag{4.7}$$

where $L'_{i,t}$ and $G'_{i,t}$ are one-step-ahead forecasts modeled with Gaussian noise ($\varepsilon \sim \mathcal{N}(1, 0.05^2)$), providing the agent with imperfect but informative estimates of its own energy position.

4.3.2 Action Spaces

Prosumers choose discrete price-quantity pairs from fixed menus. A prosumer with net surplus posts a sell order (ask) with quantity bounded by the forecast surplus: $0 \leq q_{i,t}^{\text{ask}} \leq \max(0, \hat{G}_{i,t} - \hat{L}_{i,t})$. A prosumer with net deficit posts a buy order (bid) bounded similarly. Storage-equipped prosumers additionally choose charge, discharge, or idle decisions subject to the SOC and power-rate constraints in Equations 4.4–4.5. Consumers choose purchase quantities at the grid tariff. All actions must satisfy physical feasibility constraints enforced by the environment.

4.3.3 Transition and Clearing

The transition function T is deterministic given the stochastic environment inputs. At each slot, the CDA clears by greedily matching the highest buy price to the lowest sell price, executing trades, and updating inventories. The grid agent settles any residual demand or surplus at the retail tariff.

4.4 LLM-Based Fairness Shaping

4.4.1 Fairness Metrics

After each trading slot, the LLM processes a compact ledger of the cleared trades and returns three normalized fairness scores in $[0, 1]$:

1. **Fairness-to-Grid (FTG):** Measures the degree to which demand is met through local P2P trades rather than grid procurement. FTG equals 1 when all demand is met via P2P and 0 when demand is entirely grid-supplied. This metric incentivizes local energy self-sufficiency and reduces grid dependency.

2. **Fairness-Between-Sellers (FBS):** Measures the equality of trading volume across active sellers. FBS equals 1 when quantities are evenly shared among all sellers and decreases as any single seller dominates the market. This metric prevents market concentration.

3. **Fairness-of-Pricing (FPP):** Measures the tightness of clearing prices around the market median. FPP equals 1 when all clearing prices are identical and decreases as price dispersion increases. This metric discourages price manipulation and ensures consistent pricing across transactions.

The expansion from two metrics (FTB, FBS in Chapter 3) to three reflects the richer structure of the CDA market, where pricing dynamics and grid dependency become first-order concerns that the simpler two-stage mechanism did not expose.

4.4.2 Shaped Reward Function

For each prosumer i, the slot-level reward integrates raw economic profit with the three fairness signals:

$$R_{i,t} = \pi_{i,t} + \lambda_{\text{grid}}(e) \cdot \beta_{\text{grid}} \cdot \text{FTG}_t + \lambda_{\text{price}}(e) \cdot \beta_{\text{price}} \cdot \text{FPP}_t + \lambda_{\text{peer}}(e) \cdot \beta_{\text{peer}} \cdot \text{FBS}_t \cdot \left(\frac{q_{i,t}^{\text{sold}}}{\sum_{k=1}^{m} q_{k,t}^{\text{sold}}} \right) \tag{4.8}$$

where $\pi_{i,t}$ is the raw hourly profit (revenue from sales minus cost of purchases), $\beta_\bullet$ are fixed scaling factors that control the relative magnitude of each fairness bonus, and $\lambda_\bullet(e)$ are episode-dependent ramping coefficients that follow a three-phase schedule:

$$\lambda_\bullet(e) = \begin{cases} 0 & \text{if } e < e_\bullet^{\text{start}} \\ \dfrac{e - e_\bullet^{\text{start}}}{e_\bullet^{\text{full}} - e_\bullet^{\text{start}}} & \text{if } e_\bullet^{\text{start}} \leq e < e_\bullet^{\text{full}} \\ 1 & \text{if } e \geq e_\bullet^{\text{full}} \end{cases} \tag{4.9}$$

This three-phase schedule (inactive, linear ramp, full strength) generalizes the linear curriculum from Chapter 3 by allowing each fairness dimension to be introduced at different

training stages. In practice, λ_{grid} activates first (encouraging P2P participation), followed by λ_{peer} (balancing seller volumes), and finally λ_{price} (stabilizing prices). This staggered introduction prevents the agents from being overwhelmed by simultaneous fairness objectives.

4.4.3 Raw Reward Functions

The raw economic rewards, before fairness shaping, differ by agent type. For prosumer i, the hourly profit is:

$$\pi_{i,t} = \sum_{(i \to \kappa)} p \cdot q - \sum_{(\kappa \to i)} p \cdot q \tag{4.10}$$

representing revenue from energy sold minus cost of energy purchased, summed over all trades involving agent i at slot t. The total raw return over the episode is $r_i^{\text{raw}} = \sum_{t=0}^{H-1} \pi_{i,t}$. For pure consumer j, the hourly cost is $\chi_{j,t} = \sum_{(\kappa \to j)} p \cdot q$, with total raw return $r_j^{\text{raw}} = -\sum_{t=0}^{H-1} \chi_{j,t}$.

4.5 Training and Policy Architecture

All agents are trained using Proximal Policy Optimization (PPO) [16] with independent policies, consistent with the IPPO approach validated in Chapter 3. Gradient clipping and the PPO clip ratio ensure policy-update stability across the multi-slot horizon. The policy network takes the local observation vector (Equation 4.7) as input and outputs a probability distribution over discrete price-quantity actions. A separate value network estimates the state-value function for advantage computation.

The key architectural design choice is that the policy class and LLM-guided reward structure remain unchanged across all three experimental scenarios. Scaling from 5 participants to 13 participants to a mixed-asset real-world community requires no modifications to the network architecture, prompt structure, or fairness integration mechanism. This architectural invariance is what enables scalable fairness shaping: the framework is trained on a small pilot and transferred to larger deployments without redesign.

4.6 Experimental Validation

The framework is validated across three progressively complex scenarios, each testing a different aspect of scalability.

4.6.1 Case 1: 30-Day Pilot Community

The pilot consists of 3 prosumers and 2 consumers trading over a 30-day horizon (720 hourly slots). This small-scale setting verifies that the extended CDA mechanism with three fairness metrics converges to equitable outcomes. Results show that 54% of energy exchanges occur through P2P trades, with the remaining 46% served by the grid. Prosumer rewards stabilize within the first training phase, and fairness metrics reach strong levels: FTG in the range of 0.80–0.85, FBS approximately 0.90, and FPP consistently above 0.95. Figure 4.1 presents the training dynamics for this scenario.

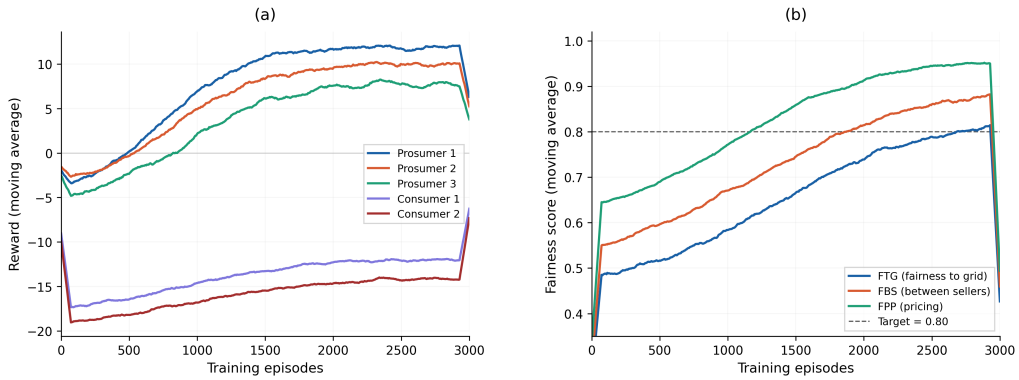

Figure 4.1: Training dynamics for the 30-day pilot community (3 prosumers, 2 consumers). Agent rewards converge within the early training phase. Fairness scores reach target levels: FTG 0.80–0.85, FBS $\approx$ 0.90, FPP $>$ 0.95.

4.6.2 Case 2: 90-Day Scaled Community

The scaled community increases participation to 10 prosumers and 3 consumers over a 90-day horizon (2,160 hourly slots), testing whether fairness properties are maintained as the market grows. The P2P trade share remains similar at 53.6%, with grid trades comprising 46.4%. FPP stays above 0.95, FBS near 0.90, and FTG improves to 0.72, reflecting the increased difficulty of achieving full grid independence with more participants. Consumers reduce their electricity bills by 27–28% compared to grid-only procurement, demonstrating meaningful economic benefit. Figure 4.2 presents the results.

4.6.3 Case 3: 90-Day Real-World Community

The most demanding scenario uses a European residential dataset with a heterogeneous asset mix: 3 battery-equipped PV prosumers, 7 PV-only prosumers, and 3 consumers. The 90-day

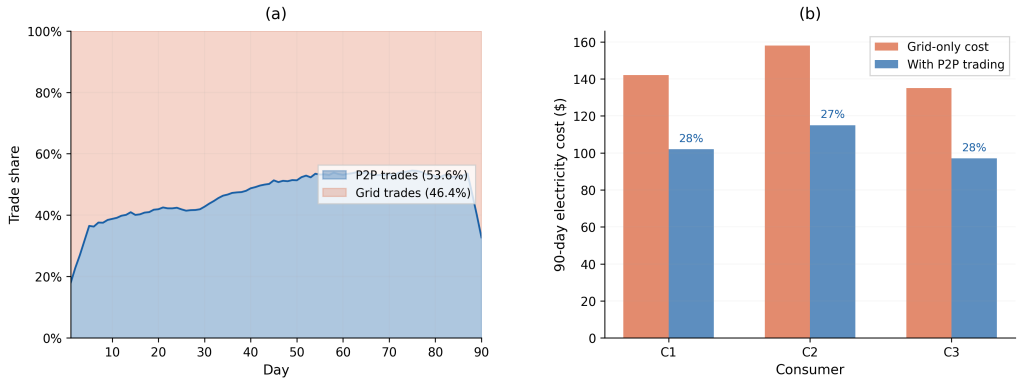

Figure 4.2: Results for the 90-day scaled community (10 prosumers, 3 consumers). P2P trade share remains at 53.6%. Consumer bills are reduced by 27–28% relative to grid-only procurement.

simulation tests whether policies trained on synthetic profiles generalize to real consumption and generation patterns without retraining. Battery-equipped prosumers achieve the highest profits by strategically timing charge/discharge decisions around peak pricing periods. PV-only prosumers maintain positive margins through direct energy sales. Consumers sustain bill reductions exceeding 30% compared to grid-only procurement. Grid viability is preserved throughout, with approximately $379 in net grid revenue over the 90-day period. Figure 4.3 presents the results.

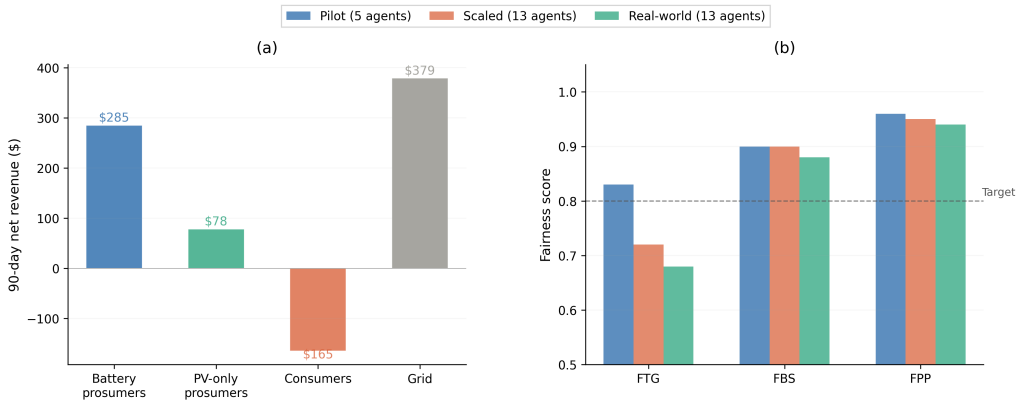

Figure 4.3: Results for the 90-day real-world community using European residential data. Battery-equipped prosumers achieve highest profits, consumers save over 30%, and grid viability is maintained with $379 net revenue.

4.6.4 Computational Scalability

The 90-day real-world simulation completed in approximately three hours on a standard 12-core CPU, demonstrating that the framework is computationally tractable for practical deployment. The LLM evaluation cost scales linearly with the number of trading slots (one LLM call per slot during training), and the IPPO training cost scales linearly with agent count because each agent maintains an independent policy without centralized critic computation.

4.7 Sensitivity Analysis

To assess robustness under varying operating conditions, sensitivity analyses were conducted by perturbing two key environmental parameters: solar PV availability ($\pm 20\%$) and aggregate electricity demand ($\pm 10\%$). Under reduced PV availability, the P2P trade share decreases as prosumers have less surplus to sell, but fairness metrics remain within acceptable bounds because the LLM critic adapts its evaluation to the reduced trading volume. Under increased demand, grid procurement increases proportionally, but consumer bill savings are maintained because the CDA clearing mechanism adjusts prices to reflect scarcity. These results indicate that the framework is robust to the range of operating conditions expected in real distribution systems.

Figure 4.4 presents the sensitivity analysis results.

4.8 Comparison with Chapter 3

Table 4.1 summarizes the key differences between the proof-of-concept framework in Chapter 3 and the scalable extension presented in this chapter. Every simplifying assumption from the initial study has been relaxed, and the LLM-guided fairness mechanism continues to produce equitable outcomes across all tested conditions.

4.9 Summary

This chapter extended FairMarket-RL from a proof of concept to a scalable framework validated under realistic P2P electricity market conditions. The continuous double auction mechanism with partial observability and discrete actions replaced the simplified two-stage trading of Chapter 3. Three fairness metrics (FTG, FBS, FPP) replaced two, reflecting the richer dynamics of realistic markets. Realistic residential load and PV profiles with stochastic weather mod-

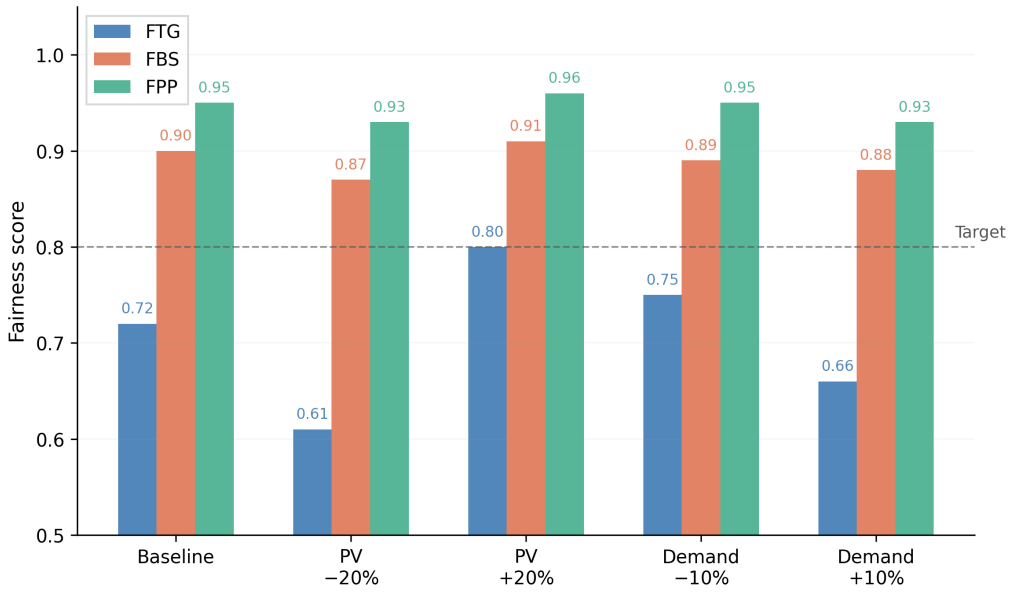

Figure 4.4: Sensitivity of fairness and economic metrics to ±20% PV availability and ±10% aggregate demand perturbations. Fairness and grid viability are maintained across all conditions.

ulation replaced static demand sampling. Battery energy storage with physical constraints added a new dimension of strategic decision-making. Across all three validation scenarios, the framework shifted energy exchanges toward local P2P trades, lowered consumer costs by 27–30% relative to grid-only procurement, sustained strong fairness across participants, and preserved grid viability.

The central finding is architectural: the same LLM-guided reward design and IPPO policy class that worked for a 3-agent pilot transferred without modification to a 13-agent community with real-world data and battery storage. This demonstrates that scalable fairness shaping through LLM-based critics is achievable in practice, not merely in controlled experiments.

Chapters 3 and 4 have demonstrated that LLMs can shape what *RL agents learn to value, producing fairer outcomes in energy trading at both pilot and community scale. However, for real-world deployment, grid operators also need to understand* why *autonomous agents make their decisions. Can the LLM companion paradigm be extended from behavior shaping to decision explanation? Chapter 5 addresses this question by applying LLMs to a fundamentally different task: generating natural language explanations of RL-based voltage control decisions, completing the trustworthiness triad of fairness, scalability, and explainability.*

Table 4.1: Comparison of the proof-of-concept (Chapter 3) and scalable (this chapter) FairMarket-RL frameworks.

Dimension	**Chapter 3 (Proof of Concept)**	**This Chapter (Scalable)**
Market mechanism	Two-stage offer/allocate	Continuous double auction
Observability	Full (global state)	Partial (local forecasts only)
Action space	Continuous price-quantity	Discrete menus
Participants	2 sellers, 1 buyer	Up to 10 prosumers, 3 consumers
Trading horizon	Single episode	Multi-day, multi-slot (720–2160 slots)
Generation profiles	Static, sampled i.i.d.	Realistic diurnal with weather and noise
Storage	None	Battery SOC with charge/discharge
Fairness metrics	2 (FTB, FBS)	3 (FTG, FBS, FPP)
λ schedule	Single linear ramp	Staggered three-phase per metric
Validation	Single case study	Pilot $\rightarrow$ scaled $\rightarrow$ real-world data

Chapter 5

XRL-LLM: Explainable Reinforcement Learning Framework for Voltage Control

---

Publication status. This chapter is based on the following manuscript:

> S. Jadhav, B. Sevak, and V.-H. Bui, "XRL-LLM: Explainable Reinforcement Learning Framework for Voltage Control," *Energies*, 19(7), 1789, 2026.
> *Available at:* `https://www.mdpi.com/1996-1073/19/7/1789`

Research question addressed. *RQ3: Can LLMs generate physically grounded, actionable explanations of RL voltage control decisions that are superior to template-based and numerical baselines?*

Author contributions. Shrenik Jadhav: Conceptualization, Methodology, Software, Formal analysis, Visualization, Data curation, Writing (original draft and review). Birva Sevak: Methodology, Software, Investigation, Visualization, Writing (review). Van-Hai Bui: Conceptualization, Methodology, Supervision, Project administration, Validation, Funding acquisition, Writing (review). (Corresponding author).

Connection to thesis. Chapters 3 and 4 demonstrated that LLMs can shape *what* RL agents learn to value, producing fairer outcomes in P2P energy trading. This chapter broadens the LLM companion paradigm to a fundamentally different role: explaining *why* RL agents make their decisions. The application domain also shifts from energy trading to voltage control

on a distribution network, demonstrating that the LLM-RL companion concept generalizes across power system tasks. Together with the preceding chapters, this work completes the three-pillar trustworthiness framework: fairness (Chapter 3), scalability (Chapter 4), and explainability (this chapter).

5.1 Introduction

Reinforcement learning agents are increasingly deployed for voltage control in power distribution networks, achieving strong performance in maintaining bus voltages within statutory limits [2, 5, 4]. However, their opaque decision-making creates a significant trust barrier: grid operators who must supervise these agents cannot understand why a particular capacitor bank was activated or a specific tap changer was adjusted, limiting their ability to verify decisions, anticipate consequences, or intervene when conditions deviate from training scenarios [11].

Explainable artificial intelligence offers a path toward addressing this opacity. Post-hoc methods such as SHAP [24] assign each input feature a Shapley value quantifying its contribution to the model's output. In the power systems domain, Zhang et al. [25] applied DeepSHAP to DRL agents for emergency control, demonstrating that Shapley-value attributions can reveal which grid state variables influence an agent's actions. However, their approach presents SHAP values only as raw numerical vectors without translating them into human-understandable language. A numerical attribution telling an operator that "Bus 17 voltage has a SHAP value of 0.342" conveys importance but not meaning: it does not explain what physical condition triggered the decision, how the chosen action addresses it, or what the operator should monitor afterward.

Recent advances in LLMs have opened new possibilities for bridging this gap. Slack et al. [30] demonstrated that LLMs can explain ML predictions through natural language conversations, and Krishna et al. [31] showed at NeurIPS 2023 that translating SHAP attributions into natural language rationales improves downstream task performance. However, these approaches have not been applied to RL policies or power system applications.

This chapter presents XRL-LLM, a framework that generates natural language explanations for RL voltage control decisions by combining game-theoretic feature attribution (KernelSHAP) with LLM reasoning grounded in power systems domain knowledge. The frame-

work is validated on the IEEE 33-bus radial distribution network using a PPO agent trained for voltage regulation, with explanation quality assessed through G-Eval [68] automated scoring, physics-grounded counterfactual verification, and human validation.

5.2 Methodology

The XRL-LLM framework transforms numerical attribution outputs from a trained RL agent into natural language explanations that power system operators can interpret and act upon. The framework consists of five components: (1) a voltage control environment built on the IEEE 33-bus radial distribution network, (2) a PPO agent trained for voltage regulation, (3) a KernelSHAP attribution pipeline that identifies the input features driving each control decision, (4) a large language model that translates those attributions into operator-readable explanations with domain-specific context, and (5) a counterfactual verification procedure that validates explanation faithfulness through physics-based simulation. Figure 5.1 illustrates the overall architecture.

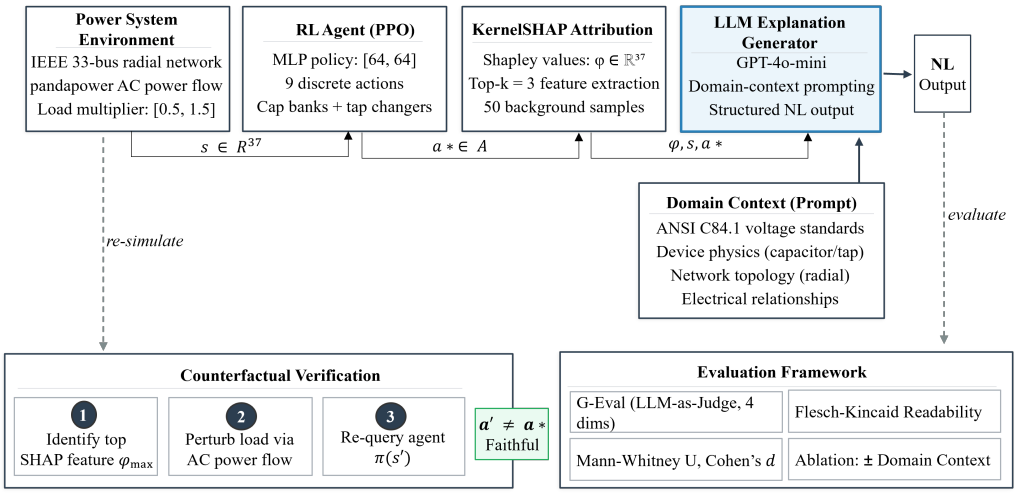

Figure 5.1: Overview of the XRL-LLM framework. The PPO agent observes bus voltages and device states from the IEEE 33-bus network, selects a control action, and KernelSHAP identifies the most influential state features. These attributions are processed by a domain-context-engineered LLM to generate operator-facing natural language explanations. A counterfactual verification branch validates explanation faithfulness through physics-based simulation.

Table 5.1: Mapping of discrete actions to physical control operations on the IEEE 33-bus network.

Action	Device	Location	Operation
0	None	—	Do nothing
1	Capacitor Bank	Bus 12	Switch ON
2	Capacitor Bank	Bus 12	Switch OFF
3	Capacitor Bank	Bus 24	Switch ON
4	Capacitor Bank	Bus 24	Switch OFF
5	Tap Changer	Bus 6	Tap UP (+1 step)
6	Tap Changer	Bus 6	Tap DOWN (-1 step)
7	Tap Changer	Bus 28	Tap UP (+1 step)
8	Tap Changer	Bus 28	Tap DOWN (-1 step)

5.2.1 Power System Environment

The test system is the IEEE 33-bus radial distribution network [19], implemented in pandapower [18]. The network consists of 33 buses connected in a radial topology with a single substation at Bus 0 serving as the slack bus. To create diverse operating conditions during training, each episode samples a load multiplier uniformly from $[0.5, 1.5]$ times the nominal demand, spanning light loading through critical loading where multiple buses experience voltage sag below the ANSI C84.1 lower limit of 0.95 p.u.

At each time step, the agent observes a state vector $\mathbf{s} \in \mathbb{R}^{37}$ containing 33 bus voltage magnitudes in per unit and 4 device state variables:

$$\mathbf{s} = \left[V_0, V_1, \ldots, V_{32}, c_{12}, c_{24}, \tau_1, \tau_2\right]^\top \tag{5.1}$$

where V_i denotes the voltage magnitude at bus i, c_{12} and c_{24} are binary indicators for the capacitor bank status at Bus 12 and Bus 24, and τ_1 and τ_2 represent normalized tap changer positions at Bus 6 and Bus 28.

The agent selects from a discrete set of nine control actions corresponding to switching operations on four voltage regulation devices: two capacitor banks (at Bus 12 and Bus 24) that inject reactive power to raise local voltage, and two on-load tap changers (OLTCs) at Bus 6 and Bus 28 that adjust transformer turns ratios. Table 5.1 maps each action index to its physical meaning.

The reward function guides the agent toward maintaining all bus voltages within the ANSI

Table 5.2: PPO training hyperparameters for the voltage control agent.

Parameter	Value
Learning rate	3×10^{-4}
Rollout length (n_{steps})	256
Mini-batch size	64
Optimization epochs per rollout	10
Discount factor (γ)	0.99
GAE parameter (λ)	0.95
Clip range (ε)	0.2
Entropy coefficient	0.02
Policy network	[64, 64] with Tanh
Value network	[64, 64] with Tanh
Random seed	42

C84.1 Range A limits of $[0.95, 1.05]$ p.u.:

$$r(\mathbf{s}, a) = -\alpha \sum_{i=1}^{33} |V_i - V_{\text{ref}}| - \beta \sum_{i=1}^{33} \mathbb{1}[V_i \notin [0.95, 1.05]] - \gamma_{\text{sw}} \cdot c(a) \tag{5.2}$$

where $V_{\text{ref}} = 1.0$ p.u. is the nominal voltage, α weights the total voltage deviation, β imposes a heavy penalty for each bus violating statutory limits, and $\gamma_{\text{sw}} \cdot c(a)$ penalizes unnecessary switching to discourage equipment wear.

5.2.2 Reinforcement Learning Agent

The agent is trained using Proximal Policy Optimization (PPO) [16] implemented through Stable-Baselines3. The policy is parameterized as a multilayer perceptron with two hidden layers of 64 neurons each using Tanh activations. A separate value network with the same architecture estimates the state-value function. The clipped PPO objective constrains the magnitude of policy updates:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta) \hat{A}_t,\ \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon) \hat{A}_t \right) \right] \tag{5.3}$$

where $r_t(\theta)$ is the probability ratio between updated and previous policies, $\hat{A}_t$ is the generalized advantage estimate, and $\varepsilon = 0.2$ is the clipping parameter. The agent is trained for 50,000 environment interactions. Table 5.2 lists the complete set of training hyperparameters.

5.2.3 SHAP Attribution Pipeline

Once the agent is trained, KernelSHAP [24] computes an attribution vector $\phi \in \mathbb{R}^{37}$ for each decision, where each ϕ_i quantifies the contribution of feature i to the probability assigned to the chosen action relative to a baseline expectation:

$$\sum_{i=1}^{37} \phi_i = f(\mathbf{s})_{a^*} - \mathbb{E}_{\mathbf{s}' \sim \mathscr{D}}[f(\mathbf{s}')_{a^*}] \tag{5.4}$$

The background dataset consists of 50 state observations from random policy rollouts, summarized to 15 representative centroids via k-means clustering. Each KernelSHAP computation uses 80 perturbation samples per explanation. The top $k = 3$ features by absolute attribution magnitude are extracted for each decision, forming the factual basis for the subsequent natural language explanation.

5.2.4 LLM Explanation Generation

The central contribution of this work is the translation of numerical SHAP attributions into natural language explanations that operators can understand and act upon. The LLM (GPT-4o-mini) receives a structured prompt containing four components: (1) the grid state with voltage magnitudes at every bus and explicit identification of violations, (2) the agent's selected action translated into its physical meaning, (3) the top three SHAP features ranked by absolute attribution value, and (4) a domain context encoding ANSI C84.1 voltage limits, the physical effect of each control device, the radial topology of the IEEE 33-bus network, and the electrical relationship between device locations and affected buses.

The domain context is essential because it enables the LLM to connect numerical attributions to physical mechanisms. Without it, the LLM can describe which features had high SHAP values but cannot explain *why* a particular bus voltage matters or *how* the chosen action will address the issue. The LLM is instructed to generate explanations structured as: (1) situation assessment describing the current voltage condition, (2) action justification linking the chosen action to SHAP-identified features through a physical causal chain, (3) expected impact on bus voltages, and (4) monitoring guidance identifying buses that remain at risk.

Two baseline methods are defined for comparison. **Raw SHAP** presents the top three attribution values directly as a ranked numerical list with bus indices and SHAP scores. **Template NLG** applies deterministic rules to convert SHAP features into sentences using predefined templates (e.g., "The primary factor is Bus 17 voltage at 0.93 p.u., which is below the 0.95 p.u. limit.").

5.2.5 Counterfactual Verification

A key challenge in evaluating explanations is determining whether the identified important features are genuinely causal drivers or merely statistical correlates. The counterfactual verification procedure addresses this through physics-grounded simulation. For each explanation, the feature with the highest absolute SHAP attribution (corresponding to the voltage at some bus j) is identified. A counterfactual scenario is constructed by modifying the load at bus j in pandapower and re-solving the full nonlinear AC power flow to obtain a physically consistent counterfactual state $\mathbf{s}'$. The agent is then queried on $\mathbf{s}'$:

$$\text{Faithful}(j) = \begin{cases} 1 & \text{if } \pi(\mathbf{s}') \neq a^* \\ 0 & \text{if } \pi(\mathbf{s}') = a^* \end{cases} \tag{5.5}$$

If the agent changes its action when the identified important feature is corrected, the SHAP attribution is confirmed as causally faithful. The critical distinction from naive perturbation is that the counterfactual state respects Kirchhoff's laws and all network constraints, making the verification physically meaningful.

5.2.6 Evaluation Framework

Explanation quality is assessed using G-Eval [68], which employs GPT-4o-mini as an automated judge scoring each explanation on a 1–5 Likert scale across four dimensions: Accuracy (factual correctness), Actionability (whether the explanation suggests concrete interventions), Completeness (coverage of relevant features, causes, and consequences), and Conciseness (information density without verbosity). Readability is measured using the Flesch-Kincaid Grade Level [71]. All comparisons are tested for statistical significance using the Mann-Whitney U test with Cohen's d effect sizes.

5.2.7 Experimental Protocol

The trained PPO agent is evaluated across four load levels: light ($0.8\times$ nominal), normal ($1.0\times$), heavy ($1.2\times$), and critical ($1.35\times$). For each load level, evaluation scenarios are generated and the agent's actions, SHAP attributions, and explanations from all three methods are recorded. Table 5.3 reports the computational cost of each pipeline component.

Table 5.3: Computational cost breakdown for each XRL-LLM pipeline component.

Component	Phase	Time	Frequency
PPO training (50k steps)	Offline	$\sim$5 min	Once
AC power flow (pandapower)	Online	$\sim$45 ms	Per step
PPO policy inference	Online	$<$1 ms	Per step
KernelSHAP (k=3, 80 samples)	Online	$\sim$3.2 s	Per decision
LLM explanation (GPT-4o-mini)	Online	$\sim$1.8 s	Per decision
Counterfactual verification	Online	$\sim$0.4 s	Per decision
Total online latency		**$\sim$5.5 s**	

Table 5.4: RL agent voltage violation reduction across load levels (50 episodes).

Load Level	Episodes	Initial Viol.	Final Viol.	Reduction (%)
Light	7	1.6	0.0	71.4
Normal	26	7.8	0.2	96.8
Heavy	8	15.4	1.0	92.7
Critical	9	22.6	3.2	86.1
Overall	**50**	**10.7**	**0.9**	**90.5**

5.3 Results and Discussion

5.3.1 RL Agent Training and Voltage Control Performance

The PPO agent was trained on the IEEE 33-bus network under stochastic load conditions ranging from 0.8$\times$ to 1.35$\times$ nominal demand. Figure 5.2 presents the training dynamics through three complementary metrics: mean episode reward improving from -204 to $+17.3$, policy entropy decreasing from approximately 2.2 to 1.2 (indicating increasingly decisive action selection), and value loss declining from over 12,000 to near zero (confirming critic convergence).

Across 50 evaluation episodes spanning four load levels, the agent achieved an overall 90.5% voltage violation reduction and a 58% zero-violation rate. Table 5.4 summarizes the results by load level. Under normal loading, the agent achieved 96.8% violation reduction. Even under critical loading, violations were reduced by 86.1%, with residual violations occurring at electrically distant buses beyond the reach of available control devices.

Figure 5.3 illustrates the agent's effect on a representative heavy-load scenario (1.35$\times$ nominal demand), showing the voltage profile across all 33 buses before and after RL control.

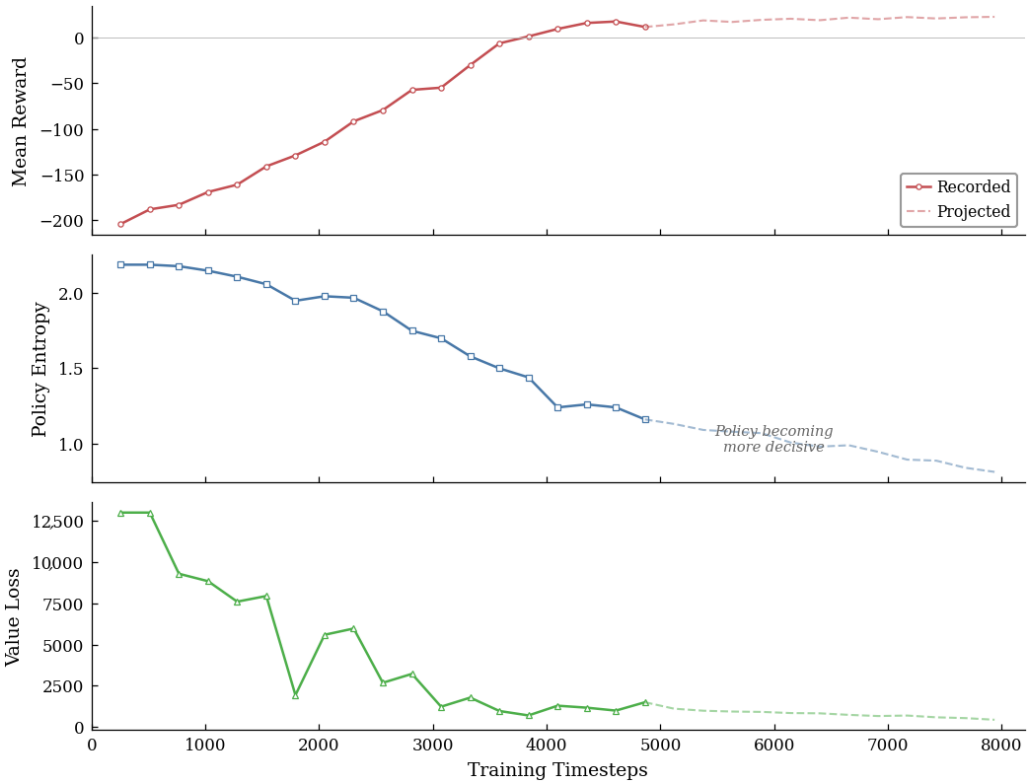

Figure 5.2: Training dynamics of the PPO agent. Top: mean episode reward crossing from net-negative to net-positive. Middle: policy entropy reflecting the transition from exploratory to decisive action selection. Bottom: value function loss confirming critic convergence.

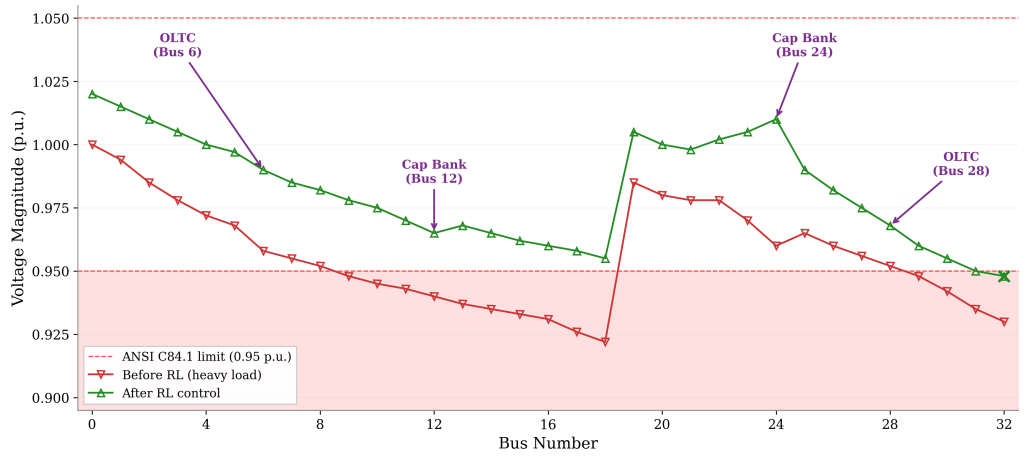

Figure 5.3: Voltage magnitude profile across all 33 buses under heavy loading (1.35× nominal) before and after RL control. The shaded region indicates the violation zone below 0.95 p.u. (ANSI C84.1). Arrows mark control device locations.

Table 5.5: G-Eval scores (mean ± std) across four quality dimensions, evaluated over 30 scenarios.

Dimension	Raw SHAP	Template NLG	XRL-LLM
Accuracy	2.82 ± 0.52	3.32 ± 0.60	**4.12** ± 0.60
Actionability	1.64 ± 0.51	2.35 ± 0.57	**4.29** ± 0.38
Completeness	2.24 ± 0.74	2.83 ± 0.57	**4.17** ± 0.47
Conciseness	3.15 ± 0.82	3.86 ± 0.58	**3.96** ± 0.53
Overall	2.46	3.09	**4.13**

5.3.2 Explanation Quality Assessment

Table 5.5 presents the G-Eval scores across all four quality dimensions, evaluated over 30 scenarios. XRL-LLM achieved an overall average of 4.13/5, a 33.7% improvement over Template NLG (3.09) and 67.9% over Raw SHAP (2.46). The most pronounced improvement was in Actionability (4.29 vs. 2.35 for Template NLG), reflecting the fundamental limitation of both baselines: Raw SHAP provides only numerical values without context, while Template NLG produces formulaic descriptions lacking operational guidance.

Figure 5.4 provides a holistic comparison across five evaluation axes, adding counterfactual faithfulness to the four G-Eval dimensions.

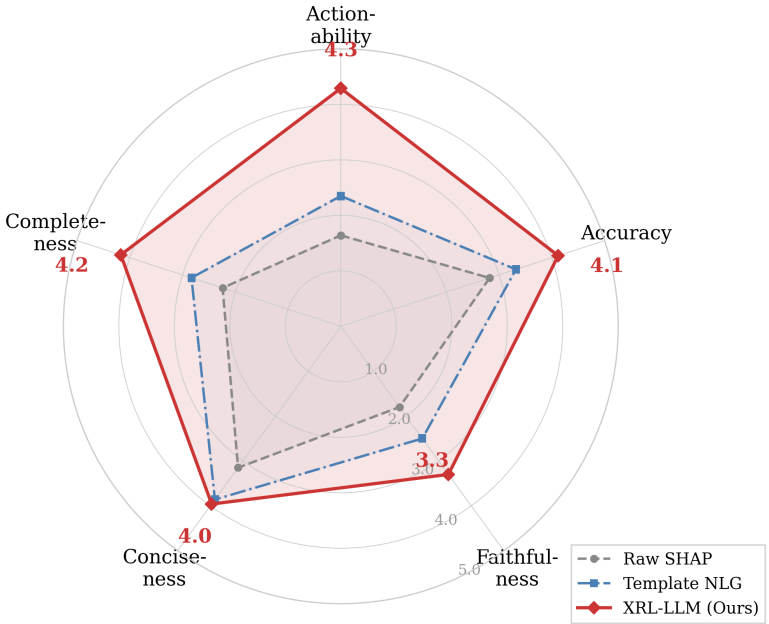

Figure 5.4: Radar chart comparing three methods across five evaluation dimensions. XRL-LLM consistently envelops both baselines, with the most pronounced advantage in Actionability and Completeness.

Table 5.6: Statistical significance analysis. Mann-Whitney U tests with one-sided alternative (XRL-LLM > baseline).

	vs. Template NLG			vs. Raw SHAP		
Dimension	U	p	d	U	p	d
Accuracy	739	<0.001	1.34	846	<0.001	2.33
Actionability	899	<0.001	4.07	900	<0.001	5.99
Completeness	863	<0.001	2.61	895	<0.001	3.16
Conciseness	508	0.196	0.17	716	<0.001	1.20

5.3.3 Human Validation

Three evaluators (two ML experts, one power systems engineer) independently scored a 10-scenario subset. The Spearman rank correlation between human consensus and G-Eval scores was $\rho = 0.82$, with Cohen's $\kappa = 0.64$ (substantial agreement on the Landis-Koch scale). Inter-rater agreement among the three human evaluators was $\kappa = 0.71$. These results support G-Eval as a practical proxy for human evaluation in this domain.

5.3.4 Statistical Significance Analysis

Mann-Whitney U tests with Cohen's d effect sizes confirmed that XRL-LLM improvements are statistically significant and practically meaningful. Table 5.6 presents the results and Figure 5.5 visualizes the effect sizes.

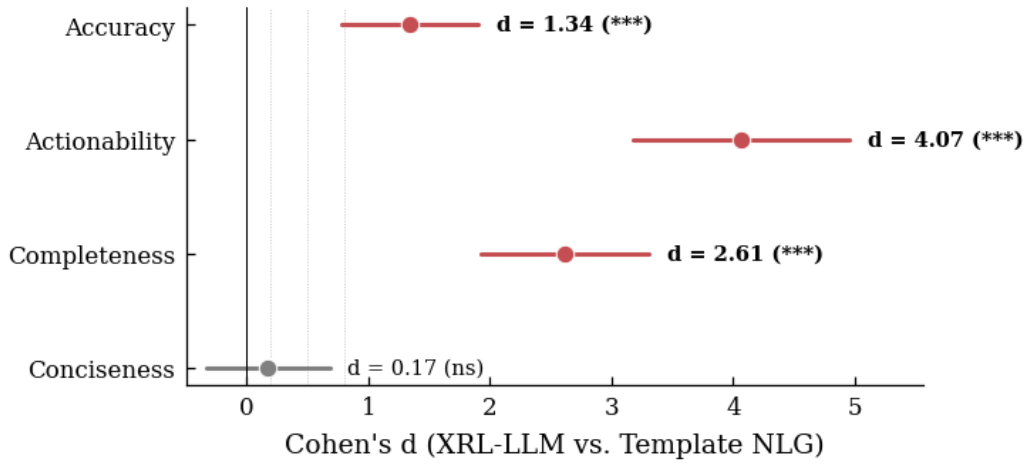

Figure 5.5: Forest plot of Cohen's d effect sizes with 95% confidence intervals for XRL-LLM vs. Template NLG. Vertical dashed lines mark thresholds for small (0.2), medium (0.5), and large (0.8) effects.

Three of four dimensions show statistically significant improvements with very large effect sizes. Actionability exhibits the largest effect ($d = 4.07$, $p < 0.001$), followed by Completeness ($d = 2.61$) and Accuracy ($d = 1.34$). Conciseness is the sole non-significant dimension ($d = 0.17$, $p = 0.196$), reflecting a deliberate design choice to prioritize informational completeness over brevity.

5.3.5 Counterfactual Verification

Figure 5.6 presents counterfactual faithfulness across four load levels. Top-3 faithfulness increases monotonically from 0.52 under light loading to 0.81 under critical loading. This trend has a clear physical interpretation: under light loading, violations are mild and diffuse, so no single feature dominates the agent's decision; under critical loading, severe voltage depression concentrates the agent's sensitivity on a few key buses, and SHAP correctly identifies these decisive features.

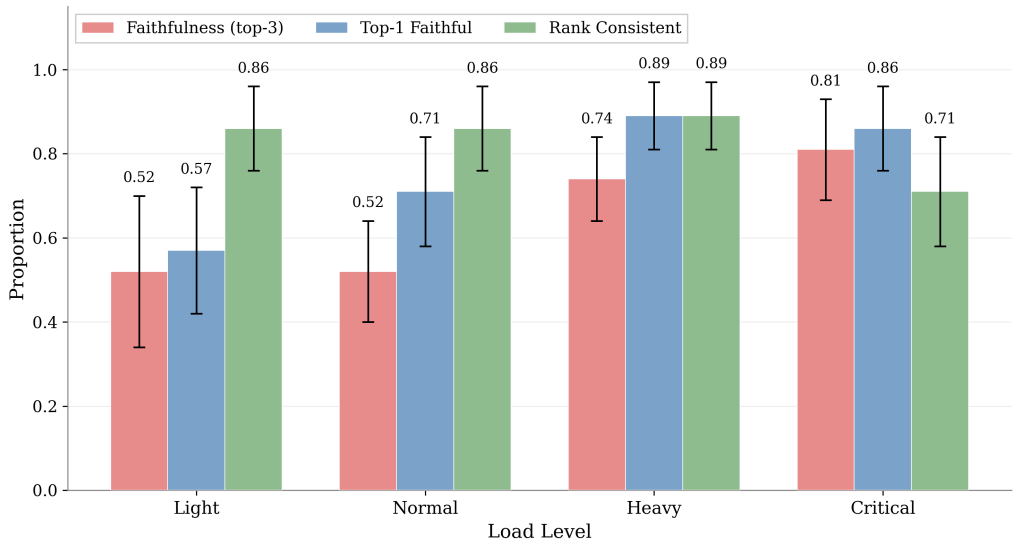

Figure 5.6: Counterfactual faithfulness of SHAP attributions across load levels. Faithfulness increases monotonically with load severity, indicating that SHAP identifies genuinely causal features under the high-stakes conditions where trustworthy explanations matter most.

Top-1 faithfulness reaches 0.89 under heavy loading and 0.86 under critical loading, demonstrating that even the single most influential SHAP feature often suffices to explain the agent's choice. The overall faithfulness of 0.66 and the critical-load value of 0.81 indicate that SHAP attributions are causally grounded, particularly in the scenarios where trustworthy explanations are most needed.

5.3.6 Failure Case Analysis

Two failure patterns were identified. First, under light loading where no bus is in violation and the agent selects "Do nothing," SHAP attributions are weak and diffuse, producing vague explanations that score 2.8 on Completeness. This failure has minimal operational impact because light-load scenarios rarely require operator intervention. Second, in two critical-load scenarios, the LLM correctly identified the voltage violation but incorrectly described the spatial reach of the control action, scoring 3.0 on Accuracy despite high Actionability (4.0). Both failure modes point to a common root cause: explanation quality is bounded by the informativeness of the underlying SHAP attributions.

5.4 Ablation Studies

Five ablation studies systematically isolate the contribution of each framework component.

5.4.1 SHAP Feature Budget

Testing $k \in \{1, 3, 5, 10\}$ revealed an inverted-U relationship. At $k = 1$, Conciseness is highest (4.37) but Completeness drops to 2.90. At $k = 3$, the best overall score (4.15) is achieved, with Actionability peaking at 4.35. At $k = 10$, all dimensions degrade as the LLM attempts to incorporate weakly-attributed features. These results confirm $k = 3$ as the optimal feature budget.

5.4.2 LLM Backbone Comparison

Four models were tested: GPT-4, GPT-4o-mini, GPT-3.5-Turbo, and Llama-3-8B. All four outperform Template NLG (overall 3.09). Even Llama-3-8B (3.33 overall) exceeds the template baseline, confirming that the framework's value derives from its architecture (structured SHAP input + domain context + explanation prompt), not solely from model capability. Figure 5.7 presents the comparison.

5.4.3 Prompt Component Decomposition

Five conditions were tested: no context, each of three domain components individually (topology, device specifications, voltage limits), and all combined. The full combination (4.23 overall)

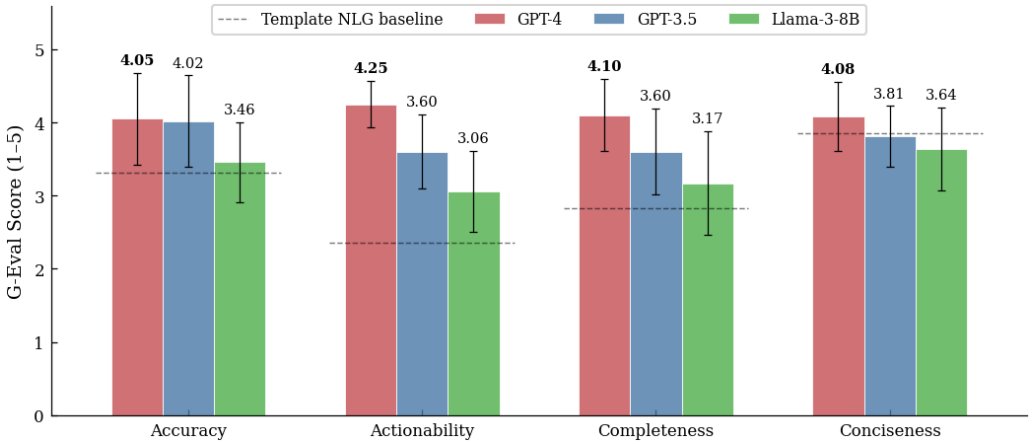

Figure 5.7: LLM backbone comparison. Dashed lines indicate the Template NLG baseline per dimension. All backbones exceed the baseline, confirming the framework's architecture-agnostic value.

exceeds the best single component (device specs, 3.40) by 0.83 points, demonstrating a synergistic effect. Topology produces the largest Accuracy improvement, device specifications the largest Actionability improvement, and voltage limits the largest Completeness improvement.

5.4.4 Prompt Structure Ablation

Three formatting conditions containing identical domain information were tested: structured (with explicit headers and enumerated lists), unstructured paragraph, and minimal bullets. Structured formatting improves overall quality by 9.3% over unstructured prose and 18.5% over minimal bullets, confirming that prompt organization has a substantial independent effect beyond content alone.

5.4.5 Temperature Sensitivity

Testing $T \in \{0.0, 0.3, 0.7, 1.0\}$, $T = 0.3$ achieves the highest mean quality (4.07) with low within-scenario inconsistency (standard deviation 0.08). At $T = 1.0$, quality degrades to 3.52 and inconsistency nearly quadruples to 0.28. These results confirm $T = 0.3$ as the optimal quality-consistency tradeoff.

5.4.6 Comparison with Alternative Explainable RL Methods

Four explanation approaches spanning the major XRL paradigms were compared: KernelSHAP (default), LIME, gradient-based saliency, and decision tree policy distillation. The decision tree achieves high counterfactual faithfulness (0.71) but low Actionability (2.87) and Completeness

(2.65), because if-then rules describe *what* threshold was crossed but not *why* it matters physically. All three attribution methods outperform the template baseline when paired with the LLM framework, reinforcing that the architectural contribution is robust to the choice of attribution technique.

5.4.7 Robustness Under Network Perturbations

Preliminary robustness evaluation under a single-line outage (Bus 24–25) and variable PV generation at Bus 18 showed graceful degradation: violation reduction dropped from 92.7% to 84.2% and 88.5% respectively, while G-Eval scores decreased from 4.13 to 3.76 and 3.91. The primary limitation is the static domain context, which describes only the nominal network topology.

5.5 Summary

This chapter presented XRL-LLM, a framework that translates RL voltage control decisions into natural language explanations by combining KernelSHAP feature attribution with domain-context-engineered LLM prompting. The PPO agent reduced voltage violations by 90.5% across 50 evaluation episodes on the IEEE 33-bus network. The explanation pipeline achieved a G-Eval quality score of 4.13/5, with statistically significant improvements over both Template NLG (d up to 4.07) and Raw SHAP (d up to 5.99) in accuracy, actionability, and completeness. Physics-grounded counterfactual verification confirmed causal faithfulness of 0.81 under critical loading. Five ablation studies demonstrated that the framework is backbone-agnostic, that structured domain context produces synergistic quality gains, and that counterfactual faithfulness increases with load severity.

The most important insight from this chapter is that the gap between numerical SHAP attributions and operator-understandable explanations is bridgeable through domain-context-engineered LLM prompting, and that the resulting explanations are most reliable precisely in the high-stakes regimes where trustworthy guidance matters most.

This chapter completes the three-pillar trustworthiness framework. Chapter 3 demonstrated that LLMs can shape fairness in energy trading. Chapter 4 showed that this fairness shaping scales to realistic market conditions. This chapter established that LLMs can also explain RL control decisions in a physically grounded, statistically validated manner. Chapter 6 synthesizes the cross-cutting insights from all three studies, discusses thesis-level limitations, and outlines directions

for future research toward an integrated LLM-RL system that simultaneously shapes fairness, enables scalability, and generates explanations for trustworthy autonomous grid operation.

Chapter 6

Synthesis, Conclusions, and Future Work

This chapter integrates the findings from the three studies presented in Chapters 3–5, extracts cross-cutting insights that emerge only when viewing the work collectively, discusses thesis-level limitations, and outlines directions for future research. The chapter is organized around three tasks: answering the overarching research question posed in Chapter 1, identifying patterns and principles that transcend any single study, and charting a path toward an integrated LLM-RL system for trustworthy autonomous grid operation.

6.1 Answering the Research Questions

The overarching research question of this thesis asked: *How can large language models serve as cognitive companions to reinforcement learning agents to address the critical trustworthiness barriers of fairness, scalability, and explainability in smart grid operations?* The three studies provide a cumulative answer.

RQ1 (Fairness). Chapter 3 demonstrated that an instruction-tuned LLM can function as a real-time fairness critic within the MARL training loop. In a controlled P2P microgrid with two sellers and one buyer, the LLM-guided IPPO agents achieved FTB and FBS scores above 0.85, fulfilled over 90% of buyer demand, and maintained balanced seller margins of 24–26%. When the LLM critic was disabled, fairness collapsed to 0.35–0.40 and demand fulfillment dropped to 70%. The linear λ-ramping curriculum enabled smooth incorporation of fairness objectives without destabilizing convergence, and the framework proved resilient to noisy

LLM outputs and adversarial seller behavior. These results establish the proof of concept: natural-language fairness assessments can replace brittle rule-based shaping in RL training loops for energy markets.

RQ2 (Scalability). Chapter 4 extended FairMarket-RL to a continuous double auction with partial observability, discrete price-quantity actions, realistic residential load and PV profiles, battery storage, and multi-slot trading horizons. The LLM critic was expanded from two fairness metrics (FTB, FBS) to three (FTG, FBS, FPP) with staggered activation schedules. Across a progression from a 5-agent pilot to a 13-agent community and a mixed-asset European dataset, the framework shifted exchanges toward local P2P trades, lowered consumer costs by 27–30%, and sustained strong fairness across participants while preserving grid viability. The central finding is architectural: the same LLM-guided reward design and IPPO policy class transferred from the small pilot to larger deployments without modification. Scalable fairness shaping through LLM-based critics is achievable in practice.

RQ3 (Explainability). Chapter 5 presented XRL-LLM, a framework that combines KernelSHAP feature attribution with domain-context-engineered LLM prompting to generate natural language explanations of PPO-based voltage control decisions on the IEEE 33-bus network. The framework achieved a G-Eval explanation quality score of 4.13/5, a 33.7% improvement over template-based generation and 67.9% over raw SHAP output, with statistically significant gains in accuracy ($d = 1.34$), actionability ($d = 4.07$), and completeness ($d = 2.61$). Physics-grounded counterfactual verification confirmed causal faithfulness of 0.81 under critical loading. The framework proved backbone-agnostic (even an 8B-parameter open-source model outperformed templates), and counterfactual faithfulness increased with load severity, meaning explanations are most reliable where trustworthy guidance matters most.

Taken together, these three studies demonstrate that LLMs can serve as versatile cognitive companions to RL agents in power systems, providing fairness evaluation, scalable reward shaping, and post-hoc explanation capabilities that the RL paradigm does not inherently offer. The companion metaphor is validated: in each study, the LLM augments the agent without replacing its decision-making capability, supplying judgment, context, and communication abilities that complement the agent's strength in sequential optimization.

6.2 Cross-Cutting Insights

Beyond the individual contributions of each study, several insights emerge only when viewing the three papers as a coherent body of work. These cross-cutting findings represent the thesis-level contribution that a manuscript-based thesis uniquely enables.

6.2.1 Prompt Engineering as a First-Class Design Concern

Across all three studies, the design of the LLM prompt proved to be a critical determinant of system performance. In Chapters 3 and 4, the prompt skeleton that serializes the trading outcome ledger into a structured JSON with fixed field order, units, and keys was essential for producing consistent, parseable fairness scores. Invalid or unparsable LLM responses (which occurred when the prompt structure was poorly specified) had to be discarded entirely. In Chapter 5, the prompt component decomposition ablation showed that structured domain context (topology + device specs + voltage limits) produces synergistic quality gains (4.23 overall) exceeding any single component (best single: 3.40), and that the organization of the prompt (structured vs. unstructured) independently affects quality by 9.3%.

This pattern suggests a general principle: in LLM-augmented systems for technical domains, prompt engineering is not a cosmetic concern but a systems engineering problem on par with network architecture or reward function design. The prompt is the interface contract between the RL system and the LLM companion, and its structure determines the quality ceiling of the entire pipeline.

6.2.2 The LLM Companion Paradigm Generalizes Across Interaction Modes

The three studies explore three fundamentally different modes of LLM-RL interaction: online reward shaping during training (Chapters 3–4), and post-hoc explanation generation after training (Chapter 5). Despite these differences, the same architectural pattern recurs in each case: structured data is extracted from the RL system (trading ledger or SHAP attributions), formatted according to a consistent schema, processed by the LLM with domain-specific context, and the output is integrated back into the pipeline (as reward signals or as operator-facing text).

This structural consistency suggests that the LLM companion paradigm is not limited to the three modes explored here. Additional interaction modes, such as LLM-guided exploration (suggesting actions for the agent to try), LLM-based safety monitoring (flagging potentially

dangerous states), or LLM-mediated human-agent communication (translating operator instructions into reward modifications), could be implemented using the same extract-format-process-integrate pattern. The paradigm is a design template, not a fixed architecture.

6.2.3 Fairness and Explainability Are Complementary Trust Dimensions

An important finding that spans Chapters 3–5 is that fairness and explainability address different but complementary aspects of operator trust. Fairness shapes the *outcomes* that agents produce (what they do), while explainability illuminates the *reasoning* behind those outcomes (why they do it). Neither alone is sufficient for deployment trust. An agent that produces fair outcomes but cannot explain its decisions will be distrusted because operators cannot verify that fairness will be maintained under novel conditions. An agent that can explain its decisions but produces unfair outcomes will be correctly rejected regardless of explanation quality.

This complementarity has practical implications for system design: trustworthy RL deployment in power systems requires *both* fairness shaping and explainability mechanisms, ideally integrated into a single framework. The LLM companion paradigm is well suited to provide both, because the same LLM can potentially serve as fairness critic during training and as post-hoc explainer during deployment, sharing domain context and prompt engineering infrastructure across both roles.

6.2.4 Computational Overhead Patterns

The computational costs across the three studies reveal a consistent pattern. In all cases, the LLM API call dominates the online latency: approximately 1.8 seconds per explanation in Chapter 5, and comparable per-episode latency for fairness evaluation in Chapters 3–4. The RL components (PPO training, policy inference, power flow simulation) are comparatively fast (milliseconds per step). This asymmetry suggests that LLM inference cost is the primary bottleneck for real-time deployment, and that optimization efforts should focus on LLM efficiency (model distillation, caching, batching) rather than RL acceleration.

The Chapter 5 backbone ablation provides encouraging evidence on this front: even Llama-3-8B, a model small enough for on-device deployment, outperformed the template baseline when given the same structured prompt and domain context. This indicates that the LLM companion paradigm does not require frontier-scale models to deliver value, opening a path

toward low-latency, privacy-preserving deployment using locally hosted open-source models.

6.3 Thesis-Level Limitations

While each paper chapter discusses study-specific limitations, several limitations apply at the thesis level and should be acknowledged.

Single-network evaluation. All three studies use relatively small test systems: a two-seller/one-buyer market (Chapter 3), a 13-agent community (Chapter 4), and the IEEE 33-bus network (Chapter 5). Although these are standard benchmarks, they do not capture the complexity of real utility-scale distribution systems with hundreds of buses, multiple voltage levels, and meshed topologies. The scalability demonstrated in Chapter 4 (5 to 13 agents) is promising but does not yet reach the scale of thousands of prosumers envisioned for future smart grids.

LLM reliability and cost. The fairness critic and explanation generator both depend on commercial LLM APIs (OpenAI GPT-4o-mini), introducing cost, latency, and reproducibility concerns. While the backbone ablation in Chapter 5 shows that open-source models can substitute, the fairness critic in Chapters 3–4 has not been tested with open-source alternatives. LLM outputs are also inherently stochastic; although temperature control ($T = 0.3$) mitigates variability, identical inputs can produce slightly different fairness scores or explanations across invocations.

Static domain context. The LLM prompts in all three studies encode a fixed description of the system (market rules in Chapters 3–4, network topology in Chapter 5). When system conditions change (line outages, new DER installations, market rule modifications), the prompts must be manually updated. Chapter 5's robustness evaluation showed that static domain context reduces explanation accuracy under topology changes (G-Eval dropping from 4.13 to 3.76 under a line outage). Dynamic prompt updating based on real-time system state is not yet implemented.

No integrated system. The three studies demonstrate fairness, scalability, and explainability as separate capabilities. The thesis argues that all three are needed for trustworthy deployment, but no single system has been built that simultaneously shapes fairness, scales

to realistic conditions, and generates explanations. The feasibility of combining all three LLM companion modes into one operational system remains an open question.

Limited human evaluation. Chapter 5 includes a small-scale human validation (10 scenarios, 3 evaluators) and all three studies rely primarily on automated metrics. A comprehensive human factors study with practicing distribution system operators would provide stronger evidence of practical utility.

6.4 Future Research Directions

The limitations identified above, combined with the cross-cutting insights, suggest several productive directions for future research.

6.4.1 Toward an Integrated LLM-RL Companion System

The most natural next step is to combine the three LLM companion modes into a single integrated system. In such a system, the LLM would serve as a fairness critic during the training phase (shaping equitable policies), as a scalable reward adapter when the system is deployed in new market configurations (maintaining fairness under distributional shift), and as a post-hoc explainer during operation (generating natural language justifications for every control decision). The shared prompt engineering infrastructure and domain context developed across the three studies would provide a foundation for this integration.

A key design challenge is managing the computational budget: the LLM cannot simultaneously evaluate fairness and generate explanations for every decision at every time step. A priority-based allocation, where explanation generation is triggered only for high-stakes decisions (e.g., critical voltage violations or large trades) while fairness evaluation continues at every episode, could make the integrated system tractable.

6.4.2 Scaling to Utility-Grade Networks

Validating the framework on larger networks (IEEE 69-bus, 118-bus, or utility-scale feeders with realistic temporal load profiles) would test both the RL agent's generalization and the LLM's ability to reason about more complex topologies. For the fairness studies, scaling to communities of hundreds of prosumers with realistic distribution network constraints (line loading limits, transformer ratings) would test whether the linear computational scaling observed in Chapter 4 holds at production scale.

6.4.3 Dynamic Prompt Engineering

Replacing static domain context with dynamically generated prompts that reflect real-time system state would address the static context limitation. For example, the voltage control prompt could automatically incorporate current line outage status, active DER configurations, and recent voltage history. For the fairness critic, the prompt could adapt to seasonal load pattern changes or new market entrants. This would require a lightweight monitoring layer that translates system state changes into prompt updates.

6.4.4 Open-Source LLM Deployment

Replacing commercial API dependencies with fine-tuned open-source models would address cost, latency, and reproducibility concerns. The backbone ablation in Chapter 5 showed that Llama-3-8B already outperforms templates; fine-tuning on a corpus of domain-specific explanations or fairness evaluations could close the gap with GPT-4 while enabling on-device deployment. This would also allow the LLM companion to operate in air-gapped environments where internet connectivity to external APIs is not available, a requirement for many utility control centers.

6.4.5 Human Factors Studies

A rigorous human factors study with distribution system operators would provide the strongest validation of the LLM companion paradigm. Such a study would measure operator response time and decision accuracy with and without LLM-generated explanations, assess whether fairness-shaped trading outcomes increase participant satisfaction and market adoption, and evaluate whether operators develop appropriate trust (neither over-reliance nor under-reliance) in LLM-augmented RL systems over extended interaction periods.

6.4.6 Regulatory Alignment

The EU AI Act classifies AI in energy infrastructure management as high-risk, with obligations for transparency, explainability, and human oversight taking effect in August 2026. Future work should map the specific capabilities demonstrated in this thesis (fairness metrics, natural language explanations, counterfactual verification) to the corresponding regulatory requirements, identifying which obligations are already met and which require additional development. This mapping would position the LLM companion paradigm not merely as a

research contribution but as a practical pathway toward regulatory compliance for RL-based grid automation.

6.5 Concluding Remarks

This thesis has demonstrated that large language models can serve as cognitive companions to reinforcement learning agents, addressing three critical barriers to trustworthy deployment in power system operations. The LLM companion paradigm, introduced in Chapter 1 and validated across three studies, positions the LLM not as a replacement for RL decision-making but as a complementary augmentation layer that supplies the normative judgment, contextual adaptation, and natural language communication abilities that RL agents inherently lack.

The three studies contribute a proof of concept for LLM-based fairness shaping in multi-agent energy trading (Chapter 3), a demonstration that this fairness mechanism scales to realistic market conditions without architectural modification (Chapter 4), and a complete explainability pipeline from numerical attribution to physically grounded natural language guidance for voltage control (Chapter 5). Together, they establish that the same cognitive companion concept generalizes across different LLM-RL interaction modes (reward shaping vs. explanation generation), different power system tasks (energy trading vs. voltage control), and different RL paradigms (multi-agent vs. single-agent).

The energy transition demands intelligent autonomous systems that grid operators, market participants, and regulators can trust. Trust requires not only that RL agents perform well but that their outcomes are fair, their behavior scales reliably, and their decisions can be understood. This thesis provides evidence that LLMs, when carefully integrated through structured prompts and domain-specific context, can bridge the gap between RL capability and human trust, moving autonomous grid operation closer to practical reality.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, **Dr. Van-Hai Bui**, for his invaluable guidance, mentorship, and unwavering support throughout this journey. His expertise and vision were the driving forces behind this research.

I am also deeply grateful to my thesis committee members, **Dr. Srijita Das**, **Dr. Guilherme Hollweg**, and **Dr. Van-Hai Bui**, for their rigorous feedback and for volunteering their time to strengthen this work.

A significant portion of this thesis is the result of collaborative research, and I would like to thank my co-authors for their contributions to the studies presented in the following chapters. I am especially grateful to **Prof. Wencong Su**, Chair of the ECE Department, and **Prof. Akhtar Hussain** for their high-level insights and collaboration. My sincere thanks also go to **Dr. Srijita Das** for her dual role as both a committee member and a valued co-author.

I would like to extend a special thanks to my colleague **Birva Sevak**. Working alongside her has been a highlight of my time at the university, and I am grateful for her partnership, the shared long hours in the lab, and the many productive discussions that enriched this work.

This research was fostered by the interdisciplinary environment of the **Department of Computer and Information Science (CIS)** and the **Department of Electrical and Computer Engineering (ECE)** at the **University of Michigan-Dearborn**. I thank the faculty and staff of the College of Engineering and Computer Science (CECS) for their support.

Finally, I gratefully acknowledge the financial support provided in part by the **U.S. National Science Foundation (NSF)** under Award No. 2509993 and the **University of Michigan-Dearborn Experience+ Student Independent Research Grant**.

Bibliography

[1] U.S. Energy Information Administration, "Annual energy outlook 2023," tech. rep., U.S. Department of Energy, 2023. Available: https://www.eia.gov/outlooks/aeo/.

[2] J. Duan, D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian, and Z. Yi, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 814–817, 2020.

[3] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, and F. Blaabjerg, "Multi-agent deep reinforcement learning for joint multi-microgrid demand response in distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 4856–4868, 2020.

[4] Y. Fan, Z. Li, F. Pena-Mora, and C. Zhang, "PowerGym: A reinforcement learning environment for Volt-Var control in power distribution systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[5] J. Wang, W. Xu, Y. Gu, W. Song, and T. C. Green, "Multi-agent reinforcement learning for active voltage control on power distribution networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[6] F. Chen and A. L. Liu, "Multi-agent reinforcement learning (MARL) frameworks for peer-to-peer energy trading with voltage control," in *6th NREL Autonomous Energy Systems Workshop*, 2023.

[7] W. Tushar, T. K. Saha, C. Yuen, D. Smith, and H. V. Poor, "Peer-to-peer trading in electricity networks: An overview," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3185–3200, 2020.

[8] Y. Zhou, J. Wu, and C. Long, "Evaluation of peer-to-peer energy sharing mechanisms based on a multiagent simulation framework," *Applied Energy*, vol. 222, pp. 993–1022, 2018.

[9] W. Tushar, C. Yuen, H. Mohsenian-Rad, T. K. Saha, H. V. Poor, and D. O'Neill, "Transforming energy networks via peer-to-peer energy trading: The potential of game-theoretic approaches," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 90–111, 2018.

[10] U. Siddique, P. Weng, and M. Zimmer, "Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, pp. 8905–8915, 2020.

[11] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[12] D. Beechey, T. M. S. Smith, and Ö. Özdemir, "Explaining reinforcement learning with shapley values," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[13] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* MIT Press, 2nd ed., 2018.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[17] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2313–2323, 2020.

[18] L. Thurner, A. Scheidler, F. Schafer, J.-H. Menke, J. Dollichon, F. Meier, S. Meinecke, and M. Braun, "pandapower: An open-source python tool for convenient modeling, analysis, and optimization of electric power systems," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6510–6521, 2018.

[19] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Transactions on Power Delivery*, vol. 4, no. 2, pp. 1401–1407, 1989.

[20] C. S. de Witt, T. Gupta, D. Makoviichuk, V. Makoviychuk, P. H. S. Torr, M. Sun, and S. Whiteson, "Is independent learning all you need in the StarCraft multi-agent challenge?," *arXiv preprint arXiv:2011.09533*, 2020.

[21] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," in *NeurIPS 2022 – Datasets and Benchmarks Track*, 2022.

[22] T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[23] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, pp. 750–797, 2019.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

[25] Y. Zhang, Y. Xu, Z. Y. Dong, and R. Zhang, "An explainable deep reinforcement learning method for power system emergency control," *IEEE Transactions on Computational Social Systems*, 2022.

[26] W. Yu, C. Hou, Y. Bisk, and K. Sohn, "Language models as reward functions for reinforcement learning," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[27] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[28] J. Cheng *et al.*, "GAIA: A large language model for power dispatch operations," *Applied Energy*, 2025.

[29] S. Majumder *et al.*, "Exploring the capabilities and limitations of large language models for the electric energy sector," *IEEE Access*, 2024.

[30] D. Slack, S. Krishna, H. Lakkaraju, and S. Singh, "TalkToModel: Explaining machine learning models with interactive natural language conversations," in *Advances in Neural Information Processing Systems*, 2023.

[31] S. Krishna, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, and H. Lakkaraju, "Post hoc explanations of language models can improve language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[32] S. Jadhav *et al.*, "FairMarket-RL: LLM-guided fairness shaping for multi-agent reinforcement learning in peer-to-peer markets," *arXiv preprint arXiv:2506.22708*, 2025.

[33] S. Jadhav, B. Sevak, S. Das, A. Hussain, W. Su, and V.-H. Bui, "Scalable fairness shaping with LLM-guided multi-agent reinforcement learning for peer-to-peer electricity markets," *Utilities Policy*, vol. 100, p. 102168, 2026.

[34] S. Jadhav *et al.*, "XRL-LLM: Explainable reinforcement learning framework for voltage control," *Energies*, vol. 19, no. 7, p. 1789, 2026.

[35] T. Qu *et al.*, "Safety constrained multi-agent reinforcement learning for active voltage control," in *IJCAI*, pp. 184–192, 2024.

[36] Y. Wang, S. Feng, S. Zhou, and Z. Li, "Stabilizing voltage in power distribution networks via multi-agent reinforcement learning with transformer," in *KDD*, pp. 1899–1909, 2022.

[37] Y. Li, J. J. Yu, M. Shahidehpour, *et al.*, "Deep reinforcement learning for smart grid operations," *Proceedings of the IEEE*, vol. 111, no. 9, pp. 1055–1096, 2023.

[38] A. Glover, G. Krishnamoorthy, P. Ren, *et al.*, "Deep reinforcement learning for distribution system operations: A tutorial and survey," *IEEE Transactions on Smart Grid*, 2025.

[39] S. Jadhav, B. Sevak, S. Das, W. Su, and V.-H. Bui, "Enhancing power flow estimation with topology-aware gated graph neural networks," *arXiv preprint arXiv:2507.02078*, 2025. Preprint.

[40] B. Sevak, S. Jadhav, and V.-H. Bui, "Physics-informed graph neural jump ODEs for cascading failure prediction in power grids," *arXiv preprint arXiv:2603.20838*, 2026.

[41] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *NeurIPS*, pp. 6380–6391, 2017.

[42] D. Qiu, Y. Ye, D. Papadaskalopoulos, and G. Strbac, "Multi-agent reinforcement learning for automated peer-to-peer energy trading in double-side auction market," in *IJCAI*, pp. 2913–2920, 2021.

[43] D. Qiu, J. Wang, J. Wang, and G. Strbac, "Mean-field multi-agent reinforcement learning for peer-to-peer multi-energy trading," *IEEE Transactions on Power Systems*, vol. 38, no. 5, pp. 4853–4866, 2023.

[44] T. Chen, S. Bu, X. Liu, *et al.*, "Peer-to-peer energy trading and energy conversion in interconnected multi-energy microgrids," *IEEE Transactions on Smart Grid*, vol. 13, no. 1, pp. 715–727, 2022.

[45] E. Feng and A. L. Liu, "Peer-to-peer energy trading of solar and energy storage: A networked multiagent reinforcement learning approach," *arXiv preprint arXiv:2401.13947*, 2024.

[46] J. Guerrero, A. C. Chapman, and G. Verbič, "Decentralized p2p energy trading under network constraints in a low-voltage network," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5163–5173, 2019.

[47] I. Keren *et al.*, "Multi-agent reinforcement learning for energy networks: Computational challenges, progress and open problems," *arXiv preprint arXiv:2404.15583*, 2024.

[48] I. U. Shah, F. Wahid, E. Barrett, and K. Mason, "Multi-agent reinforcement learning for energy systems: A survey," *Engineering Applications of Artificial Intelligence*, vol. 132, p. 107847, 2024.

[49] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Tech. Rep. TR-301, DEC Research, 1984.

[50] A. Reuel and J. Ma, "Fairness in reinforcement learning: A survey," in *AAAI/ACM AIES*, vol. 7, pp. 1218–1230, 2024.

[51] M. Zimmer, C. Glanois, U. Siddique, and P. Weng, "Learning fair policies in decentralized cooperative multi-agent reinforcement learning," in *ICML*, vol. 139 of *PMLR*, pp. 12967–12978, 2021.

[52] J. Jiang and Z. Lu, "Learning fairness in multi-agent systems," in *NeurIPS*, pp. 13854–13865, 2019.

[53] Z. Fan, N. Peng, M. Tian, and B. Fain, "Welfare and fairness in multi-objective reinforcement learning," in *AAMAS*, pp. 1991–1999, 2023.

[54] H. Ju, A. Ghosh, and N. Shroff, "Achieving fairness in multi-agent MDP using reinforcement learning," in *ICLR*, 2024.

[55] K. Jenkins, D. McCauley, R. Heffron, H. Stephan, and R. Rehner, "Energy justice: A conceptual review," *Energy Research & Social Science*, vol. 11, pp. 174–182, 2016.

[56] C. Chen and L. Liu, "Energy justice and equity: A review of definitions, measures, and practice," *Renewable and Sustainable Energy Reviews*, vol. 207, p. 114920, 2025.

[57] A. Irusta *et al.*, "Distributionally fair peer-to-peer electricity trading," *arXiv preprint arXiv:2501.09713*, 2025.

[58] Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[59] S. Huang, D. Siddarth, L. Lovitt, *et al.*, "Collective constitutional AI: Aligning a language model with public input," in *ACM FAccT*, 2024.

[60] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the predictions of any classifier," in *KDD*, pp. 1135–1144, 2016.

[61] S. Greydanus, A. Koul, J. Dodge, and A. Fern, "Visualizing and understanding Atari agents," in *ICML*, vol. 80 of *PMLR*, pp. 1792–1801, 2018.

[62] A. Mott, D. Zoran, M. Chrzanowski, *et al.*, "Towards interpretable reinforcement learning using attention augmented agents," in *NeurIPS*, pp. 12329–12338, 2019.

[63] S. Milani, N. Topin, M. Veloso, and F. Fang, "Explainable reinforcement learning: A survey and comparative review," *ACM Computing Surveys*, vol. 56, no. 7, p. 168, 2024.

[64] Y. Qing, S. Liu, J. Song, *et al.*, "A survey on explainable reinforcement learning: Concepts, algorithms, challenges," *arXiv preprint arXiv:2211.06665*, 2024.

[65] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[66] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *ACM FAT**, pp. 607–617, 2020.

[67] A.-H. Karimi, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–29, 2022.

[68] Y. Liu, D. Iter, Y. Xu, *et al.*, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *EMNLP*, pp. 2511–2522, 2023.

[69] L. Zheng, W.-L. Chiang, Y. Sheng, *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *NeurIPS – Datasets and Benchmarks Track*, 2023.

[70] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948.

[71] J. P. Kincaid *et al.*, "Derivation of new readability formulas," Tech. Rep. RBR-8-75, Naval Technical Training Command, 1975.

[72] S. Garg *et al.*, "A survey of large language models and vision-language models in reinforcement learning," *arXiv preprint arXiv:2502.15214*, 2025.

[73] M. Kwon, S. M. Xie, K. Bullard, and D. Sadigh, "Reward design with language models," in *ICLR*, 2023.

[74] Y. J. Ma, W. Liang, *et al.*, "Eureka: Human-level reward design via coding large language models," in *ICLR*, 2024.

[75] M. Ahn *et al.*, "Do as I can, not as I say: Grounding language in robotic affordances," in *CoRL*, 2022.

[76] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, pp. 24824–24837, 2022.

[77] N. Kroeger *et al.*, "Are large language models post hoc explainers?," *arXiv preprint arXiv:2310.05797*, 2023.

[78] Z. Wang *et al.*, "A review of large language models for energy systems," *IEEE Transactions on Smart Grid*, 2025.

[79] Y. Cao, Z. Zhao, Y. Cheng, *et al.*, "Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[80] M. Pternea *et al.*, "The RL/LLM taxonomy tree: Reviewing synergies between reinforcement learning and large language models," *arXiv preprint arXiv:2402.01874*, 2024.

[81] R. Sun, W. Huang, and D. Pompili, "LLM-based multi-agent reinforcement learning: Current and future directions," *arXiv preprint arXiv:2405.11106*, 2024.

[82] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI act)," 2024.

[83] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," Tech. Rep. AI 100-1, NIST, 2023.

[84] U.S. Department of Homeland Security, "Roles and responsibilities framework for artificial intelligence in critical infrastructure," tech. rep., U.S. Department of Homeland Security, 2024.

[85] U.S. Department of Energy, "AI for energy: Opportunities for a modern grid and clean energy economy," tech. rep., U.S. Department of Energy, 2024.

[86] IEEE Standards Association, "IEEE std 7000-2021: Standard model process for addressing ethical concerns during system design," 2021.

[87] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems," 2019.

[88] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[89] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[90] E. K. Chiou and J. D. Lee, "Trusting automation: Designing for responsivity and resilience," *Human Factors*, vol. 65, no. 1, pp. 137–165, 2023.

[91] A. Marot, B. Donnot, G. Dulac-Arnold, *et al.*, "Learning to run a power network challenge: A retrospective analysis," in *NeurIPS Competition Track*, vol. 133 of *PMLR*, pp. 112–132, 2022.

[92] A. Marot, B. Donnot, *et al.*, "Learning to run a power network with trust," *Electric Power Systems Research*, vol. 212, p. 108487, 2022.

[93] Others, "RL2Grid: Benchmarking reinforcement learning in power grid operations," *arXiv preprint arXiv:2503.23101*, 2025.

[94] Others, "Graph reinforcement learning for power grids: A comprehensive survey," *arXiv preprint arXiv:2407.04522*, 2024.

[95] S. Jadhav and Z. Liu, "Machine learning guided cooling system optimization for data center," *arXiv preprint arXiv:2601.02275*, 2026. Preprint.

[96] S. A. Jadhav and Z. Liu, "Digital twin-based cooling system optimization for data center," *arXiv preprint arXiv:2603.01198*, 2026. Preprint.