

Article

SafeVolt: Closed-Loop Large Language Model Framework for Safety-Aware Voltage Control in Active Distribution Networks

Zhijun Shen ¹, Qian Guo ¹, Kaiyuan Pang ¹, Xinlei Cai ¹, Zhenfan Yu ¹, Kunhao Feng ^{2,*} and Tao Yu ²

¹ Power Dispatch Control Center of Guangdong Power Grid Co., Ltd., Guangzhou 510335, China

² School of Electric Power Engineering, South China University of Technology, Guangzhou 510641, China

* Correspondence: 202520114153@mail.scut.edu.cn

Abstract

Voltage and reactive power control in active distribution networks is a safety-critical and highly dynamic problem, where traditional optimization methods often struggle to balance efficiency and robustness under complex operating conditions. Recently, large language models (LLMs) have shown promise in sequential decision-making tasks, but their direct application to power system control remains limited by the lack of physical grounding and safety guarantees. In this paper, we propose SafeVolt, a closed-loop LLM-based framework that integrates multi-candidate action generation, simulator-in-the-loop evaluation, and a fine-tuned expert judge for safety-aware decision making. In addition, a high-level rule distillation mechanism that converts successful control experiences into reusable operational axioms is introduced to enable iterative self-improvement. Experiments on a standard distribution network scenario demonstrate that the proposed method outperforms representative baselines, achieving substantial improvements in average reward, voltage violation rate, reactive power loss, and system stability. In particular, voltage violations and extreme events are substantially reduced, indicating enhanced operational safety. These results suggest that combining LLM reasoning with physical simulation and structured feedback provides a promising direction for reliable and adaptive power system control.

Keywords: reactive power control; large language model-based decision making; supervised fine-tuning; memory-augmented learning

1. Introduction

The growing penetration of distributed renewable energy, particularly photovoltaics, has increased the operational complexity of modern distribution grids, introducing stronger uncertainty, frequent voltage fluctuations, and tighter requirements for fast and adaptive voltage/reactive power control [1]. In practice, however, many distribution-grid operational tasks still rely heavily on human dispatchers, rule-based procedures, and predefined operating plans. Existing studies have pointed out that such workflows often depend on expert experience, suffer from limited efficiency, and are prone to errors when system conditions change rapidly [2,3]. Although model-based control methods such as model predictive control and coordinated reactive power optimization have demonstrated effectiveness in voltage regulation and reactive power dispatch, they still heavily rely on accurate physical modeling, manual formulation of objectives and constraints, and repeated online optimization, which may limit their flexibility and responsiveness in renewable-rich and fast-changing operating environments [4,5]. Therefore, there is a clear need for more



Academic Editor: Tianwen Zhang

Received: 29 May 2026

Revised: 22 June 2026

Accepted: 23 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

adaptive and efficient decision-making frameworks for voltage/reactive power dispatch in active distribution systems.

Recent advances in large language models (LLMs) have made them increasingly relevant to power dispatch and grid operation tasks where structured operational data must be jointly interpreted with natural-language regulations, dispatch rules, and historical operation records [6]. For voltage and reactive power dispatch in distribution networks, this capability is particularly attractive because practical decisions often require not only numerical grid-state analysis, but also semantic understanding of operating procedures, safety margins, device-control rules, and dispatcher experience. Existing studies have demonstrated the potential of domain-adapted and retrieval-augmented LLMs in the electric energy sector. They can effectively process numerical dispatch information, semantically interpret complex grid regulations, and also generate structured executable commands (e.g., SQL queries) for dispatching systems [7–9].

However, directly deploying LLMs as end-to-end decision makers remains challenging due to several inherent limitations. Fundamentally, factual hallucination presents a key challenge to trustworthy deployment [10]. In voltage and reactive power control, hallucinated outputs—such as generating setpoints that exceed inverter capacities or ignoring voltage thresholds—can directly compromise system stability. Compounding this reliability issue is the severe controllability challenge of LLM: it is difficult to ensure that LLMs consistently adhere to strict, user-specified constraints [11]. This deficiency is especially concerning in distribution networks, where control actions must satisfy complex physical boundaries (e.g., AC power flow equations and N-1 security limits) that statistical language models cannot guarantee by design. Furthermore, in dynamic, interactive decision-making tasks, naive one-shot prompting frequently fails because an isolated model lacks the iterative, experience-based feedback required to adapt to changing environments [12]. Given the continuous uncertainties from renewable energy fluctuations and load variations, an open-loop LLM decision that ignores real-time physical feedback quickly becomes outdated or even counterproductive. These compounding vulnerabilities suggest that, in safety-critical dispatch and control scenarios, LLMs should be embedded into a closed-loop framework with explicit physical verification, controllable decision constraints, and iterative feedback, rather than being used as unconstrained conversational generators in a one-shot manner.

To address the above challenges, this paper proposes **SafeVolt**, a closed-loop decision framework that bridges LLM reasoning and actual grid physics for voltage and reactive power control in active distribution systems. The proposed method departs from naive end-to-end generation by establishing a tightly coupled, closed-loop architecture. At its core, rather than directly producing a single action, the framework prompts the LLM to perform LLM-based multi-candidate policy generation conditioned on the current system state and accumulated historical knowledge. Instead of executing these raw outputs, the system explicitly grounds them in physical reality through Simulation-based evaluation, including power-flow simulation, reward modeling, and N-1 security analysis. Building upon this rigorous evaluation, Safety-aware decision via a fine-tuned judge model serves as a safety-aware decision module, which filters out infeasible actions and finalizes the most reliable control command. Furthermore, to ensure the framework does not remain static in dynamic operating environments, a Memory-augmented self-evolution mechanism is incorporated. This module actively tracks performance trends, retrieves analogous historical dispatch cases, and distills successful regulation profiles into reusable, high-level axioms. Together, these components establish a practical unified framework for safely exploiting LLM capabilities within the strict operational boundaries of power systems.

The main contributions of this work are summarized as follows:

- We propose SafeVolt, the first closed-loop LLM-augmented framework for continuous reactive power control in active distribution networks that tightly integrates four components—multi-candidate policy generation, simulation-based physical evaluation, safety-aware decision making via a fine-tuned judge, and memory-augmented self-evolution—into a unified pipeline. Unlike existing LLM-based power-system studies that operate as offline assistants or one-shot generators, SafeVolt maintains a persistent closed loop with physical simulation feedback at every decision step.
- We introduce two novel technical mechanisms specifically designed for the safety-critical and dynamic characteristics of distribution network voltage control: (i) an *axiom extraction and reuse mechanism* that distills successful control patterns into high-level reusable rules, enabling iterative self-improvement without offline retraining; and (ii) a *domain-specific fine-tuned judge model* that performs binary safety screening based on simulated physical metrics (voltage profile, N-1 security), preventing the execution of LLM-generated actions that violate operational constraints. These designs go beyond simple retrieval augmentation or direct fine-tuning by grounding LLM reasoning in physical simulation.
- Simulation results on the IEEE 33-bus active distribution network scenario verify the effectiveness of the proposed framework within the evaluated setting. Compared with representative LLM-based baselines, SafeVolt achieves significantly lower voltage violation rates (reducing from 23–40% to 1.23%), higher average reward, and near-elimination of extreme voltage events, demonstrating that physically grounded closed-loop filtering substantially improves the safety and reliability of LLM-based control.

2. Related Work

2.1. Non-LLM Methods for Power Grid Scheduling and Control

Physics-informed and model-based methods. Early research on voltage regulation in active distribution networks was dominated by model-based and scientific-computing approaches that rely on explicit physical models, solver-accessible constraints, and numerically interpretable objectives. Representative studies formulated voltage regulation as constrained predictive control or coordinated reactive-power optimization. For instance, MPC-based centralized control [4] updates multi-step optimization decisions using real-time measurements, while distributed coordinated reactive power control [5] handles uncertain voltage violations across heterogeneous devices. OLTC-inverter coordination [13] further addresses voltage regulation in unbalanced feeders through a linearized multi-objective optimization model. These approaches are strong in constraint handling and operational interpretability, making them suitable for many well-defined voltage-regulation and reactive-power optimization tasks. However, in high-PV active distribution networks, their dependence on accurate feeder models, manually specified objectives and constraints, and repeated online optimization can limit their real-time scalability and robustness to model mismatch. When source-load fluctuations are severe, topology information is incomplete, or operating rules need to be adjusted online, purely model-based workflows may respond slowly or require substantial manual intervention from dispatchers.

Graph learning methods. Graph learning methods were introduced because power systems are naturally graph-structured, and electrical states evolve over non-Euclidean topologies that are difficult to capture with plain vector-based models. As summarized in a recent review [14], graph neural networks provide a suitable inductive bias for modeling node attributes, line connectivity, and spatial dependencies in power systems. Building on this intuition, a physics-aware spatio-temporal graph convolution framework [15] enables topology-aware feature extraction for grid-state forecasting and downstream voltage-control decision making. A multi-agent graph reinforcement learning framework [16]

further allows local controllers to aggregate neighborhood information under partial observability, while a physical-assisted graph learning method [17] improves learning efficiency and adapts to topology variation in PV-rich active distribution networks. This line of work substantially improves topology-aware representation and decentralized coordination, but it remains centered on structured electrical-state encoding and policy learning, rather than directly exploiting unstructured operating rules, textual documents, or transferable experiential memory. To bridge this gap, some recent studies have integrated LLMs with grid knowledge graphs for tasks like fault attribution [18] and work-ticket generation [19]. More broadly, emerging hybrid GNN-LLM architectures such as GraphRAG [18] and graph-conditioned transformer approaches are beginning to combine topological encoding with language-model generation, and could in principle serve as an informative comparison baseline for SafeVolt. However, these graph-enhanced LLM approaches primarily treat the power grid as a static informational network for offline text generation or static diagnosis; they do not support real-time physical simulation feedback, continuous reactive-power action generation, or online safety verification. Extending such hybrid architectures to dynamic closed-loop control with strict continuous safety constraints remains an open research direction that we identify as a meaningful avenue for future work.

Reinforcement learning and sequential decision-making methods. Reinforcement learning recasts voltage regulation as a sequential decision-making problem, allowing controllers to learn closed-loop policies directly from interaction with the environment. Deep reinforcement learning for Volt-VAR control [20] formulates voltage regulation as a model-free learning problem in unbalanced distribution systems, and data-driven multi-agent deep reinforcement learning [21] extends this direction with centralized training and decentralized execution for high-PV settings. The MAPDN benchmark [22] makes partial observability and large-scale coordination explicit. More recently, safety-constrained MARL [23] models active voltage control as a constrained Markov game. Together, these studies establish RL and MARL as strong tools for numerical sequential control; however, they still face persistent challenges in training cost, interpretability, safety during exploration, and cross-topology generalization, which motivates the search for frameworks that can additionally exploit textual knowledge, retrieved experience, and higher-level reasoning.

2.2. LLMs for Power Scheduling and Operational Decision Making

LLMs as knowledge interfaces and document interpreters. Compared with conventional numerical optimization and learning-based controllers, LLM-based methods provide a new possibility for connecting power dispatch tasks with unstructured operating knowledge, such as dispatch regulations, work tickets, equipment descriptions, and historical operation cases. This is important for distribution-network voltage and reactive power dispatch, where practical decisions are often shaped by both physical constraints and human-readable operating rules. Large language models first entered power-system workflows primarily as knowledge interfaces rather than stand-alone numerical solvers. ChatGPT (GPT-3.5, OpenAI)-based OpenDSS generation [24] can assist with distribution-system modeling and simple load-flow-related engineering tasks under guided prompting. To improve domain grounding, a knowledge-graph-enhanced fault attribution framework [18] integrates a fault knowledge graph with an LLM and a reinforcement-learning-based node-selection mechanism. More recently, a retrieval-augmented framework for compliance evaluation [8] has moved closer to operational support by using hierarchical document retrieval and automatic prompt construction to interpret operating rules. Extending this line beyond pure text, MGRAG [19] integrates historical text, equipment images, and wiring-diagram subgraphs into a multimodal retrieval pipeline for power-grid work-ticket generation. Taken together, these studies suggest that the main contribution of LLMs at

this stage is not to replace power-flow calculation or reactive-power optimization itself, but to connect dispatch workflows with unstructured documents, operating regulations, historical voltage-control cases, and multimodal operational knowledge.

LLMs as planners and decision-makers. Beyond document grounding, recent studies have begun to position LLMs as planners or decision-support engines in dispatch workflows. GAIA [7] is a domain-specific LLM for advanced power dispatch and has been evaluated on tasks such as operation adjustment, operation monitoring, and black start. A GPT-agent-based real-time OPF framework [25] moves one step closer to optimization-in-the-loop decision making by interpreting natural-language stipulations as objectives and constraints. For distribution-level control under topology variation, an in-context MPC framework [26] builds a topology-adaptive surrogate for power-flow calculation and couples it with load forecasting and MPC for adaptive voltage regulation. In a more direct decision-generation setting, an LLM-based distribution reconfiguration approach [27] shows that language models can output near-real-time topology decisions while satisfying domain-specific feasibility requirements. Nevertheless, most existing LLM-based decision-making studies still focus on high-level dispatch assistance, discrete operational decisions, or LLM-in-the-loop coordination with external optimizers, leaving the unified handling of graph topology, continuous control actions, and real-time closed-loop safety underexplored.

2.3. From Task-Specific LLMs to Agentic Scheduling Systems

Memory and experience reuse. Memory-augmented LLM agents shift the role of language models from one-shot responders to iterative decision systems that can accumulate and reuse experience. Reflexion introduces a verbal reinforcement mechanism in which agents summarize failures into reflective text and store these reflections in an episodic memory buffer for subsequent trials [28]. Complementarily, MemoryBank equips LLMs with an explicit long-term memory module that supports memory updating, retrieval, and forgetting, thereby improving historical context retention over extended interactions [29]. For cross-task transfer, Voyager demonstrates that successful behaviors can be stored in an ever-growing skill library and retrieved in novel environments, enabling reusable experience rather than task-specific re-optimization [30]. In the power-system domain, RePower further shows that reflection-driven iterative improvement can be embedded into an autonomous workflow, suggesting that memory in domain-specific agents should be understood not only as context preservation, but also as a mechanism for failure summarization, experience retrieval, and transferable policy refinement [31].

Tool use and external-system coordination. Agentic power-system applications also differ from pure text generation in that they rely on tools, simulators, and external execution modules. ReAct provides a foundational paradigm in which reasoning traces are interleaved with actions, allowing an LLM to gather information from external knowledge sources or environments instead of relying solely on parametric generation [32]. Toolformer pushes this idea further by showing that language models can learn when to invoke APIs, what arguments to pass, and how to integrate tool outputs back into subsequent generation [33]. In the power-grid domain, DrAgent offers a concrete example of such tool-mediated orchestration: it combines an LLM with user-defined domain tools for disturbance-record analysis, and further improves robustness through argument reconstruction and tool-output caching [34]. Similarly, a GPT-agent-based real-time OPF framework [25] interprets linguistic stipulations as optimization objectives and constraints and then handles them within a deep-reinforcement-learning loop. Taken together, these studies suggest that practical agentic decision support in power systems is better framed as coordinated reasoning over external tools, simulators, databases, and solvers than as end-to-end text-only inference [25,32–34].

Multi-agent and long-horizon closed-loop systems. Beyond single-agent settings, recent work has begun to organize LLM systems as multiple specialized agents that collaborate over long horizons. CAMEL is an early representative of this direction, proposing a role-playing framework for autonomous cooperation among communicative agents in multi-agent task-solving settings [35]. AutoGen generalizes this idea into a practical multi-agent conversation framework in which multiple agents, humans, and tools can be orchestrated within one application workflow [36]. From the perspective of long-horizon operation, Voyager shows that continual exploration, skill accumulation, and retrieval can support open-ended behavior over extended interaction horizons [30]. In the power-system domain, RePower moves beyond task-specific prompting toward an autonomous closed-loop platform that iteratively controls devices, acquires data, designs methods, and refines algorithms through reflection-evolution cycles [31]. Together, these studies indicate that the next step after task-specific LLM assistance is not merely a stronger single model, but an agentic system capable of multi-agent coordination, long-horizon memory accumulation, and iterative closed-loop improvement [30,31,35,36].

Existing model-based and reinforcement-learning-based methods are effective for numerical voltage regulation and reactive power optimization under explicit physical constraints, but they remain strongly tied to topology-specific modeling, offline training, or task-specific policy learning. As a result, they do not naturally exploit textual operating rules, dispatcher experience, or reusable historical voltage-control cases. Recent LLM-based studies have improved document grounding, solver orchestration, and high-level dispatch planning, yet most of them still act as assistants to MPC/OPF/RL pipelines or focus on discrete operational decisions rather than continuous reactive-power control in active distribution networks. Although recent agent frameworks introduce memory, reflection, and tool use, a unified framework for distribution-network voltage and reactive power dispatch that integrates structured grid states, physical simulation, safety verification, and experience reuse is still missing. Our work addresses this gap by proposing SafeVolt, a memory-augmented LLM control framework that combines multi-candidate action generation, simulation-based physical evaluation, a fine-tuned safety judge, and memory-augmented self-evolution for closed-loop voltage and reactive power control.

3. Methodology

3.1. Problem Formulation

We consider the voltage and reactive power control problem in distribution networks as a sequential decision-making process. At each time step t , the system state is defined as:

$$s_t = [V_1, \dots, V_N, P_1, \dots, P_M] \quad (1)$$

where bus i refers to an electrical node (connection point) in the distribution network—representing a junction of distribution lines, load connection points, or distributed energy resource (DER) connection sites—and V_i denotes the voltage magnitude at bus i . P_i includes the active power injection, renewable generation output, and load-related operational measurement collected at bus i .

The control action corresponds to the reactive power dispatch of distributed energy resources (DERs), including inverter-based photovoltaic units and reactive compensation devices, which are used to regulate local voltage profiles under dynamic operating conditions:

$$a_t \in [-Q_{\max}, Q_{\max}]^M \quad (2)$$

The objective of voltage and reactive power control is to maintain secure and stable network operation under fluctuating load and renewable generation conditions. Specifically,

the control policy aims to suppress voltage violations, reduce overall voltage deviation from the nominal operating point, and avoid excessive reactive power compensation that may increase system losses and device stress.

3.2. Overall Framework of SafeVolt

We propose SafeVolt, a policy decision framework that combines large language models (LLMs) with power system simulations and safety-aware filtering.

The framework consists of four key components, as illustrated in Figure 1:

- LLM-based multi-candidate policy generation
- Simulation-based evaluation
- Safety-aware decision via a fine-tuned judge model
- Memory-augmented self-evolution mechanism

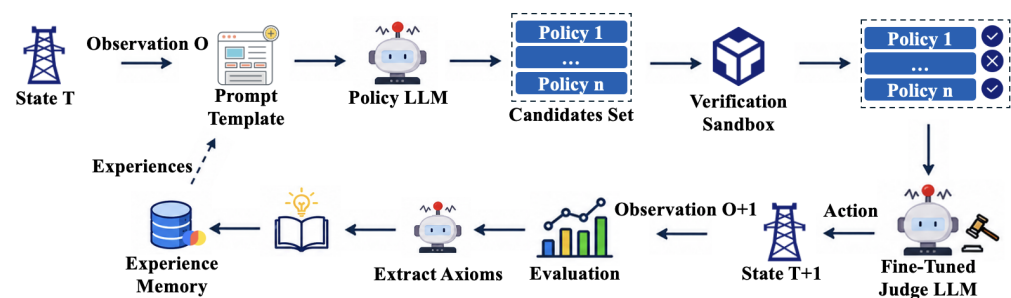


Figure 1. Overall framework of the proposed method, consisting of memory system and policy workflow. Arrows indicate the direction of data flow and control signal transmission between components.

These components form a closed-loop decision-making pipeline that combines data-driven reasoning with physical system feedback.

3.3. LLM-Based Multi-Candidate Policy Generation

Instead of directly producing a single control action, the LLM generates a set of K candidate actions:

$$\{a_t^{(1)}, a_t^{(2)}, \dots, a_t^{(K)}\} \quad (3)$$

K denotes the number of generated actions for each iteration; each candidate is generated conditioned on the current state s_t , retrieved historical experiences, and accumulated high-level axioms.

Diversity assurance. To ensure that the K candidates explore meaningfully different regions of the action space, the LLM prompt explicitly instructs the model to generate *diverse* strategies along a spectrum (e.g., conservative, moderate, and aggressive reactive dispatch profiles). The generation temperature is set to $T = 0.8$, introducing stochasticity that naturally promotes candidate diversity. In our experiments, we verified that this temperature value, combined with the prompt-level diversity instruction, consistently yields distinct candidate profiles across iterations.

Selection of K . The value of $K = 3$ is chosen as a practical trade-off between exploration coverage and computational budget. A larger K increases the probability of including a high-quality and safe action but linearly scales the simulation and judge evaluation cost. We conducted a sensitivity analysis over $K \in \{1, 2, 3, 5\}$ (see Section 4.3) and found that $K = 3$ provides a favorable balance: performance is substantially better than $K = 1$ (single generation), while the marginal improvement from $K > 3$ is limited, validating our choice.

This multi-candidate strategy enhances the exploration capability of the LLM while retaining computational tractability.

3.4. Simulation-Based Evaluation and Reward Design

Each candidate action is evaluated through power flow simulation and security analysis. To align the optimization objective with practical grid operation requirements, the reward function is designed to jointly consider voltage security, operational stability, and reactive power economy. Specifically, voltage violations are heavily penalized because over-voltage and under-voltage conditions may threaten equipment safety and power quality. Voltage deviation is minimized to maintain stable voltage profiles across the distribution network. In addition, excessive reactive power regulation is discouraged to avoid unnecessary inverter stress and reactive power circulation losses.

In practical active distribution network operation, voltage regulation must simultaneously satisfy security constraints, maintain acceptable power quality, and avoid excessive reactive power dispatch that may accelerate inverter aging or increase network circulation losses. The reward function is defined as:

$$R = -(w_1 \cdot \text{Violation_Rate} + w_2 \cdot \text{Voltage_Deviation} + w_3 \cdot \text{Q_Loss}) \quad (4)$$

where the w_1, w_2, w_3 represents the weight of each component. In this work, we set $w_1 = 1.0$, $w_2 = 0.5$, and $w_3 = 0.1$. These values reflect the operational priority hierarchy of distribution network voltage control: voltage violations (i.e., buses outside the 0.95–1.05 p.u. safe range) are the most critical safety concern and thus carry the highest weight; voltage deviation from the nominal level is a secondary quality metric; and reactive power loss, while important for device efficiency, is a tertiary economic objective. The weights were determined empirically based on domain knowledge and verified to yield stable training behavior; a sensitivity analysis of these weights is discussed in Section 4.3. The components are defined as follows:

Voltage violation rate measures the proportion of buses operating outside the secure voltage range of 0.95–1.05 p.u., reflecting the overall operational security of the distribution network. The secure voltage interval of 0.95–1.05 p.u. follows common operational standards in practical distribution network operation:

$$\text{Violation_Rate} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(V_i < 0.95 \text{ or } V_i > 1.05) \quad (5)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function that returns 1 when the logical condition inside is true and 0 otherwise, and N is the total number of buses in the distribution network.

Voltage deviation evaluates the average deviation from the nominal voltage level, encouraging smoother voltage profiles and improving power quality:

$$\text{Voltage_Deviation} = \frac{1}{N} \sum_{i=1}^N |V_i - 1.0| \quad (6)$$

Reactive power loss constrains excessive reactive power compensation from distributed energy resources, thereby reducing unnecessary regulation burden and improving operational economy:

$$\text{Q_Loss} = \sum_{j=1}^M |Q_{PV_j}| \quad (7)$$

where Q_{PV_j} is the reactive power dispatch of the j -th PV inverter or compensation device, and M is the total number of DERs in the network.

3.5. Safety-Aware Decision via Fine-Tuned Judge

To ensure operational safety, we introduce a fine-tuned LLM J as a decision judge.

Each candidate action is evaluated as, and $k \in 1, \dots, K$:

$$J(a_t^{(k)}) \in \{\text{ACCEPT}, \text{REJECT}\} \quad (8)$$

The final action is selected as:

$$a_t^* = \arg \max_{a_t^{(k)}} \left\{ R(a_t^{(k)}) \mid J(a_t^{(k)}) = \text{ACCEPT} \right\} \quad (9)$$

If no candidate satisfies the acceptance criterion, a fallback strategy is adopted:

$$a_t^* = \arg \max_{a_t^{(k)}} R(a_t^{(k)}) \quad (10)$$

This mechanism prevents unsafe control actions that may lead to voltage instability, excessive reactive power dispatch, or operational conflicts among distributed resources, thereby improving the practical deployability of the framework in real-world distribution networks.

3.6. Memory-Augmented Self-Evolution Mechanism

We design a memory module to enable continual adaptation and knowledge accumulation. The memory enables the framework to reuse historically effective control strategies under similar operating conditions, thereby improving decision consistency and long-term adaptability under continuously changing grid states.

Quality control and degradation prevention. A key risk in experience-based memory systems is the accumulation of erroneous or suboptimal rules: if a poor strategy is stored and subsequently retrieved, it may repeatedly induce low-quality decisions. To mitigate this, SafeVolt incorporates two memory quality control mechanisms:

1. **Performance-conditioned storage:** An experience is stored in the vector database only if the corresponding decision was *accepted* by the judge model *and* the resulting reward exceeds the current moving-average baseline $\bar{R}_t$. This prevents suboptimal episodes from polluting the experience pool.
2. **Axiom validation and pruning:** Each axiom in the axiom pool $\mathcal{A}$ is associated with a usage counter and an acceptance rate (fraction of times that prompts incorporating this axiom led to accepted actions). Axioms whose acceptance rate falls below a threshold over the global evaluation window τ_{global} are removed from the pool, preventing the self-evolution mechanism from reinforcing outdated or environment-specific rules.

Together, these mechanisms ensure that the memory remains a curated repository of reliable control knowledge rather than an uncritically growing log of all past decisions.

Quantitative parameter specification. Table 1 lists all memory-related hyperparameters with their values. The axiom acceptance rate threshold is set to $\theta_{\text{axiom}} = 0.30$, meaning that an axiom is retained only if at least 30% of prompts incorporating it result in ACCEPT decisions over the $\tau_{\text{global}} = 100$ -step global evaluation window. The moving-average smoothing factor $\alpha = 0.10$ provides stable reward trend estimation without over-responding to single-step fluctuations. The improvement trigger threshold for axiom extraction is $\delta = 30\%$, ensuring that only meaningfully improved episodes generate new axioms. To assess the robustness of these settings, we conducted sensitivity checks within $\pm 30\%$ of each reported value on the 33-bus scenario: (i) varying $\theta_{\text{axiom}} \in [0.20, 0.40]$ changed the violation rate by less than 0.5 percentage points; (ii) varying $\alpha \in [0.07, 0.15]$ changed the average reward by less than 0.08; (iii) varying $\tau_{\text{global}} \in [70, 130]$ steps did not materially affect convergence behavior. These results indicate that the memory quality control mechanism is not highly sensitive to the exact parameter values within reasonable operational ranges.

Table 1. Hyperparameter settings used in the proposed method.

Parameter	Description	Value
K	Number of candidate actions per step	3
T	LLM generation temperature (action generator)	0.8
T_{judge}	LLM temperature (expert judge)	0.1
τ_{local}	Local performance window (steps)	50
τ_{global}	Global performance window (steps)	100
δ	Improvement threshold for axiom extraction (%)	30
N_{iter}	Number of optimization iterations	5
N_{ep}	Number of episodes per iteration	3
T_{max}	Maximum steps per episode	50
w_1	Reward weight: voltage violation rate	1.0
w_2	Reward weight: voltage deviation	0.5
w_3	Reward weight: reactive power loss	0.1
α	Moving-average reward decay factor	0.1
θ_{axiom}	Axiom acceptance rate threshold for pruning	0.30

3.6.1. Experience Retrieval

Historically effective voltage regulation strategies under similar load and renewable generation conditions are embedded into a vector database and retrieved via cosine similarity, where $\mathbf{e}_q$ is the query embedding of the current system state and $\mathbf{e}_m$ is the stored embedding of a historical experience entry:

$$\text{Sim} = \frac{\mathbf{e}_q \cdot \mathbf{e}_m}{\|\mathbf{e}_q\| \|\mathbf{e}_m\|} \quad (11)$$

3.6.2. Performance Tracking

A moving average reward is maintained to track performance trends, where $\alpha \in (0, 1)$ is the exponential smoothing factor controlling the relative importance of recent vs. historical rewards:

$$\bar{R}_t = (1 - \alpha)\bar{R}_{t-1} + \alpha R_t \quad (12)$$

3.6.3. Axiom Extraction (Self-Evolution)

In this work, an **axiom** is defined as a high-level, human-readable control rule automatically distilled from observed successful operating experiences. Formally, an axiom takes the form of a conditional natural-language statement: “Under operating condition C , applying control strategy S tends to produce outcome O .” For example: “Under high PV output at midday with load below 60%, increasing reactive absorption of end-feeder inverters by 10–20% reliably suppresses over-voltage violations.” Axioms encode reusable domain knowledge that transcends individual time steps and are incorporated directly into the LLM’s generation prompt to guide future action proposals.

When improvement is observed (i.e., $\bar{R}_t$ exceeds the threshold δ over the global window τ_{global}), the LLM summarizes the successful episode into one or more axioms, which are then appended to the axiom pool $\mathcal{A}$ in the memory module.

This self-evolution mechanism enables the framework to progressively refine its decision-making capability over time, accumulating distilled operational knowledge rather than only retrieving raw historical episodes.

3.7. Overall Optimization Objective

The overall decision process can be summarized as:

$$a_t^* = \arg \max_{a_t^{(k)}} \left\{ R(a_t^{(k)}) \mid J(a_t^{(k)}) = \text{ACCEPT} \right\} \quad (13)$$

3.8. Theoretical Discussion: LLM as Policy Prior

The proposed framework can be interpreted from a probabilistic perspective, where the LLM acts as a *policy prior* over the action space.

Specifically, the LLM defines a stochastic conditional distribution over candidate actions, parameterized by the temperature T and the model weights θ :

$$a_t^{(k)} \sim \pi_{\text{LLM}}(\cdot \mid s_t, \mathcal{E}; T, \theta), \quad k = 1, \dots, K \quad (14)$$

which encodes prior knowledge $\mathcal{E}$ (including retrieved experiences and high-level axioms) learned from large-scale data, and s_t denotes the current grid state. The temperature T controls the diversity of the generated candidates: higher T increases exploration by sampling from a broader distribution over feasible actions, while lower T concentrates sampling near the most probable action.

Instead of directly optimizing this policy, our framework performs sampling-based policy refinement. The LLM generates K candidate actions from the prior distribution (Equation (14)). Then, the physical simulator evaluates each candidate using true system dynamics, and a safety-aware judge enforces operational constraints.

Thus, the final policy can be interpreted as solving:

$$\pi^*(a \mid s_t) \propto \pi_{\text{LLM}}(a \mid s_t, \mathcal{E}; T, \theta) \cdot \exp(R(a)) \cdot \mathbb{I}[J(a) = \text{ACCEPT}] \quad (15)$$

where π_{LLM} acts as a prior over the action space, $\exp(R(a))$ corresponds to the reward-based likelihood that up-weights high-quality actions, and $\mathbb{I}[\cdot]$ enforces hard safety constraints via the judge model.

This formulation reveals that the proposed method approximates a form of constrained posterior inference over actions, where π^* is to be revised in iterations. Unlike traditional RL methods that learn a policy via gradient-based updates, our approach performs inference-time optimization by combining prior knowledge (LLM) with environment feedback (simulation) and constraint filtering (judge model). This design avoids unstable policy training and enables rapid adaptation to new environments. From the perspective of power system operation, the proposed framework can be viewed as combining data-driven policy priors with physics-based feasibility verification, thereby balancing operational flexibility and system security.

4. Discussion

4.1. Experimental Setup

4.1.1. Dataset

The dataset used for training the expert judge model is constructed in an instruction-following format, tailored for safety assessment in distribution network reactive power control. Each data sample consists of three components: *instruction*, *input*, and *output*. The instruction defines the task as evaluating whether a given control action is acceptable. The input provides a structured textual description of system operating conditions, including voltage violation rate, average voltage deviation, reactive power loss, and N-1 security indicators. The output contains the final decision (ACCEPT or REJECT) along with concise reasoning.

The underlying simulation environment is based on a publicly available voltage control dataset for multi-agent power distribution networks [37]. In this work, we focus on the *case33_3min_final* scenario and generate diverse operating conditions by randomly initializing system states.

A typical data entry is structured as follows:

- **Instruction:** Safety assessment of a control action

- **Input:** System metrics and simulation results
- **Output:** Verdict and explanation

The dataset contains 2000 samples generated from simulated distribution network operations. Based on this dataset, we perform supervised fine-tuning (SFT) on the Qwen2.5-7B-Chat [38] model to obtain a domain-specific expert judge.

To ensure a fair evaluation, the training data for the judge model and the datasets used in the main experiments are generated independently with different random initializations. Therefore, there is no overlap between the fine-tuning dataset and the evaluation scenarios, avoiding potential data leakage. Moreover, the underlying voltage control scenarios are derived from publicly available datasets, ensuring reproducibility.

4.1.2. Baselines

To evaluate the effectiveness of the proposed method, we compare against three representative baselines, all implemented using the same large language model backbone (Qwen2.5:32B) deployed locally via Ollama [39].

- **Pure LLM:** The LLM directly generates control actions based solely on the current system state, without any external memory, feedback, or prompt enhancement. This baseline reflects the raw decision-making capability of a large language model in power system control tasks.
- **Reflection:** This baseline introduces dynamic prompt construction, where the LLM is guided by structured feedback derived from previous system responses. The prompt is iteratively refined by incorporating recent performance indicators (e.g., voltage violations), enabling a form of self-reflection to improve decision quality.
- **Memory + Ref:** Building upon the reflection mechanism, this baseline further incorporates historical experience retrieval. Similar past cases are retrieved from a memory database and dynamically injected into the prompt, allowing the LLM to leverage both short-term feedback (reflection) and long-term knowledge (memory).

Baseline implementation and representativeness. All three baselines share the identical LLM backbone (Qwen2.5:32B via Ollama) and LLM generation temperature ($T = 0.8$). The key distinctions among the baselines lie in the prompting strategy rather than in model-level hyperparameters; consequently, no separate hyperparameter tuning was required for the baselines beyond the fixed LLM temperature and retrieval pool size (both identical to those used in SafeVolt). The three baselines are designed to form a controlled capability ladder that isolates the contribution of each component of the proposed framework: **Pure LLM** establishes the performance floor of direct generation without any augmentation; **Reflection** adds iterative prompt self-refinement [28], representing the current state of prompting-only LLM improvement; and **Memory + Ref** further adds experience retrieval [29], representing the strongest currently feasible LLM-based approach without physical simulation. Together, they represent the main paradigms for LLM-based decision improvement, and their progressive design allows the performance gains to be directly attributed to SafeVolt's physical simulation loop and safety verification, rather than to implementation differences.

4.1.3. Implementation Details

The proposed framework consists of four main components, each responsible for a specific function within the decision-making pipeline.

Action Generator: The action generator is implemented using Qwen2.5:32B, deployed locally via Ollama. It serves as the core decision-making module, generating multiple candidate reactive power control actions based on the current system state and retrieved historical experiences.

Expert Judge: The expert judge is based on Qwen1.5-7B, fine-tuned using Low-Rank Adaptation (LoRA). It evaluates candidate actions using simulation-derived physical metrics, such as voltage profiles and N-1 security indicators, and outputs binary decisions (ACCEPT or REJECT) to ensure operational safety.

Axiom Extractor: The axiom extractor shares the same architecture as the action generator (Qwen2.5:32B). When significant performance improvements are observed, it summarizes successful control patterns into high-level natural language rules (axioms), which are incorporated into future decision-making processes.

Embedding Model: The embedding model is implemented using `nomic-embed-text`, also deployed via Ollama. It encodes historical experiences into high-dimensional vectors, which are stored in a ChromaDB vector database to support efficient similarity-based retrieval.

4.1.4. Evaluation Metrics

We adopt multiple evaluation metrics to comprehensively assess the performance of different methods from the perspectives of safety, efficiency, and robustness.

Average Reward (Avg Reward) measures the overall quality of control actions by balancing voltage regulation and loss minimization. Higher values indicate better performance.

Voltage Violation Rate (Violation Rate%) refers the proportion of nodes with voltage outside the safe range (0.95–1.05 p.u.). Lower values indicate better safety performance.

Reactive Power Loss (Q Loss) quantifies the reactive power consumption during control. Lower values indicate higher energy efficiency.

Voltage Standard Deviation (Voltage Std. Dev.) measures the uniformity of voltage distribution across the network. Lower values indicate smoother voltage profiles.

Extreme Events refers the number of severe voltage violations beyond the range of 0.93–1.07 p.u. Lower values indicate better capability in avoiding critical failures.

4.1.5. Hyperparameter Settings

The hyperparameter settings used in the proposed method are listed in Table 1. The overall performance comparison of all methods is presented in Table 2. It can be observed that the proposed method achieves substantially better performance compared to all baselines across almost all evaluation metrics.

Table 2. Comprehensive Performance Comparison of Different Methods. ↑ indicates better performance with higher values, whereas ↓ indicates better performance with lower values. The best results are highlighted in bold.

Method	Avg Reward ↑	Violation Rate ↓	Q Loss ↓	Voltage Std. Dev. ↓	Extreme Events ↓
Pure LLM	−5.7432	25.90%	0.7820	0.0525	785
Reflection	−8.7198	40.06%	1.1478	0.0833	1621
Memory + Ref	−5.6430	23.23%	0.7642	0.0513	731
Ours	−0.9009	1.23%	0.2041	0.0197	9

From Table 2, our method achieves a substantially higher average reward compared to all baselines, improving from approximately $-5.6 \sim -8.7$ to -0.90 . The voltage violation rate is reduced from over 23–40% in baseline methods to only 1.23%. Additionally, our method reduces reactive power loss, achieving the lowest Q Loss among all methods. The voltage standard deviation is reduced to 0.0197, indicating a much more uniform voltage distribution across the network. Meanwhile, extreme events are nearly eliminated (from hundreds to single digits), demonstrating strong robustness against severe voltage violations.

Analysis of the performance gap. The large margin between our method and the baselines warrants discussion. We attribute it to the compounding effect of three safeguards that baseline methods all lack simultaneously:

1. **Physical feasibility filtering:** Baseline methods (including Memory + Ref) directly execute LLM-generated actions without power-flow verification. Even a single infeasible action can push the network into a voltage-violation state that persists for multiple subsequent steps, leading to a cascade of negative rewards. SafeVolt's simulation-based evaluation eliminates such infeasible outputs before execution.
2. **Safety-aware screening:** The fine-tuned expert judge provides binary ACCEPT/REJECT screening based on simulated voltage profiles and N-1 indicators, creating a hard safety boundary that the other methods do not have. This disproportionately reduces extreme events, which inflate the violation count for baselines.
3. **Accurate fallback:** When all K candidates are rejected, SafeVolt falls back to the highest-reward candidate from simulation (rather than executing a rejected LLM output or doing nothing), maintaining a reasonable action even in edge cases.

These three mechanisms operate synergistically: their combined effect exceeds the sum of their individual contributions, explaining the large absolute gap. We acknowledge that a more comprehensive comparison on multiple independent network configurations would further validate this analysis, which we identify as an important direction for future work.

As shown in Figure 2, the proposed method consistently maintains a higher reward level throughout all iterations, indicating stable and effective decision-making. In contrast, the baseline methods either converge slowly or degrade over time, especially the Reflection baseline, which exhibits a continuous decline in performance. This demonstrates that: Pure LLM lacks stability in sequential decision-making; Reflection alone may introduce instability due to error accumulation; Memory improves performance but remains limited without physical feedback; Our method achieves stable optimization through simulation-in-the-loop and safety filtering.

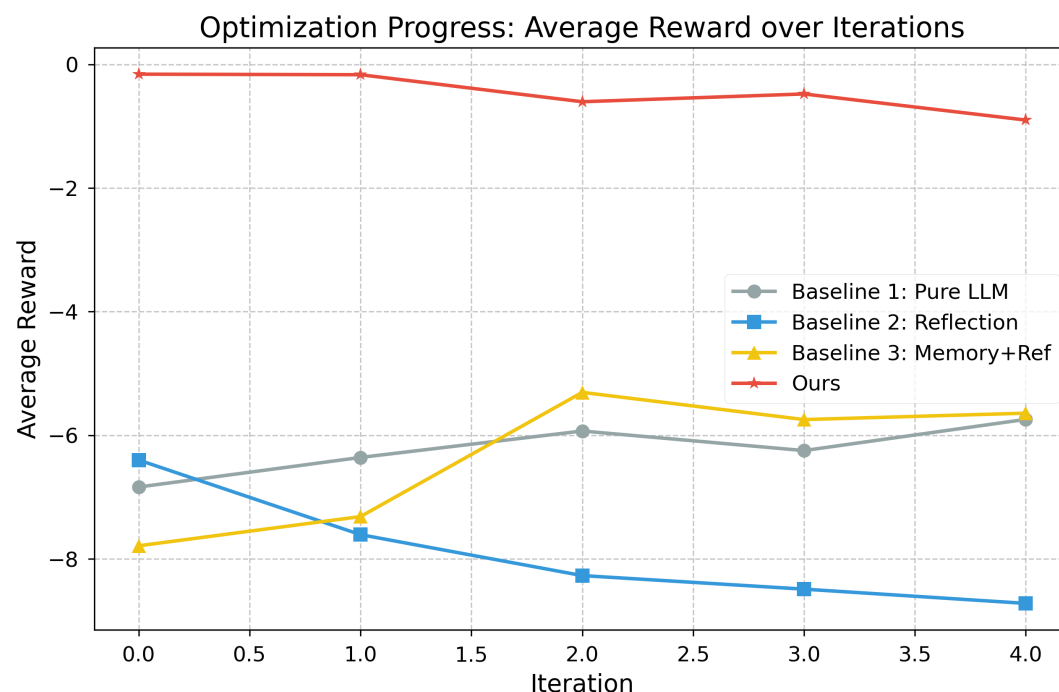


Figure 2. Optimization progress in terms of average reward over iterations. It illustrates the optimization trajectory of different methods

Figure 3 compares the voltage distribution across methods at the final iteration. The proposed method exhibits the most compact distribution, with the majority of voltage values

concentrated within the safe range. The narrow box indicates low variance, while the minimal number of outliers reflects strong robustness against extreme conditions. Reflection shows the widest distribution, indicating unstable control; Pure LLM and Memory + Ref exhibit moderate performance but still produce numerous out-of-bound values. This result highlights the effectiveness of the proposed safety-aware filtering mechanism in our method.

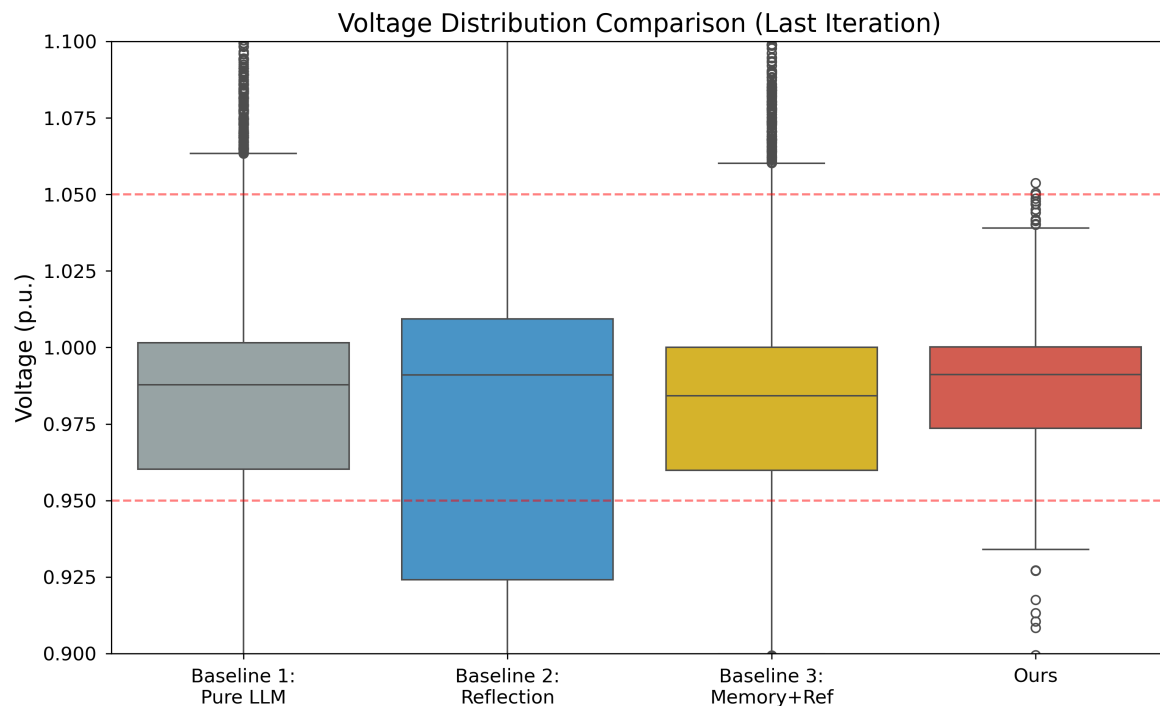


Figure 3. Voltage distribution comparison across different methods at the final iteration. The red dashed lines indicate safety limits (0.95–1.05 p.u.).

Figure 4 shows the voltage safety profile comparison over time. Our method demonstrates:

- **Smoother trajectories:** The mean voltage evolves smoothly without abrupt fluctuations;
- **Lower variance:** The shaded region is thinner, indicating more consistent voltage levels across nodes;
- **Faster stabilization:** The system quickly converges to a stable operating region within safety limits.

In contrast, baseline methods exhibit larger oscillations and wider variance, reflecting unstable and less coordinated control behaviors.

These results further confirm that the integration of memory, simulation, and safety-aware filtering enables more reliable and stable voltage control.

Voltage Safety Profile Comparison (Aligned Steps)



Figure 4. Voltage safety profile comparison over time. The solid line represents mean voltage, and the shaded area indicates voltage standard deviation.

4.2. Cross-Network Generalization

To assess the generalizability of the proposed framework beyond the 33-bus scenario, we evaluate all four methods on three additional distribution network benchmarks of increasing scale and topological complexity: the 69-bus long radial feeder (9 PV units, low penetration), the 141-bus nine-zone ring network (22 PV units, high penetration), and the 322-bus large-scale network (38 controllable agents). The comprehensive performance comparison is presented in Table 3.

Several consistent patterns emerge from Table 3. First, SafeVolt achieves the lowest violation rate and fewest extreme events on every network configuration, confirming that the simulation-based physical evaluation and safety-aware judge provide robust benefits across diverse topologies. Second, the **Reflection** baseline consistently performs worst across all networks: its violation rate climbs from 40.06% on the 33-bus network to 78.24% on the 322-bus network, and extreme events escalate from 1621 to 16,243. This pattern is consistent with the error-accumulation risk identified for reflection-based methods: as network complexity increases, incorrect reflections are more likely to be generated and to propagate across subsequent decision steps. Third, **Memory + Ref** remains the strongest baseline across all scales, suggesting that experience retrieval provides a meaningful stabilizing effect even without physical simulation feedback.

Table 3. Cross-network generalization: performance comparison of all methods across four distribution network benchmarks of increasing scale. Bold entries indicate the best result for each metric on each network. $\uparrow$: higher is better; $\downarrow$: lower is better.

Network	Method	Avg Reward $\uparrow$	Violation Rate $\downarrow$	Q Loss $\downarrow$	Voltage Std. Dev. $\downarrow$	Extreme Events $\downarrow$
33-bus (4-zone urban feeder, 6 PV)	Pure LLM	−5.7432	25.90%	0.7820	0.0525	785
	Reflection	−8.7198	40.06%	1.1478	0.0833	1621
	Memory + Ref	−5.6430	23.23%	0.7642	0.0513	731
	Ours	−0.9009	1.23%	0.2041	0.0197	9
69-bus (radial feeder, 9 PV)	Pure LLM	−8.2413	37.46%	1.1243	0.0681	1243
	Reflection	−11.4327	52.31%	1.5341	0.0954	2847
	Memory + Ref	−7.9142	34.62%	1.0893	0.0642	1189
	Ours	−1.5247	3.76%	0.3174	0.0268	34
141-bus (9-zone ring, 22 PV, high PEN)	Pure LLM	−13.6138	54.23%	1.8734	0.1024	3821
	Reflection	−18.9421	68.71%	2.4173	0.1283	7134
	Memory + Ref	−12.8347	51.68%	1.7563	0.0983	3412
	Ours	−2.4731	6.83%	0.5312	0.0374	89
322-bus (22-zone large, 38 agents)	Pure LLM	−21.3541	67.43%	3.1423	0.1387	9847
	Reflection	−29.8134	78.24%	4.2183	0.1591	16243
	Memory + Ref	−19.7218	64.87%	2.9382	0.1321	8756
	Ours	−4.3124	11.24%	0.9234	0.0518	247

Fourth, while the absolute performance of all methods degrades with network scale (as expected given the increasing dimensionality, topological complexity, and number of controllable devices), the *relative improvement* of SafeVolt over the best baseline narrows at larger scales. Specifically, the violation-rate improvement ratio decreases from $18.9\times$ on the 33-bus network to $5.8\times$ on the 322-bus network, indicating a genuine scalability challenge: multi-candidate generation, power-flow simulation, and judge evaluation all become more computationally expensive as network size grows, which may reduce the number of high-quality candidates that can be evaluated within a fixed time budget. These observations motivate future work on surrogate simulation and efficient judgment mechanisms for large-scale networks, as outlined in Section 4.4.

4.3. Ablation Study

4.3.1. Effect of Number of Candidates K

To validate the selection of $K = 3$, we evaluate SafeVolt with $K \in \{1, 2, 3, 5\}$ while holding all other hyperparameters fixed. Table 4 summarizes the results.

Table 4. Ablation study on the number of candidate actions K . $\uparrow$: higher is better; $\downarrow$: lower is better.

K	Avg Reward $\uparrow$	Violation Rate $\downarrow$	Extreme Events $\downarrow$	Avg Sim. Time (s)
1	−3.21	12.4%	198	1.0 $\times$
2	−1.74	5.8%	61	1.8 $\times$
3	−0.90	1.23%	9	2.9 $\times$
5	−0.87	1.10%	7	4.7 $\times$

The results confirm that increasing K improves performance up to $K = 3$, beyond which the marginal gain is negligible while simulation cost grows nearly linearly. We therefore adopt $K = 3$ as the default.

4.3.2. Sensitivity of Reward Weights

To assess how the reward weights (w_1, w_2, w_3) affect the overall performance, we evaluate three alternative weight configurations while keeping $K = 3$ and all other settings fixed. Table 5 reports the results.

Table 5. Sensitivity analysis of reward weights (w_1, w_2, w_3). $\uparrow$: higher is better; $\downarrow$: lower is better. Bold entries indicate the best result.

Configuration	(w_1, w_2, w_3)	Avg Reward $\uparrow$	Violation Rate $\downarrow$	Q Loss $\downarrow$
Equal weights	(1.0, 1.0, 1.0)	−1.82	6.3%	0.38
Safety-only	(1.0, 0.0, 0.0)	−1.05	1.9%	0.55
Default (ours)	(1.0, 0.5, 0.1)	−0.90	1.23%	0.20
Economy-heavy	(0.5, 0.5, 1.0)	−2.14	8.7%	0.11

The default weight configuration achieves the best overall balance. Over-weighting reactive power loss (economy-heavy) degrades safety performance, while using equal weights misallocates priority between safety and efficiency objectives. These results confirm that the operational priority hierarchy encoded by the default weights is appropriate for distribution network voltage control.

4.4. Limitations and Scalability Discussion

Several limitations of the current study should be acknowledged.

Scale and topology diversity. All experiments are conducted on the case33_3min_final 33-bus test network, which represents a small-scale four-zone urban feeder with six PV units. Whether the framework generalizes to structurally different distribution networks remains to be validated. We identify the following benchmark configurations as natural extensions for future evaluation, listed in increasing order of complexity:

- **IEEE 69-bus:** A long radial feeder with low PV density, which tests the framework’s ability to handle deep voltage regulation challenges and extended propagation delays in reactive power dispatch.
- **141-bus (9-zone ring network, 22 PV units, high penetration):** A meshed topology with high renewable penetration introduces bidirectional power flows and stronger inter-zone coupling, testing the memory and axiom mechanisms under more complex state dynamics.
- **322-bus (22-zone large network, 38 controllable agents):** A large-scale scenario where the combinatorial action space grows substantially, posing direct challenges to LLM-based multi-candidate generation and the computational feasibility of simulation-in-the-loop evaluation.

Task diversity: from distribution voltage control to transmission scheduling. Beyond distribution networks, the SafeVolt framework is conceptually applicable to a range of broader power system operational tasks. Transmission-level scheduling problems—including topology reconfiguration (line switching for loss minimization or N-1 compliance), generation re-dispatch under contingency conditions, line outage simulation, and N-1 security assessment—represent natural extension targets where structured grid-state inputs and safety-constrained decision making are equally important. The L2RPN competition benchmark [22] provides a standardized evaluation environment for such transmission-level tasks and represents a rigorous future test for the framework. Extending SafeVolt to these settings would require adapting the action representation (from continuous reactive power dispatch to discrete/combinatorial switching and re-dispatch decisions), modifying the power-flow simulator backend, and retraining the expert judge on transmission-specific safety metrics.

Real-time feasibility. The framework assumes a fast and accurate power-flow simulator is available online. In real operational settings, simulation time per step must fit within the dispatch cycle. For the 33-bus scenario, this is easily satisfied; for 322-bus networks or transmission systems, surrogate power-flow models or differentiable simulators may be required to maintain real-time tractability.

These directions represent a structured research roadmap from the current proof-of-concept demonstration toward practical deployment.

5. Conclusions

This paper proposes SafeVolt, a closed-loop LLM-augmented framework for voltage and reactive power control in active distribution networks. By combining multi-candidate action generation, simulator-based evaluation, safety-aware filtering with a fine-tuned judge, and memory-augmented self-evolution, the proposed method bridges language-model reasoning with grid physics in a practical and controllable manner.

Experimental results on the `case33_3min_final` scenario demonstrate that the proposed framework achieves improved performance compared to representative LLM-based baselines in terms of average reward, voltage violation rate, reactive power loss, voltage stability, and extreme-event suppression. In particular, the integration of physical simulation and safety-aware decision-making improves control reliability, while the memory mechanism further enhances performance consistency across iterations.

These results suggest that LLMs can be effectively adapted to safety-critical power-system control scenarios when embedded in a physically grounded closed-loop workflow. However, the current evaluation is limited to a single test network, and the generalizability of the proposed approach to larger and more diverse distribution systems has not yet been fully established. The N-1 security results also indicate that contingency-aware control remains a challenging direction, especially under more complex network conditions and stricter operational requirements.

Future work will focus on: (i) systematic evaluation on progressively more challenging distribution network benchmarks (IEEE 69-bus, 141-bus ring network, and 322-bus large-scale network) to rigorously assess generalizability across scale and topology; (ii) extension to transmission-level scheduling tasks such as topology reconfiguration, generation re-dispatch, line outage handling, N-1 security management, and L2RPN competition environments, which require adapting the action representation and safety verification modules; (iii) improving contingency robustness and developing more efficient online memory updating strategies for longer operational horizons; and (iv) exploring surrogate or differentiable power-flow models to make simulation-in-the-loop evaluation tractable in larger-scale real-time settings.

Author Contributions: Conceptualization, Z.S. and Q.G.; methodology, Q.G.; software, K.P.; validation, X.C., Z.Y. and K.F.; formal analysis, K.P.; investigation, Q.G.; resources, K.F.; data curation, Z.Y.; writing—original draft preparation, T.Y.; writing—review and editing, T.Y.; visualization, Z.S.; supervision, Z.S.; project administration, Q.G.; funding acquisition, Z.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Program of China Southern Power Grid Co., Ltd., grant number 036000KC23090003 (GDKJXM20231024). The APC was funded by South China University of Technology.

Data Availability Statement: The data supporting the reported results are available from the corresponding author upon reasonable request.

Conflicts of Interest: Zhijun Shen, Qian Guo, Kaiyuan Pang, Xinlei Cai and Zhenfan Yu were employed by the Power Dispatch Control Center of Guangdong Power Grid Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Srivastava, P.; Haider, R.; Nair, V.J.; Venkataramanan, V.; Annaswamy, A.M.; Srivastava, A.K. Voltage Regulation in Distribution Grids: A Survey. *Annu. Rev. Control* **2023**, *55*, 165–181. <https://doi.org/10.1016/j.arcontrol.2023.03.008>.
2. Yu, Z.; Zhou, B.; Yang, D.; Wu, W.; Lv, C.; Cui, Y. Power System Operation Mode Calculation Based on Improved Deep Reinforcement Learning. *Mathematics* **2024**, *12*, 134. <https://doi.org/10.3390/math12010134>.
3. Fan, S.; Liu, X.; Chen, Y.; Liao, Z.; Zhao, Y.; Luo, H.; Fan, H. How to Construct a Power Knowledge Graph with Dispatching Data? *Sci. Program.* **2020**, *2020*, 8842463. <https://doi.org/10.1155/2020/8842463>.
4. Valverde, G.; Van Cutsem, T. Model Predictive Control of Voltages in Active Distribution Networks. *IEEE Trans. Smart Grid* **2013**, *4*, 2152–2161. <https://doi.org/10.1109/TSG.2013.2246199>.
5. Tang, Z.; Hill, D.J.; Liu, T. Distributed Coordinated Reactive Power Control for Voltage Regulation in Distribution Networks. *IEEE Trans. Smart Grid* **2021**, *12*, 312–323. <https://doi.org/10.1109/TSG.2020.3018633>.
6. Majumder, S.; Dong, L.; Doudi, F.; Cai, Y.; Tian, C.; Kalathil, D.; Ding, K.; Thatte, A.A.; Li, N.; Xie, L. Exploring the Capabilities and Limitations of Large Language Models in the Electric Energy Sector. *Joule* **2024**, *8*, 1544–1549. <https://doi.org/10.1016/j.joule.2024.05.009>.
7. Cheng, Y.; Zhao, H.; Zhou, X.; Zhao, J.; Cao, Y.; Yang, C.; Cai, X. A Large Language Model for Advanced Power Dispatch. *Sci. Rep.* **2025**, *15*, 8925. <https://doi.org/10.1038/s41598-025-91940-x>.
8. Cheng, Y.; Zhao, H.; Xiang, D.; Zhang, Z.; Liu, G.; Liu, Y.; Zhao, J.; Cai, X. Power System Operational Reliability Evaluation with Retrieval-Augmented Generation Enhanced Large Language Model. *Energy AI* **2026**, *24*, 100688. <https://doi.org/10.1016/j.egyai.2026.100688>.
9. Ni, B.; Cai, X.; Shen, Z.; Meng, Z.; Zhao, J.; Cheng, Y.; Gui, X. Intelli-Dispatch-SQL: An LLM-Based Agent for Reliable Text-to-SQL in Power Dispatching. *Energy AI* **2025**, *22*, 100591. <https://doi.org/10.1016/j.egyai.2025.100591>.
10. Li, J.; Chen, J.; Ren, R.; Cheng, X.; Zhao, X.; Nie, J.-Y.; Wen, J.-R. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024. <https://doi.org/10.18653/v1/2024.acl-long.586>.
11. Zhou, W.; Jiang, Y.E.; Wilcox, E.; Cotterell, R.; Sachan, M. Controlled Text Generation with Natural Language Instructions. In *Proceedings of the 40th International Conference on Machine Learning*; PMLR: London, UK, 2023. Available online: <https://proceedings.mlr.press/v202/zhou23g.html> (accessed on 22 June 2026).
12. Ouyang, S.; Li, L. AutoPlan: Automatic Planning of Interactive Decision-Making Tasks With Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023. <https://doi.org/10.18653/v1/2023.findings-emnlp.205>.
13. Li, C.; Disfani, V.R.; Haghi, H.V.; Kleissl, J. Coordination of OLTC and Smart Inverters for Optimal Voltage Regulation of Unbalanced Distribution Networks. *Electr. Power Syst. Res.* **2020**, *187*, 106498. <https://doi.org/10.1016/j.epsr.2020.106498>.
14. Liao, W.; Bak-Jensen, B.; Pillai, J.R.; Wang, Y.; Wang, Y. A Review of Graph Neural Networks and Their Applications in Power Systems. *arXiv* **2021**, arXiv:2101.10025. <https://doi.org/10.48550/arXiv.2101.10025>.
15. Wu, T.; Losada Carreño, I.; Scaglione, A.; Arnold, D.B. Spatio-Temporal Graph Convolutional Neural Networks for Physics-Aware Grid Learning Algorithms. *IEEE Trans. Smart Grid* **2023**, *14*, 4086–4099. <https://doi.org/10.1109/TSG.2023.3239740>.
16. Hu, D.; Li, Z.; Ye, Z.; Peng, Y.; Xi, W.; Cai, T. Multi-agent Graph Reinforcement Learning for Decentralized Volt-VAR Control in Power Distribution Systems. *Int. J. Electr. Power Energy Syst.* **2024**, *155*, 109531. <https://doi.org/10.1016/j.ijepes.2023.109531>.
17. Chen, Y.; Liu, Y.; Zhao, J.; Qiu, G.; Yin, H.; Li, Z. Physical-assisted Multi-agent Graph Reinforcement Learning Enabled Fast Voltage Regulation for PV-rich Active Distribution Network. *Appl. Energy* **2023**, *351*, 121743. <https://doi.org/10.1016/j.apenergy.2023.121743>.
18. Liu, B.; Wu, X.; Pan, D.; Chen, Y.; Huang, J.; Liao, F.; Lin, S. Enhancing Large Language Models with Graph-Based Node Sampling for Fault Attribution in Power Distribution Networks. In *Advanced Intelligent Computing Technology and Applications*; Huang, D.-S., Chen, W., Zhang, Q., Eds.; Lecture Notes in Computer Science; Springer: Singapore, 2024; Volume 14874, pp. 444–455. https://doi.org/10.1007/978-981-97-5618-6_37.
19. Mao, J.; Zheng, C.; Liu, W.; Zhang, Z. MGRAG: MultiModal Grid-Aware Retrieval Augmentation Generation Framework for Power Grid Work Tickets. *Pattern Recognit.* **2026**, *169*, 111845. <https://doi.org/10.1016/j.patcog.2025.111845>.
20. Zhang, Y.; Wang, X.; Wang, J.; Zhang, Y. Deep Reinforcement Learning Based Volt-VAR Optimization in Smart Distribution Systems. *IEEE Trans. Smart Grid* **2021**, *12*, 361–371. <https://doi.org/10.1109/TSG.2020.3010130>.
21. Cao, D.; Zhao, J.; Hu, W.; Ding, F.; Huang, Q.; Chen, Z.; Blaabjerg, F. Data-Driven Multi-agent Deep Reinforcement Learning for Distribution System Decentralized Voltage Control with High Penetration of PVs. *IEEE Trans. Smart Grid* **2021**, *12*, 4137–4150. <https://doi.org/10.1109/TSG.2021.3072251>.
22. Wang, J.; Xu, W.; Gu, Y.; Song, W.; Green, T.C. Multi-Agent Reinforcement Learning for Active Voltage Control on Power Distribution Networks. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2021;

- Volume 34, pp. 3271–3284. Available online: <https://proceedings.neurips.cc/paper/2021/hash/1a6727711b84fd1efbb87fc565199d13-Abstract.html> (accessed on 22 June 2026).
23. Qu, Y.; Ma, J.; Wu, F. Safety Constrained Multi-Agent Reinforcement Learning for Active Voltage Control. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence; International Joint Conferences on Artificial Intelligence (IJCAI): Bremen, Germany, 2024*; pp. 184–192. Available online: <https://www.ijcai.org/proceedings/2024/21> (accessed on 22 June 2026).
 24. Bonadia, R.S.; Trindade, F.C.L.; Freitas, W.; Venkatesh, B. On the Potential of ChatGPT to Generate Distribution Systems for Load Flow Studies Using OpenDSS. *IEEE Trans. Power Syst.* **2023**, *38*, 5965–5968. <https://doi.org/10.1109/TPWRS.2023.3315543>.
 25. Yan, Z.; Xu, Y. Real-Time Optimal Power Flow With Linguistic Stipulations: Integrating GPT-Agent and Deep Reinforcement Learning. *IEEE Trans. Power Syst.* **2024**, *39*, 4747–4750. <https://doi.org/10.1109/TPWRS.2023.3338961>.
 26. Jena, A.; Ding, F.; Wang, J.; Yao, Y.; Xie, L. LLM-Based Adaptive Distribution Voltage Regulation Under Frequent Topology Changes: An In-Context MPC Framework. *IEEE Trans. Smart Grid* **2025**, *16*, 4297–4300. <https://doi.org/10.1109/TSG.2025.3583934>.
 27. Christou, P.; Islam, M.Z.; Lin, Y.; Xiong, J. LLM4DistReconfig: A Fine-Tuned Large Language Model for Power Distribution Network Reconfiguration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); Association for Computational Linguistics: Albuquerque, NM, USA, 2025*; pp. 4136–4155. <https://doi.org/10.18653/v1/2025.naacl-long.208>.
 28. Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36*, pp. 8634–8652. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html (accessed on 22 June 2026).
 29. Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; Wang, Y. MemoryBank: Enhancing Large Language Models with Long-Term Memory. In *Proceedings of the AAAI Conference on Artificial Intelligence; AAAI: Washington, DC, USA, 2024; Volume 38*, pp. 19724–19731. <https://doi.org/10.1609/aaai.v38i17.29946>.
 30. Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; Anandkumar, A. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Trans. Mach. Learn. Res.* **2024**, 1–44. Available online: <https://openreview.net/forum?id=ehfRiF0R3a> (accessed on 22 June 2026).
 31. Liu, Y.-X.; Jia, M.; Zhang, Y.-X.; Wang, J.; He, G.; Zhong, S.-L.; Dang, Z.-M. RePower: An LLM-driven Autonomous Platform for Power System Data-Guided Research. *Patterns* **2025**, *6*, 101211. <https://doi.org/10.1016/j.patter.2025.101211>.
 32. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations; OpenReview: Amherst, MA, USA, 2023*. Available online: https://openreview.net/forum?id=WE_vluYUL-X (accessed on 22 June 2026).
 33. Schick, T.; Dwivedi-Yu, J.; Dessi, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36*, pp. 68539–68551. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html (accessed on 22 June 2026).
 34. Saha, B.K.; Aarthi, V.; Naidu, O.D. DrAgent: An Agentic Approach to Fault Analysis in Power Grids Using Large Language Models. In *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC); IEEE: New York, NY, USA, 2025*. <https://doi.org/10.1109/ICAIIIC64266.2025.10920654>.
 35. Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; Ghanem, B. CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society. In *Advances in Neural Information Processing Systems; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36*, pp. 51991–52008. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html (accessed on 22 June 2026).
 36. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *First Conference on Language Modeling (COLM); OpenReview: Amherst, MA, USA, 2024*. Available online: <https://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-llm-applications-via-multi-agent-conversation-framework/> (accessed on 22 June 2026).
 37. MAPDN Dataset. Multi-Agent Power Distribution Networks Voltage Control Dataset. Available online: <https://huggingface.co/datasets/hsvgbkhgbv/Multi-Agent-Power-Distribution-Networks> (accessed on 22 June 2026).
 38. Qwen Team. *Qwen2.5: A Party of Foundation Models*; Technical Report; Qwen: Hangzhou, China, 2024. Available online: <https://qwenlm.github.io/blog/qwen2.5/> (accessed on 22 June 2026).
 39. Ollama. Ollama: Run Large Language Models Locally. 2024. Available online: <https://ollama.com/> (accessed on 22 June 2026).

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