ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

A Modular Architecture for Power-Sector Large Models Integrating Domain Knowledge and Physical Constraints

WeiXiang Qiao[1,*], *Jing Niu*[1], *Ke Shi*[1], *Kaibo Wang*[2] *and JiWei Jin*[2]

[1] Power Dispatching Control Center of Guizhou Power Grid Co., Ltd., GuiZhou, China
[2] Power Dispatching Control Center of Zunyi Power Supply Bureau of Guizhou Power Grid Co., Ltd., GuiZhou, China

SUMMARY: *Control-room application requires a language model to read operation instructions, match terms with buses, branches, generators and ratings; return the action that remains feasible according to AC power flow equations. In short, such general large-language models are good at producing fluent Dispatching or contingency Explanation but prone to recommending set-Point changes that break the dynamic equilibrium of active and passive power balance, exceeding branches' thermal Limits, or citing rules unrelated to the Current Topology. Based on the integration of the five audit components in this paper's modular power sector large model: domain knowledge router, topology-aware graph adapter, physics-constraint verification layer, solver bridge and response repair module. Benchmarking links to public transmission case, PGLib-style AC-OPF case, GEFCom-derived load and renewable profiles; solver trace; operation rule fragment; equipment description; task-specific question-and-answer pair. Dispatch reasoning with feasibility repair, state-estimation explanation, and contingency diagnosis are evaluated against a general LLM, retrieval-augmented LLM, supervised fine-tuned LLM, graph-adapted LLM, and physics-checked variant. The full architecture reached a 91.7% feasible-answer rate, a 1.43% mean OPF cost gap, a 0.0042 p.u. normalized power-flow residual, a 0.0069 p.u. voltage-magnitude MAE, and 89.1% macro-F1 for contingency diagnosis. Ablation results assign different responsibilities to each module: Physical verification eliminates most of the infeasible Responses; The graph adapter provides a larger boost when Topology is perturbed. Finally, Sensitivity Analysis has shown a relationship with the calibration between Knowledge Coverage and Physical loss-weighting is over-estimated to slow down repairs And text content incompleteness. This kind of result is in agreement with the large model of language generation, network representation, solver-and-communicator, and physical test mentioned above from the power industry.*

KEYWORDS: *Large-power domain models; Knowledge-based domains; Physical constraints; Retrieval-enhanced generation; Graph adapters; AC Power flows.*

1 Introduction

Modern Dispatch Centers and distribution-operation Platforms have received the same data from sensors alarms,_solver's results, outages Tickets, Market Rules, Equipment Documents, Operator Requests, etc. A single corrective-action question may mention a corridor name, while the relevant numerical evidence appears as voltage residuals, branch-loading percentages, or generator-limit flags. Therefore, there is a problem of objects being bound: The model needs to

*18286126843@163.com
https://doi.org/10.65102/is20261027
Published: 30 April, 2026

know that a line label in an instruction, a row in a power flow table, and a sentence in an operating rule all represent the same active-grid element.

Recently, power system AI work is moving towards dispatching assistance, event interpretation, technical documents being used by operators, as well as operator-friendly knowledge management [1, 2]. Large language models bring flexible interactions and code generation functions; grid-based foundation-model research focuses on the description of topological states, time-series data features, weather-driven uncertainties, simulated operation records, etc. [3, 4]. This improvement increases the input options available to the model but still fails to ensure that the re-dispatch recommendation or contingency interpretation provided is indeed valid within this activated system.

Dispatch reasoning has a separation of linguistics from engineering validity. A recommendation that is not only natural but also violates active or reactive balance, voltage limit, thermal rating, ramp limit, generator capability boundary constraints, etc. [5]. The evaluation of a power-sector model must therefore include feasible-answer rate, power-flow residual, OPF cost gap, repair success, latency, and robustness under load, renewable, and topology changes. Security and trust risks increase as follows: Prompt injection, unsafe tool call, manipulated operating data, ambiguous device reference may all generate dangerous recommendations. Incorrect terminal recognition or an outdated branch assessment are sufficient to invalidate a logically sound conclusion.

Here, the studied architecture comprises five interacting parts. The domain knowledge router translates technical terms, rules fragments and device abbreviations into grid entities. The topology-aware graph adapter is an active bus-branch-generator state [6, 7]. The solver bridge performs power flow, OPF and contingency analyses with insufficient cached evidence. The PhysicsConstraintVerifier finds imbalance, voltage, thermal, limit-generator and action conflict violations. Edit un-support device references, numeric value and recommended action in the response repair module to provide a corrected answer finally.

The validation scheme is based on operability rather than text similarity exclusively. Tasks include: Dispatch Reasoning with Feasibility Repair; State Estimation Explanation; Contingency Diagnosis under Renewable and Topology Uncertainty. The reported evidence includes AC feasibility, OPF cost gap, voltage error, power-flow residual, grounding after topology or rating changes, repair depth, and response latency. Public grid cases, Time Series Profiles, Solver Trace and Technical-Domain Text are converted into linked Samples, Ablations divide the impacts of Knowledge Grounding, Topology Adaptation, Physical Verification, Solver Calls and Repair.

2 Methods

2.1 Data, Knowledge, and Grid-State Object Construction

The benchmark is derived from four linked sources: transmission cases, time-varying operating profiles, technical-domain text and solver-derived labels. The IEEE39-terminal, IEEE118-port, IEEE300-terminal and several chosen PGlib-OPF cases provide buses, branches, generators, costs and limits to build graphs and perform AC feasibility tests [8-10]. Power-flow and OPF tags are created using a third-party Open-Source Environment that matches the pandapower and MATPOWER data standards. Load, wind, and solar trajectories are rescaled from public forecasting benchmarks, including GEFCom-style profiles [11, 12]. The text corpus includes equipment descriptions, rule fragments, event templates, solver-log summaries, and task-specific instructions-replies pairs.

After removing the faulty power flow results and duplicates, there are now approximately

132,000 operating points for reference. Use a 70-10-20 split of the scenario to form the train, validation and test sets without sharing perturbation seeds. The load Scaling range is 0.60-1.30 times the nominal demand. Around 10% to 80 per cent were from renewable sources; it would vary with the weather. Topological stress is an N-1 outage, generator outage; however, there are restrictions on N-2 scenarios. State-estimation noise integrates Gaussian analog noise and sparse gross errors; The gross-error rate is less than 8 per cent.

Solver record is contained in every sample. An OPF or power-flow run contributes feasible set points, binding line constraints, voltage-limit status, generator-limit status, residuals, and convergence flags. The explanatory trace records the violated Devices, tested Corrective Actions, and Accept/Reject results. These traces enter the natural-language supervision stage by matching devices with identifiers and aliases in graphs; therefore, a correct sentence cannot be associated correctly with an incorrect topology.

Table 1: Data Sources and Sample Construction.

Layer	Objects	Construction and scale	Role in the model
Grid topology	Buses, branches, generators, ratings, admittance terms	IEEE 39/118/300 and selected PGLib-style AC cases; topology perturbation through N-1 and limited N-2 events	Provides graph tokens and physical constraints
Temporal profiles	Load, wind, solar, ramping patterns	GEFCom-style profiles rescaled to 132,000 operating points with renewable penetration from 10% to 80%	Creates operating diversity and stress cases
Solver labels	AC power flow, OPF cost, residuals, overloaded elements	Power flow and OPF runs filtered for convergence and duplicate low-variation profiles	Supplies labels, feasibility flags, and repair reports
Technical corpus	Rules, ratings, equipment descriptions, event templates, solver-log text	18,200 evidence chunks with object aliases and source tags	Supports retrieval grounding and explanation
Instruction set	Dispatch, state-estimation, and contingency questions	68,500 filtered instruction-response pairs plus negative preference samples	Trains task responses and physical preference ranking

Before training or inference, each case will be converted into a uniform Sample Object. When the source case provides data for bus, branch, generator, load or transformer in the graph field. The document field contains proof-of-origin markers tagged with objects' alias information. The label field records the dispatch target, state-estimation target, contingency class, feasibility flag and solver residual. The representation makes the model be limited to the stored domains which were seen in the activity-grid situation. Each sample which is marked by time index contains four constituent parts: one graph object, one document aggregation, one task direction, and the corresponding labels. The graph object does the recording of node features, edge attributes, and topology-status flags. The set of labels contains physical goal objects, text goal objects, and feasibility condition states. Downstream modules are permitted to only read these recorded fields, which therefore limits the possibility that an unrecorded state variable or an unverified device identifier can be generated in the response.

The domain evidence was indexed using a limited ontology of buses, Branches, Generators, Transformers, Loads, Measurements, Ratings, Contingencies, Constraints and Corrective Actions. Each class stores aliases and required attributes. Firstly resolve the subject in the instruction, and subsequently collect pieces of evidence tagged by provenance. Therefore, an overload query is connected to the branch rating, topological state, Redispatch evidence and permissible corrective measures before generating the response.

Quality control removes solver failures unrelated to actionable operating states, aliases that cannot be mapped to grid objects, and repeated operating points across data splits. The scenario metadata record load level, renewal rate, topology status and contingency types. The retained sample is a series of linked cyber-physical records, not individual prompts.

Engineering units are listed explicitly. Voltage is stored in p.u., angle in radians, power in base-MVA p.u., line loading as a percentage of thermal rating, and cost as a normalized objective value. Branch and Generator states are Boolean. The graph adapter obtains scaled quantities and learns; The solver bridge maintains original units throughout a feasible check.

Instruction-response pairs are produced through determinist templates, solver-log translation and expert-style paraphrasing. Solver tracks record the Bus ID, overloading Branch, Binding limit and tested correction results. Invalid responses are eliminated from the positive training set, and those still valid become negative preference samples.

The hard-case pool is stored in the test Design. Includes: renewable-heavy voltage-support failures; Topology change without any modifications to the name of the lines; State estimation residual due to topological error rather than bad data. Only one working condition can execute multiple Task Forms simultaneously without going beyond the training-test line.

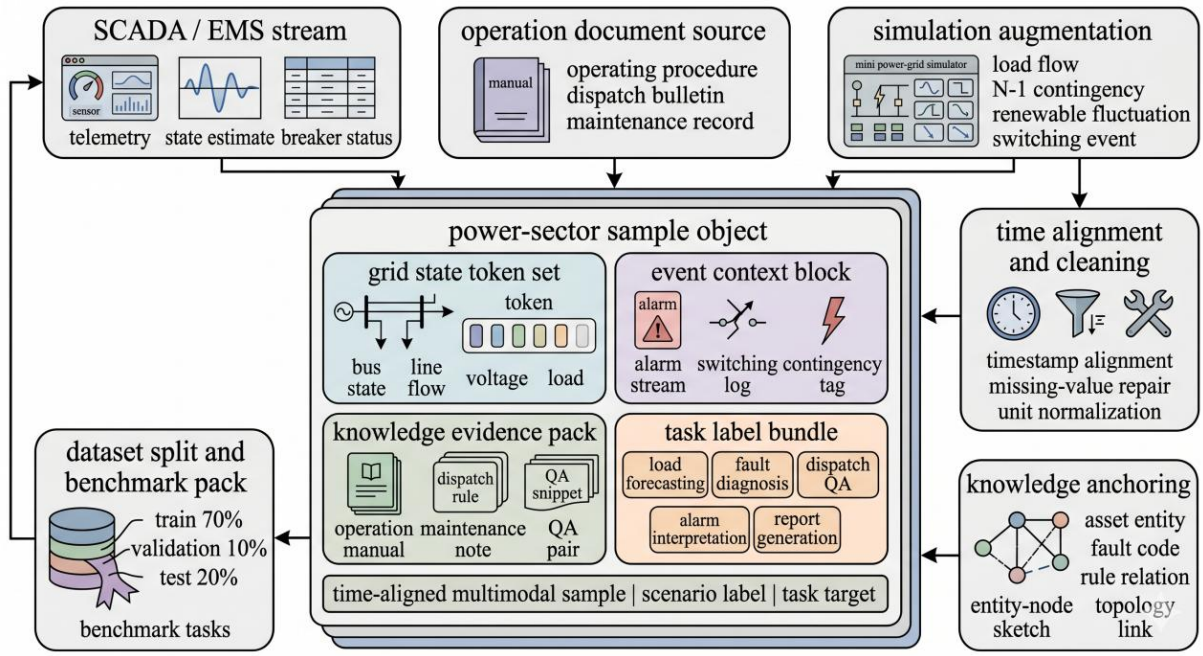


Figure 1: Power-sector Data Organization and Task Mapping Mechanism.

Figure 1 connects the four evidence sources through lines, states, documents and labels of Objects. The key relation is alias-controlled Binding; Technical documents and solver trails are linked to the current Bus, Branch, Generator or Contingency at which a change affects the Model. This boundary can eliminate two typical failures of citation: citing rules outside the active topology and resolving a network state without matching operators' terms.

2.2 Model Architecture and Knowledge-Physics Coupling

A transformer decoder is selected for the language basis, and in addition to this, there are five power-specific sub-networks. The knowledge router provides rules and equipment proof. Graph Adapter Converts The Active Grid Into Bus, Branch And Generator Tokens. The Physics Layer validates power-flow residuals and operates-limit faults. When the cached record cannot satisfy the solver criteria after calling power flow, OPF, and contingency operations. Edit the answer through a violation report and evidence trail in the repair module. Based on four identifiable lines of inquiry: instruction, obtained evidence; The graph representation that knows topology and the physical check state were together used as condition inputs to generate the response. To every single task, the module which generates responses got the instruction from the operator, took out the domain-related evidence, the topology information that is encoded by graph, and the report that verifies the physical state. These input data were recorded together with the produced answer, therefore response mistakes can be traced back to specific origins, including retrieval does not match, graph coding mistake, physical verification does not pass, or not complete on-site check proof during the process of task carrying out.

The knowledge router performs object parsing before dense retrieval. Extracts object mentions, task types, physical quantities, and requested output forms; Then filters out candidates from the active grid status. Only the evidence that matches aliases and operating conditions is approved. The router will also give inadmissible-action evidence, such as a locked branch or generator ramp violation, and therefore deny an action that has been rejected by this device.

Graph Adapter refers to active networks in terms of Graph Tokens. Bus Tokens are Voltage Injection, Load, Generation Zones, Measurement Quality Attributes. Branch Tokens bear Impedance, State, Temperature Rating, Load Information and Terminal-Bus Data. Message Passing takes place in a Bus-Branch Graph and conducts local pooling around restricted nodes. Contingency analysis updates edge-status masks and neighbourhood embeddings for tokens entering the decoder during this process.

Physical-layer verification includes checking for a power balance failure of candidates, voltage-limit violations, thermal Limit Violations generators' limits conflicts; The check verifies that the suggested actions correspond to the elements in the active admittance matrix or within the permitted range of operation. AC power-flow residual is utilized as the main numerical confirmation signal before answer acceptance carries out. The verification module has recomputed the active-power and reactive-power flows that are contained in the candidate voltage magnitudes, phase angles, and bus-admittance items. These recomputed flows have after that been compared with the candidate active and reactive power injections that exist at every bus. The residual carries out the summary of the mismatch in the whole network, and the voltage-limit violations are hence included as one extra penalty item. This numerical result was recorded into the physical-checking document, hence permitting unworkable replies to be refused or fixed before they are sent back to the operator.

When the cached state cannot verify a candidate action, it calls the solver bridge. The dispatch question may cause AC power flow, OPF or local feasible re-establishment upon proposing a set-point adjustment. Contingency Questions trigger an N-1 check through branch-loading and voltage-qualified items. Residual Ranking, Meter-Quality Metrics Return state estimation questions; Numerical verification is still at the solver level; language generation has not been realised yet.

Solver calls are limited to claims that can change operating feasibility: generator set-point changes, switching actions, set-point adjustments, violation assertions, and contingency-related corrective actions. Routine definitions, event summaries, and cached-residuals avoid the entire

OPF process. Returns convergence state, limit violation, residuals' absolute values, changed devices number, as well as the minimum correction value given by the solver.

Receiving the candidate answer, evidence bundle and verification report from the repair module. Only spans that contradict the evidence or physical verification are changed; this includes numbers, devices' IDs, action suggestions, etc. No more than four repairs at any time. Finally, the answer is accepted when its feasibility score has reached a certain level or it cannot find any feasible actions within this scope of constraints in the solver's report.

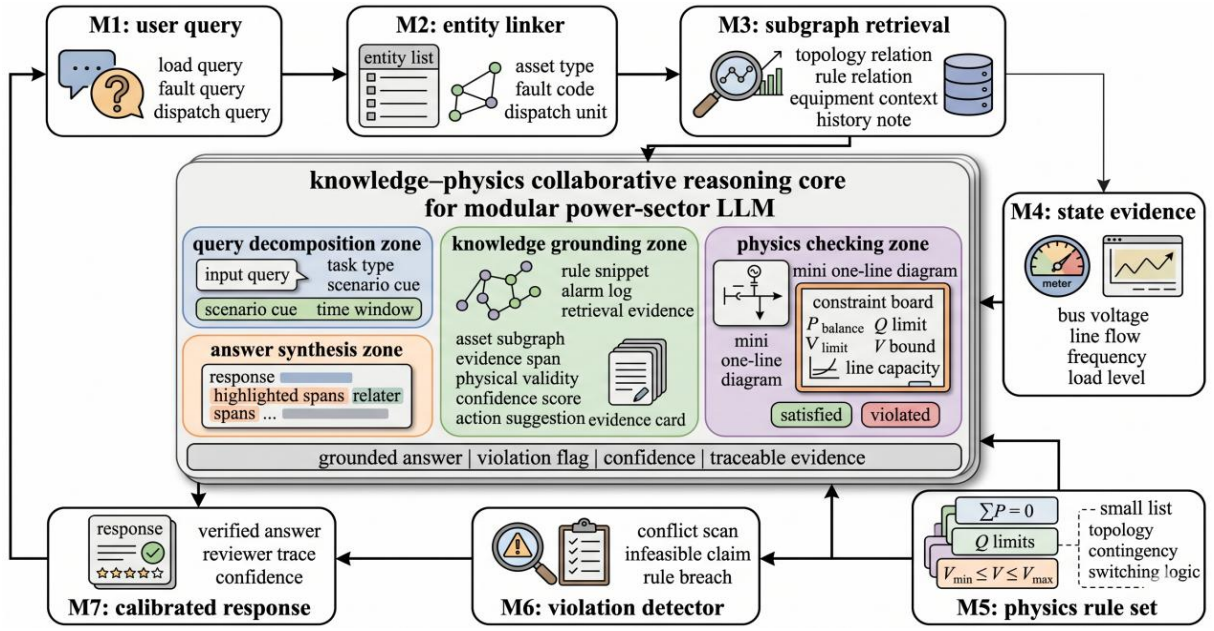


Figure 2: Knowledge-Physics collaborative reasoning Architecture.

Figure 2 presents the knowledge-physics coupling used during inference. The router and graph adapter provide evidence and active-topology tokens to the decoder; The physical layer and solver bridge verify the candidate answer before releasing it. The repair path closes the gap of rule evidence, network state and finally taken actions.

Only compact summary data is swapped by the graph adapter with the physical layer. Adapters send neighbourhood embeddings, affected-device lists, and topology change indicators. Physical layer outputs the residuals, violation categories, and responsible devices. Evidence chunks retain source tags and object identifiers, graph tokens remain aligned with bus, branch, and generator indices, and repair records store each changed span with its reason.

2.3 Training Objective, Baselines, and Evaluation Protocol

Four-stage training will be conducted. First, the language backbone is continued on technical text and solver logs. Secondly, a supervised training task uses filtered instruction-response pairs; Thirdly, low-rank and quantisation-aware adapters have been added to the routers and responses in this paper. The fourth option is: Feasible answers and Lower-Residual Answers in the physical preference-ranking Stage. Retrieval-Augmented Generation provides document grounding ([17]) and keeps prompts task-specific ([18]).

Based on drawing physics-informed and graph-structured power-system learning, such as [19-23], as well as learning-based OPF and reinforcement-learning evaluation models [24, 25]. Inside Language-Grounded Decision Support, the above ideas have been applied. Task-balanced sampling avoids dispatch case domination in supervised training. Hard negatives are preserved if they mention that the correct device or rule has violated some physical restriction.

The training target that we set has combined supervised-response learning, retrieval grounding, graph-alignment supervision, physical-residual control, and repair-ranking preference. The learning method with supervised responses uses labeled examples to train the model that it can generate answers which conform to task requirements. The retrieval-based grounding makes the answer be limited within the retrieved domain evidence. The supervision of graph alignment causes the generated reply to keep consistent with the graph representation that has topology awareness. The control of physical residual gives penalty to responses which break the feasibility of AC power flow. The preference of repair ranking further guided the model to select corrected responses that have lower feasibility risk. These component parts were appraised one by one, therefore the obtainments from physical possibility, evidence base, graph coherence, and language polishing could be separated in the ablation experiment results.

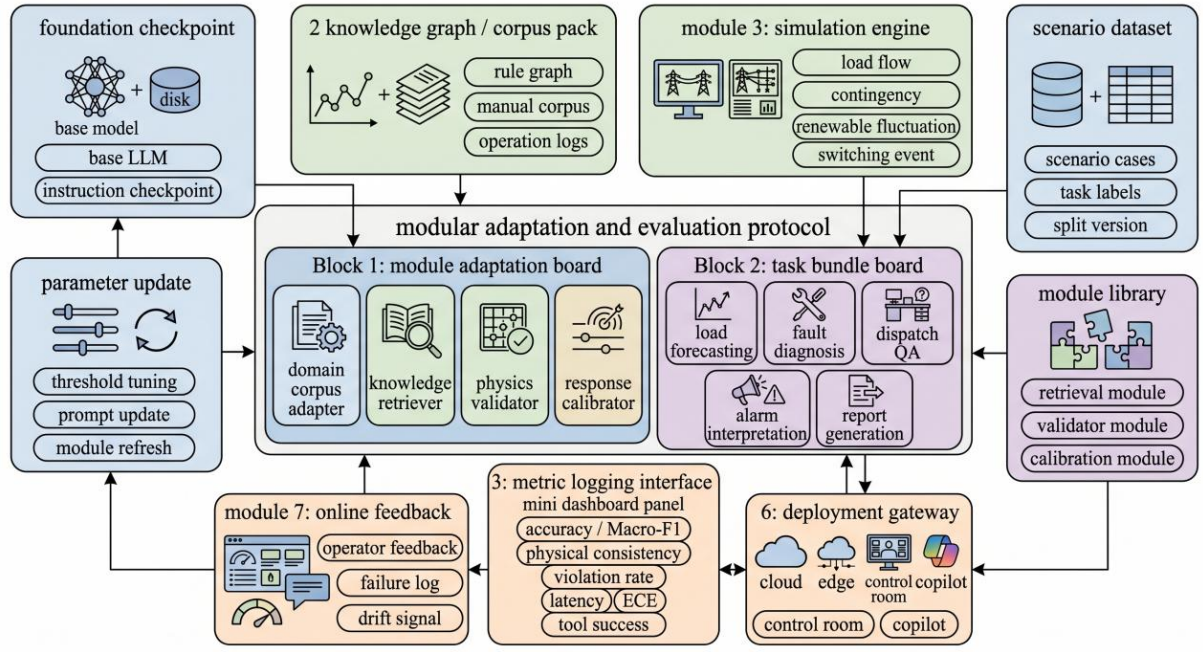
Table 2 lists the comparison Set. The general LLM receives the prompt and numerical context as text. RAG-LLM retrieves additional documents and lacks graph tokens or physical attestations. The fine-tuned LLM uses the supervised instruction set without explicit feasibility verification. Graph-constructed LLM receives a topological graph token, but it will not trigger the solver. Physics-checked LLMs verify the generated answer upon decoding, and this check is not included in the repair process. Proposed MPSLM combines the entire five modules. One-modulator ablation tests the separation of languagescoring, topologicalrepresentation, solverscaling verification, physical testing and correction recovery modules.

Table 2: Baseline method(s) and Ablation variant(s).

Model	Knowledge router	Graph adapter	Physical check	Solver bridge	Repair loop
General LLM	No	No	No	No	No
RAG-LLM	Yes	No	No	No	No
Fine-tuned LLM	No	No	No	No	No
Graph-adapted LLM	No	Yes	No	No	No
Physics-checked LLM	Partial	No	Yes	No	No
Proposed MPSLM	Yes	Yes	Yes	Yes	Yes
Ablation variants	One module removed at a time	One module removed at a time	One module removed at a time	One module removed at a time	One module removed at a time

Evaluations of three tasks' families. Dispatch reasoning returns corrective actions or no-action decisions and is scored by feasible-answer rate, OPF cost gap, power-flow residual, line and voltage violations, and latency. State-estimation explanation is scored by voltage-magnitude MAE, residual-ranking accuracy, grounding rate, and explanation consistency. Contingency Diagnosis: Macro F1 score, device identification accuracy, Physical action validity, Latency.

The test Design includes normal unseen cases and stress-test cases, etc. Stress cases use 60%–80% renewable penetration, 1.15–1.30 load scaling, and topology perturbations affecting up to 15% of switchable lines. Figure 3 shows the valid route for separating the training-time filter from the test-time repair.



Figures 3: Modular Adaptation Deployment and Evaluation Procedure.

Figure 3 distinguishes between the two processes commonly merged. During training, infeasible responses will be excluded from the positive set or converted into preference pairs. In test, unreasonable output will not be repaired immediately but sent directly to the model's repair cycle. Feasible-answer rate is therefore measured before manual correction, while repair effectiveness is measured after model-side revision.

Retrieval Depth, Graph-Token Length, Physical-Loss Weight and Repair Depth are chosen from the Validation Set. The test division will not be reported initially. Each variant runs with five random seeds, and figures report means with scenario-level bootstrap bands.

Implementations use a 7-BC-size Transformer decoder combined with low-rank adaptors for attention and feedforward projection operations. The graphs Adapter and retrieval Router were independently trained; then jointly calibrated. Only when there is a numerical control action, state value change, contingency label set to trigger solution execution.

Metrics are expressions of operations. A feasible option should not generate any violations after power flow verification. The OPF cost difference relative to the solver benchmark. Power-flow residual is reported in p.u., voltage error uses bus-voltage MAE, and macro-F1 accounts for imbalanced contingency classes. Grounded answers must match with the active Grid Object of cited evidence.

Failure analysis categorises as follows: missing grounding; incorrect limits; Power flow misalignment; action conflicts; Unclear reasoning. Latency includes retrieval, graph-token construction, physical verification, solver invocations and repair costs. After post-acceptance of main numeric indicators, separate individual evaluation will be conducted in this part without repairing;

3 Results and Discussion

3.1 Cross-Task Performance and Robustness under Operating Stress

The first Result Block checks if the Architecture can enhance dispatch Reasoning, State estimation Explanation, and Contingency Diagnosis in a similar scenario Split. Table 3 shows

the raw data. As shown in Figure 4, normalize the higher-scored items as feasible answer rate, macro-F1 score, grounded answer rate; Then Normalise the low-scored item:OPF cost gap,voltage error,power-flow residual,lateness etc.. This combination maintains smooth, but unattainable outcomes are not scored as excellent results.

Table 3: Quantitative Comparison of the Three Evaluation Tasks.

Model	Feasible answer rate / %	OPF cost gap / %	PF residual / p.u.	Voltage MAE / p.u.	Contingency macro-F1 / %	Grounded answer / %	Latency / s
General LLM	61.4	5.82	0.0198	0.0214	71.3	56.2	0.76
RAG-LLM	73.8	4.10	0.0145	0.0169	78.2	72.9	1.24
Fine-tuned LLM	77.1	3.58	0.0132	0.0144	80.2	67.6	0.88
Graph-adapted LLM	82.6	2.75	0.0101	0.0111	84.6	71.5	1.36
Physics-checked LLM	86.9	2.14	0.0068	0.0098	85.5	76.4	1.77
Proposed MPSLM	91.7	1.43	0.0042	0.0069	89.1	84.6	1.82

The proposed MPSLM reports the strongest metric combination in Table 3: 91.7% feasible-answer rate, 1.43% OPF cost gap, 0.0042 p.u. residual, 0.0069 p.u. voltage MAE, 89.1% contingency macro-F1, and 84.6% grounded-answer rate. The feasible answer rates reached 73.8%, 77.1%, 82.6%, 86.9%, and exceeded 90% respectively by Feasibility for a general LLM (LML), RAG-LLM), finely tuned LLM) adapted in physics-checked LLM). Overall, it is about 5% above an ideal state and lags for less than half a second.

Figure 4 shows the normalised cross-task profiles for all baselines and modular alternatives.

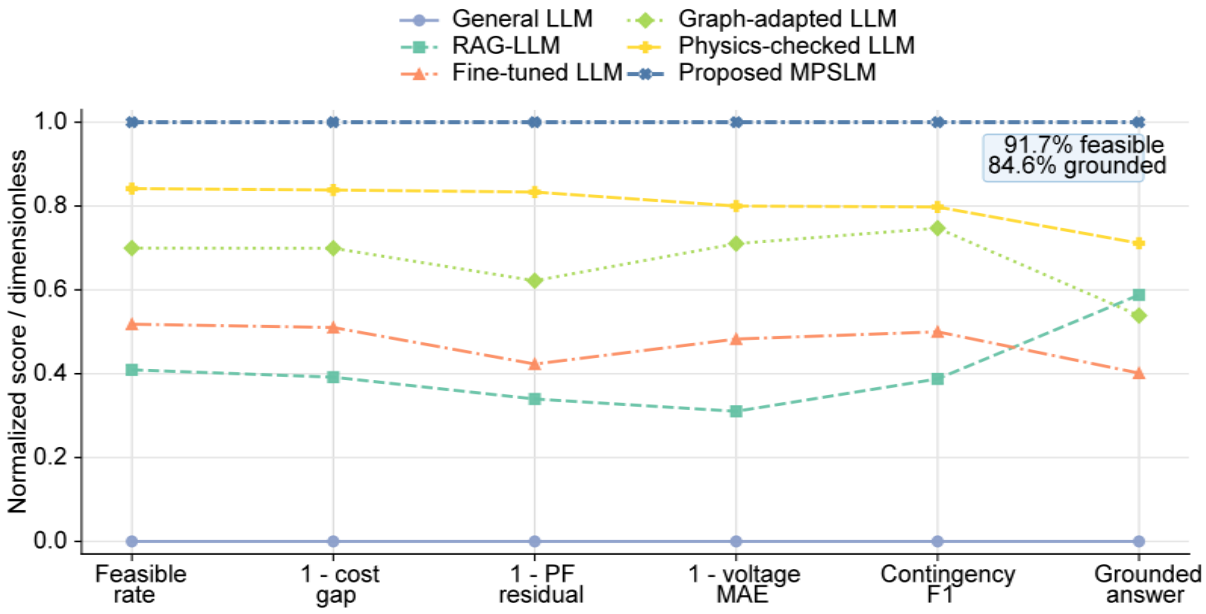


Figure 4: Cross-task Normalised Performance.

As shown in Figure 4, all base lines fail in different links of operation. The general LLM is weakest on residual and grounding. RAG-LLM improves evidence retrieval but keeps infeasible actions. Fine-tuning improves response format without reliable provenance. Graph Adaptation Reduces Voltage Error and Improves Contingency F1. Physics checking removes more infeasible outputs but lacks a graph- and evidence-conditioned repair step. Because all of the following data sources have been included in a single-validation path: retrieved evidence,

active topology, solver output, and repair decision.

Numerical deviation exceeds the specified range of deviations in grid operation standards. The proposed model's 0.0069 p.u. voltage MAE is 35.5% lower than the graph-adapted variant and 59.2% lower than RAG-LLM. Its 0.0042 p.u. residual is 51.2% lower than the physics-checked baseline and 71.0% lower than RAG-LLM. Contingency labels on their own cannot constitute a response until the voltages or flows associated with them have been verified as well.

Figure 5 Tests the same variants under renewable penetration and topology change.

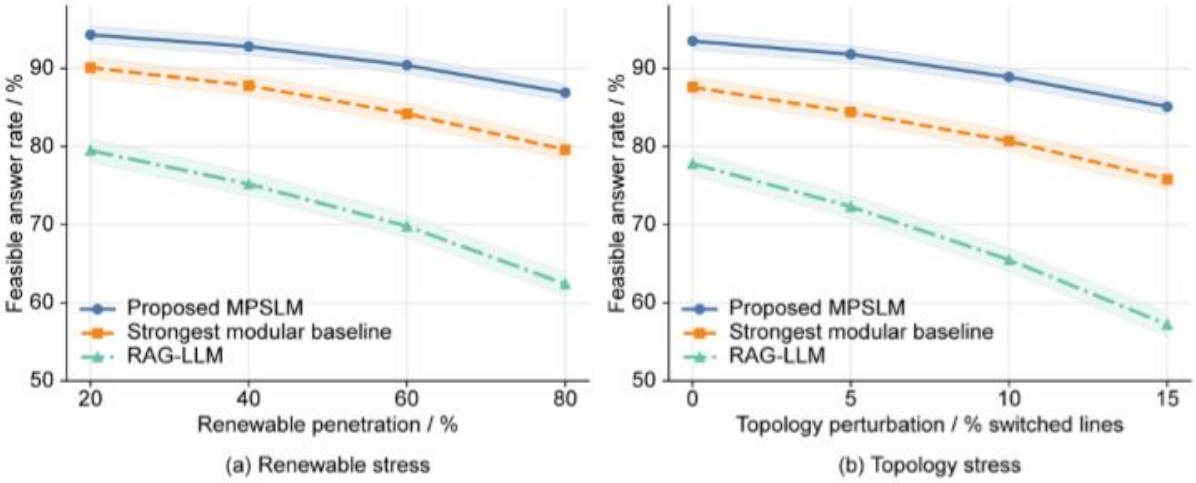


Figure 5: Robust to renewable and topology stress.

When renewable penetration rises from 20% to 80%, the proposed model's feasible-answer rate decreases from 94.3% to 86.9%. Physics-checked baseline decreases from 90.1% to 79.6%; RAG-LLM also shows a drop, from 79.5% to 62.4%. At 80% renewable penetration, the full architecture keeps a 7.3-point margin over the strongest baseline and a 24.5-point margin over RAG-LLM.

Topological Perturbation Produces the Same Ranking. With 15% switchable-line perturbation, the proposed model keeps 85.1% feasibility, while the graph-adapted baseline reaches 75.8% and RAG-LLM reaches 57.2%. RAG-LLM fails because retrieved evidence is not updated with branch status. Graph adaptation can help the model understand active topology; however, it fails to identify thermal or voltage conflicts during the solver verification process. All components in the complete model have been activated; therefore, the largest margin has not decreased.

Deployment is narrow and based on metrics. The average LLM has a 5.82 per cent OPP costs, plus a small residual of only 0.0198 pu. RAG-LLM reaches 72.9% grounded-answer rate but keeps a 0.0145 p.u. residual and weak stress feasibility. The gains of the full model come from changing the validation route rather than lengthening the prompt.

3.2 Module Ablation and Physical-Constraint Sensitivity

The second result module examines how effective the individual components are separately. In each of them, the removed component is different; thus, there will be various configurations for this experiment. Figure 6 shows the original values in a heat map to make feasible, cost, residual, voltage error, grounding and latency comparable among different units.

Figures Six identifies the Physical Verification Layer as having a large proportion of feasible protection.

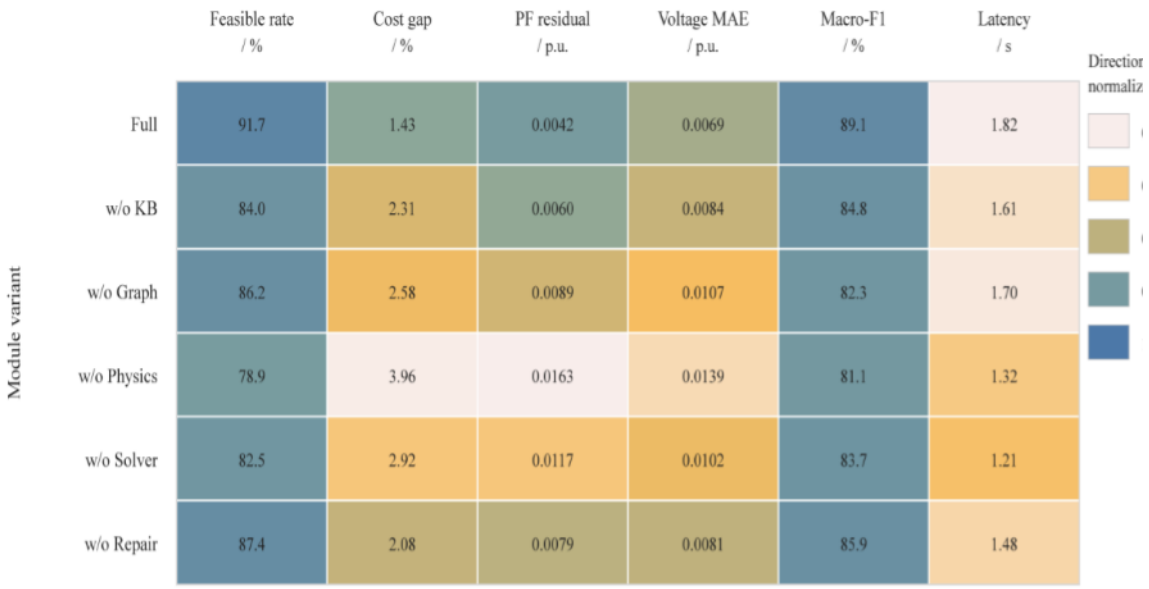


Figure 6: Module Ablation Matrix.

In Figure 6, removing the physical layer lowers feasible-answer rate from 91.7% to 78.9%, raises OPF cost gap from 1.43% to 3.96%, and increases residual from 0.0042 p.u. to 0.0163 p.u. Text supervision cannot learn the constraint of equality or inequality at this stage reliably. Removing the graph adapter lowers feasible-answer rate to 86.2%, raises cost gap to 2.58%, increases voltage error to 0.0107 p.u., and lowers contingency macro-F1 to 82.3%. The Knowledge Router mainly affects the grounded-answer rate, dropping from 84.6% to 71.3%, and also increases costs by a relatively high amount of 2.31. Removing the solver bridge leaves cached cases partly covered but reduces feasibility to 82.5% and raises residual to 0.0117 p.u. Removing repair leaves many near-feasible candidates unresolved, with 87.4% feasibility and 0.0079 p.u. residual.

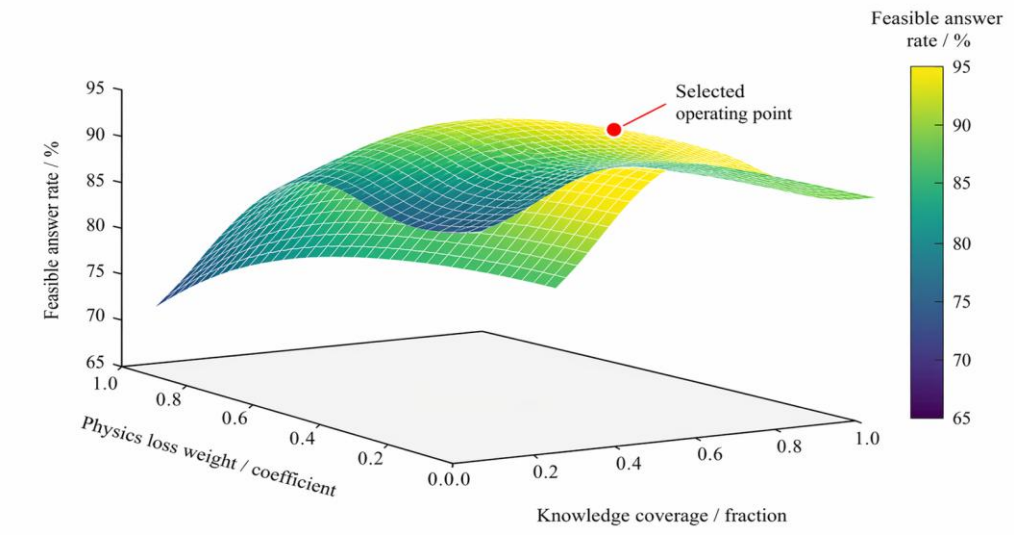


Figure 7: Three-Dimensional Response Surface

Figure 7 places knowledge coverage and physical-loss weight on the same response surface. The selected operating point lies near 0.82 knowledge coverage and 0.68 physical-loss weight, where feasible-answer rate is close to 91.7%. Coverage gains increase rapidly from 0.20 to 0.60,

then level off thereafter. Physical-loss weight is non-monotonic: a small value restricts the answers poorly; when larger, it hinders repairs and weakens textuality. On the Surface, an increase in physical load does not lead to a corresponding gain in recognitional strength but can result in non-feasibility conclusions after expanding coverage areas.

3.3 Efficiency, Case Analysis, Error Sources, and Deployment Implications

Figure 8 presents the cost of increased validation. Feasibility at 82.4 per cent and delay of only 0.92 seconds. At repair depth 3, feasibility reaches 91.7% with 1.82 s latency. Depth-4 only adds an additional 0.4% in terms of practicality and reaches 92.1%; The Latency is increased to 2.38s. The usable operating range is at depths of 2 and 3.

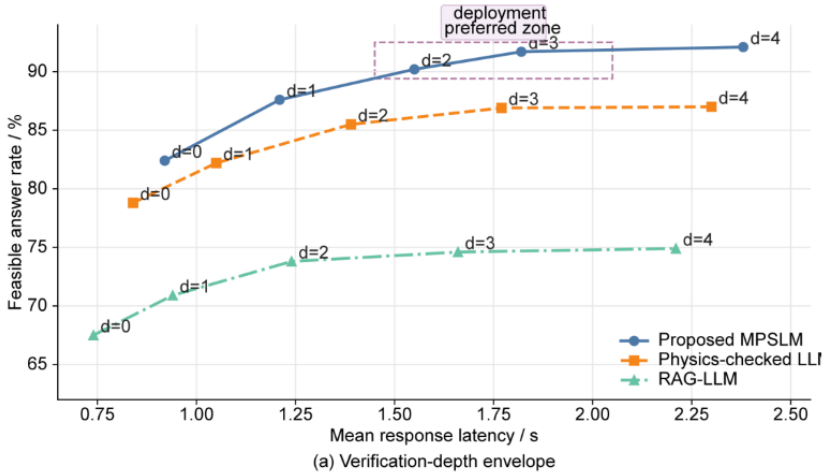


Figure 8: Accuracy-Latency Deployment Envelope.

Figure 8 shows the knee point at depth 3. Depth 2 gives 88.6% feasibility at 1.36 s, while depth 3 gives 91.7% at 1.82 s. Depth-4 is only somewhat feasible, and latency has exceeded two seconds. The physics-checked baseline increases to 86.9%, exceeding that of the full model, but still lower than it; Repair is not based on graph tokens or retrieved evidence. RAG-LLM increases from 67.5% to 74.9%; Retrieval plus verification cannot be expected to turn an infringement into a corrected act.

Figure 9 reports a 24-hour dispatch case under high renewable penetration.

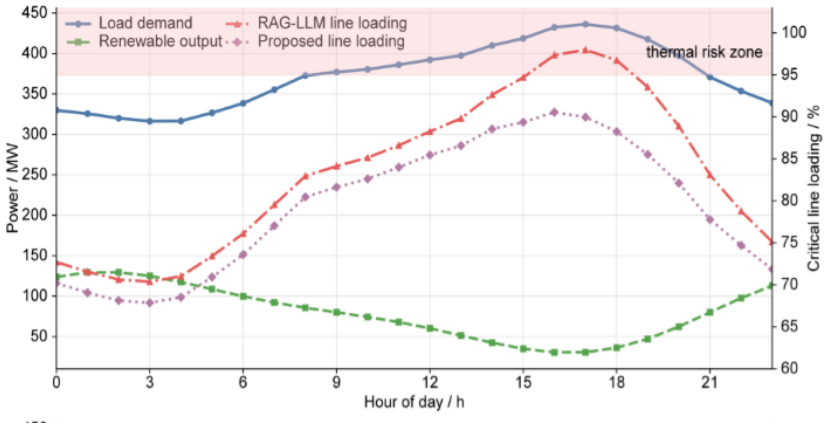


Figure 9: Twenty-four-hour Dispatch Case.

In Figure 9, hours 17–20 combine rising load with falling renewable output. RAG-LLM finds the appropriate thermal level, yet it suggests an upgrade for the generator that exceeds the critical threshold by 5%, thus increasing the load in another area. The proposed model first produces a similar candidate, then detects the violation, evaluates an alternative through the solver bridge, and revises the action. Finally, it will name the key line, report on load reduction, link together generator selection with verified transfer paths.

Figure 10 Separates residuals' Error Sources and Post-Improvement Effectivity under different Operating Scenarios.

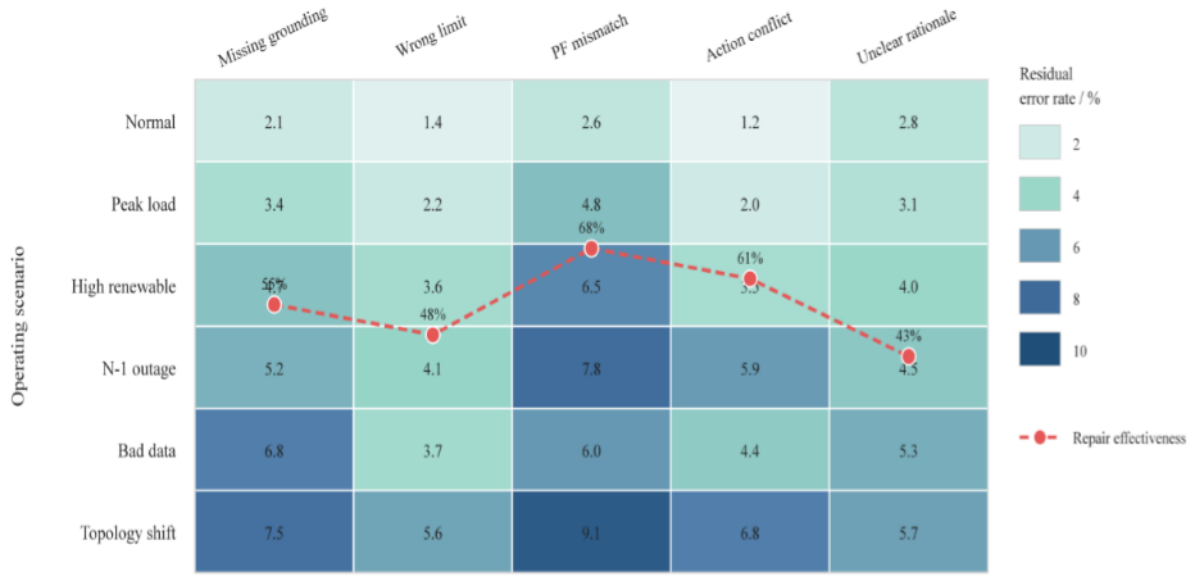


Figure 10: Error-source Distribution and Repair Effectiveness.

Residual errors concentrate in topology-shift and N-1 outage cases. Topological shift, Power Flow Discrepancy: 9.1%; Action Conflict Rate: 6.8%. N-1 failure outage value is 7.8%, N-2 fault outage is 5.9%. Repair is most effective for power-flow mismatch, with 68% of such errors corrected, because many cases can be resolved through redispatch or reactive-support adjustment. Unfounded explanation has the least improvement rate of 43%.

Residuals separated as matrices of two types. Physical risk manifests as a residual discrepancy or an action conflict, and requires solver interactions, topology blocking masks, etc. The communication-risk case, while lacking a clear basis and unclear cause, needs to retrieve more evidence, manage aliases, etc., for handling. The reduction of ordinary-scenario residuals by repair is 97.8%, while the drop in Action Conflicts is 96.4%; The residual offset exceeded 7.8% after topology transformation; therefore, path transfer needs to be inspected by humans first.

Most of the conditional latencies for the verification still fall within a tolerable range. Explanatory-type Tasks Usually Avoid Calling Solvers. Dispatch and emergency tasks call solvers for answers with numbers of actions, states, contingency labels or device-attached claims. Most accepted answers use two or three verification steps and remain below two seconds. Severe topology-shift cases are better flagged for review than forced into a fast response.

See Figures 4-10 for the places where architecture assists and those areas with lingering risks. The entire model enhances cross-task performance, maintains greater adaptability to renewable power supply and topology changes, acquires recognisable modules; needs joint calibration between evidence coverage and physical strength; can serve as an objective validity

range. The rest of the error concentrates on topology-switch and out-of-service scenarios; therefore, it can be used as an assistant for operators without being considered fully autonomous closed-loop controller.

4 Conclusion

Developed a modular architecture for power-sector large-scale models to ensure that the domain evidence, active topology, solver output, physical check, and response repair functions are independent audit units. On the reported benchmark, the full model reached 91.7% feasible-answer rate, 1.43% OPF cost gap, 0.0042 p.u. power-flow residual, 0.0069 p.u. voltage error, and 89.1% contingency macro-F1. These valuations show that a power system language Answer must have object binding and AC-whether check to ensure its application in Dispatching and emergency planning purposes.

Firstly, the data layer provides an unchanging reference for running this model. Convert public grid cases, time-series profiles, technical documents and solver traces to graph, state, evidence and label types. Link the retrieved rules with buses, branches, generators, ratings, etc., and reduce mismatches between operators' language descriptions of power system states and numeric grids.

Secondly, the ablated experiments have varying module functions. Physical verification shows the smallest change in unrealistic results. The graph Adapter Optimises topological transmission and condition monitoring. The solver bridge and repair module transform near-feasible candidates into revised suggestions. The combined route outperforms the prompt expansion, retrieval only and unsupervised fine-tuning separately.

Third, the scope remains limited to steady-state AC feasibility, dispatch reasoning, state-estimation explanation, and contingency diagnosis. The reference does not include transient stability, protection coordination, market-clearing functions, electromagnetic-transient processes, or live-utility telemetry systems. Therefore, this architecture is considered to be verifiable but not necessarily autonomously Grid Control.

Further validation should add dynamic simulation, protection-rule checking, cyber-security filtering, telemetry governance, tool-call authorization, operator-facing audit logs, and field tests with utility data. In the future evaluation reports, provide feasibility grounding for repairs and data latencies to reflect multiple operational risks separately.

Funding

This work was supported by Research and Application of Key Technologies for Improving the Reliability of Distribution Network Protection Devices.

About the Author

WeiXiang Qiao, male, Han ethnicity, born in January 1993. Graduated from Guizhou University with a bachelor's degree and a bachelor's degree in engineering. Currently serves as a relay protection specialist in the Power Dispatching Control Center of Guizhou Power Grid Co., Ltd., main research direction being power system and automation. Email: 18286126843@163.com.

Jing Niu, female, Ethnic Han, born in October 1984, graduated from Guizhou University with a postgraduate education and a master's degree. She currently serves as the Relaying Protection Specialist at the Power Dispatching and Control Center of Guizhou Power Grid Co., Ltd., main research direction is power system and its automation. Email:

15286085692@163.com.

Ke Shi, female, Ethnic Hui, born in September 1993, graduated from Newcastle University with a master's degree. Currently serves as a relay protection specialist in the Power Dispatching Control Center of Guizhou Power Grid Co., Ltd., main research direction being power system and automation; Email:18786755904@163.com.

KaiboWang, male, Han ethnicity, born in January 1989. Graduated from Guizhou University with a Bachelor of Engineering degree. Currently serves as Relay Protection Specialist at the Power Dispatching Control Center of Zunyi Power Supply Bureau, Guizhou Power Grid Co., Ltd., with primary research focus on Power Systems and Automation. Email: wangkb202507@163.com.

Jiwei Jin, male, Ethnic Han, born in March 1993, graduated from North China Electric Power University with a bachelor's degree in engineering. He currently serves as the Relaying Protection Specialist at the Power Dispatching and Control Center of Zunyi Power Supply Bureau, Guizhou Power Grid Co., Ltd. His main research direction is power system and its automation. Email:15120121743@163.com.

References

[1] Yao, Q., Fang, F., Chen, Y., et al. (2025). AI large models for power system: A survey and outlook. IET Smart Energy Systems, 1(1), 3-21.

[2] Majumder, S., Dong, L., Doudi, F., et al. (2024). Exploring the capabilities and limitations of large language models in the electric energy sector. Joule, 8(6), 1544-1549.

[3] Hamann, H. F., Gjorgiev, B., Brunschwiler, T., et al. (2024). Foundation models for the electric power grid. Joule, 8(12), 3245-3258.

[4] Huang, C., Li, S., Liu, R., et al. (2024). Large foundation models for power systems. In 2024 IEEE Power & Energy Society General Meeting (PESGM) (pp. 1-5).

[5] Cheng, Y., Zhao, H., Zhou, X., et al. (2025). A large language model for advanced power dispatch. Scientific Reports, 15, 8925.

[6] Ruan, J., Liang, G., Zhao, H., et al. (2024). Applying large language models to power systems: Potential security threats. IEEE Transactions on Smart Grid, 15(3), 3333-3336.

[7] Tu, S., Zhang, Y., Zhang, J., et al. (2024). PowerPM: Foundation model for power systems. In Advances in Neural Information Processing Systems (Vol. 37, pp. 115233-115260).

[8] Klamkin, M., Tanneau, M., & Van Hentenryck, P. (2025). PGLearn: An open-source learning toolkit for optimal power flow. arXiv, arXiv:2505.22825.

[9] Varbella, A., Amara, K., Gjorgiev, B., et al. (2024). PowerGraph: A power grid benchmark dataset for graph neural networks. In Advances in Neural Information Processing Systems (Vol. 37, Datasets and Benchmarks Track).

[10] Babaeinejadsarookolaee, S., Birchfield, A. B., Christie, R. D., et al. (2019). The power grid library for benchmarking AC optimal power flow algorithms. arXiv, arXiv:1908.02788.

[11] Thurner, L., Scheidler, A., Schäfer, F., et al. (2018). pandapower—An open-source Python tool for convenient modeling, analysis, and optimization of electric power systems. IEEE Transactions on Power Systems, 33(6), 6510-6521.

[12] Hong, T., Pinson, P., Fan, S., et al. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. International Journal of Forecasting, 32(3), 896-913.

[13] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998-6008).

[14] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877-1901).

[15] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.

[16] Dettmers, T., Pagnoni, A., Holtzman, A., et al. (2023). QLoRA: Efficient finetuning of quantized LLMs. In Advances in Neural Information Processing Systems (Vol. 36, pp. 10088-10115).

[17] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459-9474).

[18] Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (Vol. 35, pp. 24824-24837).

[19] Huang, B., & Wang, J. (2023). Applications of physics-informed neural networks in power systems: A review. IEEE Transactions on Power Systems, 38(1), 572-588.

[20] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707.

[21] Ngo, Q.-H., Nguyen, B. L. H., Vu, T. V., et al. (2024). Physics-informed graphical neural network for power system state estimation. Applied Energy, 358, 122602.

[22] Zamzam, A. S., & Sidiropoulos, N. D. (2020). Physics-aware neural networks for distribution system state estimation. IEEE Transactions on Power Systems, 35(6), 4347-4356.

[23] Liao, W., Bak-Jensen, B., Pillai, J. R., et al. (2022). A review of graph neural networks and their applications in power systems. Journal of Modern Power Systems and Clean Energy, 10(2), 345-360.

[24] Pan, X., Chen, M., Zhao, T., et al. (2021). DeepOPF: Deep neural networks for optimal power flow. In BuildSys '21: The 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (pp. 1-10).

[25] Biagioni, D., Zhang, X., Wald, D., et al. (2022). PowerGridworld: A framework for multi-agent reinforcement learning in power systems. In e-Energy'22: ACM International Conference on Future Energy Systems (pp. 565-570).