

Benchmarking Multi-Agent Coordination for Energy Planning

Altan Ulaş Zöhre
Izmir Institute of Technology,
Department of Computer Engineering
Izmir, Türkiye
auzohre@gmail.com

Deniz Sakaroglu
Izmir Institute of Technology,
Department of Computer Engineering
Izmir, Türkiye
denisakaroglu@gmail.com

Abstract

Multi-agent coordination has become a common design choice for complex planning tasks, yet systematic and reproducible evaluation of coordination mechanisms under feasibility targets remains limited in energy planning settings. This paper introduces an episode-based benchmark for country-level energy planning built on OWID-derived electricity indicators. Each episode trains a demand forecaster on historical years and evaluates planning decisions on a held-out period. Plans are represented by four normalized decision variables—renewable investment, storage investment, demand response, and carbon tax—and are assessed using a deterministic evaluation engine that reports cost, emission, and security outcomes. Feasibility is incorporated through soft constraints, where budget exceedance, emission exceedance, and security shortfall are aggregated into a continuous violation measure and applied as a smooth penalty to the overall score.

Six planning strategies are evaluated under identical episode construction and forecasting: a deterministic baseline, a single-shot LLM planner, a candidate-based multi-agent refinement approach, an LLM-coordinated multi-agent variant with a security-focused veto safeguard, and two ablations that remove veto or collapse iterative interaction. Across 15 country episodes, candidate-based multi-agent refinement achieves the highest mean score (0.825), while coordination-heavy variants incur substantially higher runtime and message counts, highlighting a clear quality-latency trade-off. The findings suggest that lightweight candidate selection can provide strong feasibility-aware performance without requiring expensive coordination mechanisms.

I. INTRODUCTION

Energy planning requires decisions that balance competing objectives such as economic cost, environmental impact, and supply reliability. Even when standardized country-level indicators are available, translating historical signals into actionable plans remains challenging because decisions are constraint-driven, multi-objective, and often benefit from iterative refinement rather than one-shot selection. This motivates evaluation settings that capture not only the final plan quality but also feasibility behavior and the practical overhead of producing a plan.

In parallel, the evaluation of large language models has increasingly shifted from static question answering toward agentic behaviors, including multi-step decision-making, tool use, and coordination. Multi-agent designs are frequently adopted to decompose objectives into specialist roles (e.g., cost, emissions, security) and then integrate proposals through a coordination mechanism. However, coordination itself introduces design choices—communication patterns, aggregation rules, and reliability safeguards—that can substantially affect outcomes. Without controlled benchmarks, it is difficult to determine whether improvements arise from the underlying model, the coordination protocol, or simply increased interaction budget.

A central challenge in this context is the trade-off between planning quality, feasibility, and computational cost. Coordination-heavy approaches may improve constraint control but can increase latency through additional LLM calls and longer inference chains, while lighter mechanisms may be more efficient but less robust. A benchmark that reports both outcome metrics (cost, emissions, security) and overhead metrics (messages, rounds, runtime) is therefore essential for interpreting multi-agent planning behavior.

This paper introduces an episode-based benchmark for country-level energy planning built on OWID-derived electricity indicators. Episodes are constructed with a fixed train-test split, a shared forecasting stage, and a deterministic evaluation engine that produces cost, emission, and security outcomes while incorporating feasibility through soft constraints. Multiple planning strategies are compared under identical episode construction and prediction components, including deterministic baselines, single-shot LLM planning, candidate-based multi-agent refinement, an LLM-coordinated multi-agent variant with a security-focused veto safeguard, and targeted ablations that isolate the effects of veto and iterative interaction. The resulting benchmark enables reproducible analysis of coordination mechanisms and supports systematic study of the quality-feasibility-latency trade-off in multi-agent energy planning.

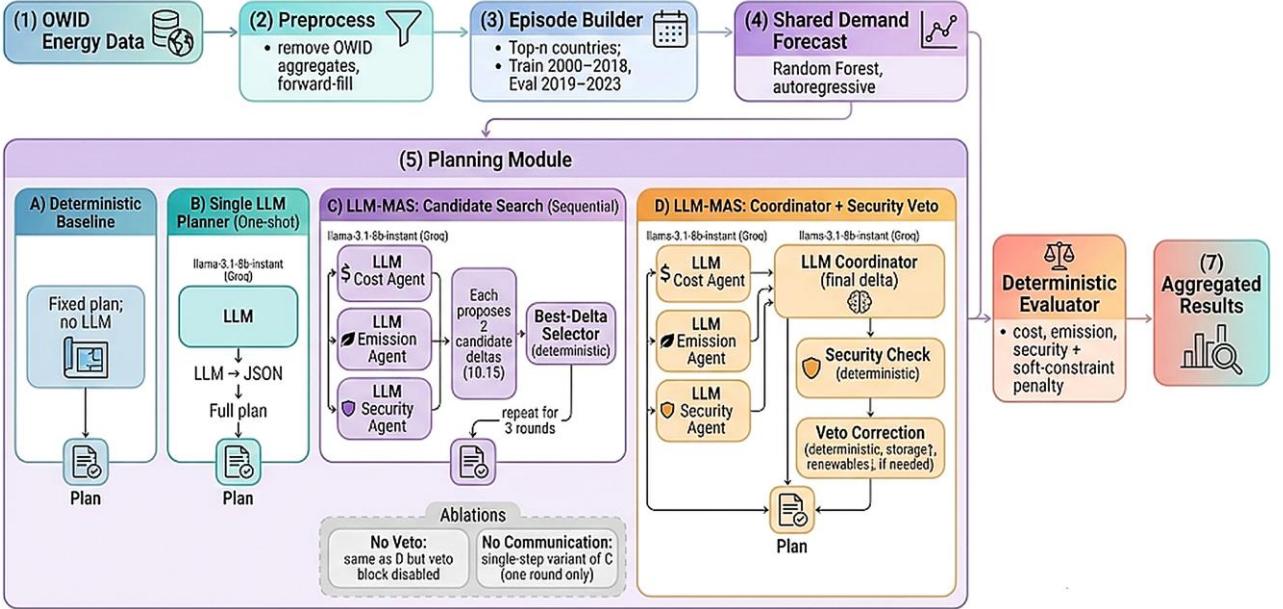


Figure 1: Method overview

II. RELATED WORK

Country-level energy indicators are frequently used to construct reproducible evaluation settings for energy planning and transition analysis. Public repositories such as OWID provide standardized signals—including electricity demand, carbon intensity, low-carbon and fossil generation shares, and net import indicators—supporting consistent preprocessing across countries and years and enabling comparable experimental protocols [1]. Such standardized datasets are particularly useful for benchmark construction because they reduce data acquisition ambiguity and allow controlled variations in episode definition and evaluation horizons.

In parallel, evaluation of large language models has expanded beyond static question answering toward agentic behaviors such as multi-step planning, iterative decision-making, tool use, and coordination under constraints. Agentic benchmarks highlight that many failures emerge only in interactive settings, motivating evaluation designs that explicitly account for iteration, feedback, and constraint satisfaction rather than relying solely on single-shot accuracy measures [2]. Survey work on LLM evaluation further emphasizes that reliable conclusions depend on clearly defined tasks and metrics that capture practical dimensions such as feasibility and efficiency in addition to output quality [3].

Recent surveys of LLM-based autonomous agents systematize architectural components including role specialization, planning loops, memory, and multi-agent collaboration patterns [4]. A consistent observation is that multi-agent performance depends not only on the base model but also on coordination choices—communication protocols, aggregation rules, and reliability mechanisms—which can substantially change outcomes even when the same model is used [4]. This motivates controlled ablations that isolate coordination components, such as removing explicit safeguards or reducing interaction depth, to attribute gains to specific design decisions.

Within the energy domain, agentic LLM frameworks have been explored for load scheduling and home energy management, often using an LLM as an orchestrator to propose actions under user objectives and operational constraints [5]. Related directions propose multi-agent systems that combine information extraction from unstructured building data with coordinated decision recommendations for energy optimization [6]. Additional work investigates LLM-assisted workflows for building energy model development and debugging, illustrating that decomposition into specialized roles can improve reliability and productivity for engineering tasks [7]. While these efforts demonstrate feasibility, systematic benchmarking of coordination mechanisms with explicit feasibility targets and overhead reporting remains limited, particularly in country-level settings derived from standardized indicators.

III. METHOD

The benchmark evaluates energy planning at the country level using annual electricity indicators derived from the OWID energy dataset. Preprocessing removes aggregate OWID entries and standardizes the year field to ensure temporal consistency. Countries are retained only if they provide sufficient coverage and enough non-missing electricity demand observations to support forecasting and evaluation, and the final episode set is formed by ranking eligible countries by demand data availability and selecting the top n (15 in the reported configuration). To preserve continuity in the time series, missing values in electricity demand and carbon intensity are handled conservatively through forward-filling; remaining missing carbon intensity values are replaced with a high default to avoid optimistic emissions estimates.

Each country episode is defined using a fixed train–evaluation split to ensure controlled comparison across methods. The training window spans 2000–2018 and supplies the historical data for demand forecasting, while the evaluation window spans 2019–2023 and is used to assess planning outcomes. Episode metadata includes baseline

context extracted from the last training year (e.g., low-carbon share, fossil share, net import share) together with feasibility targets that define a budget cap, a maximum emission threshold, and a minimum security requirement. These feasibility targets are held constant across all methods.

All planning approaches operate on an identical two-stage pipeline. First, evaluation-year demand is forecast using a shared RandomForestRegressor trained on simple lagged features, where the input combines the target year and the previous-year demand and the target is current-year demand. Forecasts are generated autoregressively over the evaluation horizon by feeding the latest predicted demand into the subsequent step. When an episode does not provide enough training data to fit a stable model, a fallback predictor returns the mean of the training demand. Forecast quality is reported via MAPE, and the same forecasting routine is applied across all methods to isolate differences attributable to planning behavior.

Second, candidate plans are assessed through a lightweight deterministic evaluation engine designed to capture the core trade-offs in energy planning while enabling reproducible comparisons. A plan is parameterized by four normalized decision variables—renewable investment, storage investment, demand response, and carbon tax—each clipped to [0,1]. Given the demand forecast, the engine computes cost, emissions, and security using consistent rules: demand response reduces effective demand; renewable investment and carbon tax reduce carbon intensity; storage improves security; and high renewable deployment without sufficient storage induces a grid penalty that increases cost and reduces reliability. Rather than rejecting infeasible solutions, feasibility is modeled through soft constraints by converting budget exceedance, emission exceedance, and security shortfall into normalized violation terms and aggregating them into a continuous violations measure. The final benchmark score rewards lower cost and emissions and higher security, while applying a smooth penalty that decreases the score as violations increase, thereby discouraging infeasible plans while preserving a graded comparison signal.

The benchmark is designed to compare single-system and multi-agent planning architectures under identical evaluation conditions. Single-system planners produce a complete plan in one step, either deterministically via a fixed policy or through a single LLM call that outputs a full plan in a strict JSON schema under explicit feasibility instructions. In contrast, multi-agent planners decompose the objective into role-specialized agents—cost, emissions, and security—where each agent is implemented as an LLM and produces bounded updates rather than an entire plan. Specifically, each agent proposes two candidate deltas that adjust one or more plan variables within $[-0.15, +0.15]$, with the additional requirement that at least one parameter change is non-zero. This bounded-update formulation supports conservative refinement and reduces instability from large plan jumps.

Two multi-agent coordination mechanisms are evaluated. In the sequential candidate-search approach, the system starts from a shared initial plan and executes a fixed number of rounds. Within each round, the cost, emission, and security agents are queried sequentially; after each agent call, the proposed deltas are applied individually, clipped to valid ranges, and scored by the deterministic evaluator, and the delta that yields the highest score is selected to update the

plan before the next agent step. This yields a conservative local-search procedure in which coordination is realized through deterministic selection grounded in the benchmark score. In the coordinator-based approach, the three agents first propose candidates and an additional coordinator—also implemented as an LLM—produces a single final delta conditioned on the current plan, current evaluation metrics, and the candidate pool. To improve feasibility under security targets, a deterministic veto-like safeguard is included: when security enters a risky region, coordination is bypassed and a corrective adjustment increases storage investment (and reduces renewables when necessary) to restore feasibility, followed by an additional security re-check that can trigger further correction if a hard threshold remains unmet. Two ablations isolate the contribution of these coordination components by disabling the veto while retaining coordination, or by collapsing iterative interaction into a single proposal step.

LLM usage is restricted to structured proposal generation to ensure stable execution and reproducibility. All LLM calls are served through Groq’s OpenAI-compatible API using the llama-3.1-8b-instant model with fixed decoding settings (temperature = 0.25 and a 350-token response limit) and a bounded retry policy (up to 2 retries). Outputs are required to conform to strict JSON schemas and are parsed using robust extraction; values are validated against the plan schema and clipped to allowed bounds. When parsing fails, conservative fallback updates are used to prevent runtime failures and to ensure that measured differences primarily reflect architectural and coordination choices rather than formatting artifacts or out-of-range outputs.

IV. EVALUATION AND RESULTS

Evaluation is conducted on a set of 15 country episodes constructed from the OWID energy dataset. Aggregate OWID entries are removed and annual series are preprocessed to ensure temporal continuity in key signals. Countries are eligible only if they provide sufficient historical coverage and enough non-missing electricity demand observations to support forecasting and evaluation. Eligible countries are ranked by demand data availability, and the top n are selected; in the reported configuration, $n=15$, yielding 15 episodes. Each episode uses a fixed temporal split with 2000–2018 as the training window and 2019–2023 as the evaluation window. Episode feasibility targets are held constant across methods through a budget cap, an emission threshold, and a minimum security requirement.

All methods share an identical forecasting module to isolate differences attributable to planning behavior. Demand is forecast over 2019–2023 using a RandomForestRegressor trained on lagged features (year and previous-year demand) with an autoregressive rollout. Forecasting error is reported using MAPE. Planning quality is assessed through a deterministic evaluation engine that reports cost, emission, and security outcomes for a proposed plan, together with a continuous violations measure derived from soft constraints (budget exceedance, emission exceedance, and security shortfall). The overall score reflects cost–emission–security preferences and applies a smooth penalty that increases with total violations, discouraging infeasible solutions while preserving a graded comparison signal.

Method	Score	Forecast MAPE	Cost	Emission	Security	Violations	Messages	Rounds	Vetos	Runtime (s)
MAS: Candidate Search (Sequential)	0.825129	4.377175	0.066667	0.325234	0.909500	0.000015	9.000	3.0	0.000	34.773
MAS: LLM Coordination + Security Veto	0.814956	4.377175	0.112994	0.307037	0.916833	0.000352	12.133	3.0	0.133	41.067
Ablation: No Communication (Single-step)	0.814954	4.377175	0.112333	0.307340	0.916833	0.000808	1.000	1.0	0.000	11.671
Ablation: No Veto (LLM Coordination)	0.812206	4.377175	0.125000	0.299403	0.912500	0.000793	12.000	3.0	0.000	44.907
Deterministic Baseline	0.811204	4.377175	0.136000	0.293990	0.916000	0.000000	0.000	1.0	0.000	0.337
Single-LLM Planner (One-shot)	0.792399	4.377175	0.192733	0.282110	0.922667	0.003400	1.000	1.0	0.000	2.085

Table 1: Benchmark results across methods

In addition to outcome metrics, practical overhead is explicitly reported. Interaction and computational costs are measured using the number of LLM messages, planning rounds, veto interventions (where applicable), and wall-clock runtime in seconds. Results are aggregated across episodes by reporting mean values per method, enabling direct comparison of performance and overhead.

Table 1 summarizes the mean benchmark results across 15 countries. The strongest mean score is achieved by the sequential candidate-search multi-agent approach, indicating that lightweight iterative refinement with objective candidate selection can provide robust performance under feasibility targets. Coordination-heavy variants that rely on an additional LLM coordinator exhibit substantially higher message counts and runtime, highlighting a clear quality-latency trade-off. The no-veto ablation shows that removing the reliability safeguard can increase feasibility risk, while the single-shot LLM planner demonstrates that one-step plan generation may yield strong values on individual components yet still underperform overall when constraint violations accumulate.

V. CONCLUSION

This paper presented a benchmark for energy planning that enables systematic evaluation of multi-agent coordination mechanisms under feasibility-aware scoring. Using OWID-derived country-level electricity indicators, the benchmark operationalizes planning as an episode-based task with a shared forecasting stage and a deterministic evaluation engine that reports cost, emissions, and security outcomes while incorporating feasibility through soft constraints. This design supports reproducible comparison of planning strategies while explicitly accounting for both outcome quality and practical overhead.

Across 15 country episodes, results indicate that lightweight multi-agent refinement via candidate search achieves the strongest mean score among the evaluated methods. In contrast, coordination-heavy variants that introduce an additional LLM coordinator incur substantially higher message counts and runtime, illustrating a clear quality-latency trade-off. Reliability mechanisms such as the security-focused veto contribute to controlling feasibility risk in coordinated settings, while ablations suggest that removing safeguards or collapsing interaction depth can alter the balance between score improvements and computational

cost. Overall, the findings emphasize that stronger coordination is not uniformly beneficial; the value of coordination depends on whether its feasibility gains justify added interaction and latency.

Several limitations should be noted. The evaluation engine is intentionally lightweight and abstracts many real-world power-system dynamics; consequently, metric values should be interpreted primarily as benchmark signals for method comparison rather than as operational prescriptions. In addition, LLM-based methods are sensitive to prompting and decoding configurations, and measured runtimes depend on external API conditions and resource availability.

Future work includes expanding the benchmark to broader episode sets (more countries and richer indicators), adopting more realistic cost–emission–security models calibrated to domain parameters, and testing additional coordination protocols and model families. Practical scalability is also an important direction: GPU and resource availability can constrain coordination-heavy multi-agent designs, motivating research on more efficient coordination schemes, caching, and hybrid optimization baselines that improve feasibility-aware performance without incurring prohibitive latency.

REFERENCES

- [1] Our World in Data, “Energy,” Our World in Data. [Online]. Available: <https://ourworldindata.org/energy>
- [2] X. Liu et al., “AgentBench: Evaluating LLMs as Agents,” arXiv preprint arXiv:2308.03688, 2023, doi: 10.48550/arXiv.2308.03688.
- [3] Y. Chang et al., “A Survey on Evaluation of Large Language Models,” arXiv preprint arXiv:2307.03109, 2023, doi: 10.48550/arXiv.2307.03109.
- [4] L. Wang et al., “A survey on large language model based autonomous agents,” Frontiers of Computer Science, vol. 18, no. 6, art. no. 186345, 2024, doi: 10.1007/s11704-024-40231-1.
- [5] R. El Makroum, S. Zwickl-Bernhard, and L. Kranzl, “Agentic AI Home Energy Management System: A Large Language Model Framework for Residential Load Scheduling,” arXiv preprint arXiv:2510.26603, 2025.
- [6] T. Xiao and P. Xu, “Exploring automated energy optimization with unstructured building data: A multi-agent based framework leveraging large language models,” Energy and Buildings, vol. 322, art. no. 114691, 2024, doi: 10.1016/j.enbuild.2024.114691.
- [7] L. Zhang, V. Ford, Z. Chen, and J. Chen, “Automatic building energy model development and debugging using large language models agentic workflow,” Energy and Buildings, vol. 327, art. no. 115116, 2025, doi: 10.1016/j.enbuild.2024.115116.