FEMS, (2026)

Frontiers in Emerging Multidisciplinary Sciences (Open Access)

Volume 03, Issue 08, August 2026
Publish Date: 19-08-2026
DOI: https://doi.org/10.64917/fems/Volume03Issue08-05 PageNo.111-117

ScaleGen: A Combinatorial Generative LLM Framework for Scalability-Constrained Decision Optimization

Muhammad Hamza Khan
Department of Artificial Intelligence Institute of Computing and Emerging Technologies, Islamabad, Pakistan Sri Lanka

Ayesha Malik
Department of Computer Science and Machine Learning National University of Digital Sciences, Lahore, Pakistan

ABSTRACT

Scalability-constrained decision optimization requires decision systems to generate useful alternatives while simultaneously respecting computational, operational, and resource limitations. Conventional optimization approaches often depend on predefined candidate spaces, whereas generative large language models (LLMs) can construct semantically diverse decision alternatives but may produce infeasible or computationally expensive solutions. This paper proposes ScaleGen, a combinatorial generative LLM framework designed to integrate generative reasoning with explicit scalability constraints. The framework combines candidate generation, constraint representation, combinatorial evaluation, feasibility filtering, and final decision selection into a structured optimization pipeline. Its theoretical foundation is informed by research on machine learning, quantum machine learning, classification, and computational decision processes. Studies on quantum methods demonstrate the value of structured computational representations for complex learning problems (Ablayev et al., 2020a; Ablayev et al., 2020b), while quantum machine-learning applications illustrate the potential of advanced computational paradigms for classification and large decision spaces (Bhavsar et al., 2023; Yun et al., 2022). The proposed ScaleGen architecture extends these principles into an LLM-centered optimization setting by treating generated decisions as combinatorial objects subject to explicit scalability constraints. The framework emphasizes controllable generation rather than unrestricted text generation, thereby providing a conceptual bridge between generative AI and constrained optimization. The analysis indicates that ScaleGen can potentially improve candidate diversity, constraint awareness, and decision adaptability, although its effectiveness remains dependent on constraint encoding, evaluation reliability, computational overhead, and empirical validation.

KEYWORDS: Generative Large Language Models; Scalability-Constrained Optimization; Combinatorial Optimization; Decision Optimization; Constraint-Aware Generation; Quantum Machine Learning; Generative AI; Intelligent Decision Systems

INTRODUCTION

1.1 Background

Human Modern decision systems increasingly operate in environments where the number of possible alternatives grows rapidly with the number of decision variables. In such environments, generating a decision is not sufficient; the decision must also remain computationally manageable, operationally feasible, and consistent with predefined scalability requirements. This creates a fundamental tension between solution diversity and solution tractability. A system capable of generating many alternatives may increase the probability of identifying a high-quality decision, but unrestricted generation can simultaneously increase evaluation cost and produce infeasible candidates.

Large language models offer a potentially useful mechanism for addressing this problem because they can transform natural-language requirements into structured candidate decisions. However, conventional LLM generation is primarily probabilistic and semantic rather than explicitly optimization-oriented. Consequently, an LLM may generate alternatives that appear reasonable while violating resource, latency, capacity, or complexity constraints. A scalable decision framework therefore requires a mechanism that places generation inside an explicit combinatorial and constraint-aware optimization process.

The concept of ScaleGen is motivated by the broader development of machine-learning methods capable of processing increasingly complex decision spaces. Research on quantum machine-learning methods emphasizes computational tools and structured approaches for machine-learning problems (Ablayev et al., 2020a). The corresponding work on quantum classification algorithms further demonstrates how specialized computational representations can be employed for classification problems (Ablayev et al., 2020b). These perspectives are relevant to ScaleGen because scalability requires the decision space to be represented and processed systematically rather than treated as unstructured output.

Recent research also illustrates the application of machine learning to complex astronomical problems. Machine-learning approaches have been used for asteroid detection (Chaitanya Prasad et al., 2020) and asteroid-group identification (Carruba et al., 2019). Such applications demonstrate that machine learning can support classification and pattern-discovery tasks in domains characterized by large and heterogeneous search spaces. Similarly, supervised quantum machine learning has been examined for classification of potentially hazardous asteroids (Bhavsar et al., 2023), illustrating the relationship between advanced computational methods and high-dimensional decision or classification environments.

1.2 Problem Statement

The central problem addressed by ScaleGen is the absence of a unified mechanism through which an LLM can generate diverse decision candidates while maintaining explicit scalability constraints throughout the generation and selection process. A purely generative architecture may produce redundant, infeasible, or excessively complex alternatives. Conversely, a conventional constrained optimizer may satisfy formal constraints but lack the semantic flexibility required to construct novel alternatives from ambiguous or natural-language requirements.

The challenge is therefore to construct a hybrid mechanism in which generation and optimization are not independent stages but interconnected operations. The proposed framework treats scalability as a first-class property of the generation process.

1.3 Research Objectives

The primary objective is to formulate a conceptual architecture for combinatorial generative decision optimization using an LLM. The study specifically seeks to develop a constraint-aware candidate-generation mechanism, define a scalable combinatorial evaluation process, establish a feasibility-filtering layer, and formulate a decision-selection mechanism capable of balancing utility, diversity, feasibility, and computational cost.

1.4 Scope and Significance

ScaleGen is positioned as a general framework rather than a domain-specific optimizer. Its architecture can theoretically support resource allocation, infrastructure configuration, service selection, scheduling, portfolio construction, and other problems in which decisions consist of multiple interacting components. The significance of the framework lies in connecting the semantic generation capabilities of LLMs with explicit optimization requirements.

The conceptual direction is particularly aligned with the scalability-oriented principles described by Ramamurthy et al. (2026), whose ScalePulse framework addresses combinatorial LLM processing under scalability constraints. The present ScaleGen framework extends this conceptual direction toward decision optimization by treating generated alternatives as candidate solutions that must be evaluated against explicit feasibility and scalability criteria (Ramamurthy et al., 2026).

LITERATURE REVIEW

The supplied literature provides several complementary foundations for ScaleGen. Ablayev et al. (2020a) examine quantum tools for machine-learning problems, emphasizing computational mechanisms for representing and processing learning tasks. Their work is relevant to ScaleGen because scalability-constrained decision optimization similarly requires a structured representation of a potentially large computational search space. Rather than generating alternatives independently, ScaleGen organizes candidate decisions into a structured combinatorial representation.

Ablayev et al. (2020b) extend this perspective to quantum classification algorithms. Classification is fundamentally a selection problem: candidate observations must be mapped to meaningful classes according to computationally defined criteria. ScaleGen adopts a related principle but changes the objective from class assignment to decision selection. Generated candidates are evaluated against feasibility and utility criteria, producing an optimization-oriented selection mechanism rather than a conventional classifier.

The systematic literature review by García et al. (2022) provides broader evidence of the development of quantum machine learning and its applications. The importance of this work for ScaleGen is conceptual rather than technological. It demonstrates that emerging computational paradigms are increasingly evaluated according to how their computational structures can address machine-learning problems. ScaleGen similarly proposes an architectural structure in which LLM generation is embedded within a computational decision pipeline rather than used as an unrestricted conversational mechanism.

Bhat et al. (2022) provide a broader discussion of quantum-computing fundamentals, implementations, and applications. Their work reinforces the importance of computational architecture when addressing advanced learning and optimization tasks. For ScaleGen, this supports the argument that model capability alone is insufficient; the surrounding computational architecture determines whether generated outputs can be transformed into usable decisions.

The work of Bhavsar et al. (2023) demonstrates supervised quantum machine learning for potentially hazardous asteroid classification. This application is important because it illustrates the use of advanced computational learning methods for classification problems involving complex astronomical objects. Chaitanya Prasad et al. (2020) similarly investigate machine learning for asteroid detection, while Carruba et al. (2019) examine machine-learning identification of asteroid groups. Collectively, these studies demonstrate that machine learning can identify patterns and categories within complex search spaces, but they do not directly address LLM-based combinatorial decision generation.

Aftosmis et al. (2019) focus on simulation-based height-of-burst mapping for asteroid airburst damage prediction. Although the application differs from ScaleGen, it demonstrates the importance of computational modeling when decision consequences depend on multiple interacting variables. Such simulation-oriented reasoning is relevant to ScaleGen's candidate-evaluation layer, where generated alternatives may require structured assessment before selection.

Yun et al. (2022) examine projection-valued-measure-based quantum machine learning for multiclass classification. Their work highlights the importance of formal representations when multiple possible outcomes must be distinguished. ScaleGen applies a related principle to combinatorial decision alternatives by requiring generated candidates to pass structured evaluation rather than relying exclusively on linguistic plausibility.

IBM's Quantum Decade report provides an industry-oriented perspective on the anticipated development and applications of quantum computing (IBM, 2024). For ScaleGen, the relevance lies in the broader recognition that emerging computational architectures must be evaluated according to their ability to address practical computational constraints.

The principal gap across the supplied literature is therefore clear. Existing studies address machine learning, classification, quantum computation, astronomical detection, and computational simulation, but they do not provide a complete LLM-centered architecture for generating and selecting combinatorial decisions under explicit scalability constraints. ScalePulse provides a closely related conceptual foundation for scalability-constrained LLM processing (Ramamurthy et al., 2026), but ScaleGen focuses specifically on converting generative capacity into optimized decision alternatives. This distinction establishes the theoretical position of the proposed framework.

METHODOLOGY

3.1 Research Design

This study adopts a framework-design methodology. Rather than claiming experimental performance without an empirical dataset or implementation, the research develops a technically specified architecture for scalability-constrained generative decision optimization. The methodology defines the decision representation, generation mechanism, constraint model, candidate evaluation procedure, selection strategy, and scalability controls.

The framework assumes an optimization problem represented as:

$$
D^*=\arg\max_{D\in\mathcal{D}} U(D)
$$

subject to:

$$
C_i(D)\leq \tau_i,\quad i=1,\ldots,m
$$

where D represents a candidate decision, $\mathcal{D}$ represents the decision space, $U(D)$ represents decision utility, $C_i(D)$ represents the i-th scalability or operational constraint, and τ_i represents its permitted threshold.

ScaleGen introduces an additional generation function:

$$
D_j=G_{\theta}(X,S,P)
$$

where G_{θ} is the LLM generator, X is the problem context, S represents scalability constraints, and P represents generation policies. The output is not immediately accepted as a decision. Instead, it enters a structured evaluation pipeline.

3.2 ScaleGen Architecture

The proposed architecture contains five principal layers: requirement interpretation, combinatorial candidate generation, constraint validation, candidate scoring, and decision selection.

The requirement interpretation layer converts natural-language requirements into structured variables. For example, a cloud-resource planning request could specify computational capacity, latency requirements, budget limits, and acceptable architectural components. The LLM converts these requirements into a machine-readable decision schema.

The combinatorial generation layer creates candidate decisions by recombining permissible components. Instead of generating one unrestricted response, the system generates a controlled candidate set:

$$
\mathcal{C}=\{D_1,D_2,\ldots,D_k\}
$$

The value of combinatorial generation is that alternatives can differ in component selection, configuration, sequencing, or resource allocation. This provides greater decision-space coverage than a single generated answer.

The constraint-validation layer evaluates each candidate against explicit restrictions. A candidate may be linguistically coherent but computationally invalid. Therefore, ScaleGen assigns each candidate a feasibility indicator:

$$
F(D)=
\begin{cases}
1,& C_i(D)\leq\tau_i\ \forall i\\
0,& \text{otherwise}
\end{cases}
$$

Only candidates satisfying required hard constraints progress to final evaluation.

3.3 Multi-Criteria Candidate Scoring

ScaleGen evaluates feasible candidates using a composite score:

$$
S(D)=w_uU(D)+w_dD_v(D)-w_cC(D)-w_rR(D)
$$

where $U(D)$ represents utility, $D_v(D)$ represents candidate diversity, $C(D)$ represents computational cost, and $R(D)$ represents estimated risk or constraint exposure. The weights (w_u,w_d,w_c,w_r) allow the system to adapt to application-specific priorities.

This formulation prevents the model from selecting an alternative solely because it appears semantically strong. A highly useful but computationally excessive decision can be rejected or penalized. Conversely, an inexpensive but low-value decision should not automatically dominate the candidate pool.

3.4 Scalability Control

Scalability control is central to the framework. If k candidates are generated independently across n decision dimensions, the theoretical search space can expand exponentially. ScaleGen therefore applies progressive filtering. Initial generation may be broad, while subsequent stages reduce the candidate set according to feasibility and utility.

A simplified computational objective can be represented as:

[

\min \left(\alpha N_g+\beta N_e+\gamma L\right)

]

where (N_g) is the number of generated candidates, (N_e) is the number of evaluated candidates, (L) is processing latency, and (\alpha,\beta,\gamma) represent application-specific cost coefficients.

This architecture follows the scalability-oriented reasoning associated with ScalePulse, in which combinatorial LLM processing is explicitly considered under scalability constraints (Ramamurthy et al., 2026). ScaleGen differs by applying the constraint mechanism directly to decision alternatives and final selection.

3.5 Hypothetical Operational Example

Consider an organization selecting an infrastructure configuration subject to budget, latency, and capacity requirements. An unrestricted LLM could generate several plausible architectures, including configurations that exceed budget or require excessive computational resources. ScaleGen instead converts the requirements into constraints, generates multiple architectural combinations, removes infeasible alternatives, evaluates the remaining candidates, and selects the highest-scoring configuration.

The process can be summarized as:

[

\text{Requirements}\rightarrow

\text{Structured Constraints}\rightarrow

\text{Candidate Generation}\rightarrow

\text{Feasibility Filtering}\rightarrow

\text{Scoring}\rightarrow

\text{Optimal Decision}

]

This architecture makes the LLM a candidate-generation and reasoning component rather than the sole decision authority.

3.6 Theoretical Positioning

The theoretical contribution of ScaleGen lies in integrating three mechanisms: generative semantic reasoning, combinatorial search, and explicit constraint optimization. Quantum machine-learning studies demonstrate the importance of structured computational methods for complex learning problems (Ablayev et al., 2020a; Ablayev et al., 2020b). ScaleGen transfers this general architectural principle to LLM-based decision optimization without claiming that LLMs themselves provide quantum computational capabilities.

The framework also recognizes that classification-oriented machine-learning research, such as asteroid detection and identification, provides useful evidence for structured candidate discrimination (Carruba et al., 2019; Chaitanya Prasad et al., 2020). ScaleGen extends discrimination from object classes to feasible decision alternatives.

RESULTS

The framework analysis produces four principal findings. First, ScaleGen establishes that generative LLM capability can be separated from final decision authority. This distinction is technically important because an LLM can generate semantically meaningful alternatives without guaranteeing feasibility. By introducing a constraint-validation layer, the framework transforms generation from an open-ended language task into a controlled candidate-generation process. This provides a stronger theoretical basis for decision optimization than direct prompting alone.

Second, the architecture indicates that combinatorial generation can improve the breadth of the candidate space while progressive filtering controls computational growth. The key mechanism is not unlimited generation but staged reduction. Candidates are initially diversified, after which hard constraints eliminate infeasible alternatives and multi-criteria scoring ranks the survivors. Consequently, the framework explicitly balances exploration and computational tractability. This principle is consistent with the broader scalability orientation of ScalePulse (Ramamurthy et al., 2026).

Third, the literature synthesis indicates that structured computational representations are particularly relevant to complex decision spaces. Quantum machine-learning research emphasizes computational tools and classification structures (Ablayev et al., 2020a; Ablayev et al., 2020b), while applications to asteroid detection and classification demonstrate how machine learning can identify meaningful structures in complex data environments (Chaitanya Prasad et al., 2020; Bhavsar et al., 2023). ScaleGen generalizes this structural perspective into an LLM-driven decision pipeline.

Fourth, the proposed architecture identifies scalability as a multidimensional property rather than a single computational threshold. Candidate count, evaluation complexity, processing latency, resource consumption, and constraint violations can all affect scalability. Therefore, a decision system that merely reduces the number of generated responses may remain computationally inefficient if candidate evaluation is expensive. ScaleGen addresses this by incorporating generation, evaluation, and selection costs into the conceptual optimization process.

However, these findings are architectural rather than experimentally validated performance results. No benchmark dataset, implementation, latency measurement, accuracy evaluation, or comparative experiment was supplied. Accordingly, claims of superiority over conventional optimization or alternative LLM architectures cannot be established from the available evidence. The principal result is the development of a coherent framework capable of being empirically tested.

DISCUSSION

ScaleGen's principal theoretical implication is that generative AI and constrained optimization should not necessarily be treated as competing paradigms. Generative models provide flexible exploration of a large semantic decision space, whereas optimization mechanisms provide formal control over feasibility and objective quality. Their integration therefore creates a layered decision architecture in which generation supports exploration and optimization governs selection.

The framework also clarifies a limitation of treating LLM outputs as final decisions. Natural-language plausibility does not imply mathematical feasibility. A candidate may satisfy an apparent business requirement while violating a quantitative constraint. ScaleGen addresses this weakness by separating candidate construction from candidate validation. This separation is especially important in environments where computational cost, capacity, latency, or risk cannot be inferred reliably from language alone.

The relationship with existing machine-learning research further strengthens this interpretation. Studies of asteroid detection and group identification demonstrate the usefulness of machine learning for identifying structured relationships within complex problem spaces (Carruba et al., 2019; Chaitanya Prasad et al., 2020). Research involving potentially hazardous asteroid classification similarly illustrates the role of advanced computational methods in complex classification environments (Bhavsar et al., 2023). Nevertheless, these systems generally operate on predefined data and classification objectives rather than generating new combinatorial decisions. ScaleGen therefore represents a conceptual transition from recognition-oriented machine learning toward generation-oriented decision optimization.

The quantum-machine-learning literature introduces another relevant comparison. Ablayev et al. (2020a) and Ablayev et al. (2020b) emphasize formal computational approaches for machine-learning problems, while Yun et al. (2022) investigate structured quantum learning for multiclass classification. ScaleGen does not replicate these computational methods; instead, it adopts the broader architectural lesson that complex learning problems benefit from explicit representations and controlled computational operations. Bhat et al. (2022) similarly reinforces the importance of computational architecture when evaluating emerging technologies.

The major practical trade-off concerns candidate diversity versus computational cost. Increasing the number of generated candidates may improve the probability of finding a strong solution but increases evaluation requirements. Excessive filtering, conversely, can remove novel but valuable alternatives. Consequently, ScaleGen requires adaptive control over generation breadth, constraint strictness, and evaluation depth.

Another limitation is LLM reliability. The framework assumes that generated candidates can be represented accurately and evaluated by external constraint mechanisms. If the LLM misinterprets a requirement, the resulting candidate set may be systematically biased. Future empirical implementations should therefore test robustness under ambiguous requirements, adversarial inputs, changing constraints, and high-dimensional decision spaces.

Finally, the framework should be interpreted as a research architecture rather than evidence of empirically demonstrated optimization superiority. The ScalePulse research provides an important conceptual basis for scalability-constrained LLM processing (Ramamurthy et al., 2026), but ScaleGen requires independent benchmarking to establish improvements in solution quality, computational efficiency, diversity, and constraint satisfaction. This distinction is essential for maintaining methodological validity.

CONCLUSION

This paper proposed ScaleGen, a combinatorial generative LLM framework for scalability-constrained decision optimization. The framework integrates natural-language requirement interpretation, controlled candidate generation, formal constraint validation, multi-criteria evaluation, and final decision selection. Its central contribution is the separation of generative exploration from optimization-based decision authority.

The literature indicates that structured computational representations, machine-learning classification, and emerging computational paradigms provide relevant theoretical foundations for complex search and decision problems. ScaleGen extends these principles into an LLM-oriented architecture while maintaining explicit scalability controls.

The framework's principal limitation is the absence of empirical implementation and benchmark validation. Future research should implement ScaleGen across representative combinatorial optimization tasks and compare it with conventional optimization, direct LLM prompting, and hybrid decision systems. Evaluation should include solution quality, constraint satisfaction, candidate diversity, latency, token consumption, computational cost, and robustness. Such experiments would determine whether the proposed architecture can translate its theoretical advantages into measurable improvements in scalable intelligent decision-making.

REFERENCES

1. Ablayev F, Ablayev M, Huang JZ, et al. On quantum methods for machine learning problems part I: Quantum tools. Big Data Mining and Analytics. 2020; 3(1): 41-55. doi: 10.26599/bdma.2019.9020016
2. Ablayev F, Ablayev M, Huang JZ, et al. On quantum methods for machine learning problems part II: Quantum classification algorithms. Big Data Mining and Analytics. 2020; 3(1): 56-67. doi: 10.26599/bdma.2019.9020018
3. Aftosmis MJ, Mathias DL, Tarano AM. Simulation-based height of burst map for asteroid airburst damage prediction. Acta Astronautica. 2019; 156: 278-283. doi: 10.1016/j.actaastro.2017.12.021
4. Bhavsar R, Jadav NK, Bodkhe U, et al. Classification of Potentially Hazardous Asteroids Using Supervised Quantum Machine Learning. IEEE Access. 2023; 11: 75829-75848. doi: 10.1109/access.2023.3297498
5. Bhat HA, Khanday FA, Kaushik BK, et al. Quantum Computing: Fundamentals, Implementations and Applications. IEEE Open Journal of Nanotechnology. 2022; 3: 61-77, 2022, doi: 10.1109/OJNANO.2022.3178545
6. Carruba V, Aljbaae S, Lucchini A. Machine-learning identification of asteroid groups. Monthly Notices of the Royal Astronomical Society. 2019; 488(1): 1377-1386. doi: 10.1093/mnras/stz1795
7. Chaitanya Prasad LVR, Reddy TAS, Kashi B. Asteroid Detection using Machine Learning Algorithm. Communications of BAO. 2020; 67(2).
8. García DP, Cruz-Benito J, García-Peñalvo FJ. Systematic Literature Review: Quantum Machine Learning and its Applications. Available online: https://arxiv.org/abs/2201.04093 (accessed on 2 February 2024).
9. IBM. Quantum Decade.Available online: https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/quantum-decade (accessed on 15 February 2024).
10. Yun WJ, Baek H, Kim J. Projection Valued Measure-based Quantum Machine Learning for Multi-Class Classification.ADS Abstract Service. 2022; 2. doi:10.48550/arXiv.2210.16731
11. K. Ramamurthy, N. Bellamkonda and N. Amanmadov, "ScalePulse: Combinatorial Llm Framework for Scalability Constraint," SoutheastCon 2026, Huntsville, AL, USA, 2026, pp. 1-6, doi: 10.1109/SoutheastCon63549.2026.11476075
12. Geo Philip, Paulson, Robotics-Enabled Sustainable Construction Management: A Socio-Technical Framework for Operational Efficiency, Digital Integration, and Sustainability Performance. Available at SSRN: https://ssrn.com/abstract=6845167 or http://dx.doi.org/10.2139/ssrn.6845167