REDUNDANCY MEETS SYNERGY: DEPENDENCY-AWARE EXPERT SELECTION FOR MOE VIA SUBMODULAR OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While Mixture-of-Experts (MoE) models effectively scale model capacity through sparse activation, their deployment is often bottlenecked by prohibitive memory requirements. Extracting a compact subset of experts presents a promising solution. However, existing expert selection heuristics predominantly rely on Top-k ranking, which isolates the evaluation of individual experts and ignores the intricate inter-expert dependencies introduced by the MoE gating network. In this paper, we propose DS-MoE, a theoretically grounded framework that redefines expert selection via difference-of-submodular (DS) optimization. By analyzing the second-order Taylor expansion of the loss degradation, we reveal functional duality within expert combinations: redundancy (where experts encode overlapping representations) and synergy (where experts provide complementary error cancellation). To navigate this duality, we mathematically decouple redundancy reduction from synergy maximization by formulating the selection objective as a DS function. Furthermore, we devise a tailored majorization-minimization (MM) algorithm with provable guarantees to efficiently identify the optimal expert subset. Extensive experiments demonstrate that DS-MoE effectively preserves indispensable expert combinations, achieving superior performance compared to the state-of-the-art baselines.

1 INTRODUCTION

The pursuit of scaling laws in large language models (LLMs) Kaplan et al. (2020); Shi et al. (2024); Hoffmann et al. (2022); Clark et al. (2022) has driven a fundamental paradigm shift from dense neural networks to Mixture-of-Experts (MoE) architectures Shazeer et al. (2017); Liu et al. (2024b); Fedus et al. (2022). In contrast to traditional training paradigms that activate all parameters for every input, the MoE architecture is built on the principle of conditional computation. It relies on a gating mechanism to dynamically route each input token to a sparse subset of experts (e.g., Top-k) Fedus et al. (2022). This selective execution effectively decouples model capacity from computational cost, allowing MoE architectures to scale to the trillion-parameter regime Lepikhin et al. (2021) and unlock massive knowledge capacity without a proportional increase in computing overhead Li et al. (2025); Liu et al. (2024a).

Despite its significance, the theoretical understanding of expert contributions to training performance remains in its infancy. Existing expert selection Zhou et al. (2022); Puigcerver et al. (2024); Tan et al. (2024) heuristics evaluate experts in strict isolation, assigning routing probabilities independently and thus ignoring intricate inter-expert dependencies Chi et al. (2022), as detailed in Appendix B. Theoretically, this strict independence assumption equates to approximating the loss landscape's Hessian matrix as purely diagonal. Consequently, prior works fail to distinguish between functional redundancy (where experts encode overlapping representations) and synergy (where specific expert combinations provide complementary error cancellation). This oversight inevitably leads to suboptimal routing: redundant experts monopolize capacity, while combinatorially indispensable but individually negligible synergistic components are discarded. Therefore, expert selection must transcend heuristic scalar sorting and be formulated as a rigorous optimization problem. However, transitioning from isolated scoring to collective evaluation inherently introduces an intractable combinatorial optimization problem. Constrained by the immense discrete space, prior works

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

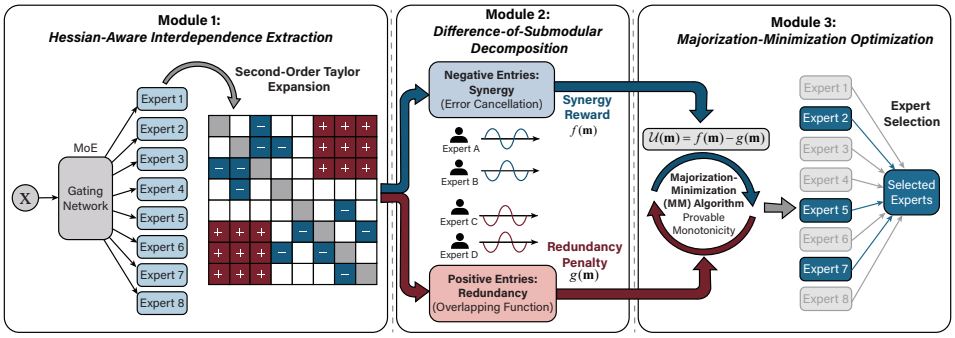


Figure 1: Overview of the DS-MoE framework.

inevitably default to simplistic and retreat to independent scoring heuristics Zhou et al. (2022); Puigcerver et al. (2024); Tan et al. (2024). Thus, there is a critical need for a framework that can systematically optimize inter-expert synergy and mitigate redundancy without succumbing to combinatorial explosion.

To bridge this gap, we propose DS-MoE, a theoretically grounded MoE framework that redefines expert selection via difference-of-submodular (DS) optimization Iyer & Bilmes (2012). As depicted in Figure 1, DS-MoE anchors its selection criterion in the second-order Taylor expansion of the loss degradation, effectively bridging the gap between continuous loss curvature and discrete expert selection. This formulation reveals that the MoE Hessian serves as a precise record of expert interdependence, manifesting fundamental duality Hassibi & Stork (1992): positive off-diagonal entries correspond to redundancy (penalizing overlapping functionality), while negative entries capture synergy (rewarding complementary error cancellation). To harness this duality, we decouple redundancy reduction from synergy maximization by formulating the selection objective as a DS optimization problem. Subsequently, we devise a tailored majorization-minimization (MM) algorithm Hunter & Lange (2004); Mairal (2013) with provable guarantees to efficiently identify the optimal expert combination.

Our contributions are summarized as follows: i) We theoretically prove that the MoE Hessian is a rigorous metric for characterizing inter-expert interdependence, where its off-diagonal entries precisely quantify the inter-expert functional redundancy and synergy; ii) We cast the key parameter optimization of expert selection as a DS problem, thereby decomposing the collective impact of expert combinations into a controllable trade-off between penalizing redundancy and rewarding synergy; iii) We develop a tailored MM algorithm with guaranteed monotonic convergence to efficiently identify the optimal expert subset; iv) Extensive experiments demonstrate that DS-MoE consistently outperforms state-of-the-art baselines, and strikes a superior accuracy-efficiency trade-off.

2 Methodology

2.1 Preliminaries & Problem Formulation

Standard MoE mechanism. We begin by formalizing the MoE architecture Jacobs et al. (1991). An MoE layer comprises a set of N expert networks, denoted as $E = \{E_1, \ldots, E_N\}$. For an input token $x \in \mathbb{R}^d$, a gating network $g(\cdot)$ maps x to a sparse routing vector $G(x) \in \mathbb{R}^N$ via a learnable projection matrix $W_g \in \mathbb{R}^{N \times d}$, selecting the subset of Top-k experts with the highest response values. The gating weights are derived by normalizing the projected logits over the selected subset:

$$G(x) = \mathrm{Softmax}(W_g x; \mathcal{T}(x)), \tag{1}$$

where $\mathcal{T}(x) = \left\{ i \;\middle|\; \sum_{j=1}^{N} \mathbb{I}(G_j(x) > G_i(x)) < k \right\}$ represents the set of selected expert indices, $\mathbb{I}(\cdot)$ denotes the indicator function, and $\mathrm{Softmax}(\cdot; \mathcal{T}(x))$ applies exponential normalization strictly to indices $i \in \mathcal{T}(x)$ while zeroing out others. The layer output y is computed as the weighted sum of the active experts:

$$y = \sum_{i \in \mathcal{T}(x)} G_i(x) \cdot E_i(x), \tag{2}$$

where $G_i(x)$ is the i-th element of $G(x)$ that represents the routing score (i.e., routing probability) assigned to expert E_i, and $E_i(x)$ denotes the output representation generated by the expert network.

Limitations of Top-k mechanism. While the Top-k mechanism effectively filters experts based on gating logits, it relies on a limiting independence assumption: the utility of an expert is evaluated solely via its isolated routing score $G_i(x)$. This greedy heuristic introduces a critical blind spot regarding inter-expert dependencies Chi et al. (2022). Specifically, it fails to distinguish between functional redundancy and synergy. Consequently, the standard routing protocol risks discarding experts that appear individually negligible but are combinatorially indispensable. By reducing expert selection to a series of independent scalar decisions rather than a combinatorial optimization Puigcerver et al. (2024), current routing protocols fail to exploit the complementary interactions required to drive loss reduction beyond the capacity of isolated experts.

Formalizing Dynamic Expert Selection. To address these limitations and transition from heuristic-based routing to a rigorous optimization framework, we introduce a binary selection vector $\mathbf{m} \in \{0,1\}^N$ to explicitly control expert activation during training. Here, m_i denotes the binary state of the i-th expert, where $m_i = 1$ indicates that the expert is active for model training, and $m_i = 0$ implies exclusion. Consequently, the output of the MoE layer is reformulated as:

$$y(x;\mathbf{m}) = \sum_{i=1}^{N} m_i \cdot G_i(x) \cdot E_i(x). \tag{3}$$

We frame the dynamic expert selection as a minimization problem. We posit that a fully activated model represents the theoretical upper bound of the network's representational capacity. Our objective is to identify the optimal sparse selection vector $\mathbf{m}$ that minimizes the expected loss degradation relative to the fully activated baseline, subject to a strict cardinality constraint dictated by storage limitations (i.e., the model can accommodate a maximum of k experts, $\|\mathbf{m}\| = k$):

$$\min_{\mathbf{m}\in\{0,1\}^N} \quad \underbrace{\mathbb{E}_{x\sim\mathcal{D}}\left[\mathcal{L}(y(x;\mathbf{m})) - \mathcal{L}(y(x;\mathbf{1}))\right]}_{\text{Loss Degradation } \Delta\mathcal{L}} \quad \text{s.t.} \quad \|\mathbf{m}\| = k. \tag{4}$$

Optimization Challenges. Directly solving Eqn. equation 4 presents formidable obstacles. The discrete cardinality constraint $\|\mathbf{m}\| \le k$ induces a combinatorial search space of size $\binom{N}{k}$, rendering the problem NP-hard Hartmanis (1982). In practice, brute-force forward propagation is computationally intractable due to the prohibitive latency of repeatedly retrieving expert parameters. To overcome these bottlenecks, we derive an efficient analytical proxy to quantify the collective importance of expert combinations by analyzing the second-order Taylor expansion LeCun et al. (1989); Dong et al. (2017) of the loss degradation in the following section.

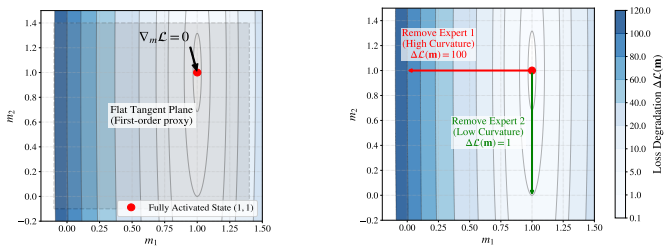
2.2 Hessian-Aware Interdependence Extraction

Second-Order Taylor Expansion of Loss Degradation. To circumvent the computational intractability of combinatorial search in Eqn. equation 4, we reformulate the dynamic expert selection problem by analyzing the model's behavior around its fully activated state ($\mathbf{m} = \mathbf{1}$). Since the fully activated state represents the theoretical upper bound of the network's representational capacity, selecting a sparse subset of experts can be mathematically modeled as a finite deviation from this optimum. By taking the second-order Taylor expansion of the loss degradation around the fully activated state, we have:

$$\Delta\mathcal{L}(\mathbf{m}) = \mathcal{L}(\mathbf{m}) - \mathcal{L}(\mathbf{1}) \approx \underbrace{(\mathbf{m}-\mathbf{1})^\top \nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}}_{\text{First-order Term}} + \underbrace{\frac{1}{2}(\mathbf{m}-\mathbf{1})^\top \mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}(\mathbf{m}-\mathbf{1})}_{\text{Second-order Term}}, \tag{5}$$

where $\nabla_{\mathbf{m}}\mathcal{L}$ and $\mathbf{H}_{\mathbf{m}} = \nabla^2_{\mathbf{m}}\mathcal{L}$ are the gradient vector and Hessian matrix of the loss function with respect to $\mathbf{m}$, respectively.

Necessity of Second-Order Curvature. While Eqn. equation 5 incorporates the first-order gradient term, relying solely on this term is insufficient for expert selection due to the vanishing gradient problem on convergence. In a dynamic training regime, experts naturally settle into local stationary points, causing their first-order gradients to diminish ($\|\nabla_{\mathbf{m}}\mathcal{L}\| \to \mathbf{0}$). This induces a severe linearization error: the approximation fits a flat tangent plane to the loss landscape, failing to capture the



(a) Gradients (First-Order) (b) Hessian (Second-Order)

Figure 2: First-order gradients versus second-order Hessian. (a) First-order proxy fits a flat tangent plane at local stationary points, failing to distinguish between experts. (b) Hessian captures steep curvature, revealing the massive penalty of removing expert 1 ($\Delta\mathcal{L}(\mathbf{m}) = 100$) versus expert 2 ($\Delta\mathcal{L}(\mathbf{m}) = 1$).

steep curvature penalties associated with dropping critical experts. To illustrate this limitation, we consider a simplified MoE landscape $\Delta\mathcal{L}(\mathbf{m}) = 100(m_1 - 1)^2 + (m_2 - 1)^2$, as shown in Figure 2. At the fully activated state $(1, 1)$, the gradient vanishes ($\nabla_{\mathbf{m}}\mathcal{L} = \mathbf{0}$), rendering the first-order term completely uninformative and incorrectly implying that both experts are equally expendable. For this reason, directly evaluating the impact of expert removal reveals a significant disparity: removing expert 1 (high curvature) incurs a substantial loss penalty of $100(0 - 1)^2 = 100$, whereas removing expert 2 (low curvature) yields a marginal penalty of only 1. Since the Hessian matrix $\mathbf{H}$ explicitly encapsulates this critical curvature distinction, retaining the second-order term in Eqn. equation 5 is imperative for establishing a robust selection criterion across all convergence states.

Hessian Matrix Derivation. Directly computing the exact high-dimensional Hessian matrix $\mathbf{H_m}$ is computationally prohibitive for LLM MoE architectures, as calculating second-order derivatives requires double back-propagation through massive expert networks. To circumvent this bottleneck, we employ the Gauss-Newton approximation Martens (2020); Singh & Alistarh (2020) to estimate local curvature using only first-order Jacobian information:

$$[\mathbf{H_m}]_{ij} \approx \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\frac{\partial\mathcal{L}}{\partial m_i}\right)\left(\frac{\partial\mathcal{L}}{\partial m_j}\right)^\top\right]. \tag{6}$$

Recalling the MoE output in Eqn. equation 3, the gradient with respect to the binary state m_i is given by $\frac{\partial\mathcal{L}}{\partial m_i} = \left(\frac{\partial y}{\partial m_i}\right)^\top \nabla_y\mathcal{L} = G_i(x)(E_i(x)^\top\nabla_y\mathcal{L})$. We substitute this into Eqn. equation 6, yielding the following analytical decomposition:

$$[\mathbf{H_m}]_{ij} \approx \mathbb{E}_x\left[\underbrace{G_i(x)G_j(x)}_{\text{Routing Co-activation}} \cdot \underbrace{(E_i(x)^\top\nabla_y\mathcal{L})(E_j(x)^\top\nabla_y\mathcal{L})^\top}_{\text{Gradient Alignment}}\right]. \tag{7}$$

This formulation decouples the expert interaction into topological and geometric components. The first term on the right-hand side, $G_i(x)G_j(x)$, acts as a topological gate that is non-zero only when experts i and j are co-activated for the same input x, thereby capturing sparse routing co-occurrence. The second term captures geometric correlations, where $E_i(x)^\top\nabla_y\mathcal{L}$ represents the functional contribution of expert i, quantifying the projection of its output onto the gradient to indicate its effectiveness in reducing the training loss. Consequently, when two co-activated experts exhibit identically signed projections (i.e., driving the loss descent along the consistent optimization trajectory), it implies functional redundancy and results in a positive curvature penalty ($[\mathbf{H_m}]_{ij} > 0$). Conversely, oppositely signed projections imply complementary error cancellation where one expert provides reverse compensation for the other, yielding a negative curvature reward ($[\mathbf{H_m}]_{ij} < 0$) that signifies functional synergy.

2.3 DIFFERENCE-OF-SUBMODULAR DECOMPOSITION

Problem Reformulation and Utility Function. With the analytical form of $[\mathbf{H_m}]_{ij}$ established, we substitute it into the loss degradation approximation to construct a tractable optimization objective. Our goal is to find the binary selection vector $\mathbf{m} \in \{0,1\}^N$ that minimizes the loss degradation, which is mathematically equivalent to maximizing the retained utility $\mathcal{U}(\mathbf{m}) = -\Delta\mathcal{L}(\mathbf{m})$. As derived in Appendix C.1, by expanding the quadratic term and discarding the constant term independent of $\mathbf{m}$, we reformulate the maximization of the retained utility as follows:

$$\max_{\mathbf{m}\in\{0,1\}^N} \mathcal{U}(\mathbf{m}) = \sum_{i=1}^{N} m_i d_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij}, \quad \text{s.t.} \quad \|\mathbf{m}\| = k, \tag{8}$$

where $d_i = \sum_{j=1}^{N}[\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij} - [\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i$ represents the individual utility of expert E_i. Here, $-[\nabla_{\mathbf{m}}\mathcal{L}]_i$ captures the standalone contribution of E_i toward minimizing the loss, while $\sum_{j=1}^{N}[\mathbf{H_m}]_{ij}$ aggregates the pairwise coupling effects between E_i and all other experts. The quadratic term $-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j [\mathbf{H_m}]_{ij}$ explicitly governs the combinatorial interaction among the selected experts. Eqn. equation 8 casts the original loss minimization as a cardinality-constrained quadratic binary optimization (QBO) problem, thereby decoupling the individual capability of each expert from their inter-expert dependencies.

Difference-of-Submodular Decomposition. While the cardinality-constrained QBO problem in equation 8 provides an explicit structural interpretation, directly maximizing the objective $\mathcal{U}(\mathbf{m})$ is computationally intractable, as proved to be an NP-hard problem in Appendix C.2. This bottleneck stems from the indefiniteness of the MoE Hessian matrix $\mathbf{H_m}$, which contains both positive entries (indicating redundancy) and negative entries (indicating synergy). This mixture of conflicting combinatorial forces renders the objective landscape highly non-convex and non-submodular, preventing the use of standard greedy or convex optimization solvers.

To overcome this bottleneck, we isolate these conflicting coupling effects by decomposing the Hessian $\mathbf{H_m}$ into its non-negative components. We define $\mathbf{H_m^+}$ and $\mathbf{H_m^-}$ such that $[\mathbf{H_m^+}]_{ij} = \max(0, [\mathbf{H_m}]_{ij})$ and $[\mathbf{H_m^-}]_{ij} = \max(0, -[\mathbf{H_m}]_{ij})$. By substituting $\mathbf{H_m} = \mathbf{H_m^+} - \mathbf{H_m^-}$ into the utility function $\mathcal{U}(\mathbf{m})$, we reformulate the objective as the difference between a gain function $f(\mathbf{m})$ and a cost function $g(\mathbf{m})$:

$$\max_{\mathbf{m}\in\{0,1\}^N} \mathcal{U}(\mathbf{m}) = \underbrace{\left(\sum_{i=1}^{N} m_i d_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j [\mathbf{H_m^-}|_{\mathbf{m}=\mathbf{1}}]_{ij}\right)}_{f(\mathbf{m})} - \underbrace{\left(\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j [\mathbf{H_m^+}|_{\mathbf{m}=\mathbf{1}}]_{ij}\right)}_{g(\mathbf{m})}, \text{ s.t. } \|\mathbf{m}\| = k, \tag{9}$$

where the gain function $f(\mathbf{m})$ represents the cumulative positive utility, incorporating individual expert capability via d_i and functional synergy arising from complementary interactions captured by $\mathbf{H_m^-}$; the cost function $g(\mathbf{m})$ quantifies the functional redundancy, acting as a penalty term to discourage the co-selection of overlapping experts via $\mathbf{H_m^+}$. The non-negativity of the decomposed components $\mathbf{H_m^+}$ and $\mathbf{H_m^-}$ imparts a mathematical property: both the gain function $f(\mathbf{m})$ and the cost function $g(\mathbf{m})$ are monotonically non-decreasing and strictly supermodular (see proof in Appendix C.3). This guarantee confirms that our objective $\mathcal{U}(\mathbf{m})$ is strictly the difference of two supermodular functions. Since the negation of a supermodular function is submodular, the objective $f(\mathbf{m}) - g(\mathbf{m})$ is algebraically equivalent to $(-g(\mathbf{m})) - (-f(\mathbf{m}))$. This fundamental equivalence converts the originally intractable indefinite quadratic maximization into a canonical DS optimization problem.

2.4 MAJORIZATION-MINIMIZATION OPTIMIZATION.

MM Algorithm. To efficiently solve the maximization problem in equation 9, we employ the MM algorithm. The core principle is to iteratively maximize a tractable modular surrogate function $\mathcal{Q}(\mathbf{m}; \mathbf{m}')$, which serves as a tight global lower bound for the objective $\mathcal{U}(\mathbf{m})$ at the given selection vector $\mathbf{m}'$. To construct this surrogate for $\mathcal{U}(\mathbf{m}) = (-g(\mathbf{m})) - (-f(\mathbf{m}))$, we apply the *Modular-Modular* procedure Iyer & Bilmes (2012) to establish a tight upper bound for the second $-f(\mathbf{m})$ (equivalent to finding a lower bound for the gain function $f(\mathbf{m})$) and lower bound for the first submodular term $-g(\mathbf{m})$ (equivalent to finding a upper bound for the cost function $g(\mathbf{m})$).

Lower Bound for Gain Function. Given that the negative gain term $-f(\mathbf{m})$ is submodular, it is upper-bounded by its first-order Taylor expansion at $\mathbf{m}'$. Inverting this bounding relationship mathematically yields a tight lower bound $L_f(\mathbf{m}; \mathbf{m}')$ for the gain function $f(\mathbf{m})$:

$$f(\mathbf{m}) \geq f(\mathbf{m}') + \sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}')]_i (m_i - m_i') \triangleq L_f(\mathbf{m}; \mathbf{m}'), \tag{10}$$

where the gradient $[\nabla_{\mathbf{m}} f(\mathbf{m}')]_i = d_i + \sum_{j=1}^{N} [\mathbf{H}_{\mathbf{m}}^{-}|_{\mathbf{m}=\mathbf{1}}]_{ij} m_j'$ captures the synergy gain of expert E_i relative to the selection vector $\mathbf{m}'$.

Upper Bound for Cost Function. To derive a upper bound $U_g(\mathbf{m})$ for cost function $g(\mathbf{m})$, we leverage the arithmetic equality for binary variables ($m_i m_j = \frac{1}{2}(m_i^2 + m_j^2) - \frac{1}{2}(m_i - m_j)^2 = \frac{1}{2}(m_i + m_j) - \frac{1}{2}(m_i - m_j)^2$). This reformulates the cost function $g(\mathbf{m})$ into a linear component and a strictly concave residual. By applying the first-order Taylor expansion to the concave residual around $\mathbf{m}^{(t)}$, we obtain a tight upper bound (with detailed proof in Appendix C.4):

$$g(\mathbf{m}) \leq \sum_{i=1}^{N} m_i \left[\sum_{j=1}^{N} [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i' + m_j' \right) \right] + C_g' \triangleq U_g(\mathbf{m}; \mathbf{m}'), \tag{11}$$

where $C_g' = \frac{1}{4} \sum_{i=1}^{N} \sum_{j=1}^{N} (m_i' - m_j')^2 [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij}$ is a state-dependent constant independent of the selection vector $\mathbf{m}$, and $\sum_{j=1}^{N} [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i' + m_j' \right)$ gives the context-aware redundancy penalty representing the the potential functional overlap of expert E_i relative to the entire selected expert subset, which heavily penalizes an expert for overlapping with currently active experts, while forgiving overlaps with inactive one.

By synthesizing the above bounding functions, we construct the tractable modular surrogate $\mathcal{Q}(\mathbf{m}; \mathbf{m}')$ for $\mathcal{U}(\mathbf{m})$ as follows:

$$\mathcal{Q}(\mathbf{m}; \mathbf{m}') \triangleq L_f(\mathbf{m}; \mathbf{m}') - U_g(\mathbf{m}; \mathbf{m}') \leq \mathcal{U}(\mathbf{m}). \tag{12}$$

Based on the surrogate function $\mathcal{Q}(\mathbf{m}; \mathbf{m}')$ in Eqn. equation 12, we establish the iterative optimization procedure. Let $\mathbf{m}^{(t)}$ denote the expert selection vector at the t-th iteration. By setting the reference point $\mathbf{m}' = \mathbf{m}^{(t)}$ and discarding constant terms independent of $\mathbf{m}$, the optimization problem at the $(t+1)$-th iteration reduces to a cardinality-constrained linear maximization (with detailed proof in Appendix C.5):

$$\mathbf{m}^{(t+1)} = \underset{\mathbf{m}:\|\mathbf{m}\| \leq k}{\arg\max} \sum_{i=1}^{N} m_i S_i^{(t)}, \tag{13}$$

where $S_i^{(t)} = [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i - \sum_{j=1}^{N} [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)} \right)$ represents the importance score of E_i, which quantifies the trade-off between its synergy gain and redundancy penalty.

This linear formulation transforms the intractable NP-hard combinatorial problem into a simple Top-k sorting. Consequently, the optimal subset $\mathbf{m}^{(t+1)}$ is determined by selecting the Top-k experts with the highest scores. The complete procedure of the proposed MM algorithm is outlined in Appendix D.1, which provides the following theoretical convergence guarantee:

Theorem 1 (Convergence Properties) *The sequence of expert selection vectors $\{\mathbf{m}^{(t)}\}_{t \geq 0}$ generated by the MM algorithm yields a monotonically non-decreasing objective value. Specifically, for any t-th iteration, the following inequality holds:*

$$\mathcal{U}(\mathbf{m}^{(t+1)}) \geq \mathcal{U}(\mathbf{m}^{(t)}), \quad \forall t \geq 0. \tag{14}$$

Given that the feasible region of $\mathbf{m}$ is a finite discrete hypercube, the MM algorithm is guaranteed to converge in a finite number of iterations.

Proof. See Appendix C.6.

Table 1: Experimental results on test accuracy (%) (↑) and convergence rounds (↓). We evaluate DS-MoE and four baselines on Switch-Base-32, Switch-Base-64, and DeepSeek-MoE-16B.

Model	Method	Test Accuracy (%) ↑					Convergence Rounds ↓				
		AGNews	PIQA	HellaSwag	MMLU	Avg.	AGNews	PIQA	HellaSwag	MMLU	Avg.
Switch-Base-32	Rand-MoE	87.3 ± 0.3	68.1 ± 0.8	65.9 ± 0.7	39.8 ± 1.1	65.3 ± 0.7	84 ± 6	26 ± 3	28 ± 4	34 ± 5	43.0 ± 4.5
	S-MoE	92.6 ± 0.2	74.3 ± 0.5	71.8 ± 0.4	46.4 ± 0.6	71.3 ± 0.4	59 ± 3	15 ± 2	17 ± 2	19 ± 3	27.5 ± 2.5
	SEER-MoE	93.1 ± 0.3	75.2 ± 0.4	72.7 ± 0.5	46.9 ± 0.5	72.0 ± 0.4	49 ± 3	12 ± 1	13 ± 2	17 ± 2	22.8 ± 2.0
	DiEP	93.3 ± 0.2	75.8 ± 0.3	73.2 ± 0.4	47.5 ± 0.3	72.5 ± 0.3	48 ± 2	12 ± 2	12 ± 1	15 ± 2	21.8 ± 1.8
	DS-MoE	$\mathbf{94.7 \pm 0.1}$	$\mathbf{78.6 \pm 0.2}$	$\mathbf{75.6 \pm 0.2}$	$\mathbf{51.8 \pm 0.3}$	$\mathbf{75.2 \pm 0.2}$	$\mathbf{32 \pm 1}$	$\mathbf{8 \pm 1}$	$\mathbf{10 \pm 1}$	$\mathbf{11 \pm 2}$	$\mathbf{15.3 \pm 1.3}$
Switch-Base-64	Rand-MoE	87.6 ± 0.4	68.7 ± 0.6	66.1 ± 0.9	40.3 ± 1.2	65.7 ± 0.8	82 ± 7	23 ± 4	25 ± 3	26 ± 5	39.0 ± 4.8
	S-MoE	92.8 ± 0.3	74.8 ± 0.6	72.1 ± 0.5	46.8 ± 0.8	71.6 ± 0.6	55 ± 4	13 ± 2	14 ± 3	16 ± 2	24.5 ± 2.8
	SEER-MoE	93.2 ± 0.2	76.3 ± 0.5	73.3 ± 0.4	47.9 ± 0.6	72.7 ± 0.4	47 ± 2	11 ± 2	13 ± 1	14 ± 3	21.3 ± 2.0
	DiEP	93.5 ± 0.2	76.5 ± 0.3	73.9 ± 0.4	48.1 ± 0.5	73.0 ± 0.4	45 ± 3	10 ± 1	11 ± 2	13 ± 1	19.8 ± 1.8
	DS-MoE	$\mathbf{94.9 \pm 0.1}$	$\mathbf{78.9 \pm 0.2}$	$\mathbf{75.8 \pm 0.3}$	$\mathbf{52.4 \pm 0.4}$	$\mathbf{75.5 \pm 0.3}$	$\mathbf{30 \pm 2}$	$\mathbf{7 \pm 1}$	$\mathbf{8 \pm 1}$	$\mathbf{10 \pm 1}$	$\mathbf{13.8 \pm 1.3}$
DeepSeek-MoE-16B	Rand-MoE	86.1 ± 0.5	69.8 ± 0.9	68.2 ± 0.8	32.8 ± 1.3	64.2 ± 0.9	138 ± 9	33 ± 6	35 ± 5	39 ± 7	61.3 ± 6.8
	S-MoE	92.1 ± 0.4	75.2 ± 0.7	73.4 ± 0.6	39.1 ± 0.9	70.0 ± 0.7	85 ± 6	20 ± 3	18 ± 4	22 ± 3	36.3 ± 4.0
	SEER-MoE	92.3 ± 0.3	77.1 ± 0.5	74.6 ± 0.6	40.3 ± 0.7	71.1 ± 0.5	77 ± 4	17 ± 3	17 ± 2	18 ± 3	32.3 ± 3.0
	DiEP	92.9 ± 0.2	77.6 ± 0.4	75.1 ± 0.5	42.1 ± 0.6	71.9 ± 0.4	73 ± 3	15 ± 2	16 ± 3	17 ± 2	30.3 ± 2.5
	DS-MoE	$\mathbf{94.1 \pm 0.2}$	$\mathbf{81.2 \pm 0.3}$	$\mathbf{77.9 \pm 0.2}$	$\mathbf{45.1 \pm 0.4}$	$\mathbf{74.6 \pm 0.3}$	$\mathbf{45 \pm 2}$	$\mathbf{12 \pm 1}$	$\mathbf{13 \pm 1}$	$\mathbf{14 \pm 2}$	$\mathbf{21.0 \pm 1.5}$

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Tasks and Datasets. We conduct extensive experiments across three tasks: i) *Natural Language Understanding* on AGNews Zhang et al. (2015) dataset; ii) *Commonsense Reasoning* on PIQA Bisk et al. (2020) and HellaSwag Zellers et al. (2019) datasets; and iii) *Advanced Multi-domain Reasoning* on MMLU Hendrycks et al. (2021) dataset. Comprehensive dataset descriptions and experimental setups are detailed in Appendix D.2.

Models. We evaluate the proposed DS-MoE framework on three representative large-scale MoE architectures: *Switch-Base-32*, *Switch-Base-64* Fedus et al. (2022), and *DeepSeek-MoE-16B* Dai et al. (2024). The implementation details are shown in Appendix D.3.

Baselines. We compare DS-MoE against four representative baselines under consistent cardinality constraints: i) *SEER-MoE* Muzio et al. (2024): A state-of-the-art sparse expert efficiency framework that dynamically enforces expert sparsity by introducing a differentiable regularization penalty on the gating distribution; ii) *DiEP* Bai et al. (2025): A expert selection strategy that utilizes gradient-based evaluation to dynamically identify and drop redundant experts with non-uniform sparsity across layers; iii) *Standard MoE (S-MoE)* Shazeer et al. (2017): The conventional routing mechanism that independently activates a fixed number of experts for each token based purely on the highest localized gating probabilities; and iv) *Random MoE (Rand-MoE):* A data-independent protocol where the subset of active experts is sampled uniformly at random during model training, strictly adhering to the required cardinality constraints without utilizing any routing probabilities or gradient signals.

3.2 MAIN RESULTS

Natural Language Understanding. Table 1 shows DS-MoE consistently outperforms all baselines on the AGNews dataset, achieving over 94% test accuracies across Switch-Base-32, Switch-Base-64, and DeepSeek-MoE-16B. This performance consistently surpasses both standard Top-k routing (S-MoE) and advanced baselines (SEER-MoE, DiEP). Furthermore, by penalizing functional redundancy via positive Hessian entries, DS-MoE drastically expedites model training. On DeepSeek-MoE-16B, DS-MoE converges in just 45 rounds—reducing S-MoE's required 100 rounds by more than half, while maintaining a substantial speed advantage over both DiEP and SEER-MoE.

Commonsense Reasoning. Table 1 shows that DS-MoE exhibits significant superiority on complex reasoning tasks PIQA and HellaSwag. The first-order methods like DiEP struggle in this regime due to the lack of capturing inter-expert dependencies. By contrast, DS-MoE exploits second-order interactions to identify synergistic expert combinations, reaches 78.6% and 75.6% test accuracy on Switch-Base-32, and outperforms the standard S-MoE as well as SEER-MoE and DiEP. Beyond accuracy improvements, DS-MoE accelerates model convergence for these tasks by 27% to 46% compared to all baselines.

Advanced Multi-Domain Reasoning. Table 1 highlights DS-MoE's significant advantage on the MMLU benchmark. While SEER-MoE and DiEP marginally improve accuracy over the isolated

Table 2: Robustness and Scalability Analysis on DeepSeek-MoE-16B. The training performance versus (a) expert sparsity k with $N = 64$ and (b) expert pool size N with $k = 2$.

(a) Performance versus expert sparsity k

Model	k	Method	Test Accuracy (%) ↑					Convergence Rounds ↓				
			AGNews	PIQA	HellaSwag	MMLU	Avg.	AGNews	PIQA	HellaSwag	MMLU	Avg.
DeepSeek-16B	1	Rand-MoE	83.2 ± 0.6	66.8 ± 1.2	65.9 ± 0.9	28.9 ± 1.5	61.2 ± 1.1	162 ± 14	59 ± 7	56 ± 9	57 ± 11	83.5 ± 10.3
		S-MoE	90.1 ± 0.5	73.4 ± 0.9	72.1 ± 0.7	37.7 ± 1.2	68.3 ± 0.8	118 ± 9	26 ± 4	23 ± 5	29 ± 6	49.0 ± 6.0
		SEER-MoE	90.9 ± 0.4	75.7 ± 0.7	73.2 ± 0.8	39.2 ± 0.9	69.8 ± 0.7	112 ± 6	19 ± 4	19 ± 3	21 ± 4	42.8 ± 4.3
		DiEP	91.6 ± 0.3	75.9 ± 0.6	73.8 ± 0.7	40.8 ± 0.8	70.5 ± 0.6	104 ± 5	18 ± 3	17 ± 4	18 ± 3	39.3 ± 3.8
		DS-MoE	$\mathbf{93.2 \pm 0.3}$	$\mathbf{79.8 \pm 0.4}$	$\mathbf{76.3 \pm 0.3}$	$\mathbf{43.9 \pm 0.5}$	$\mathbf{73.3 \pm 0.4}$	$\mathbf{66 \pm 3}$	$\mathbf{12 \pm 2}$	$\mathbf{13 \pm 2}$	$\mathbf{15 \pm 3}$	$\mathbf{26.5 \pm 2.5}$
	2	Rand-MoE	86.1 ± 0.5	69.8 ± 0.9	68.2 ± 0.8	32.8 ± 1.3	64.2 ± 0.9	138 ± 9	33 ± 6	35 ± 5	39 ± 7	61.3 ± 6.8
		S-MoE	92.1 ± 0.4	75.2 ± 0.7	73.4 ± 0.6	39.1 ± 0.9	70.0 ± 0.7	85 ± 6	20 ± 3	18 ± 4	22 ± 3	36.3 ± 4.0
		SEER-MoE	92.3 ± 0.3	77.1 ± 0.5	74.6 ± 0.6	40.3 ± 0.7	71.1 ± 0.5	77 ± 4	17 ± 3	17 ± 2	18 ± 3	32.3 ± 3.0
		DiEP	92.9 ± 0.2	77.6 ± 0.4	75.1 ± 0.5	42.1 ± 0.6	71.9 ± 0.4	73 ± 3	15 ± 2	16 ± 3	17 ± 2	30.3 ± 2.5
		DS-MoE	$\mathbf{94.1 \pm 0.2}$	$\mathbf{81.2 \pm 0.3}$	$\mathbf{77.9 \pm 0.2}$	$\mathbf{45.1 \pm 0.4}$	$\mathbf{74.6 \pm 0.3}$	$\mathbf{45 \pm 2}$	$\mathbf{12 \pm 1}$	$\mathbf{13 \pm 1}$	$\mathbf{14 \pm 2}$	$\mathbf{21.0 \pm 1.5}$
	3	Rand-MoE	86.7 ± 0.4	70.6 ± 0.7	69.4 ± 0.6	34.1 ± 1.1	65.2 ± 0.7	125 ± 7	30 ± 4	31 ± 4	33 ± 5	54.8 ± 5.0
		S-MoE	92.4 ± 0.3	75.9 ± 0.5	73.9 ± 0.5	40.0 ± 0.7	70.6 ± 0.5	76 ± 5	18 ± 2	17 ± 3	19 ± 2	32.5 ± 3.0
		SEER-MoE	92.7 ± 0.2	77.3 ± 0.4	75.1 ± 0.4	41.3 ± 0.6	71.6 ± 0.4	75 ± 3	16 ± 2	15 ± 1	17 ± 2	30.8 ± 2.0
		DiEP	93.1 ± 0.1	78.0 ± 0.3	75.7 ± 0.3	42.6 ± 0.4	72.4 ± 0.3	68 ± 2	13 ± 1	14 ± 2	16 ± 1	27.8 ± 1.5
		DS-MoE	$\mathbf{94.5 \pm 0.1}$	$\mathbf{81.4 \pm 0.2}$	$\mathbf{78.2 \pm 0.1}$	$\mathbf{45.4 \pm 0.2}$	$\mathbf{74.9 \pm 0.2}$	$\mathbf{38 \pm 1}$	$\mathbf{11 \pm 1}$	$\mathbf{12 \pm 1}$	$\mathbf{12 \pm 1}$	$\mathbf{18.3 \pm 1.0}$

(b) Performance versus expert pool size N

Model	N	Method	Test Accuracy (%) ↑					Convergence Rounds ↓				
			AGNews	PIQA	HellaSwag	MMLU	Avg.	AGNews	PIQA	HellaSwag	MMLU	Avg.
DeepSeek-16B	32	Rand-MoE	84.3 ± 0.8	67.9 ± 1.2	66.8 ± 1.0	31.2 ± 1.7	62.6 ± 1.1	153 ± 12	41 ± 8	51 ± 8	45 ± 9	72.5 ± 8.7
		S-MoE	90.8 ± 0.5	72.5 ± 0.8	73.7 ± 0.9	38.0 ± 1.3	68.8 ± 0.8	115 ± 9	22 ± 4	24 ± 5	28 ± 4	47.3 ± 5.0
		SEER-MoE	91.4 ± 0.4	75.6 ± 0.7	73.7 ± 0.8	39.3 ± 0.9	70.0 ± 0.7	108 ± 7	19 ± 4	19 ± 4	21 ± 5	41.8 ± 4.2
		DiEP	91.8 ± 0.4	76.2 ± 0.7	74.3 ± 0.6	41.2 ± 1.0	70.9 ± 0.6	103 ± 5	18 ± 3	17 ± 4	17 ± 4	38.8 ± 3.5
		DS-MoE	$\mathbf{93.1 \pm 0.3}$	$\mathbf{80.0 \pm 0.5}$	$\mathbf{77.2 \pm 0.4}$	$\mathbf{44.0 \pm 0.7}$	$\mathbf{73.6 \pm 0.4}$	$\mathbf{69 \pm 4}$	$\mathbf{13 \pm 2}$	$\mathbf{14 \pm 2}$	$\mathbf{15 \pm 3}$	$\mathbf{27.8 \pm 2.7}$
	48	Rand-MoE	85.3 ± 0.6	69.0 ± 1.2	67.6 ± 0.8	32.1 ± 1.4	63.5 ± 1.0	144 ± 10	37 ± 6	41 ± 7	43 ± 8	66.3 ± 7.8
		S-MoE	91.5 ± 0.5	74.5 ± 0.7	73.0 ± 0.8	38.6 ± 1.0	69.4 ± 0.7	106 ± 6	20 ± 4	23 ± 4	23 ± 4	43.0 ± 4.5
		SEER-MoE	91.9 ± 0.4	76.5 ± 0.5	74.2 ± 0.7	39.9 ± 0.9	70.6 ± 0.5	102 ± 5	18 ± 3	17 ± 3	19 ± 3	39.0 ± 3.5
		DiEP	92.4 ± 0.2	77.0 ± 0.6	74.8 ± 0.6	41.7 ± 0.8	71.5 ± 0.5	76 ± 4	17 ± 3	17 ± 3	17 ± 2	31.8 ± 3.0
		DS-MoE	$\mathbf{93.7 \pm 0.3}$	$\mathbf{80.7 \pm 0.4}$	$\mathbf{77.6 \pm 0.4}$	$\mathbf{44.6 \pm 0.5}$	$\mathbf{74.2 \pm 0.3}$	$\mathbf{52 \pm 3}$	$\mathbf{13 \pm 2}$	$\mathbf{12 \pm 1}$	$\mathbf{14 \pm 2}$	$\mathbf{22.8 \pm 2.3}$
	64	Rand-MoE	86.1 ± 0.5	69.8 ± 0.9	68.2 ± 0.8	32.8 ± 1.3	64.2 ± 0.9	138 ± 9	33 ± 6	35 ± 5	39 ± 7	61.3 ± 6.8
		S-MoE	92.1 ± 0.4	75.2 ± 0.7	73.4 ± 0.6	39.1 ± 0.9	70.0 ± 0.7	85 ± 6	20 ± 3	18 ± 4	22 ± 3	36.3 ± 4.0
		SEER-MoE	92.3 ± 0.3	77.1 ± 0.5	74.6 ± 0.6	40.3 ± 0.7	71.1 ± 0.5	77 ± 4	17 ± 3	17 ± 2	18 ± 3	32.3 ± 3.0
		DiEP	92.9 ± 0.2	77.6 ± 0.4	75.1 ± 0.5	42.1 ± 0.6	71.9 ± 0.4	73 ± 3	15 ± 2	16 ± 3	17 ± 2	30.3 ± 2.5
		DS-MoE	$\mathbf{94.1 \pm 0.2}$	$\mathbf{81.2 \pm 0.3}$	$\mathbf{77.9 \pm 0.2}$	$\mathbf{45.1 \pm 0.4}$	$\mathbf{74.6 \pm 0.3}$	$\mathbf{45 \pm 2}$	$\mathbf{12 \pm 1}$	$\mathbf{13 \pm 1}$	$\mathbf{14 \pm 2}$	$\mathbf{21.0 \pm 1.5}$

S-MoE (which yields sub-optimal accuracies of 46.4%, 46.8%, 39.1%, and slow convergence of 19, 16, 22 rounds), they still fail to accurately capture complex expert functional correlation. In contrast, by the design of difference-of-submodular optimization, DS-MoE successfully decouples functional synergy and redundancy across experts. This enables DS-MoE to achieve state-of-the-art accuracies (51.8%, 52.4%, 45.1%) while minimizing convergence to 11, 10, and 14 rounds across the three architectures.

3.3 Micro-Benchmarking

Robustness to Expert Sparsity (k). Table 2a evaluates model's robustness across varying sparsity constraints ($k \in \{1, 2, 3\}$) on DeepSeek-MoE-16B. Under extreme sparsity ($k = 1$), S-MoE, SEER-MoE, and DiEP degrade significantly to 68.3%, 69.8%, and 70.5% average accuracy, respectively. In contrast, DS-MoE preserves critical expert synergies to achieve a superior 73.3% average accuracy. As the routing budget expands ($k = 3$), DS-MoE efficiently translates the added capacity into a peak 74.9% accuracy. Beyond accuracy, DS-MoE consistently accelerates convergence over all baselines. At $k = 3$, it reduces the average convergence time to merely 18.3 rounds, achieving approximately 1.5×, 1.7×, and 1.8× speedups over DiEP (27.8 rounds), SEER-MoE (30.8 rounds), and S-MoE (32.5 rounds), respectively.

Scalability Across Expert Pool Sizes (N). Table 2b evaluates training performance as the expert search space expands from $N = 32$ to 64 at a fixed sparsity ($k = 2$). While an enlarged pool increases model capacity, it triggers an exponential explosion in possible expert combinations, making optimal routing increasingly difficult for standard mechanisms. Despite this, DS-MoE consistently outperforms all baselines, improving the average accuracy from 73.6% to 74.6% at $N = 64$. Furthermore, DS-MoE maintains stable convergence; at $N = 64$, it converges in just 21.0 rounds, preserving a significant efficiency advantage over DiEP (30.3 rounds) and SEER-MoE (32.3 rounds). This confirms DS-MoE's robust capability to reliably extract synergistic expert combinations at scale.

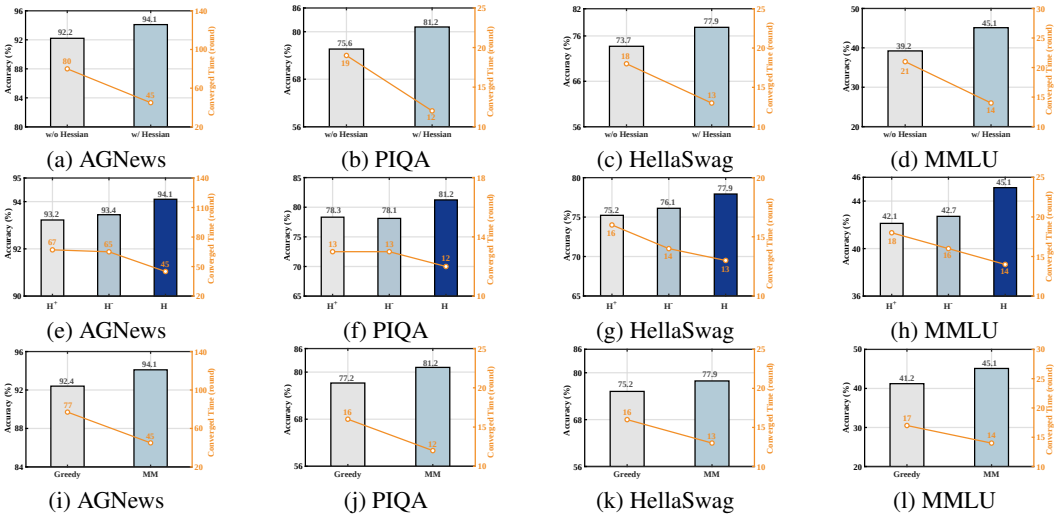


Figure 3: Comprehensive ablation study of the DS-MoE framework on DeepSeek-MoE-16B with $k = 2$ and $N = 64$. (a)-(d): Effectiveness of second-order Hessian over first-order proxies. (e)-(h): Necessity of DS decomposition. (i)-(l): Optimization superiority of the proposed MM algorithm.

3.4 Ablation Study

Study on the Impact of Hessian Matrix. Figures 3(a)-3(d) compare a purely first-order gradient proxy (w/o Hessian) against our framework utilizing the second-order Hessian matrix (w/ Hessian). The results demonstrate that integrating the Hessian matrix consistently yields higher test accuracy and a sharp reduction in convergence rounds. This validates a fundamental limitation of first-order heuristics: they fail to capture loss landscape curvature and suffer from vanishing gradients in flat regions. In contrast, formulating expert selection via the second-order Hessian explicitly captures combinatorial inter-expert dependencies, enabling accurate evaluation of expert synergies and circumventing sub-optimal local minima.

Study on the Impact of DS Decomposition. Figures 3(e)-3(h) evaluates the necessity of the DS decomposition by comparing the complete framework ($\mathbf{H}$) against isolated variants employing strictly the redundancy penalty ($\mathbf{H}^+$) or synergy reward ($\mathbf{H}^-$). The results reveal that deploying either component individually yields marginal performance gains, integrating both via the complete DS decomposition consistently achieves the highest test accuracy and fastest convergence. This confirms that mathematically decoupling redundancy reduction from synergy maximization is indispensable for fully exploiting inter-expert dependencies and optimizing routing efficacy.

Study on the Impact of MM Algorithm. Figures 3(i)-3(l) evaluate the effectiveness of the proposed MM algorithm against a standard local greedy search strategy (Greedy). We see that the greedy approach is highly susceptible to sub-optimal local minima, manifesting as degraded test accuracy and prolonged convergence times. Conversely, the MM algorithm consistently achieves superior accuracy while significantly reducing the required training rounds. These empirical gains validate the monotonic convergence guarantees intrinsic to the MM framework. By bounding the highly non-convex objective via tight surrogate functions, the MM algorithm can circumvent the sub-optimal routing inherent to heuristic greedy searches.

4 Conclusion

In this paper, we introduced DS-MoE, a theoretically grounded MoE selection framework. By leveraging the Hessian of the loss degradation, we quantify complex inter-expert dependencies, revealing fundamental duality between functional redundancy and cooperative synergy. To effectively harness this duality, we formulate the expert selection process as a DS optimization problem and develop a tailored MM algorithm with a provable convergence guarantee. Finally, we discuss the limitations of DS-MoE and outline promising directions for future research in Appendix E, and present the practical impact of DS-MoE in Appendix F.

REFERENCES

Sikai Bai, Haoxi Li, Jie Zhang, Zicong Hong, and Song Guo. DiEP: Adaptive Mixture-of-Experts Compression through Differentiable Expert Pruning. In *Proc. of the 39th Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 2025.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. PIQA: Reasoning About Physical Commonsense in Natural Language. In *Proc. of the 34th AAAI Conf. on Artif. Intell. (AAAI)*, 2020.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al. On the Representation Collapse of Sparse Mixture of Experts. In *Proc. of the 36th Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 2022.

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified Scaling Laws for Routed Language Models. In *Proc. of the 39th Int. Conf. on Mach. Learn. (ICML)*, 2022.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Hui Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingjian Yu, Y. Wu, et al. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. In *Proc. of the 62nd Annu. Meet. of the Assoc. for Comput. Linguistics (ACL)*, 2024.

Xin Dong, Shangyu Chen, and Sinno Jialin Pan. Learning to Prune Deep Neural Networks via Layer-wise Optimal Brain Surgeon. In *Proc. of the 31st Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 2017.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *J. Mach. Learn. Res.*, 23(120):1–39, 2022.

Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. MegaBlocks: Efficient Sparse Training with Mixture-of-Experts. In *Proc. of the 6th Conf. on Mach. Learn. and Syst. (MLSys)*, 2023.

Jian Gao, Zhenghang Xu, Ruizhi Li, and Minghao Yin. An Exact Algorithm with New Upper Bounds for the Maximum k-Defective Clique Problem in Massive Sparse Graphs. In *Proc. of the 36th AAAI Conf. on Artif. Intell. (AAAI)*, 2022.

Juris Hartmanis. Computers and Intractability: A Guide to the Theory of NP-Completeness. *SIAM Rev.*, 24(1):90, 1982.

Babak Hassibi and David G. Stork. Second Order Derivatives for Network Pruning: Optimal Brain Surgeon. In *Proc. of the 5th Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 1992.

Jiaao He, Jiezhong Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang. FastMoE: A Fast Mixture-of-Expert Training System. *arXiv preprint arXiv:2103.13262*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *Proc. of the 9th Int. Conf. on Learn. Representat.(ICLR)*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training Compute-Optimal Large Language Models. In *Proc. of the 36th Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 2022.

David R. Hunter and Kenneth Lange. A Tutorial on MM Algorithms. *The Amer. Statistician*, 58(1): 30–37, 2004.

Rishabh Iyer and Jeff Bilmes. Algorithms for Approximate Minimization of the Difference between Submodular Functions, with Applications. In *Proc. of the 28th Conf. on Uncertainty in Artif. Intell. (UAI)*, 2012.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive Mixtures of Local Experts. *Neural Comput.*, 3(1):79–87, 1991.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020.

Yann LeCun, John Denker, and Sara Solla. Optimal Brain Damage. In *Proc. of the 2nd Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 1989.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *Proc. of the 9th Int. Conf. on Learn. Representat. (ICLR)*, 2021.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-MoE: Scaling Unified Multimodal LLMs with Mixture of Experts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3424–3439, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2024a.

Jiacheng Liu, Peng Tang, Wenfeng Wang, Yuhang Ren, Xiaofeng Hou, Pheng Ann Heng, Minyi Guo, and Chao Li. A Survey on Inference Optimization Techniques for Mixture of Experts Models. *ACM Comput. Surv.*, 2024b.

Julien Mairal. Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization. In *Proc. of the 27th Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 2013.

James Martens. New Insights and Perspectives on the Natural Gradient Method. *J. Mach. Learn. Res.*, 21(146):1–76, 2020.

Alexandre Muzio, Alex Sun, and Churan He. Seer-MoE: Sparse Expert Efficiency through Regularization for Mixture-of-Experts. *arXiv preprint arXiv:2404.05089*, 2024.

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From Sparse to Soft Mixtures of Experts. In *Proc. of the 12th Int. Conf. on Learn. Representat. (ICLR)*, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *Proc. of the 5th Int. Conf. on Learn. Representat.(ICLR)*, 2017.

Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li. Scaling Law for Time Series Forecasting. In *Proc. of the 38th Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 2024.

Sidak Pal Singh and Dan Alistarh. WoodFisher: Efficient Second-Order Approximation for Neural Network Compression. In *Proc. of the 34th Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 2020.

Shawn Tan, Yikang Shen, Rameswar Panda, and Aaron Courville. Scattered Mixture-of-Experts Implementation. In *Proc. of the 1st Conf. on Lang. Modeling (COLM)*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can A Machine Really Finish Your Sentence? In *Proceedings of the 57th Annu. Meet. of the Assoc. for Comput. Linguistics (ACL)*, 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-Level Convolutional Networks for Text Classification. In *Proc. of the 29th Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 2015.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M. Dai, Quoc V. Le, James Laudon, et al. Mixture-of-Experts with Expert Choice Routing. In *Proc. of the 36th Annu. Conf. on Neural Inform. Process. Syst. (NeurIPS)*, 2022.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing Stable and Transferable Sparse Expert Models. *arXiv preprint arXiv:2202.08906*, 2022.

A Appendix

Contents

1 Introduction **1**

2 Methodology **2**

2.1 Preliminaries & Problem Formulation 2

2.2 Hessian-Aware Interdependence Extraction 3

2.3 Difference-of-Submodular Decomposition 5

2.4 Majorization-Minimization Optimization. 5

3 Experiments **7**

3.1 Experimental Setup 7

3.2 Main Results 7

3.3 Micro-Benchmarking 8

3.4 Ablation Study 9

4 Conclusion **9**

A Appendix **12**

B Related Work and Research Gap **13**

C Mathematical Proofs and Theoretical Derivations **13**

C.1 Derivation of the Retained Utility Function 13

C.2 Proof of NP-Hardness for Cardinality-Constrained QBO 15

C.3 Proof of Supermodularity for Decomposed Functions 16

C.4 Derivation of the Cost Function Upper Bound 18

C.5 Proof of Cardinality-Constrained Linear Maximization Problem 19

C.6 Proof of Convergence for the MM Algorithm 21

D Algorithmic Implementation and Experimental Details **23**

D.1 The Complete Majorization-Minimization Algorithm 23

D.2 Datasets and Evaluation Metrics 23

D.3 Model Architectures and Hyperparameter Settings 24

E Limitations and Future Work **25**

F Practical Impact **25**

B Related Work and Research Gap

The MoE architecture has emerged as a pivotal paradigm for scaling model capacity without incurring proportional computational overhead via sparse expert activation. Recent advancements in MoE primarily focus on routing optimizations and structural refinements:

For *routing optimizations*, existing methods are categorized into token-level Shazeer et al. (2017); Fedus et al. (2022); Lepikhin et al. (2021); Jiang et al. (2024) and expert-level routing Zhou et al. (2022); Puigcerver et al. (2024); Tan et al. (2024). Token-level routing, pioneered by Sparsely-Gated MoE Shazeer et al. (2017), employs a Top-k mechanism to dispatch each token to an expert combination based on localized routing probabilities; Switch Transformer Fedus et al. (2022) streamlined this mechanism by routing each token to only the Top-1 expert, thereby improving training efficiency; GShard Lepikhin et al. (2021) introduced automatic sharding to enable conditional computation in LLMs; Mixtral Jiang et al. (2024) demonstrated the scalability of a sparse Top-2 routing strategy within open-weight foundation models. For expert-level routing, expert choice routing Zhou et al. (2022) inverted this logic by allowing experts to selectively pull their top-scoring tokens up to a predefined capacity for enforcing strict load balancing; Soft-MoE Puigcerver et al. (2024) addressed discrete routing instabilities by allowing experts to pull continuous, weighted combinations of tokens; ScatterMoE Tan et al. (2024) optimized the expert choice mechanism by eliminating padding and redundant memory accesses during token aggregation.

For *structural refinements*, several frameworks have been proposed to enhance system execution efficiency: He et al. (2021); Gale et al. (2023); Zoph et al. (2022); Muzio et al. (2024); Dai et al. (2024): FastMoE He et al. (2021) accelerated training by optimizing communication primitives and customized routing operators; MegaBlocks Gale et al. (2023) utilized block-sparse matrix operations to support variable-capacity routing without token-dropping overhead; ST-MoE Zoph et al. (2022) incorporated auxiliary router z-loss to stabilize routing dynamics; SEER-MoE Muzio et al. (2024) introduced differentiable sparsity regularizers to refine gating distributions; and DeepSeek-MoE Dai et al. (2024) advocated for finer-grained expert partitioning combined with isolated shared experts to capture generalized knowledge.

While the aforementioned architectural and routing innovations Shazeer et al. (2017); Fedus et al. (2022); Lepikhin et al. (2021); Jiang et al. (2024); Zhou et al. (2022); Puigcerver et al. (2024); Tan et al. (2024); He et al. (2021); Gale et al. (2023); Zoph et al. (2022); Muzio et al. (2024); Dai et al. (2024) significantly improve MoE efficiency, they share a fundamental limitation: experts are evaluated and selected in absolute isolation. Whether governed by localized gating probabilities or global sparsity regularizers, existing frameworks Zhou et al. (2022); Puigcerver et al. (2024); Tan et al. (2024) compute selection scores independently for each individual expert. Consequently, these approaches inherently fail to distinguish between functional redundancy (i.e., experts encoding overlapping representation subspaces) and synergy (i.e., specific expert combinations providing complementary error compensation). To bridge this gap and explicitly model inter-expert relationships, a rigorous analytical framework is essential. In the broader context of neural network compression, quantifying such interdependencies via second-order (Hessian) information has a well-established theoretical foundation LeCun et al. (1989); Hassibi & Stork (1992). However, adapting Hessian-aware analysis to effectively quantify the complicated combinatorial dependencies within MoE architectures remains a critical, yet unexplored frontier.

C Mathematical Proofs and Theoretical Derivations

C.1 Derivation of the Retained Utility Function

Recalling the second-order Taylor expansion of the loss degradation around the fully activated state $\mathbf{m} = \mathbf{1}$ from Eqn. equation 5, we have:

$$\Delta\mathcal{L}(\mathbf{m}) \approx (\mathbf{m}-\mathbf{1})^\top \nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}} + \frac{1}{2}(\mathbf{m}-\mathbf{1})^\top \mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}(\mathbf{m}-\mathbf{1}). \tag{15}$$

Our optimization objective is to minimize this expected loss degradation, which is mathematically equivalent to maximizing the retained utility function $\mathcal{U}(\mathbf{m}) = -\Delta\mathcal{L}(\mathbf{m})$. To convert this continuous approximation into a tractable cardinality-constrained QBO problem, we expand the vector-matrix multiplications into their scalar summation forms:

$$\Delta\mathcal{L}(\mathbf{m}) = \sum_{i=1}^{N}(m_i - 1)[\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(m_i - 1)(m_j - 1)[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}. \tag{16}$$

By expanding the polynomial product $(m_i - 1)(m_j - 1) = m_i m_j - m_i - m_j + 1$ within the double summation of Eqn. equation 16, we can isolate the quadratic, linear, and constant components:

$$\Delta\mathcal{L}(\mathbf{m}) = \sum_{i=1}^{N}(m_i[\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i - [\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(m_i m_j - m_i - m_j + 1)[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}. \tag{17}$$

Given that the Hessian matrix $\mathbf{H}_{\mathbf{m}}$ in Eqn. equation 17 is strictly symmetric (i.e., $[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} = [\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ji}$ for all $i, j \in \{1, ..., N\}$), the summation over the rows is mathematically identical to the summation over the columns:

$$\begin{aligned}\Delta\mathcal{L}(\mathbf{m}) = &\sum_{i=1}^{N} m_i[\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i - \sum_{i=1}^{N}[\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_j[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}.\end{aligned} \tag{18}$$

Notably, the Hessian matrix $\mathbf{H}_{\mathbf{m}}$ is strictly symmetric, implying $[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} = [\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ji}$ for all $i, j \in \{1, \ldots, N\}$. Consequently, the summation over the rows is mathematically identical to the summation over the columns:

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N} m_j[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} &= \sum_{j=1}^{N} m_j\left(\sum_{i=1}^{N}[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}\right) = \sum_{j=1}^{N} m_j\left(\sum_{i=1}^{N}[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ji}\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} m_i[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}.\end{aligned} \tag{19}$$

Applying this symmetry property allows us to consolidate the linear terms originating from the quadratic expansion. By factoring out the binary variable m_i, we segregate the terms dependent on m from the data-independent constants:

Substituting the symmetry property from Eqn. equation 19 back into the expanded loss degradation in Eqn. equation 18 helps consolidate the linear terms originating from the quadratic expansion:

$$\begin{aligned}\Delta\mathcal{L}(\mathbf{m}) = &\sum_{i=1}^{N} m_i[\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i - \sum_{i=1}^{N}[\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} \\ &- \sum_{i=1}^{N}\sum_{j=1}^{N} m_i[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}.\end{aligned} \tag{20}$$

Then, we factor out the binary variable m_i from all linear terms in Eqn. equation 20 to segregate the terms dependent on $\mathbf{m}$ from the data-independent constants:

$$\begin{aligned}\Delta\mathcal{L}(\mathbf{m}) = &\sum_{i=1}^{N} m_i\left([\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i - \sum_{j=1}^{N}[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}\right) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} \\ &+ \underbrace{\left(\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\mathbf{H}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} - \sum_{i=1}^{N}[\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i\right)}_{\text{Constant } C}.\end{aligned} \tag{21}$$

We then substitute the consolidated form Eqn. equation 21 back into the retained utility function $\mathcal{U}(\mathbf{m}) = -\Delta\mathcal{L}(\mathbf{m})$, reversing the signs:

$$\mathcal{U}(\mathbf{m}) = \sum_{i=1}^{N} m_i \left(\sum_{j=1}^{N} [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij} - [\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i \right) - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} m_i m_j [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij} - C. \quad (22)$$

Since C is a constant independent of the decision vector $\mathbf{m}$, shifting the objective function by this scalar does not alter the optimal $\mathbf{m}^*$ of the maximization problem. Thus, it can be safely discarded.

Finally, we define the individual utility d_i to encapsulate standalone contribution of E_i toward minimizing the loss ($-[\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i$) and the aggregated pairwise coupling effects between E_i and all other experts ($\sum_{j=1}^{N} [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij}$):

$$d_i = \sum_{j=1}^{N} [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij} - [\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i. \quad (23)$$

Substituting Eqn. equation 23 into Eqn. equation 22, we obtain the exact analytical form for the retained utility presented in problem equation 8:

$$\mathcal{U}(\mathbf{m}) = \sum_{i=1}^{N} m_i d_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} m_i m_j [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij}. \quad (24)$$

C.2 Proof of NP-Hardness for Cardinality-Constrained QBO

We prove the NP-hardness of the considered problem by demonstrating a polynomial-time reduction from the decision version of the Maximum k-Clique problem Gao et al. (2022), a classical Karp's 21 NP-complete problem Hartmanis (1982).

Consider an arbitrary instance of the k-Clique problem defined on an undirected and unweighted graph $G = (V, E_G)$, where V is the set of vertices with $|V| = N$, and $E_G \subseteq V \times V$ is the set of edges. The structure of G can be algebraically represented by its symmetric adjacency matrix $\mathbf{A} \in \{0,1\}^{N \times N}$, defined as:

$$\mathbf{A}_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E_G; \\ 0, & \text{otherwise.} \end{cases} \quad (25)$$

Given a positive integer $k \leq N$, the decision problem asks whether there exists a fully connected subgraph (a clique) in G composed of exactly k vertices.

We construct a specific instance of our expert selection problem to mathematically mirror this graph G. Let the total number of candidate experts be strictly equal to the number of vertices, $N = |V|$, and set the cardinality constraint to k. We define the binary selection vector $\mathbf{m} \in \{0,1\}^N$, where the scalar $m_i = 1$ indicates that vertex i is selected into our target induced subgraph $V_{\mathbf{m}} \subseteq V$. This establishes the mapping:

$$V_{\mathbf{m}} = \{i \in V \mid m_i = 1\}, \quad \text{s.t.} \quad |V_{\mathbf{m}}| = \|\mathbf{m}\| \leq k. \quad (26)$$

To evaluate the retained utility function $\mathcal{U}(\mathbf{m})$ derived in problem equation 8, we decompose its quadratic penalty into diagonal and off-diagonal components:

$$\mathcal{U}(\mathbf{m}) = \sum_{i=1}^{N} m_i d_i - \left(\frac{1}{2} \sum_{i=1}^{N} m_i^2 [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ii} + \sum_{i<j} m_i m_j [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij} \right). \quad (27)$$

Since the decision variables are strictly binary ($m_i \in \{0,1\}$), the property $m_i^2 = m_i$ holds. Factoring in this allows us to consolidate all linear terms:

$$\mathcal{U}(\mathbf{m}) = \sum_{i=1}^{N} m_i \left(d_i - \frac{1}{2} [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ii} \right) - \sum_{i<j} m_i m_j [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij}. \quad (28)$$

To isolate the combinatorial pairwise interactions, we first construct our problem instance such that the consolidated linear coefficient vanishes, requiring $d_i = \frac{1}{2}[\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ii}$. Recalling the definition $d_i = \sum_{j=1}^{N}[\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij} - [\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i$ from Eqn. equation 23, this is achieved by specifying the initial first-order gradient vector as

$$[\nabla_{\mathbf{m}}\mathcal{L}|_{\mathbf{m}=\mathbf{1}}]_i = \sum_{j=1}^{N}[\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij} - \frac{1}{2}[\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ii} \quad \forall i \in \{1,\dots,N\}. \tag{29}$$

Simultaneously, we construct the off-diagonal entries of the Hessian matrix $\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}$ to encode the negative adjacency matrix of G:

$$[\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij} = -\mathbf{A}_{ij}, \quad \forall i,j \in \{1,\dots,N\}, i \neq j. \tag{30}$$

Constructing these matrices and gradient mappings requires $\mathcal{O}(N^2)$ operations, strictly bounding the computational transformation within polynomial time. By substituting our exact constructions into the consolidated utility equation, the linear terms sum exactly to zero. The objective reduces to the summation of off-diagonal entries, i.e., counting the edges within the selected subgraph:

$$\mathcal{U}(\mathbf{m}) = -\sum_{i<j} m_i m_j(-\mathbf{A}_{ij}) = \sum_{i<j} m_i m_j \mathbf{A}_{ij} = |E_G(V_{\mathbf{m}})|. \tag{31}$$

According to graph theory, the maximum number of edges in any simple graph bounded by $|V_{\mathbf{m}}| \leq k$ vertices is constrained by the binomial coefficient $\binom{k}{2}$. Because the objective function in Eqn. equation 31 strictly accumulates non-negative edge weights without incurring structural penalties, the maximum utility will exhaust the full cardinality budget k. Thus, the optimization objective is upper-bounded by:

$$\max_{\mathbf{m}:\|\mathbf{m}\|\leq k} \mathcal{U}(\mathbf{m}) = \max_{|V_{\mathbf{m}}|\leq k} |E_G(V_{\mathbf{m}})| \leq \binom{k}{2} = \frac{k(k-1)}{2}. \tag{32}$$

This theoretical upper bound is saturated if and only if the induced subgraph $V_{\mathbf{m}}$ is a complete graph (i.e., a k-clique). Consequently, the formulated QBO problem yields the optimal mathematical utility $U(\mathbf{m}^*) = \frac{k(k-1)}{2}$ if and only if the original graph G contains a k-clique. Given that the k-clique decision problem is classically NP-complete, any algorithmic problem capable of determining its solution must be at least as difficult. Because finding the exact solution to our cardinality-constrained QBO problem equation 8 is equivalent to solving the k-clique problem, we therefore conclude that problem equation 8 is NP-hard.

C.3 Proof of Supermodularity for Decomposed Functions

To establish the supermodularity of the decomposed functions $g(\mathbf{m})$ and $f(\mathbf{m})$, we analyze their discrete marginal gains directly on the Boolean hypercube. We first formalize the condition for supermodularity within this discrete space as follows:

Definition 1 (Supermodularity on Boolean Hypercube). A pseudo-Boolean function $h : \{0,1\}^N \to \mathbb{R}$ is supermodular if it exhibits non-decreasing marginal returns. Formally, let $\mathbf{m}^A, \mathbf{m}^B \in \{0,1\}^N$ be two binary expert selection vectors such that $\mathbf{m}^A \leq \mathbf{m}^B$ (i.e., $m_i^A \leq m_i^B$ for all $i \in \{1,\dots,N\}$, implying the set of active experts in $\mathbf{m}^A$ is a subset of that in $\mathbf{m}^B$). For any inactive expert k where $m_k^A = m_k^B = 0$, let $\mathbf{e}_k \in \{0,1\}^N$ denote the one-hot indicator vector for expert k. The supermodularity condition requires:

$$h(\mathbf{m}^A + \mathbf{e}_k) - h(\mathbf{m}^A) \leq h(\mathbf{m}^B + \mathbf{e}_k) - h(\mathbf{m}^B). \tag{33}$$

Following Definition 1 in equation 33, we evaluate the discrete marginal gains for both the gain function $f(\mathbf{m})$ and the cost function $g(\mathbf{m})$. In the subsequent analysis, we demonstrate that both functions are monotonically non-decreasing and supermodular.

Supermodularity of Cost Function $g(\mathbf{m})$. Recall the formulation of the cost function $g(\mathbf{m})$ from Eqn. equation 9:

$$g(\mathbf{m}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij}. \tag{34}$$

By definition of decomposition, the matrix $\mathbf{H}_{\mathbf{m}}^{+}$ isolates the non-negative components of the Hessian, guaranteeing that $[\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij} \geq 0$ for all i, j. Since the Hessian matrix $\mathbf{H}_{\mathbf{m}}$ consists of continuous second-order partial derivatives, it is symmetric according to Schwarz's theorem. The element-wise thresholding preserves this structural symmetry, yielding $[\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij} = [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ji}$.

To evaluate the marginal penalty of activating a new expert k (i.e., transitioning from the state $\mathbf{m}^A$ to $\mathbf{m}^A + \mathbf{e}_k$), we compute the discrete difference of the quadratic cost function in Eqn. equation 34. Expanding this formulation with respect to the increment $\mathbf{e}_k$ yields the diagonal entry representing the intrinsic self-redundancy of expert k and the off-diagonal summations that capture its cross-redundancy with the active set in $\mathbf{m}_A$:

$$g(\mathbf{m}^A + \mathbf{e}_k) - g(\mathbf{m}^A) = \frac{1}{2}\left(\sum_{j=1}^{N} m_j^A [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{kj} + \sum_{i=1}^{N} m_i^A [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ik} + [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{kk}\right). \tag{35}$$

By applying the symmetry property $[\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{kj} = [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{jk}$ into Eqn. equation 35, the two cross-term summations inside the parentheses are mathematically equivalent. Consolidating these terms yields a simplified expression for the marginal penalty:

$$g(\mathbf{m}^A + \mathbf{e}_k) - g(\mathbf{m}^A) = \frac{1}{2}[\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{kk} + \sum_{i=1, i\neq k}^{N} m_i^A [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ik}. \tag{36}$$

Following the exact same algebraic expansion to the superset selection vector $\mathbf{m}_B$ ($\mathbf{m}_A \leq \mathbf{m}_B$), we obtain its corresponding marginal penalty:

$$g(\mathbf{m}^B + \mathbf{e}_k) - g(\mathbf{m}^B) = \frac{1}{2}[\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{kk} + \sum_{i=1, i\neq k}^{N} m_i^B [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ik}. \tag{37}$$

To verify the supermodularity condition, we calculate the discrete difference between these two marginal gains of $\mathbf{m}_A$ and $\mathbf{m}_B$. Subtracting equation 36 from equation 37 eliminates the constant quadratic self-interaction term $\frac{1}{2}[\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{kk}$, leaving only the difference in the cross-redundancy summations:

$$(g(\mathbf{m}^B + \mathbf{e}_k) - g(\mathbf{m}^B)) - (g(\mathbf{m}^A + \mathbf{e}_k) - g(\mathbf{m}^A)) = \sum_{i=1, i\neq k}^{N} (m_i^B - m_i^A)[\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ik}. \tag{38}$$

Because $\mathbf{m}^A$ is a subset of $\mathbf{m}^B$ ($\mathbf{m}_A \leq \mathbf{m}_B$), the increment $(m_i^B - m_i^A)$ is non-negative for all experts i. Coupled with the inherent non-negativity of the decomposed Hessian entries ($[\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ik} \geq 0$), their product in equation 38 is strictly non-negative:

$$\sum_{i=1, i\neq k}^{N} (m_i^B - m_i^A)[\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ik} \geq 0. \tag{39}$$

This non-negativity guarantees that $g(\mathbf{m}^A + \mathbf{e}_k) - g(\mathbf{m}^A) \leq g(\mathbf{m}^B + \mathbf{e}_k) - g(\mathbf{m}^B)$, proving that the cost function $g(\mathbf{m})$ is supermodular: adding a new expert to a larger activated expert subset incurs a greater or equal redundancy penalty than adding it to a smaller subset.

Supermodularity of Gain Function $f(\mathbf{m})$. Next, we evaluate the gain function $f(\mathbf{m})$ formulated in Eqn. equation 9, which captures both the standalone utility of individual experts and functional synergy arising from complementary interactions:

$$f(\mathbf{m}) = \sum_{i=1}^{N} m_i d_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j [\mathbf{H}_{\mathbf{m}}^{-}|_{\mathbf{m}=\mathbf{1}}]_{ij}, \tag{40}$$

where the matrix $\mathbf{H}_{\mathbf{m}}^{-}$ isolates the negative components of the Hessian and inverts their signs, thereby maintaining strict non-negativity ($[\mathbf{H}_{\mathbf{m}}^{-}|_{\mathbf{m}=\mathbf{1}}]_{ij} \geq 0$) while preserving structural symmetry.

To calculate the marginal gain of activating expert k for the activated subset $\mathbf{m}_A$, we expand the function over the increment $\mathbf{e}_k$. The linear summation $\sum m_i d_i$ in equation 40 increases by the standalone capability d_k. Following the same algebraic expansion used for $g(\mathbf{m})$, the exact marginal gain evaluates to:

$$f(\mathbf{m}^A + \mathbf{e}_k) - f(\mathbf{m}^A) = d_k + \frac{1}{2}[\mathbf{H}^-_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{kk} + \sum_{i=1, i\neq k}^{N} m_i^A [\mathbf{H}^-_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ik}. \tag{41}$$

By an analogous expansion for the denser selection vector $\mathbf{m}^B$ ($\mathbf{m}^A \leq \mathbf{m}^B$), the corresponding marginal gain is given by:

$$f(\mathbf{m}^B + \mathbf{e}_k) - f(\mathbf{m}^B) = d_k + \frac{1}{2}[\mathbf{H}^-_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{kk} + \sum_{i=1, i\neq k}^{N} m_i^B [\mathbf{H}^-_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ik}. \tag{42}$$

To verify the supermodularity condition, we evaluate the discrete difference. Notably, the standalone capability d_k is a pure modular component; its marginal contribution is a fixed constant irrespective of the previously selected experts. Consequently, subtracting equation 41 from equation 42 perfectly cancels both d_k and the quadratic self-interaction term, isolating the difference in the cross-synergy summations:

$$(f(\mathbf{m}^B + \mathbf{e}_k) - f(\mathbf{m}^B)) - (f(\mathbf{m}^A + \mathbf{e}_k) - f(\mathbf{m}^A)) = \sum_{i=1, i\neq k}^{N} (m_i^B - m_i^A)[\mathbf{H}^-_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ik}. \tag{43}$$

Given $\mathbf{m}_A \leq \mathbf{m}_B$ and the non-negativity of the Hessian entries ($[H^-_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ik} \geq 0$) in equation 43, this residual summation is strictly non-negative:

$$\sum_{i=1, i\neq k}^{N} (m_i^B - m_i^A)[\mathbf{H}^-_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ik} \geq 0. \tag{44}$$

This confirms that the gain function $f(\mathbf{m})$ is strictly supermodular.

C.4 Derivation of the Cost Function Upper Bound

To derive a tight surrogate upper bound $U_g(\mathbf{m}; \mathbf{m}^{(t)})$ for the cost function $g(\mathbf{m})$, we begin with its original quadratic formulation from Eqn. equation 45:

$$g(\mathbf{m}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i m_j [\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}. \tag{45}$$

To decouple the binary cross-terms, we leverage the fundamental algebraic identity $m_i m_j = \frac{1}{2}(m_i^2 + m_j^2) - \frac{1}{2}(m_i - m_j)^2$. Given that the expert selection variables are binary ($m_i, m_j \in \{0, 1\}$), they satisfy the Boolean property $m_i^2 = m_i$. Substituting this into the objective facilitates reformulating the cost function into a linear component and a linear component $L(\mathbf{m})$ and a concave residual $C(\mathbf{m})$ via a difference-of-convex decomposition:

$$g(\mathbf{m}) = \underbrace{\frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{N}(m_i + m_j)[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}}_{\text{Linear Component } L(\mathbf{m})} + \underbrace{\left(-\frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{N}(m_i - m_j)^2[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}\right)}_{\text{Concave Component } C(\mathbf{m})}. \tag{46}$$

We can separate the summation in the linear component $L(\mathbf{m})$ in Eqn. equation 46:

$$L(\mathbf{m}) = \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{N} m_i [\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} + \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{N} m_j [\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}. \tag{47}$$

By swapping the dummy indices i and j in the second term, and leveraging the structural symmetry of the Hessian matrix $[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ji} = [\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}$, the linear component $L(\mathbf{m})$ simplifies to

$$L(\mathbf{m}) = \frac{1}{2}\sum_{i=1}^{N} m_i \left(\sum_{j=1}^{N} [\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} \right). \tag{48}$$

Since $[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} \geq 0$, the residual component $C(\mathbf{m})$ in Eqn. equation 46 is a strictly concave function. Taking the partial derivative with respect to m_k yields

$$\frac{\partial C(\mathbf{m})}{\partial m_k} = -\frac{1}{4}\left[\sum_{j=1}^{N} 2(m_k - m_j)[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{kj} + \sum_{i=1}^{N} 2(m_i - m_k)(-1)[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ik} \right] \tag{49}$$

$$= -\frac{1}{4}\left[\sum_{j=1}^{N} 2(m_k - m_j)[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{kj} + \sum_{i=1}^{N} 2(m_k - m_i)[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ki} \right]. \tag{50}$$

By applying the symmetry property $[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ik} = [\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ki}$, the two summations merge, yielding the general partial derivative with respect to m_k:

$$[\nabla_{\mathbf{m}} C(\mathbf{m}^{(t)})]_k = \frac{\partial C(\mathbf{m})}{\partial m_k} = -\sum_{j=1}^{N} [\mathbf{H}^+_{\mathbf{m}|\mathbf{m}=1}]_{kj}(m_k - m_j). \tag{51}$$

From the property of convex analysis, the strictly concave function $C(\mathbf{m})$ is globally upper-bounded by its first-order Taylor expansion at $\mathbf{m}^{(t)}$:

$$C(\mathbf{m}) \leq C(\mathbf{m}^{(t)}) + \sum_{i=1}^{N} [\nabla_{\mathbf{m}} C(\mathbf{m}^{(t)})]_i (m_i - m_i^{(t)}). \tag{52}$$

Substituting the exact gradient from Eqn. equation 51 into the Taylor expansion in Eqn. equation 52 and extracting the terms dependent on m_i, we obtain:

$$C(\mathbf{m}) \leq \sum_{i=1}^{N} m_i \left[-\sum_{j=1}^{N} [\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}(m_i^{(t)} - m_j^{(t)}) \right] + C_c^{(t)}, \tag{53}$$

where $C_c^{(t)} = C(\mathbf{m}^{(t)}) - \sum_{i=1}^{N} m_i^{(t)} [\nabla_{\mathbf{m}} C(\mathbf{m}^{(t)})]_i$ is a state-dependent constant.

Finally, by substituting the consolidated linear component $L(\mathbf{m})$ from Eqn. equation 48 and the bounded concave component from Eqn. equation 53 back into the original difference-of-convex decomposition $g(\mathbf{m}) = L(\mathbf{m}) + C(\mathbf{m})$, we factor out m_i and $[\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}$, and transform the state-dependent constant $C_c^{(t)}$ using structural symmetry to obtain the following tight upper bound:

$$g(\mathbf{m}) \leq \sum_{i=1}^{N} m_i \left[\sum_{j=1}^{N} [\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)} \right) \right] + C_g^{(t)} \triangleq U_g(\mathbf{m}; \mathbf{m}^{(t)}), \tag{54}$$

where $C_g^{(t)} = \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{N}(m_i^{(t)} - m_j^{(t)})^2 [\mathbf{H}^+_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij}$ is the state-dependent constant independent of the optimization variable $\mathbf{m}$.

C.5 Proof of Cardinality-Constrained Linear Maximization Problem

To establish the optimization procedure at the $(t+1)$-th iteration of the MM algorithm, our objective is to maximize the tractable modular surrogate function $\mathcal{Q}(\mathbf{m}; \mathbf{m}^{(t)})$, subject to the cardinality constraint $||\mathbf{m}|| \leq k$. The global surrogate function $\mathcal{Q}(\mathbf{m}; \mathbf{m}^{(t)})$ is constructed by subtracting the upper bound of the cost function $U_g(\mathbf{m}; \mathbf{m}^{(t)})$ from the lower bound of the gain function $L_f(\mathbf{m}; \mathbf{m}^{(t)})$:

$$\mathcal{Q}(\mathbf{m}; \mathbf{m}^{(t)}) = L_f(\mathbf{m}; \mathbf{m}^{(t)}) - U_g(\mathbf{m}; \mathbf{m}^{(t)}). \tag{55}$$

Recall the analytical expressions for these two bounds in Eqn. equation 10 and Eqn. equation 11. The first-order Taylor expansion provides the lower bound $L_f(\mathbf{m}; \mathbf{m}^{(t)})$ for the gain function $f(\mathbf{m})$ around the state $\mathbf{m}^{(t)}$:

$$L_f(\mathbf{m}; \mathbf{m}^{(t)}) = f(\mathbf{m}^{(t)}) + \sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i (m_i - m_i^{(t)}), \tag{56}$$

and the upper bound $U_g(\mathbf{m})$ for the cost function $g(\mathbf{m})$ is given by

$$U_g(\mathbf{m}; \mathbf{m}^{(t)}) = \sum_{i=1}^{N} m_i \left[\sum_{j=1}^{N} [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)} \right) \right] + C_g^{(t)}. \tag{57}$$

We substitute these exact expressions from Eqn. equation 56 and Eqn. equation 57 into the surrogate function $\mathcal{Q}(\mathbf{m}; \mathbf{m}^{(t)})$ in Eqn. equation 55, yielding

$$\begin{aligned} \mathcal{Q}(\mathbf{m}; \mathbf{m}^{(t)}) = & \left[f(\mathbf{m}^{(t)}) + \sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i (m_i - m_i^{(t)}) \right] \\ & - \left[\sum_{i=1}^{N} m_i \left(\sum_{j=1}^{N} [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)} \right) \right) + C_g^{(t)} \right]. \end{aligned} \tag{58}$$

After that, we expand the inner product associated with the gradient term in the lower bound $L_f(\mathbf{m}; \mathbf{m}^{(t)})$, and distribute the gradient $[\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i$ across the difference $m_i - m_i^{(t)}$:

$$\sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i (m_i - m_i^{(t)}) = \sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i m_i - \sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i m_i^{(t)}. \tag{59}$$

Substituting the expanded form from Eqn. equation 59 into our surrogate function in Eqn. equation 58 yields:

$$\begin{aligned} \mathcal{Q}(\mathbf{m}; \mathbf{m}^{(t)}) = & f(\mathbf{m}^{(t)}) + \sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i m_i - \sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i m_i^{(t)} \\ & - \sum_{i=1}^{N} m_i \left(\sum_{j=1}^{N} [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)} \right) \right) - C_g^{(t)}. \end{aligned} \tag{60}$$

Next, we regroup the terms in Eqn. equation 60 to segregate those dependent on the decision variable $\mathbf{m}$ from those that are strictly constant at iteration t:

$$\begin{aligned} \mathcal{Q}(\mathbf{m}; \mathbf{m}^{(t)}) = & \underbrace{\sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i m_i - \sum_{i=1}^{N} m_i \left(\sum_{j=1}^{N} [\mathbf{H}_{\mathbf{m}}^{+}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)} \right) \right)}_{\text{Terms dependent on } \mathbf{m}} \\ & + \underbrace{f(\mathbf{m}^{(t)}) - \sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i m_i^{(t)} - C_g^{(t)}}_{\text{Constant terms independent of } \mathbf{m}}. \end{aligned} \tag{61}$$

Let $C^{(t)}$ incorporate all terms from Eqn. equation 61 that are independent of the decision variable $\mathbf{m}$:

$$C^{(t)} = f(\mathbf{m}^{(t)}) - \sum_{i=1}^{N} [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i m_i^{(t)} - C_g^{(t)}. \tag{62}$$

Substituting the constant $C^{(t)}$ from Eqn. equation 62 into Eqn. equation 61 simplifies the formulation:

$$\mathcal{Q}(\mathbf{m};\mathbf{m}^{(t)}) = \sum_{i=1}^{N}[\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i m_i - \sum_{i=1}^{N} m_i \left(\sum_{j=1}^{N} [\mathbf{H}^{+}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)} \right) \right) + C^{(t)}. \tag{63}$$

Factoring out the binary variable m_i from the remaining summations in Eqn. equation 63 isolates the marginal contribution of each expert E_i:

$$\mathcal{Q}(\mathbf{m};\mathbf{m}^{(t)}) = \sum_{i=1}^{N} m_i \left([\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i - \sum_{j=1}^{N} [\mathbf{H}^{+}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)} \right) \right) + C^{(t)}. \tag{64}$$

We define the importance score $S_i^{(t)}$ for the i-th expert, which quantifies the trade-off between the expert's synergy gain and its redundancy penalty:

$$S_i^{(t)} = [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i - \sum_{j=1}^{N} [\mathbf{H}^{+}_{\mathbf{m}}|_{\mathbf{m}=\mathbf{1}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)} \right). \tag{65}$$

By substituting $S_i^{(t)}$ from Eqn. equation 65 into Eqn. equation 64, the complex surrogate function is reduced to an affine form:

$$\mathcal{Q}(\mathbf{m};\mathbf{m}^{(t)}) = \sum_{i=1}^{N} m_i S_i^{(t)} + C^{(t)}. \tag{66}$$

The iterative update step of the MM algorithm requires finding $\mathbf{m}^{(t+1)} = \arg\max_{\mathbf{m}} \mathcal{Q}(\mathbf{m};\mathbf{m}^{(t)})$, subject to the capacity constraint $||\mathbf{m}|| \leq k$. Since $C^{(t)}$ acts as a uniform scalar shift across the objective landscape, it does not alter the location of the maximum. Thus, maximizing $\mathcal{Q}(\mathbf{m};\mathbf{m}^{(t)})$ in Eqn. equation 66 is strictly equivalent to maximizing the linear summation alone:

$$\arg\max_{\mathbf{m}:||\mathbf{m}||\leq k} \left(\sum_{i=1}^{N} m_i S_i^{(t)} + C^{(t)} \right) = \arg\max_{\mathbf{m}:||\mathbf{m}||\leq k} \sum_{i=1}^{N} m_i S_i^{(t)}. \tag{67}$$

This rigorous deduction yields the exact formulation presented in Eqn. equation 68:

$$\mathbf{m}^{(t+1)} = \arg\max_{\mathbf{m}:||\mathbf{m}||\leq k} \sum_{i=1}^{N} m_i S_i^{(t)}. \tag{68}$$

This formulation demonstrates that the MM framework rigorously bounds the NP-hard quadratic optimization and projects it into a cardinality-constrained linear space, which can be solved via a Top-k sorting operation.

C.6 Proof of Convergence for the MM Algorithm

To prove the convergence property of the MM algorithm presented in Theorem 1, we demonstrate that the sequence of the retained utility values $\{\mathcal{U}(\mathbf{m}^{(t)})\}_{t\geq 0}$ generated by the iterative updates is monotonically non-decreasing. The MM algorithm operates by sequentially maximizing the surrogate function $\mathcal{Q}(\mathbf{m};\mathbf{m}^{(t)})$ constructed in Eqn. equation 12. To guarantee monotonic gain, this surrogate must satisfy two fundamental minorization conditions:

First, the surrogate must act as a global lower bound for the true objective function. Based on the definitions in Eqn. equation 10 and Eqn. equation 11, the first-order Taylor expansion lower-bounds the supermodular gain function ($L_f(\mathbf{m};\mathbf{m}^{(t)}) \leq f(\mathbf{m})$), and our arithmetic relaxation tightly upper-bounds the cost function ($U_g(\mathbf{m};\mathbf{m}^{(t)}) \geq g(\mathbf{m})$). Recalling the surrogate construction, subtracting these components preserves the global inequality:

$$\mathcal{Q}(\mathbf{m};\mathbf{m}^{(t)}) \leq \mathcal{U}(\mathbf{m}). \tag{69}$$

Algorithm 1: MM-based Expert Selection

Input: Hessian matrix $\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}} \in \mathbb{R}^{N\times N}$, utility vector $\mathbf{d} \in \mathbb{R}^N$, sparsity k, max iterations $T_{\max}$
Output: Optimal selection vector $\mathbf{m}^* \in \{0,1\}^N$ s.t. $\|\mathbf{m}^*\| = k$

// Step 1: Initialization
$t \leftarrow 0, \quad \mathbf{m}^{(0)} \leftarrow \mathbf{0}$; ▷ Start with empty selection
// Step 2: Hessian Decomposition
$[\mathbf{H}^+_{\mathbf{m}}]_{ij} \leftarrow \max(0, [\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij})$; ▷ Redundancy component
$[\mathbf{H}^-_{\mathbf{m}}]_{ij} \leftarrow \max(0, -[\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}]_{ij})$; ▷ Synergy component
// Step 3: MM Optimization
while $t < T_{\max}$ **do**
  ;// Construct Surrogate Function
  for $i = 1$ **to** N **do**
    $[\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i \leftarrow d_i + \sum_{j=1}^N [\mathbf{H}^-_{\mathbf{m}}]_{ij} m_j^{(t)}$; ▷ Marginal synergy gain
    $p_i^{(t)} \leftarrow \sum_{j=1}^N [\mathbf{H}^+_{\mathbf{m}}]_{ij} \left(\frac{1}{2} - m_i^{(t)} + m_j^{(t)}\right)$; ▷ Context-aware redundancy penalty
    $S_i^{(t)} \leftarrow [\nabla_{\mathbf{m}} f(\mathbf{m}^{(t)})]_i - p_i^{(t)}$; ▷ Importance score
  ;// Update Selection via Top-k Projection
  $\mathbf{m}^{(t+1)} \leftarrow$ Top-$k(\mathbf{S}^{(t)}, k)$;
  if $\mathbf{m}^{(t+1)} = \mathbf{m}^{(t)}$ **then**
    break ; ▷ Selection converged
  $t \leftarrow t + 1$;
return $\mathbf{m}^* \leftarrow \mathbf{m}^{(t)}$;

Second, the surrogate must perfectly coincide with the true objective at the current anchor state $\mathbf{m} = \mathbf{m}^{(t)}$. In the bounding formulations for both $L_f(\mathbf{m}; \mathbf{m}^{(t)})$ and $U_g(\mathbf{m}; \mathbf{m}^{(t)})$, the approximation deviations are driven by the first-order gradient terms $(m_i - m_i^{(t)})$. When evaluated exactly at the anchor state, these difference terms intrinsically evaluate to zero, neutralizing all gradient penalties. Consequently, both bounds revert to their original function: $L_f(\mathbf{m}^{(t)}; \mathbf{m}^{(t)}) = f(\mathbf{m}^{(t)})$ and $U_g(\mathbf{m}^{(t)}; \mathbf{m}^{(t)}) = g(\mathbf{m}^{(t)})$. Subtracting these yields the tightness condition:

$$\mathcal{Q}(\mathbf{m}^{(t)}; \mathbf{m}^{(t)}) = \mathcal{U}(\mathbf{m}^{(t)}). \tag{70}$$

During the update phase defined by the cardinality-constrained linear maximization in Eqn. equation 68, the algorithm unconditionally selects the new state $\mathbf{m}^{(t+1)}$ as the exact global maximizer of the surrogate function over the constrained feasible region. By the mathematical definition of a maximizer, the surrogate value cannot degrade:

$$\mathcal{Q}(\mathbf{m}^{(t+1)}; \mathbf{m}^{(t)}) \geq \mathcal{Q}(\mathbf{m}^{(t)}; \mathbf{m}^{(t)}). \tag{71}$$

By chaining these mathematical properties together, we form the standard MM induction step:

$$\mathcal{U}(\mathbf{m}^{(t+1)}) \geq \mathcal{Q}(\mathbf{m}^{(t+1)}; \mathbf{m}^{(t)}) \geq \mathcal{Q}(\mathbf{m}^{(t)}; \mathbf{m}^{(t)}) = \mathcal{U}(\mathbf{m}^{(t)}). \tag{72}$$

Extracting the outermost bounds strictly guarantees that the sequence of the retained utilities $\{\mathcal{U}(\mathbf{m}^{(t)})\}_{t=0}^{\infty}$ is monotonically non-decreasing, which aligns with the theoretical guarantee presented in Eqn. equation 14:

$$\mathcal{U}(\mathbf{m}^{(t+1)}) \geq \mathcal{U}(\mathbf{m}^{(t)}), \quad \forall t \geq 0 \tag{73}$$

This establishes the formal proof of convergence, ensuring that our iterative expert selection process consistently improves or perfectly maintains the optimal mathematical utility.

D Algorithmic Implementation and Experimental Details

D.1 The Complete Majorization-Minimization Algorithm

Algorithm 1 outlines the complete procedure of the proposed MM optimization for dynamic expert selection. The algorithm initializes the iteration counter $t = 0$ and sets the starting selection vector $\mathbf{m}^{(0)}$ to an empty state $\mathbf{0}$ (`Lines 2`). Next, the Hessian matrix $\mathbf{H_m}|_{\mathbf{m}=\mathbf{1}}$ is decomposed into non-negative functional redundancy component $\mathbf{H}_\mathbf{m}^+$ and complementary synergy component $\mathbf{H}_\mathbf{m}^-$ (`Lines 4-5`). Inside the MM optimization loop, the algorithm dynamically calculates the marginal synergy gain $\mathbf{g}^{(t)}$ for each candidate expert relative to the active subset $\mathbf{m}^{(t)}$ selected in the preceding iteration (`Lines 9`). Simultaneously, it computes a context-aware redundancy penalty $p_i^{(t)}$ based on the current selection state to evaluate functional overlap (`Lines 10`). By subtracting the redundancy penalty from the marginal synergy gain, the algorithm formulates the composite importance score $S_i^{(t)}$ (`Lines 11`). The updated expert selection vector $\mathbf{m}^{(t+1)}$ is then determined via a Top-k projection of these evaluated scores (`Lines 12`). Finally, the algorithm enforces a termination condition: if the newly selected subset precisely matches the previous one ($\mathbf{m}^{(t+1)} = \mathbf{m}^{(t)}$), the optimization has reached a stable state, prompting the loop to break and return the optimal expert selection vector $\mathbf{m}^*$ (`Lines 13-16`).

D.2 Datasets and Evaluation Metrics

Datasets. To evaluate the proposed DS-MoE framework against the state-of-the-art routing baselines, we conduct comprehensive empirical evaluations across four widely adopted datasets spanning a spectrum from fundamental semantic comprehension to advanced, knowledge-intensive reasoning. The selected benchmarks evaluate the MoE architectures across varying degrees of complexity and domain heterogeneity:

- **AGNews (Natural Language Understanding)** Zhang et al. (2015): A fundamental benchmark for text classification comprising news articles categorized into four topical domains (World, Sports, Business, Sci/Tech). We utilize this dataset to evaluate the model's capacity to capture linearly separable semantic representations, verifying whether the routing mechanism can distinctly allocate experts to non-overlapping topics without representation collapse.
- **PIQA (Physical Interaction Question Answering)** Bisk et al. (2020): A physical commonsense reasoning benchmark requiring the deduction of the most plausible continuation of a scenario governed by real-world physical laws. This task challenges the MoE framework to dynamically activate expert subsets specialized in intuitive physics and spatial-temporal logic, extending significantly beyond static factoid retrieval.
- **HellaSwag (Predictive Reasoning)** Zellers et al. (2019): A challenging dataset for evaluating grounded, adversarial natural language reasoning. Because the incorrect options in HellaSwag are adversarially generated by machines to be structurally coherent but logically flawed, this benchmark explicitly stresses the MoE's ability to disentangle spurious correlations, explicitly highlighting the necessity of synergistic expert combinations (error cancellation) to effectively resolve complex logical reasoning.
- **MMLU (Massive Multitask Language Understanding)** Hendrycks et al. (2021): A comprehensive benchmark designed to evaluate zero-shot and few-shot knowledge-intensive reasoning across 57 distinct academic and professional disciplines (e.g., STEM, humanities, social sciences). Given its extreme domain heterogeneity, MMLU is critical for stress-testing the MoE architecture. It validates the MoE framework's capacity to retain a diverse expert pool and orchestrate combinatorial expert collaboration, while effectively mitigating catastrophic forgetting and negative transfer.

Evaluation Metrics. To capture both the final generalization capability and the underlying optimization dynamics, we assess the models using two primary quantitative metrics:

- **Test Accuracy (%):** The standard Top-1 accuracy measured on the unseen test splits across all evaluated tasks. As the primary proxy for generalization efficacy, higher test accuracy

provides direct empirical validation that the proposed expert selection successfully preserves indispensable, synergistic expert representations while discarding redundant ones.

- **Convergence Efficiency (Rounds):** Measured by the number of training iterations required for the validation performance to reach a stationary accuracy threshold. Standard Top-k routing often suffers from slow convergence or oscillatory behavior due to conflicting gradient updates from functionally redundant experts. By tracking the optimization trajectory, this metric quantitatively demonstrates how the proposed framework's mitigation of redundancy translates to accelerated gradient flow and superior training stability.

D.3 Model Architectures and Hyperparameter Settings

To conduct the evaluation of the DS-MoE framework, we benchmark our approach against three representative MoE architectures: Switch-Base-32, Switch-Base-64, and DeepSeek-MoE-16B. The detailed model architecture and experimental configurations are outlined below.

Model Architectures

- **Switch-Base Variants ($N \in \{32, 64\}$):** Both Switch-Base-32 and Switch-Base-64 models adopt a symmetric sequence-to-sequence architecture, comprising 12 encoder layers and 12 decoder layers. We uniformly set the hidden dimension to $d_{\text{model}} = 768$ and the vocabulary size to 32,128. These variants employ an alternating arrangement of dense and sparse layers: specifically, half of the Transformer blocks utilize standard dense feed-forward networks (FFN) with an expansion dimension of 3,072, while the remaining blocks are replaced by sparse MoE modules integrating 32 (resp. Switch-Base-32) and 64 (resp. Switch-Base-64) experts. To ensure computational equivalence for a single forward pass, the internal expansion dimension of all individual experts is strictly locked at 3,072, ensuring parity with the standard dense FFN layers.

- **DeepSeek-MoE-16B ($N = 64$):** We instantiate the DeepSeek-MoE-16B as a 28-layer decoder-only autoregressive architecture. We set the hidden dimension $d_{\text{model}} = 2,048$ and the vocabulary size to 102,400. To stabilize shallow-layer feature representations, the foundational layer incorporates a standard dense FFN with an intermediate dimension of 10,944. The subsequent 27 layers are configured as MoE modules. Within each MoE layer, we deploy a hybrid expert configuration consisting of 64 independent routed experts with an individual dimension of 1,408.

Experimental Configurations

- **Optimization and Parameter Updating Strategy:** All experiments are executed using bfloat16 mixed precision and the 8-bit AdamW optimizer. The learning rate is uniformly set to 1×10^{-4} across all tasks. We configure a micro-batch size of 4 coupled with 4 gradient accumulation steps, yielding a batch size of 16. During fine-tuning, only the sparse components (expert networks and routers) and task-specific heads are updated, while the dense backbone remains frozen.

- **Task-Specific Setup:** We adapt the maximum sequence length based on the inherent context requirements of each specific benchmark. Specifically, the lengths are configured to 128 tokens for AGNews; 256 tokens for PIQA; and 512 tokens for HellaSwag and MMLU. During the evaluation of generative multiple-choice tasks (such as HellaSwag and MMLU), the generation configuration is strictly bounded to a maximum of 4 new tokens.

Specific Configurations for MM Algorithm. The proposed MM expert selection framework introduces several algorithm-specific hyperparameters. We set the maximum iteration threshold of $T_{\max} = 10$ and active expert capacity $k = 2$ by default unless specified otherwise.

Hardware. All experiments are conducted on a high-performance central server equipped with NVIDIA GeForce RTX 4090 GPUs.

E LIMITATIONS AND FUTURE WORK

While the DS-MoE framework establishes a rigorous mathematical foundation for optimizing expert combinations, our current analysis confines the modeling of combinatorial dependencies to intra-layer interactions. Given the sequential feature transformations inherent in deep language models, capturing inter-layer dynamics presents a critical avenue for exploration. A promising direction for future research involves investigating cross-layer expert synergies, such as approximating a global, block-diagonal Hessian proxy, to enable holistic, network-wide routing strategies that further maximize overall parameter efficiency.

F PRACTICAL IMPACT

By bridging rigorous theoretical guarantees with expert selection optimizations, the DS-MoE framework addresses critical bottlenecks in the training and deployment of large-scale MoE architectures:

- **Theoretically Grounded Routing Paradigm:** The training of industrial-scale MoE models is impeded by routing collapse and optimization instabilities, necessitating extensive hyperparameter tuning. By introducing an MM framework with guaranteed monotonic convergence, DS-MoE circumvents the heuristic unpredictability inherent in standard Top-k routing, thereby enhancing the training performance.
- **Facilitating Deployment in Resource-Constrained Edge Systems:** By optimizing representational capacity under strict cardinality constraints (e.g., a predefined expert budget k), DS-MoE drastically reduces the memory footprint and computing requirements for training and deploying LLMs. This enables resource-constrained computing entities to leverage the state-of-the-art MoE architectures without compromising performance.
- **Sustainable and Cost-Effective Training:** As demonstrated in our empirical results, decoupling functional redundancy from expert synergy resolves gradient conflicts and substantially expedites model convergence. This algorithmic efficiency translates directly to significant reductions in computing overhead and overall energy consumption, rendering the training of large-scale MoE models highly resource-efficient and economically viable.