

A MODEL ARCHITECTURE DETAILS

This appendix provides detailed mathematical formulations and implementation specifications for the HSMoE-FSA framework components. Figure A1 illustrates the complete architecture with all technical components.

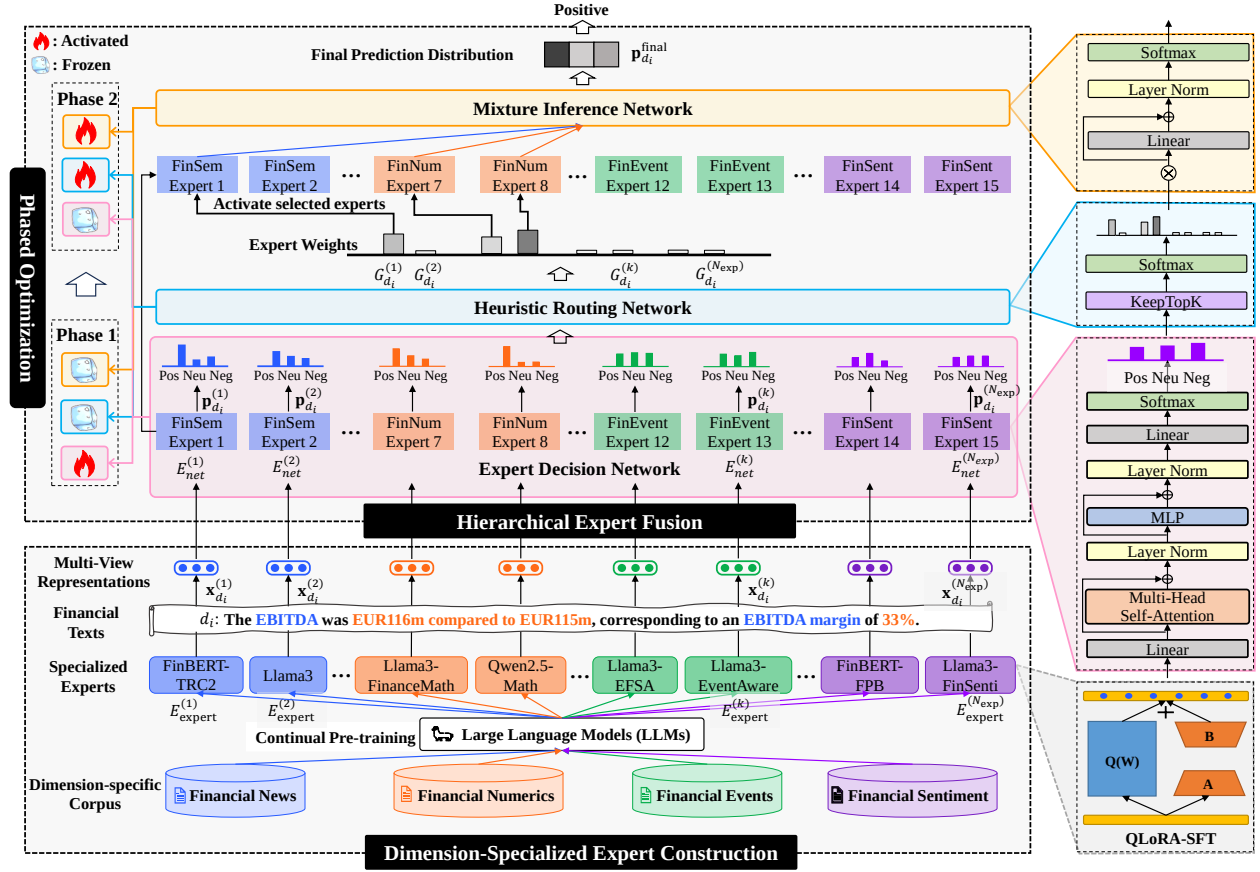


Figure A1. Detailed architecture of the HSMoE-FSA framework.

A.1 Multi-View Expert Representation Generation

The multi-view representation extraction process begins with QLoRA-based supervised fine-tuning of the expert repository. Given the foundational expert model set $\mathcal{E}_{expert} = \{E_{expert}^{(1)}, E_{expert}^{(2)}, \dots, E_{expert}^{(k)}, \dots, E_{expert}^{(N_{exp})}\}$ with N_{exp} experts, each model undergoes optimization:

$$E'_{expert}{}^{(k)} = \text{QLoRA_SFT}(E_{expert}^{(k)}, \mathcal{D}_{train}) \quad (\text{A1})$$

where QLoRA employs 4-bit quantization and trainable low-rank matrices. The last-layer hidden states from each fine-tuned expert yield:

$$\mathbf{x}_{d_i}^{(k)} = E'_{expert}{}^{(k)}(d_i) \in \mathbb{R}^{N_h \times 1} \quad (\text{A2})$$

where $\mathbf{x}_{d_i}^{(k)}$ represents the k -th expert's embedding for text d_i , and N_h is the hidden dimension. The composite multi-view representation integrates all expert perspectives:

$$\mathbf{X}_{d_i} = \left[\mathbf{x}_{d_i}^{(1)}, \mathbf{x}_{d_i}^{(2)}, \dots, \mathbf{x}_{d_i}^{(k)}, \dots, \mathbf{x}_{d_i}^{(N_{\text{exp}})} \right]^\top \in \mathbb{R}^{N_{\text{exp}} \times N_h} \quad (\text{A3})$$

A.2 Expert Decision Network Architecture

Each expert decision sub-network $E_{\text{net}}^{(k)}$ processes dimension-specific representations through five sequential stages:

Domain Projection:

$$\mathbf{x}'_{d_i}{}^{(k)} = \mathbf{W}_{\text{proj}}^{(k)} \mathbf{x}_{d_i}^{(k)} + \mathbf{b}_{\text{proj}}^{(k)} \quad (\text{A4})$$

Multi-Head Self-Attention:

$$\mathbf{x}''_{d_i}{}^{(k)} = \mathbf{W}_O \cdot \text{Concat} \left(\text{Softmax} \left(\frac{\mathbf{W}_Q^{(h)} \mathbf{x}'_{d_i}{}^{(k)} (\mathbf{W}_K^{(h)} \mathbf{x}'_{d_i}{}^{(k)})^\top}{\sqrt{d_K}} \right) \mathbf{W}_V^{(h)} \mathbf{x}'_{d_i}{}^{(k)} \right)_{h=1}^{N_{\text{head}}} \quad (\text{A5})$$

Feature Fusion with Residual Connections:

$$\mathbf{x}'''_{d_i}{}^{(k)} = \text{LayerNorm}(\mathbf{x}'_{d_i}{}^{(k)} + \mathbf{x}''_{d_i}{}^{(k)}) \quad (\text{A6})$$

$$\mathbf{x}^{iv}_{d_i}{}^{(k)} = \mathbf{W}_{m2}^{(k)} \cdot \text{GELU}(\mathbf{W}_{m1}^{(k)} \mathbf{x}'''_{d_i}{}^{(k)} + \mathbf{b}_{m1}^{(k)}) + \mathbf{b}_{m2}^{(k)} \quad (\text{A7})$$

Normalized Output:

$$\mathbf{x}^v_{d_i}{}^{(k)} = \text{LayerNorm}(\mathbf{x}'''_{d_i}{}^{(k)} + \mathbf{x}^{iv}_{d_i}{}^{(k)}) \quad (\text{A8})$$

Classification Layer:

$$\mathbf{p}_{d_i}^{(k)} = \text{Softmax}(\mathbf{W}_{\text{cls}}^{(k)} \mathbf{x}^v_{d_i}{}^{(k)} + \mathbf{b}_{\text{cls}}^{(k)}) \quad (\text{A9})$$

where $\mathbf{W}_{\text{cls}}^{(k)} \in \mathbb{R}^{N_{\text{cls}} \times N_{\text{proj}}}$ maps features to N_{cls} sentiment categories. The expert ensemble output aggregates all predictions:

$$\mathbf{P}_{d_i}^{\text{experts}} = \left[\mathbf{p}_{d_i}^{(1)}, \mathbf{p}_{d_i}^{(2)}, \dots, \mathbf{p}_{d_i}^{(k)}, \dots, \mathbf{p}_{d_i}^{(N_{\text{exp}})} \right]^\top \in \mathbb{R}^{N_{\text{exp}} \times N_{\text{cls}}} \quad (\text{A10})$$

A.3 Entropy-Based Heuristic Routing Mechanism

The routing mechanism operates through three phases:

Expert Confidence Assessment. For each expert's sentiment distribution $\mathbf{p}_{d_i}^{(k)}$, compute information entropy:

$$H_{d_i}^{(k)} = - \sum_{c=1}^{N_{\text{cls}}} \mathbf{p}_{d_i,c}^{(k)} \log \mathbf{p}_{d_i,c}^{(k)} \quad (\text{A11})$$

Lower entropy indicates higher confidence. The reliability score is computed as:

$$W_{d_i}^{(k)} = \frac{\exp(-H_{d_i}^{(k)})}{\sum_{j=1}^{N_{\text{exp}}} \exp(-H_{d_i}^{(j)})} \quad (\text{A12})$$

Top-K Sparse Selection. Select the top- K most confident experts:

$$S_{d_i}^{(k)} = \begin{cases} W_{d_i}^{(k)}, & \text{if } W_{d_i}^{(k)} \in \text{TopK}(\{W_{d_i}^{(1)}, W_{d_i}^{(2)}, \dots, W_{d_i}^{(N_{\text{exp}})}\}) \\ -\infty, & \text{otherwise} \end{cases} \quad (\text{A13})$$

Expert Weight Allocation. Normalize sparse scores to obtain final gate weights:

$$\mathbf{G}_{d_i}^{\text{heur}} = [G_{d_i}^{(1)}, G_{d_i}^{(2)}, \dots, G_{d_i}^{(k)}, \dots, G_{d_i}^{(N_{\text{exp}})}]^\top = \text{Softmax}([S_{d_i}^{(1)}, S_{d_i}^{(2)}, \dots, S_{d_i}^{(k)}, \dots, S_{d_i}^{(N_{\text{exp}})}]^\top) \quad (\text{A14})$$

Parameter-Based Routing Baseline. For comparison, the trainable routing alternative employs:

$$\mathbf{G}_{d_i}^{\text{param}} = \text{Softmax}(\text{TopK}(\mathbf{P}_{d_i}^{\text{experts}} \mathbf{W}_g)) \quad (\text{A15})$$

where $\mathbf{W}_g \in \mathbb{R}^{N_{\text{cls}} \times 1}$ is a learnable parameter matrix.

A.4 Mixture Inference Network Formulation

The final prediction synthesizes expert decisions through three stages:

Expert Decision Fusion:

$$\mathbf{p}_{d_i}^{\text{weighted}} = (\mathbf{P}_{d_i}^{\text{experts}})^\top \cdot \mathbf{G}_{d_i}^{\text{heur}} \quad (\text{A16})$$

Semantic Space Projection:

$$\mathbf{z}_{d_i} = \mathbf{W}_{\text{mixer}} \mathbf{p}_{d_i}^{\text{weighted}} + \mathbf{b}_{\text{mixer}} \quad (\text{A17})$$

Residual Enhancement and Final Prediction:

$$\mathbf{p}_{d_i}^{\text{final}} = \text{Softmax} \left(\text{LayerNorm} \left(\mathbf{p}_{d_i}^{\text{weighted}} + \mathbf{z}_{d_i} \right) \right) \quad (\text{A18})$$

where $\mathbf{p}_{d_i}^{\text{final}}$ represents the calibrated sentiment probability distribution preserving multi-dimensional financial semantics.

B COMPUTATIONAL COMPLEXITY ANALYSIS

The computational complexity of the HSMoE-FSA framework is predominantly concentrated in the training phase, which is performed only once. During training, we employ QLoRA technology to fine-tune N_{exp} dimension-specialized experts, resulting in an overall training complexity of approximately $O(N_{\text{exp}} \times T_{\text{train}} \times L \times d^2)$, where T_{train} denotes the number of training iterations, L represents the sequence length, and d is the hidden dimension of the base model. The QLoRA approach reduces memory overhead by introducing low-rank adapters with trainable parameters of $O(N_{\text{exp}} \times 2 \times d \times r)$, where $r \ll d$ is the low-rank dimension, achieving approximately 75% memory reduction compared to full fine-tuning. Once this one-time training phase is completed, all expert parameters remain frozen for subsequent inference tasks.

In contrast, the inference phase incurs substantially lower computational cost as it requires no parameter updates. The primary computation involves the entropy-based routing mechanism, which evaluates expert confidence and performs Top-K selection with complexity $O(N_{\text{exp}} \times N_{\text{cls}})$, where N_{cls} is the number of sentiment classes. The subsequent weighted fusion of the selected K experts introduces only $O(K \times N_{\text{cls}})$ additional operations. These routing computations are orders of magnitude smaller than the training phase complexity, making the inference-time cost comparable to single-expert models while maintaining multi-dimensional analytical capabilities.

To further illustrate the performance-to-complexity trade-off of our proposed HSMoE-FSA framework, we present in Table A1 a detailed comparison of Weighted F1-scores across different expert configurations. The results clearly demonstrate that increasing the number of experts yields consistent performance gains across all three benchmark datasets. Combined with the fact that the additional computational cost arises only during the one-time training phase, these findings support the practicality and effectiveness of incorporating multiple specialized experts within our hierarchical mixture-of-experts design.

Table A1. Impact of expert count on Weighted F1-score across benchmark datasets. Single-expert models (1 expert) represent baseline LoRA-adapted configurations, while ablation configurations (9–13 experts) demonstrate the contribution of different expert modules. The full 15-expert HSMoE-FSA achieves the best performance, validating that increased expert collaboration justifies the added training complexity.

Dataset	Expert Count	Model	Weighted Avg F1 (%)
FPB	1 expert	Llama3-FinanceMath	86.70
	1 expert	Llama3-EventAware	86.89
	1 expert	Llama3-EFSA	87.08
	1 expert	Llama3-FinSenti	88.22
	9 experts	HSMoE-FSA w/o FinSem	89.33
	10 experts	HSMoE-FSA w/o FinNum	89.16
	13 experts	HSMoE-FSA w/o FinEvent	89.50
	13 experts	HSMoE-FSA w/o FinSent	89.42
	15 experts	HSMoE-FSA	89.67
FiQA-SA	1 expert	Llama3-FinanceMath	88.99
	1 expert	Llama3-EventAware	88.35
	1 expert	Llama3-EFSA	88.43
	1 expert	Llama3-FinSenti	89.53
	9 experts	HSMoE-FSA w/o FinSem	92.07
	10 experts	HSMoE-FSA w/o FinNum	92.00
	13 experts	HSMoE-FSA w/o FinEvent	92.02
	13 experts	HSMoE-FSA w/o FinSent	91.93
	15 experts	HSMoE-FSA	92.29
TFNS	1 expert	Llama3-FinanceMath	90.25
	1 expert	Llama3-EventAware	89.35
	1 expert	Llama3-EFSA	90.06
	1 expert	Llama3-FinSenti	91.32
	9 experts	HSMoE-FSA w/o FinSem	91.92
	10 experts	HSMoE-FSA w/o FinNum	92.14
	13 experts	HSMoE-FSA w/o FinEvent	92.10
	13 experts	HSMoE-FSA w/o FinSent	91.94
	15 experts	HSMoE-FSA	92.10