Filtered not Mixed: Filtering-Based Online Gating for Mixture of Large Language Models

Raeid Saqur, Anastasis Kratsios, Florian Krach, Yannick Limmer, Jacob-Junqi Tian, John Willes, Blanka Horvath, Frank Rudzicz

ICLR
Singapore
April 24-28, 2025

Introduction

- MoE models are state-of-the-art in LLMs for *static tasks* (Mixtral, Gemini, DBRX, ...)
- don't leverage the *temporal structure* of time-series data
- i.e., at every step we get feedback about the performance of each expert
- instantly feeding this back to the gating mechanism we derive an online adaptive MoE model, the *Mixture-of-Experts Filter (MoE-F)*

Problem formulation

- N pre-trained expert models $F = (f^{(1)}, \dots, f^{(N)})$
- we assume that at any time t one of them is the true best expert specified by $w_t \in \{0, 1\}^N$ with $|w_t| = 1$
- w is the unobserved signal process, assumed to be a hidden Markov process

Problem formulation

- N pre-trained expert models $F = (f^{(1)}, \ldots, f^{(N)})$
- we assume that at any time t one of them is the true best expert specified by $w_t \in \{0,1\}^N$ with $|w_t| = 1$
- w is the unobserved signal process, assumed to be a hidden Markov process
- we observe the target process Y (the time series to be predicted), assumed to be

$$Y_t = Y_0 + \underbrace{\int_0^t w_s^\top F(x_{[0,s]})\, ds}_{\text{Best Expert Estimate}} + \underbrace{\int_0^t dW_s}_{\text{Idiosyncratic Residual Noise}} \tag{1}$$

- problem of selecting the best expert is to filter w from observations of Y, F
- this is a continuous-time finite state-space stochastic filtering problem (Wonham, 1964)
- we leverage the Wonham-Shiryaev filter, which is a closed-form recursive solution

MoE Filtering Algorithm

- 2 step approach applied at any time step t, using MSE or BCE loss function ℓ

MoE Filtering Algorithm

- 2 step approach applied at any time step t, using MSE or BCE loss function ℓ

Step 1: Optimal Parallel Filtering

- for each expert $f^{(n)}$ compute its running performance $\ell_t^{(n)} = \ell(Y_t, f^{(n)}(x_{[0:t-1]}))$
- and consider the filtering problem of optimally estimating w_t given its loss $\ell_t^{(n)}$
- solve these N stochastic filtering problems (in parallel) with the Wonham-Shiryaev filter
- this yields the estimates $\pi_t^{(n)} = \big(\mathbb{P}\big(w_t = e_i \mid \mathcal{F}_t^{(n)}\big)\big)_{i=1}^N$, where $\mathcal{F}_t^{(n)} = \sigma\{\ell_s^{(n)}\}_{0 \le s \le t}$

MoE Filtering Algorithm

- 2 step approach applied at any time step t, using MSE or BCE loss function ℓ

Step 1: Optimal Parallel Filtering

- for each expert $f^{(n)}$ compute its running performance $\ell_t^{(n)} = \ell(Y_t, f^{(n)}(x_{[0:t-1]}))$
- and consider the filtering problem of optimally estimating w_t given its loss $\ell_t^{(n)}$
- solve these N stochastic filtering problems (in parallel) with the Wonham-Shiryaev filter
- this yields the estimates $\pi_t^{(n)} = \left(\mathbb{P}\left(w_t = e_i \mid \mathcal{F}_t^{(n)}\right)\right)_{i=1}^N$, where $\mathcal{F}_t^{(n)} = \sigma\{\ell_s^{(n)}\}_{0\leq s\leq t}$

Step 2: Robust Aggregation

- each filter implies an individual (a posteriori) prediction $\hat{Y}_t^{(n)} = (\pi_t^{(n)})^\top F(x_{[0:t-1]})$
- loss scores $s_n = \ell\left(Y_t, \hat{Y}_t^{(n)}\right)$ used to define aggregation weights $\bar{\pi}_t^n = \frac{e^{-\lambda s_n}}{\sum_{i=1}^N e^{-\lambda s_i}}$
- they aggregate experts into single (a posteriori) prediction $\hat{Y}_t = \sum_{n=1}^N \bar{\pi}_t^n \hat{Y}_t^{(n)}$
- aggregated estimate of signal w is $\hat{\pi}_t = \sum_{n=1}^N \bar{\pi}_t^n \pi_t^{(n)}$, used to mix experts at $t+1$

Theoretical Guarantees

Theorem 1 (informal)

The individual estimates $\pi_t^{(n)}$ of Step 1 are optimal filters (in L^2 sense) of the signal w.

Theorem 2 (informal)

The aggregation weights $\bar{\pi}_t$ optimally aggregate the loss scores under entropic regularization.

Experiment: Financial market movement (FMM)

Experiment: Financial market movement (FMM)

Dataset:

- test split of US equity market movement dataset (Saqur et al., 2024)
- features at t: market's current contextual information, i.e., relevant financial news headlines, market's financial numerics (OHLCV & technical indicators of past few days)
- labels at t: movement of \$SPY at $t + 1$ in $\{$*'Fall'*, *'Neutral'*, *'Rise'* $\}$

Experiment: Financial market movement (FMM)

Dataset:

- test split of US equity market movement dataset (Saqur et al., 2024)
- features at t: market's current contextual information, i.e., relevant financial news headlines, market's financial numerics (OHLCV & technical indicators of past few days)
- labels at t: movement of $SPY at $t + 1$ in {*'Fall'*, *'Neutral'*, *'Rise'* }

Experts:

- Llama-2, Llama-3, Mixtral, DBRX-Intruct, GPT-4o
- each expert LLM is prompted at each t to predict the label given the features

Experiment: Financial market movement (FMM)

Results:

- all single LLM experts have similar performance (none is outstanding)
- MoE-F yields large performance increase of 17% real and 48.5% relative F1 measure improvement

Metrics ↑	LLM Experts							Experts Filter
	Llama-2 7b-chat	Llama-2 70b-chat	Llama-3 8B-Instruct	Llama-3 70B-Instruct	Mixtral-8x7B Instruct-v0.1	DBRX Instruct	OpenAI GPT-4o	MoE-F (ours)
F1	0.22	0.33	0.35	0.20	0.34	0.34	0.34	**0.52**
Acc	0.27	0.37	0.39	0.30	0.33	0.34	0.37	**0.57**
Precision	0.35	0.33	0.31	0.32	0.36	0.36	0.33	**0.61**
Recall	0.27	0.37	0.39	0.30	0.33	0.34	0.37	**0.57**

References I

Raeid Saqur, Ken Kato, Nicholas Vinden, and Frank Rudzicz. Nifty financial news headlines dataset, 2024. Manuscript under review.

W Murray Wonham. Some applications of stochastic differential equations to optimal nonlinear filtering. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 2 (3):347–369, 1964.