A Centralized Mixture-of-Experts using Domain-Specialized SLMs

Term Project - Progress Presentation

2020-19352 Youngho Cho
2022-14622 Jaewon Lee

Introduction and Motivation

Our Core Question:

→ **Can multiple domain-specialized SLMs outperform a single general-purpose model?**

Why this matters:

- 🐘 **Large LMs = powerful but costly and inefficient**
- 🐝 **Small LMs = efficient and strong within specific domains**
- 🔥 **Opportunity: make them collaborate**

Goals:

- ⚡ Combine efficiency and expertise
- 🌿 Enable scalable, sustainable NLP

Proposed Architecture

Centralized Mixture-of-Experts (MoE)

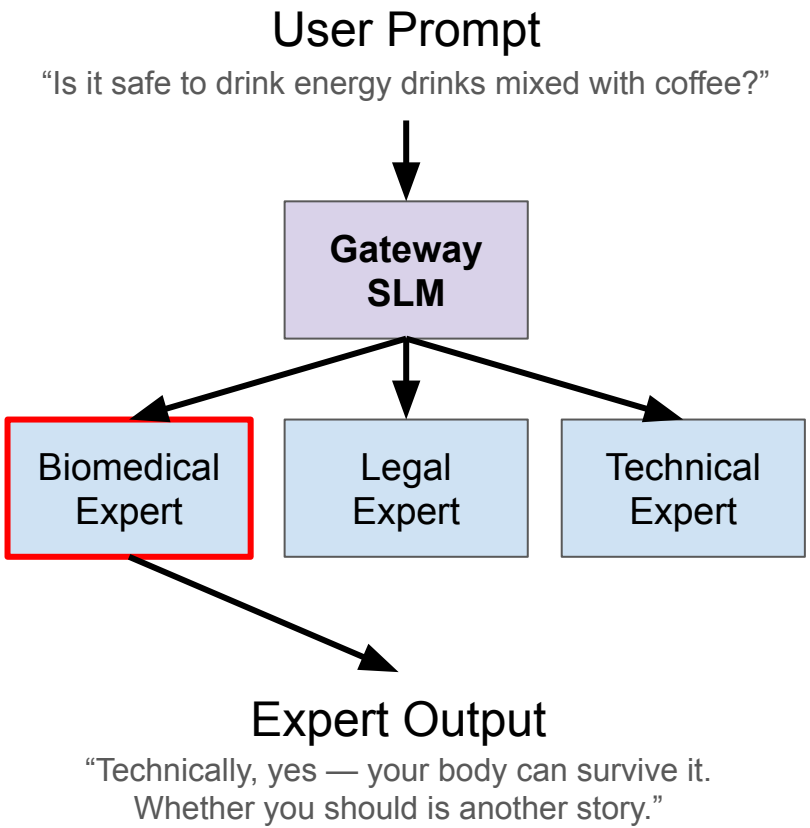
- Gateway SLM routes input to domain experts
- Experts: e.g. Biomedical, Legal, Technical
- Gateway **fuses** expert outputs into a final response

Advantages:

✓ Modular

✓ Efficient (activates only relevant experts)

✓ Scalable



Datasets & Domain Setup

Domain-Specific Datasets

Domain	Source	#Parameters	Core Function
Gateway	Gemma 2B	2B	High-speed query classification (Minimal Latency).
Biomedical	Phi-3	3.8B	SOTA medical/scientific knowledge and reasoning.
Legal	Legal-BERT	300M	Legal classification, entity recognition, and high-precision extraction.
Programming	DeepSeek-Coder	7B	SOTA code generation and mathematical/logical reasoning.

Core advantage: Dynamic Loading

Gateway Finetuning

Category	Description	Size Estimate	Rationale
Domain-Specific Queries	~1,500 queries for each of the 3 Experts (Biomedical, Legal, Programming).	**4,500 - 6,000 queries**	Establishes the core distinction between the three specialized domains.
Boundary & Ambiguous Queries	~1,500 queries that blend domains or are highly complex	**1,500 queries**	**Crucial for robust routing:** Trains the model to resolve ambiguity and make optimal choices.
General/Fallback Queries	~1,000 common, non-specialized questions	**1,000 queries**	Prepares the router for common user input and determines when to direct to a designated 'General' or 'Fallback' Expert

Evaluation Datasets

We merge these datasets into a unified evaluation dataset

Domain	SOTA Benchmarks Used	Evaluation Focus
Biomedical	MedQA, PubMedQA	**Factuality & Precision** (Contextual Q&A).
Legal	LexGLUE, CaseLaw	**Classification & F1-score** (Structural analysis).
Programming	HumanEval, GSM8K	**Functional Correctness** (Code generation & Math).
Gateway	Synthetic Data	**Classification Accuracy** on boundary queries.

Evaluation Datasets

Step	Action
1. Query Ambiguity Adjustment	**Paraphrase** benchmark questions and inject **Boundary Queries** (queries between domains).
2. Input/Output Standardization	Unify all data into a single format: **(Query, Ref_Answer, Domain_Label)**. Map non-generative outputs (e.g., Legal-BERT labels) to natural language.
3. Text Normalization	Apply standard cleaning: lowercasing, special character removal, and **tokenizer-specific filtering**.

Future Plan

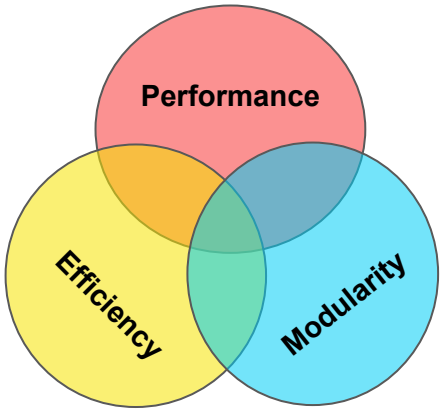
Phase (#)	Action Plan	Key Milestones
1. Soft Routing Finalization	Fine-tune Gemma 2B. Implement **Routing** logic. Pilot test system routing.	**Gateway Accuracy greater than 85%** on routing validation set.
2. Full Evaluation	Execute full **Unified Evaluation Dataset**. Calculate Domain-Specific Metrics.	**System Average Accuracy greater than 70%** across all domains (Legal, Bio, Code).
3. Comparative Analysis	Run **Baseline 1 (Centralized SLM)** and **Baseline 2 (Random Router)**. Compare performance/cost.	MoE Gateway achieves **higher accuracy** and **60% cost reduction** vs. baseline LLM API.
4. Optimization & Final Report	Optimize **Dynamic Loading/Unloading**. Finalize latency, cost, and accuracy report.	**End-to-End Latency less than 1.5 seconds** per query. Project **Final Report** submitted.

Risks & Backup Strategy

Risk	Description	Backup Strategy
R1. Routing Overfitting	Gemma 2B fails on **ambiguous, real-world queries**.	**Mitigation:** Implement **Semantic Routing** fallback using dense embedding to handle low-confidence classifications.
R2. Expert Underperformance	An Expert misses the 70% domain accuracy target.	**Mitigation:** Allocate **Buffer Time (Weeks 7-8)** for focused **LoRA Fine-Tuning** of the underperforming Expert.
R3. Latency Overrun	**Dynamic Loading** is too slow (>1.5s per query).	**Mitigation: Quantization.** Convert Expert models (7B/14B) to a lower bit precision (**4-bit GGUF**) to speed up loading and reduce VRAM footprint.

Expected Contribution

- Centralized MoE of SLMs → **higher accuracy** than single SLM
- **Lightweight** & **scalable** framework for domain specialization
- **Open-source prototype** for future NLP research



Thank you