Semantic Inference Routing Protocol (SIRP)

Motivation and Requirements

IETF 124 HotRFC

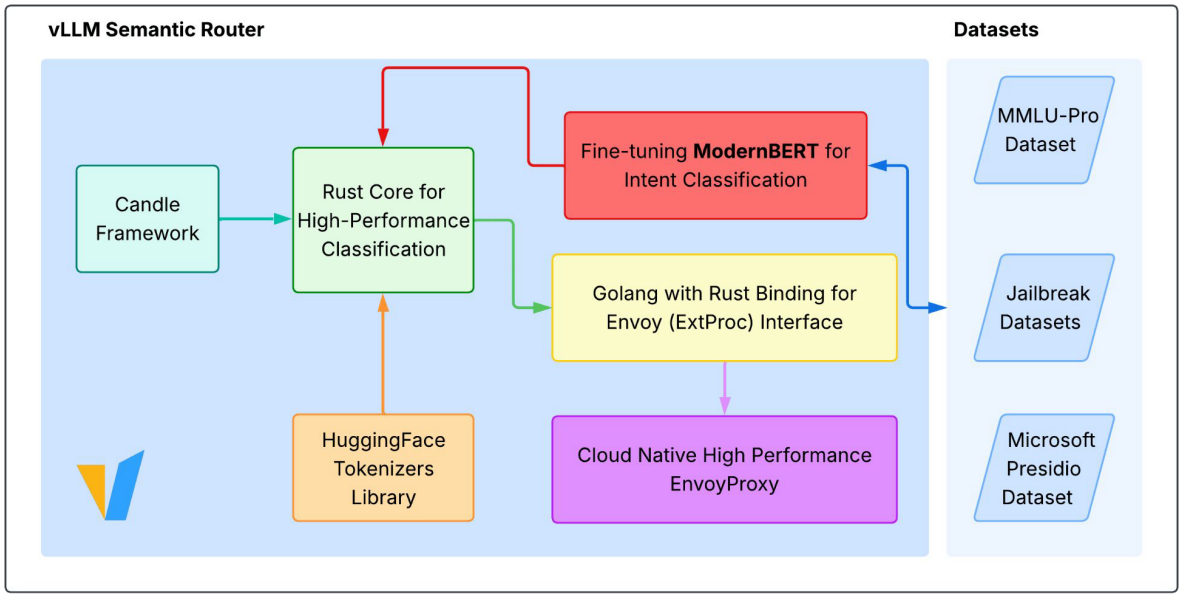
Huamin Chen
hchen@redhat.com

Red Hat

Motivation: Semantic Inference Routing Protocol (SIRP)

- Large Language Models (LLMs) demonstrate substantial accuracy gains when augmented with reasoning modes such as chain-of-thought and inference-time scaling.
- However, reasoning also incurs significant costs in inference latency and token usage
  - Environmental and financial impacts, which are unnecessary for many simple prompts.
- Semantic router classifies queries based on their reasoning requirements and selectively applies reasoning only when beneficial.
- Semantic routing achieves a 10.2 percentage point improvement in accuracy on the MMLU-Pro benchmark while reducing response latency by 47.1% and token consumption by 48.5%compared to direct inference with vLLM.
- Semantic routing offers an effective mechanism for striking a balance between accuracy and efficiency in open-source LLM serving systems.

Semantic Router Architecture



The system comprises four pillars:

- Semantic Classification:
  - Uses ModernBERT to determine routing paths.
- Smart Routing:
  - Simple queries → “fast path” inference.
  - Complex queries → “Chain-of-Thought” reasoning mode.
- High-Performance Engine:
  - Written in Rust using Hugging Face Candle, it delivers high concurrency and zero-copy inference.
- Cloud-Native Integration:
  - Works out-of-the-box with Kubernetes and Envoy via the ext_proc plugin.

Call for collaboration

- Drafts submitted in NMRG
  - https://datatracker.ietf.org/doc/html/draft-chen-nmrg-semantic-inference-routing
  - https://www.ietf.org/archive/id/draft-chen-nmrg-multi-provider-inference-api-00.html
- Looking for like minded individuals to join our community
  - https://github.com/vllm-project/semantic-router
- Collaboration on furthering the drafts
  - Use cases
- Comments and feedback