SmartRoute

Apresentação do módulo de roteamento.

Problemática

- Instabilidade das API's dos modelos conversasionais.
- Falta de roteamento automático por tipo de prompt eleva custos.

Algumas contas de padaria...

Provider	model	input cost Mtoken	output cost per Mtoken
Chatgpt	gpt-4o-mini	$0.150	$0.600
Chatgpt	gpt-4o	$2.50	$10.00
Chatgpt	o3-mini	$1.10	$4.40
Anthropic	claude-3-haiku	$0.25	$1.25
Anthropic	claude-3-5-haiku	$0.80	$4.00
Anthropic	claude-3-7-sonnet	$3.00	$15.00
DeepSeek	deepseek-chat	0.27	$1.10
DeepSeek	deepseek-reasoner	$0.55	$2.19

Cenário 1: Sem Troca Entre Modelos

Forçado a usar um único modelo (por exemplo, o *gpt-4o*) para todas as tarefas:

- **Custo fixo:** $12,50/Mtoken
  Mesmo para tarefas simples.

Cenário 2: Livre Escolha

- **Modelos baratos (90%)** i.e. *gpt-4o-mini*
  Custo: $0,15 (input) + $0,60 (output) = **$0,75/Mtoken**
- **Modelos caros (10%)** i.e. *gpt-4o*
  Custo: $2,50 (input) + $10,00 (output) = **$12,50/Mtoken**

Custo médio ponderado:

$$0.9 \times 0.75 + 0.1 \times 12.50 \approx 1.93 \, \$/\text{Mtoken}$$

Como smartRoute soluciona esses problemas?

- **Integração simplificada:** Facilita a conexão entre múltiplas APIs de IA, permitindo que cada consulta seja direcionada ao modelo mais adequado.
- **Otimização de recursos:** Envia consultas simples para modelos rápidos e consultas complexas para modelos mais elaborados, melhorando serviços como o AssessorAI.

Como smartRoute soluciona esses problemas?

- **Histórico unificado:** Armazena o histórico das conversas, garantindo continuidade mesmo ao alternar entre diferentes IAs.
- **Aumento da disponibilidade:** Melhora a performance e a confiabilidade dos serviços que dependem das respostas das diversas APIs de IA.

Funcionalidades já implementadas:

- **Roteamento inteligente:** Direciona cada requisição para o modelo de IA mais adequado (fast, mid, reasoning ou latency) conforme o tipo e a complexidade da consulta.
- **Configuração customizável:** Permite definir, via parâmetros, qual modelo usar e se a execução será em sequência ou em paralelo, adaptando-se às necessidades específicas da aplicação.

Funcionalidades já implementadas:

- **Classificação automática:** Incorpora uma IA dedicada para classificar a requisição e determinar o modelo ideal, garantindo respostas otimizadas.
- **Memória agnóstica de IA:** Armazena o histórico das conversas independentemente do modelo utilizado, permitindo uma continuidade consistente na interação.

Dúvidas?

Demonstração