Arquitetura Híbrida para Assistentes Virtuais: LLMs com Delegação de Tarefas a SLMs Especializados

Cláudio Generoso
Instituto Tecnológico de Aeronáutica / Universidade Federal de São Paulo (UNIFESP)
Praça Marechal Eduardo Gomes, 50 – Vila das Acácias, São José dos Campos – SP, Brasil
claudio.generoso@ga.ita.br

Profa. Dra. Ana Carolina Lorena
Instituto Tecnológico de Aeronáutica (ITA)
Praça Marechal Eduardo Gomes, 50 – Vila das Acácias, São José dos Campos – SP, Brasil
ana.lorena@ita.br

RESUMO

Este trabalho propõe uma arquitetura híbrida para assistentes virtuais baseada em LLMs (Large Language Models), responsáveis pela compreensão da intenção do usuário e pela delegação da tarefa a SLMs (Small Language Models) especializados, otimizados para domínios específicos. O objetivo é reduzir o custo computacional e melhorar a eficiência no atendimento automatizado, com aplicação inicial nos setores de segurança pública e defesa civil.

A arquitetura utiliza um LLM como ponto de entrada, encarregado de interpretar a solicitação, corrigir eventuais erros e estruturar o texto antes de encaminhá-la a um SLM específico. O SLM de segurança pública foi treinado para classificar relatos urbanos em categorias como roubo, furto, perturbação do sossego e dano ao patrimônio, usando vetorização TF-IDF e classificação via SVM. Já o SLM de defesa civil analisa dados de pluviômetros do CEMADEN, estações meteorológicas locais e arquivos geoespaciais (KML) de áreas de risco, utilizando regras determinísticas calibradas com especialistas e enriquecendo as respostas com documentos oficiais por meio da técnica RAG (Retrieval-Augmented Generation).

Os dados utilizados incluem aproximadamente 800 registros simulados de ocorrências urbanas, informações pluviométricas e meteorológicas em tempo real e mapas de zonas de risco fornecidos pela Defesa Civil de São José dos Campos (SP). A arquitetura foi validada em testes simulados, apresentando resultados qualitativos promissores: no domínio de segurança pública, 46 de 50 casos foram corretamente classificados por especialistas; no domínio de defesa civil, todos os cenários críticos simulados resultaram em alertas corretos. O tempo médio de resposta do sistema completo foi de aproximadamente 2 segundos, com funcionamento viável em infraestrutura computacional convencional.

A solução proposta mostrou-se escalável, auditável e adaptável a outros domínios da gestão pública, como saúde, educação e mobilidade urbana. A modularidade da arquitetura facilita a inclusão de novos SLMs e a integração com plataformas institucionais, como o sistema 156, ampliando o potencial de impacto social por meio de respostas mais rápidas, precisas e fundamentadas.

PALAVRAS CHAVE. Assistentes Virtuais; LLMs; SLMs.

https://proceedings.science/p/212402?lang=pt-br