Bagaimana cara kerja chatbot Anda?

Chatbot kami menggunakan semantic search untuk mencari referensi yang sesuai dari basis data internal. Proses ini melibatkan embedding dengan model ada-002 untuk mengubah teks menjadi vektor, sehingga memungkinkan pencocokan semantik yang lebih efektif. Kami kemudian mengintegrasikan sistem ini dengan GPT-3.5 Turbo dari OpenAI untuk menghasilkan jawaban berdasarkan referensi yang ditemukan.

Bagaimana cara embedding digunakan dalam sistem Anda?

Embedding dilakukan dengan model ada-002, yang mengonversi data teks menjadi vektor numerik. Vektor ini digunakan dalam semantic search untuk mengidentifikasi kemiripan kata atau konsep, memungkinkan sistem menemukan referensi yang paling relevan untuk pertanyaan pengguna.

Apa fungsi dari masing-masing komponen LLM dalam sistem Anda?

Kami memiliki tiga fungsi utama yang menggunakan GPT-3.5 Turbo:

Classify: Fungsi ini mengklasifikasikan tipe pertanyaan pengguna menjadi tiga kategori: absurd_question, faq_question, dan greeting. Klasifikasi ini membantu mengisolasi jenis pertanyaan agar sistem tidak memberikan jawaban yang tidak sesuai. Misalnya, pertanyaan yang absurd atau salam diproses secara khusus untuk menghindari kesalahan tanggapan.

Greeting: Fungsi ini secara khusus menangani salam dari pengguna. Dengan isolasi ini, chatbot memberikan respons yang ramah dan tepat untuk berbagai bentuk salam.

Question: Fungsi ini menjalankan semantic search pada basis data untuk menemukan referensi yang sesuai dan menggunakan hasilnya untuk menjawab pertanyaan pengguna dengan pendekatan Retrieval-Augmented Generation (RAG). Ini berarti sistem menghasilkan jawaban berdasarkan informasi yang telah ditemukan melalui pencarian semantik.

Apa kelebihan dari pendekatan yang Anda gunakan?

Pendekatan kami menggabungkan embedding untuk pencarian semantik dengan kemampuan generatif LLM. Hal ini memungkinkan chatbot memberikan jawaban yang lebih relevan dan kontekstual, serta menangani berbagai jenis pertanyaan dengan presisi yang lebih baik. Integrasi fungsi klasifikasi juga mencegah tanggapan yang tidak sesuai, meningkatkan akurasi dan kualitas interaksi pengguna.

Mengapa memilih RAG dibandingkan metode lainnya (fine-tuning, menggunakan model siap pakai, atau melatih dari awal)?

Kami memilih Retrieval-Augmented Generation (RAG) karena beberapa alasan:

Informasi yang Relevan: RAG memungkinkan penggunaan informasi terkini dan relevan dengan menggabungkan pencarian semantik dari basis data dengan kemampuan generatif model. Ini mengatasi keterbatasan model siap pakai yang mungkin tidak memiliki informasi terbaru atau spesifik.

Pembaruan Mudah: Dengan RAG, pembaruan pengetahuan hanya memerlukan pembaruan pada basis data referensi tanpa perlu melatih ulang model. Ini lebih efisien dibandingkan dengan fine-tuning, di mana setiap perubahan informasi akan memerlukan proses pelatihan ulang yang mahal dan memakan waktu.

Sumber Daya Komputasi Lebih Rendah: Melatih model dari awal membutuhkan sumber daya komputasi yang besar dan waktu yang lama. RAG mengurangi kebutuhan ini dengan menggunakan model yang sudah dilatih sebelumnya dan menambahkan lapisan pencarian informasi untuk menghasilkan jawaban yang relevan dan akurat.

Dinamis untuk Data yang Terus Diperbarui: RAG sangat berguna dalam situasi di mana data dinamis yang sering diperbarui diperlukan. Dengan hanya memperbarui basis data referensi, sistem dapat terus memberikan jawaban yang akurat tanpa perlu perubahan besar pada model inti.

Mengapa menggunakan ada-002 dan GPT-3.5 Turbo?

ada-002: Model embedding ada-002 dipilih karena performanya yang unggul dalam bahasa Indonesia. Ini memastikan bahwa pencarian semantik dan pemrosesan teks dalam bahasa Indonesia lebih akurat, sehingga relevansi hasil pencarian meningkat.

GPT-3.5 Turbo: Kami memilih GPT-3.5 Turbo dari OpenAI karena memberikan keseimbangan yang baik antara performa dan biaya. Model ini dikenal karena kualitasnya yang tinggi dalam menghasilkan teks, dan harganya lebih terjangkau dibandingkan dengan model-model yang lebih besar, menjadikannya pilihan yang ekonomis untuk kebutuhan kami.

Apa kelebihan dari pendekatan yang Anda gunakan?

Pendekatan kami menggabungkan embedding untuk pencarian semantik dengan kemampuan generatif LLM. Hal ini memungkinkan chatbot memberikan jawaban yang lebih relevan dan kontekstual, serta menangani berbagai jenis pertanyaan dengan presisi yang lebih baik. Integrasi fungsi klasifikasi juga mencegah tanggapan yang tidak sesuai, meningkatkan akurasi dan kualitas interaksi pengguna.