Pengantar Kecerdasan Buatan dan RAG

1. Apa Itu Kecerdasan Buatan

Kecerdasan buatan atau artificial intelligence (AI) adalah cabang ilmu komputer yang mempelajari cara membuat mesin melakukan tugas yang biasanya membutuhkan kecerdasan manusia, seperti mengenali gambar, memahami bahasa, dan mengambil keputusan. Istilah ini pertama kali diperkenalkan oleh John McCarthy pada konferensi Dartmouth tahun 1956.

Pembelajaran mesin (machine learning) adalah bagian dari AI yang memungkinkan komputer belajar dari data tanpa diprogram secara eksplisit. Pembelajaran mendalam (deep learning) adalah bagian dari pembelajaran mesin yang menggunakan jaringan saraf tiruan berlapis banyak.

2. Jenis Pembelajaran Mesin

Terdapat tiga jenis utama pembelajaran mesin. Pembelajaran terawasi memakai data berlabel, misalnya klasifikasi email spam. Pembelajaran tak terawasi mencari pola pada data tanpa label, misalnya pengelompokan pelanggan dengan algoritma K-Means. Pembelajaran penguatan melatih agen melalui hadiah dan hukuman, misalnya pada program yang bermain catur atau permainan video.

Jenis	Data	Contoh algoritma
Terawasi	Berlabel	Regresi linier, pohon keputusan
Tak terawasi	Tanpa label	K-Means, PCA
Penguatan	Hadiah dan hukuman	Q-Learning

3. Retrieval-Augmented Generation (RAG)

RAG adalah teknik yang menggabungkan pencarian dokumen dengan model bahasa besar. Pertama, dokumen dipecah menjadi potongan kecil yang disebut chunk. Kedua, setiap chunk diubah menjadi vektor angka yang disebut embedding dan disimpan dalam basis data vektor. Ketiga, saat pengguna bertanya, pertanyaan diubah menjadi embedding lalu dicari chunk yang paling mirip. Terakhir, chunk tersebut diberikan kepada model bahasa sebagai konteks agar jawabannya berdasar pada dokumen, bukan sekadar ingatan model.

Keunggulan RAG adalah jawaban dapat diperbarui hanya dengan mengganti dokumen, tanpa melatih ulang model. RAG juga mengurangi halusinasi, yaitu jawaban yang terdengar meyakinkan tetapi salah, karena jawaban dapat ditelusuri ke sumbernya.

4. Strategi Chunking

Ukuran chunk sangat memengaruhi kualitas pencarian. Chunk yang terlalu kecil, misalnya 200 karakter, sering kehilangan konteks sehingga jawaban terpotong. Chunk yang terlalu besar, misalnya di atas 2000 karakter, memuat banyak topik sekaligus sehingga embedding menjadi kabur dan hasil pencarian kurang tepat. Praktik umum adalah memakai ukuran 500 sampai 1000 karakter dengan overlap 10 sampai 20 persen agar kalimat di perbatasan tidak hilang.

Pemotongan sebaiknya dilakukan pada batas kalimat atau paragraf. Tabel dan daftar membutuhkan perhatian khusus karena pemotongan di tengah tabel dapat memisahkan judul kolom dari isinya.

5. Evaluasi dengan Hit@k

Hit@k mengukur proporsi pertanyaan uji yang jawabannya ditemukan di antara k hasil teratas. Misalnya, Hit@3 sebesar 0,80 berarti 80 persen pertanyaan memiliki chunk relevan di antara tiga hasil teratas. Nilai Hit@k selalu naik atau tetap ketika k diperbesar. Metrik ini sederhana namun tidak menilai urutan; untuk itu dapat dipakai Mean Reciprocal Rank (MRR) yang memberi nilai lebih tinggi bila chunk relevan berada di peringkat awal.

Pencarian semantik unggul dibanding pencarian kata kunci saat pertanyaan memakai kata berbeda dengan teks dokumen, misalnya pertanyaan tentang mobil dapat menemukan teks yang menyebut kendaraan roda empat. Kelemahannya, pencarian semantik kadang gagal pada angka, kode, atau nama spesifik yang harus cocok persis.