RM-MCP: Server AI Portabel dari Limbah Set-Top Box, dengan RAG Tertutup untuk Asisten HR UMKM

Program PPMI STEI 2026 – Riset Muda Pesat

Devi Willieam Anggara[1], Fetty Fitriyanti Lubis[1], Dion Tanjung[1], Lenny Putri Yulianti[1], Marcellino Joestiawan[1], Mohd Shafry Mohd Rahim[2], Ibrahim Salim Sulaiman Al Farsi[3]

[1] Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung [2] UTM-IRDA Digital Media Centre, Universiti Teknologi Malaysia [3] Faculty of Computing and IT, Sohar University, Oman

1 Mengenal RAG: AI yang Menjawab dengan Rujukan

RAG (Retrieval-Augmented Generation) membuat AI membuka dokumen rujukan sebelum menjawab, sehingga jawabannya lebih sesuai sumber dan lebih jarang mengarang (halusinasi). Berikut jenis-jenis RAG yang kami bandingkan; SRAG+CRAG adalah gabungan SRAG dan CRAG.

Naive RAG (RAG standar)

Tanya ▶ Cari ▶ Jawab

Mencari dokumen yang mirip dengan pertanyaan, lalu langsung dipakai menjawab. Sederhana, tetapi dokumennya tidak diperiksa.

SRAG (Structured RAG)

Beri label ▶ Cari ▶ Jawab

Potongan dokumen diberi label (topik, jenis isi, kata kunci) sebelum disimpan, agar pencarian lebih tepat.

CRAG (Corrective RAG)

Cari ▶ Nilai ▶ Koreksi ▶ Jawab

Hasil pencarian dinilai dulu: yang tepat dipakai, yang ragu atau salah dicari ulang. Versi aslinya mencari ke internet.

CRAG tertutup (usulan kami)

Cari ▶ Nilai ▶ Cari ulang ▶ Jawab

Hasil pencarian dinilai. Bila kurang tepat (28% pertanyaan), pencarian diulang dari buku yang sama, bukan internet.

2 Abstract

Penelitian ini membuat **server mini portabel** dari Set-Top Box (STB) bekas untuk menjalankan Agentic AI (AI yang bekerja mandiri). Kuncinya **RM-MCP** (RAG Multi-Model Context Protocol), standar yang menghubungkan model AI kecil dengan basis data dokumen. Pengujian memakai asisten HR UMKM berbasis satu buku rujukan, empat jenis RAG, dan 270 pertanyaan. **Hasilnya, CRAG tertutup yang kami usulkan unggul dalam tiga hal:** (1) Faithfulness (kesesuaian dengan sumber) tertinggi, 0,876 vs 0,856 pada RAG standar, setara secara statistik; (2) jawaban dapat ditelusuri ke satu buku, tanpa internet; (3) bila hasil pencarian kurang tepat (28% pertanyaan), pencarian diulang otomatis dari buku yang sama.

Keyword: Agentic AI, RAG, MCP, Edge Computing, Set-Top Box, Smart HR

3 Introduction

Kecerdasan buatan mandiri membutuhkan komputer berdaya besar dengan biaya tinggi, dan data harus dikirim ke pihak ketiga. Akibatnya, UMKM (61,07% PDB; hingga 97% tenaga kerja) sulit memanfaatkannya. Padahal, STB bekas penyedia internet (sekitar Rp200–400 ribu) tersedia murah, tetapi belum diteliti sebagai server AI.

Server murah saja belum cukup: asisten HR harus menjawab dari rujukan tepercaya. RAG standar langsung menjawab tanpa memeriksa mutu dokumen, sedangkan CRAG memeriksanya tetapi mencari ke internet sehingga sumbernya sulit ditelusuri. Karena itu, kami membangun server dari STB bekas dan mengadaptasi CRAG di uji memakai satu buku rujkan.

4 Research Method

A. Membangun server mini dari STB

1. **Siapkan STB** — Linux ringan (Armbian), ZRAM, dan Docker.
2. **Pasang perangkat lunak AI** — MCP (Model Context Protocol), n8n, Qdrant (basis data), Cloudflare Tunnel.
3. **Bandingkan STB dengan VPS** — Biaya dan kinerja; hasilnya menjadi artikel dan modul ajar.

B. Menguji RAG untuk asisten HR

1. **Siapkan buku rujukan** — Armstrong's Handbook of HRM: 1.032 potongan teks di Qdrant.
2. **Bangun empat jenis RAG** — RAG standar, SRAG, CRAG tertutup (dengan penilai T5), dan SRAG+CRAG.
3. **Ajukan 270 pertanyaan** — Rekrutmen, kinerja, dan karier.
4. **Nilai dan bandingkan** — RAGAS, dan uji statistik (Wilcoxon, TOST).

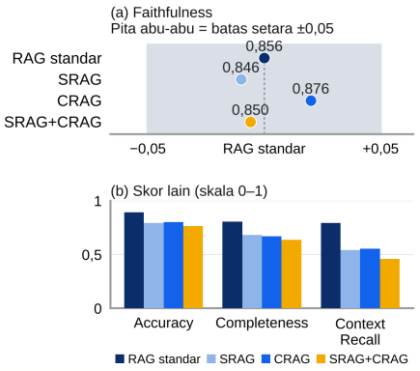
Dari mana 28%? Penilai T5 (model AI kecil yang kami latih menilai apakah teks menjawab pertanyaan) memeriksa 5 potongan teks per pertanyaan. Pada 75 dari 270 pertanyaan tak satu pun dinilai tepat, sehingga pencarian diulang. Penilai ketat: menolak 95,7% teks sebidang topik, juga 70,9% teks rujukan benar.

5 Discussion & Result

RAGAS (alat penilai otomatis) mengukur Faithfulness (porsi isi jawaban yang didukung dokumen, 0–1) dan Context Recall (kelengkapan dokumen yang ditemukan).

Uji Wilcoxon menguji apakah beda skor dua metode itu nyata atau kebetulan.

TOST (equivalence test) menguji apakah dua metode setara: selisih skor ≤ 0,05 (sekitar 1 dari 20 pernyataan); batas ditetapkan setelah data ada.



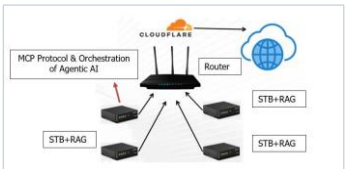
Gambar 1. Skor rata-rata 270 pertanyaan.

- **Hasil utama:** Faithfulness CRAG tertutup (0,876) lebih unggul dari RAG standar (0,856): selisih 0,02 (masuk dalam batas toleransi TOST 0,05; Wilcoxon p = 0,19).
- **Catatan:** Untuk satu buku rujukan yang terstruktur, RAG standar sudah cukup optimal—varian CRAG membutuhkan waktu proses ~2× lebih lama, memiliki skor Accuracy dan Completeness sedikit lebih rendah, serta fitur pencarian ulangnya belum memberi peningkatan skor yang signifikan.

6 Prototipe dan Rancangan



Gambar 2. Prototipe: STB bekas menggunakan power bank.



Gambar 3. Rancangan klaster STB dengan MCP jika pengetahuan terpisah.

7 Conclusion

- **Dapat dipercaya (Trusted AI):** jawaban berasal dari satu buku rujukan, sehingga mudah dicek.
- **Kualitas terjaga:** Faithfulness setara RAG standar, dengan keunggulan pencarian ulang otomatis untuk mengkorekso output akhir.
- **Murah dan ramah lingkungan:** server dari STB bekas (Rp200–400 ribu).