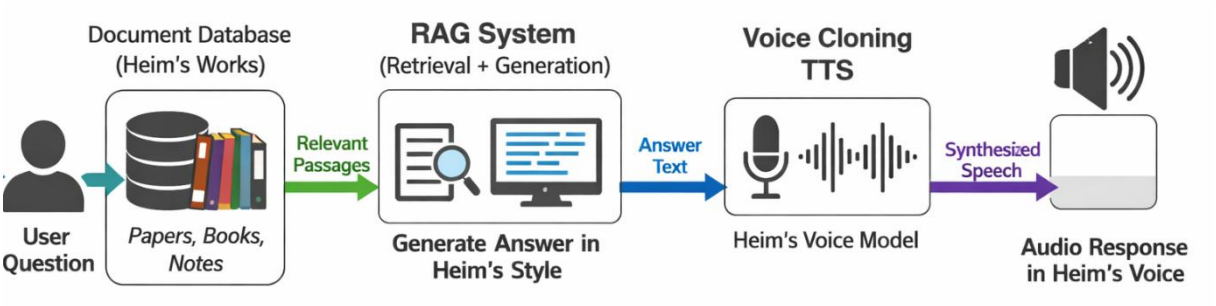
GTP5, 28.03.2026

Burkhard Heim Dialogue System Setup

Objective:
Develop a dialogue system capable of reasoning and responding **in the conceptual space of Burkhard Heim**, using his terminology, style, and internal logic.



1. Document Corpus

- Comprises all available Heim texts, papers, and notes.
- Currently **growing and evolving**, so the corpus is **not stable**.
- Needs preprocessing:
    - PDF → clean text
    - Chunking into 500–1500 token sections
    - Embedding for retrieval

2. Approaches for Language Model Integration

Option A: Retrieval-Augmented Generation (RAG)

- **How it works:** Retrieve relevant documents at runtime and feed them to the language model for answering.
- **Pros:**
    - Works with a dynamic corpus
    - Flexible, up-to-date
    - Transparent: sources visible
- **Cons:** Less “internalized” style, depends on prompt design

Option B: Fine-Tuning

- **How it works:** Train the language model on Heim texts directly.
- **Pros:**
    - Natural style and internalized reasoning
    - Less prompt engineering

- **Cons:**
    - Static: requires retraining with new documents
    - More expensive and technically involved

Recommendation: Start with **RAG**, optionally move to fine-tuning once the corpus stabilizes.

3. Voice Cloning

- Objective: Generate speech in **Heim's voice**.
- **Process:** Separate from language reasoning
    - Text output from RAG / model → TTS system → audio in Heim voice
- **Data needed:** ~30–120 minutes clean recordings (minimum 5–10 min)
- **Tools:**
    - Cloud: ElevenLabs, OpenAI Voice
    - Local: Coqui TTS, Bark, Tortoise TTS
- **Notes:** Fine-tuning the language model is **not needed** for voice replication.

4. Deployment Options

Component	Cloud-Based	Local Setup
Language Model	GPT-5 / OpenAI API, Claude	LLaMA 3/4, Mistral
Retrieval & Database	Pinecone, Weaviate, Chroma	Local Chroma/FAISS
Voice Cloning / TTS	ElevenLabs, OpenAI Voice	Coqui, Bark, Tortoise (GPU req)
Maintenance	Managed, easy updates	Manual, more control

Key Difference: Cloud = lower setup, ongoing costs; Local = higher setup, no recurring costs.

5. Cost Overview

Component	Cloud Cost Estimate	Local Cost Estimate
RAG / API usage	50–300 €/month (depending on usage)	GPU + setup one-time
Voice Cloning / TTS	~10–50 €/month for moderate use	GPU + training time one-time
Fine-Tuning (optional)	100s–1000s € per retraining	GPU + technical overhead

Notes:

- Cloud costs are recurring due to:
    - Realtime TTS generation
    - Hosting and serving the voice model
    - API-based language model access

- Local setup has one-time costs, but requires technical expertise.

Summary Workflow

Question → RAG (retrieve Heim docs) → Language Model (generate answer in Heim-style) → TTS / Voice Cloning → Audio output

Recommendation:
Start with **RAG + Cloud TTS** to rapidly prototype Heim-style dialogue. Transition to fine-tuning or local voice models as corpus stabilizes or project scales.