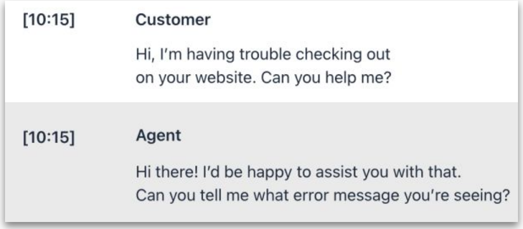
Live Transcript RAG

Dungeon Master or Call Agent CoPilot

Roman Pfeiler

Idea

Transcript of Live Conversation



[10:15] **Customer**

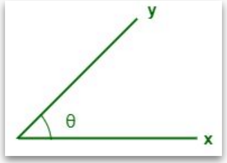
Hi, I'm having trouble checking out on your website. Can you help me?

[10:15] **Agent**

Hi there! I'd be happy to assist you with that. Can you tell me what error message you're seeing?

→

Vector Similarity with Documents



→

LLM Call if > Threshold

Use Case

Call Agent CoPilot

Supporting Call Center Agents by pulling up the right information and steps to follow to solve client problems.

- Especially useful for New Joiners and
- Changing processes and applications

Clients can now reactivate their blocked cards themself via the following Self Service page.

Dungeon Master CoPilot

Supporting Dungeon Masters leading a game of Dungeons and Dragons (D&D) by giving relevant information on the rules and the campaign setting.

- Reduces preparation time and
- Time to look up information

Falling damage is calculated as 1d6 bludgeoning damage for every 10 feet fallen, up to a maximum of 20d6.

Implementation

“Active Loop” RAG

1. Ingestion: PDFs are loaded, split into 1000-char chunks with overlap, and indexed in a local ChromaDB.
2. Triggering: As transcript text arrives, it is vectorized with the embedding model: Qwen3-Embedding-0.6B
3. Thresholding: The system calculates Cosine Similarity between the live transcript and the Knowledge Base.
    - If Similarity Score < Threshold (0.95): Trigger Retrieval.
    - Else: Ignore (classified as "chit-chat").

Fine-Tuning

- Fine-tuned "Qwen/Qwen3-Embedding-0.6B" using SentenceTransformers and MultipleNegativesRankingLoss.

LLM Call

- Cloud: Google Gemini 2.0 Flash
- 1. Local: Llama-3.2-1B via Ollama
- 2. Local: Qwen2.5-3B-Instruct

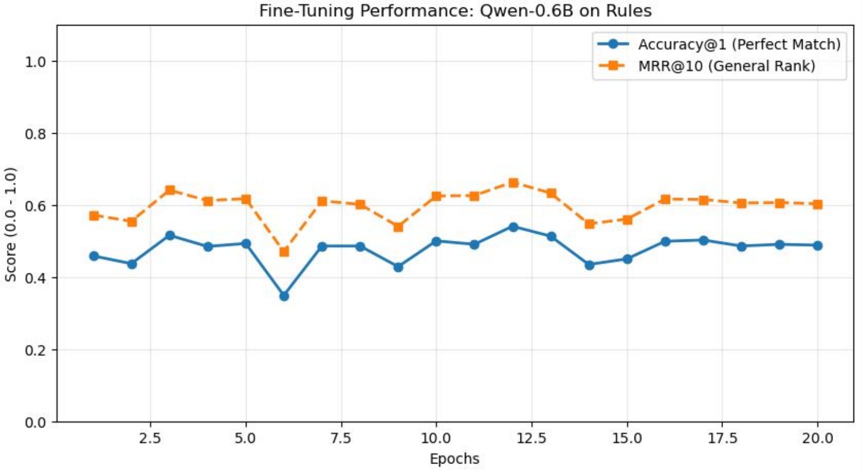
Fine - Tuning

Training Data

- Synthetic data created with LLM
- Chunks of D&D rules document to create tabletop scenarios needing that chunk as an answer

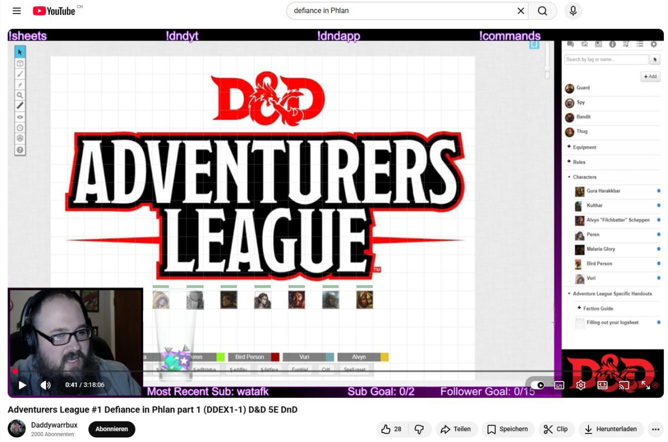
Approach

- Fine-tuned "Qwen/Qwen3-Embedding-0.6B"
- Using SentenceTransformers
- Evaluated with MultipleNegativesRankingLoss.



Result

Standard Qwen + Gemini API



Adventurers League #1 Defiance in Phlan part 1 (DDEX1-1) D&D 5E DnD

Daddywarrbux
2000 Abonnenten
Abonnieren

Details	Transcript	LLM response
Timestamp: 2:54 Response Time: 2.059	preparing to bunk down for the evening when one of free owners daughters peeks into the room she calls a few of the adventures including all of you into the hall pardon my interruption says the halfling girl named Rhys a **chap just came into the common room downstairs** and he asked me to fetch you something about some coin needing to change hands for an easy job she plays	The **hooded figure in the common room** is a member of the Harpers, disguising his voice, and may be an older half-elf with close-cropped grey hair. He has a job offer involving coin.
Timestamp: 5:53 Response Time: 1.651	you know either a human or elf but you can't really tell his identity or anything else he speaks in a low voice and he begins to talk he says I'm a member of the Harper's and I have a job for you what's that job well you see we captured a **merchant** who was going to illegally purchase a red dragon egg I want you guys to pose as	One of the characters (chosen at random) **bears enough of a resemblance to the captured merchant that the sellers** (who have never seen the merchant) can be easily fooled. The rest of the characters can act the roles of bodyguards, assistants, porters, etc.
Timestamp: 11:44 Response Time: 1.653	you to be there Oh what time are in like in like 20 minutes you have to go you have to go now okay now I haven't even told you where you're going okay there's a barn on the edge of town it's it's only about a **15 minute walk** it's a big big-ass burn you can't miss it you just head east down the road and you'll see it	The characters must hurry to the barn on the edge of town, heading east. It's a **15-minute walk** to a large barn.

Result

Run on Laptop with no dedicated GPU

Standard Qwen + Gemini API

- AVG Response Time: 1.9 sec
- Result quality: good

Fine-Tuned Qwen + Llama 1B

- AVG Response Time: 14.8 sec
- Result quality: very bad

Standard Qwen + Qwen 3B

- AVG Response Time: 1.8 sec
- Result quality: usable

Additional Changes:

- Smaller transcript & document chunk size
- Stop sending irrelevant context
- Prompt design

Discussion

Issues

- Fine-Tuning: Improve D&D specific retrieval
  - Chunk Size Mismatch
  - "Messy live Play Transcript" not reflected in training data
- LLM Model can not handle context information correctly
  - Transcript is messy, informal speech
  - Attention failure with context prompt
- Latency: Long delay with local models

Future Development

- Change to Cosine similarity from Euclidean Distance
- **Retry Fine-Tuning** for embeddings
- **Reranker** for better retrieval results
- Optimize **chunking** size
- **Intention routing** (either rules or campaign)
- **Audio Integration** for real live transcription
- **UI overlay** for the DM

GitHub: https://github.com/RomanPfeiler/DM-CoPilot/tree/main