YouTube Videos of Medfluencer Channels as Source of Medical Information

From Advanced Webscraping to Downstream Application

Heidelberg University

Supervisor: Marina Walther

Student: Jonas Gann

31.07.2024

What are the most common causes of lower back pain and what measures can I take to alleviate this discomfort?

The most common causes of lower back pain are usually functional disorders of the muscles, ligaments, tendons and, above all, the fascia. Contrary to popular belief, disc problems are not the most common cause. Back pain is often caused by muscle tension, which leads to irritation of the pain receptors ...

...

Project Description

Sportklinik Hellersen

Ihr Sportarzt

Uniklinik Tübingen

RAG

QA

What are the most common causes of lower back pain and what measures can I take to alleviate this discomfort?

The most common causes of lower back pain are usually functional disorders of the muscles, ligaments, tendons and, above all, the fascia. Contrary to popular belief, disc problems are not the most common cause. Back pain is often caused by muscle tension, which leads to irritation of the pain receptors …

Data from verified Medical YouTube Channels

Large Language Model

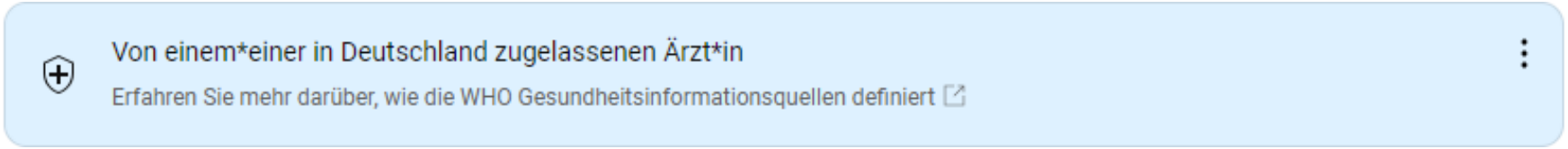
Application

How can medical data from YouTube be used for ML applications?

Data Retrieval

Goal: Make QA-System usable in every medical field

Allgemeinmedizin, Anästhesiologie, Anatomie, Arbeitsmedizin, Augenheilkunde, Biochemie, Chirurgie, Allgemeine Chirurgie, Gefäßchirurgie, Herzchirurgie, Kinderchirurgie, Orthopädie und Unfallchirurgie, Plastische, Rekonstruktive und Ästhetische Chirurgie, Thoraxchirurgie, Viszeralchirurgie, Frauenheilkunde und Geburtshilfe, Gynäkologische Endokrinologie und Reproduktionsmedizin, Gynäkologische Onkologie, Spezielle Geburtshilfe und Perinatalmedizin, Hals-Nasen-Ohrenheilkunde, Haut- und Geschlechtskrankheiten, Humangenetik, Hygiene und Umweltmedizin, Innere Medizin ... (63 medical fields)



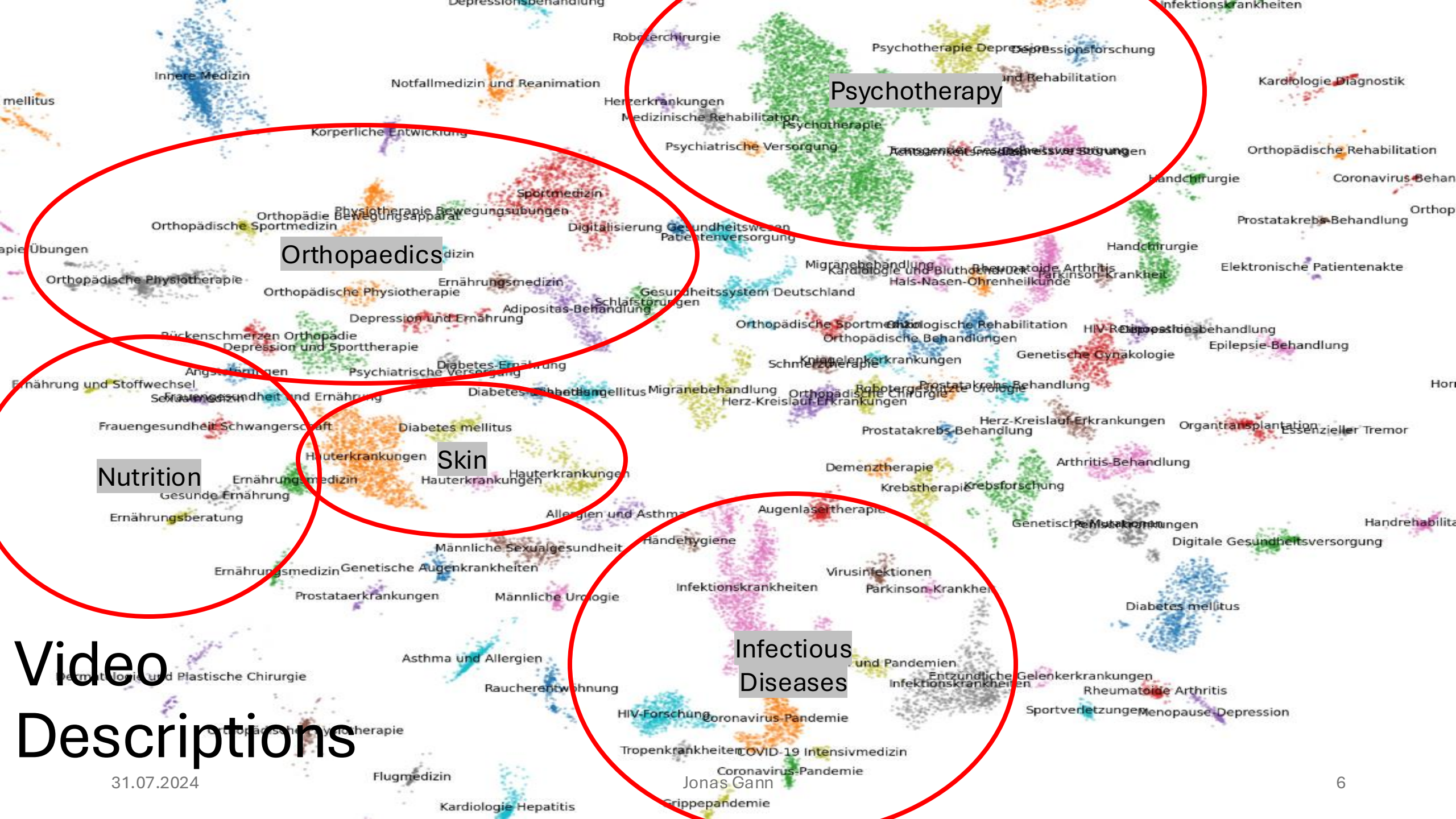
Channels (total: **362**, 0.3 MB)
- name: str
- description: str

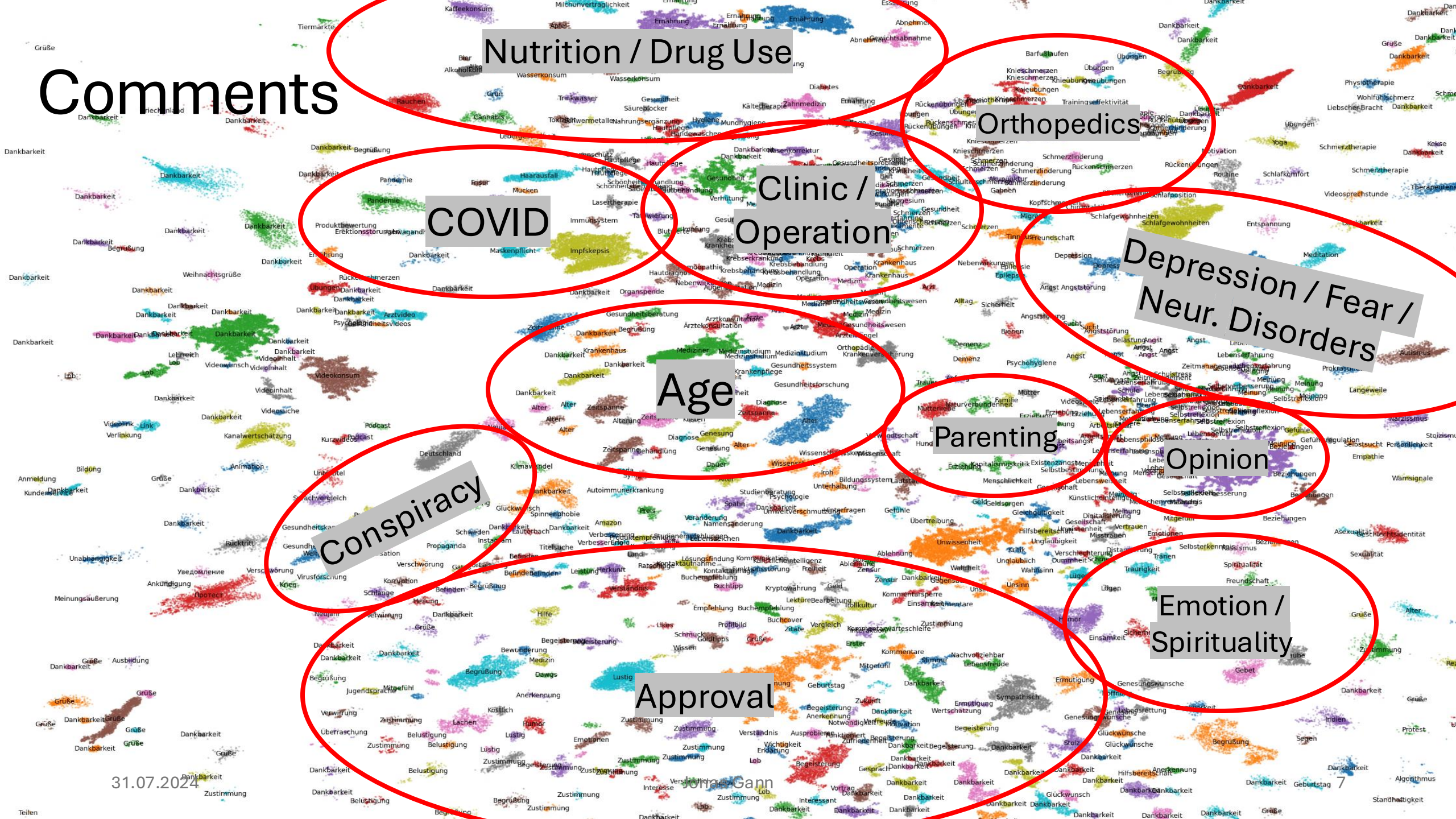
Videos (total: **94.422**, 803 MB)
- id: str
- title: str
- description: str
- transcript: str

Comments (total: **998.721**, 312 MB)
- id: str
- video_id: str
- text: str

Data Analysis: Semantic Clustering

1. Embed Video Descriptions
2. Reduce Embedding Dimension to 2 **(UMAP)**
3. Cluster Points by cosine similarity **(DBSCAN)**
4. Retrieve Transcriptions of Videos for each Cluster
5. Remove all words not part of medical keyword dataset **(MESH)**
6. Sort words by frequency
7. Ask LLM to infer topic label from top 15 words for each cluster


Psychotherapy
Orthopaedics
Nutrition
Skin
Infectious Diseases
Video Descriptions


Comments
Nutrition / Drug Use
Orthopedics
COVID
Clinic / Operation
Depression / Fear / Neur. Disorders
Age
Parenting
Opinion
Conspiracy
Emotion / Spirituality
Approval

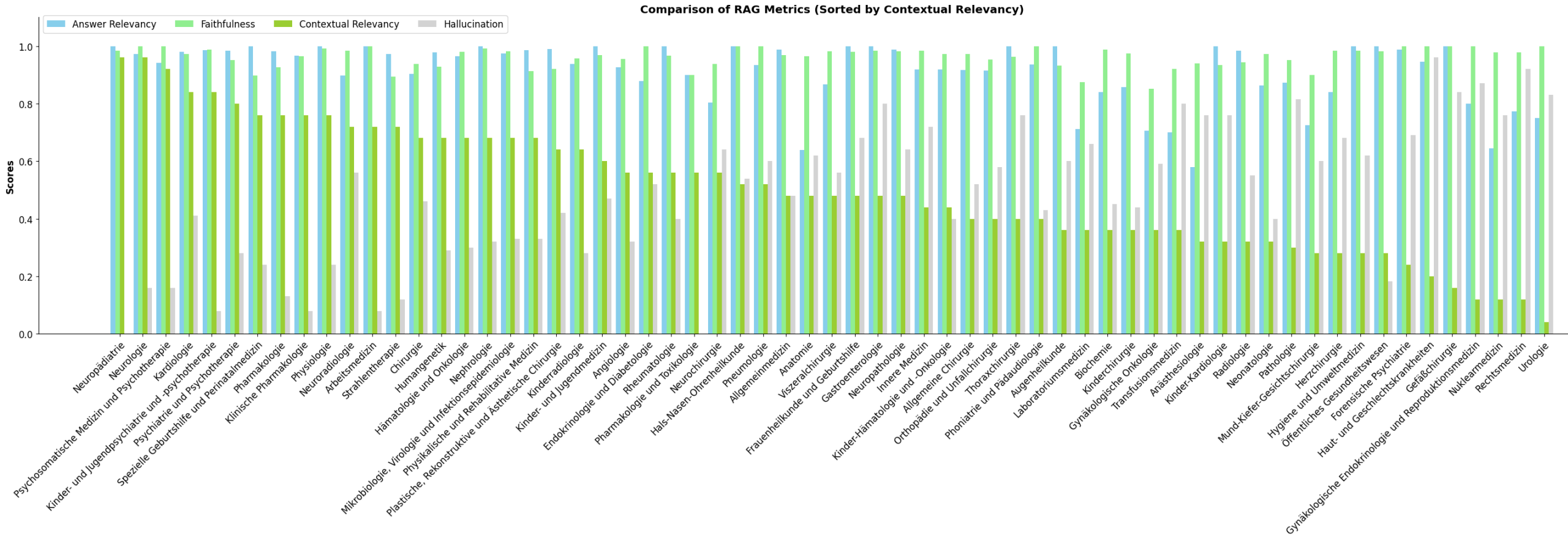
RAG System

1. Chunking of text
2. Embedding of chunks (en-de sentence transformer)
3. Persisting in Vector Storage (pinecone)
4. Retriever (top_k=20)
5. Reranking (en-de cross-encoder, top_k=5) - optional
6. Response Synthesizer (llm=Claude 3.5 Sonnet)

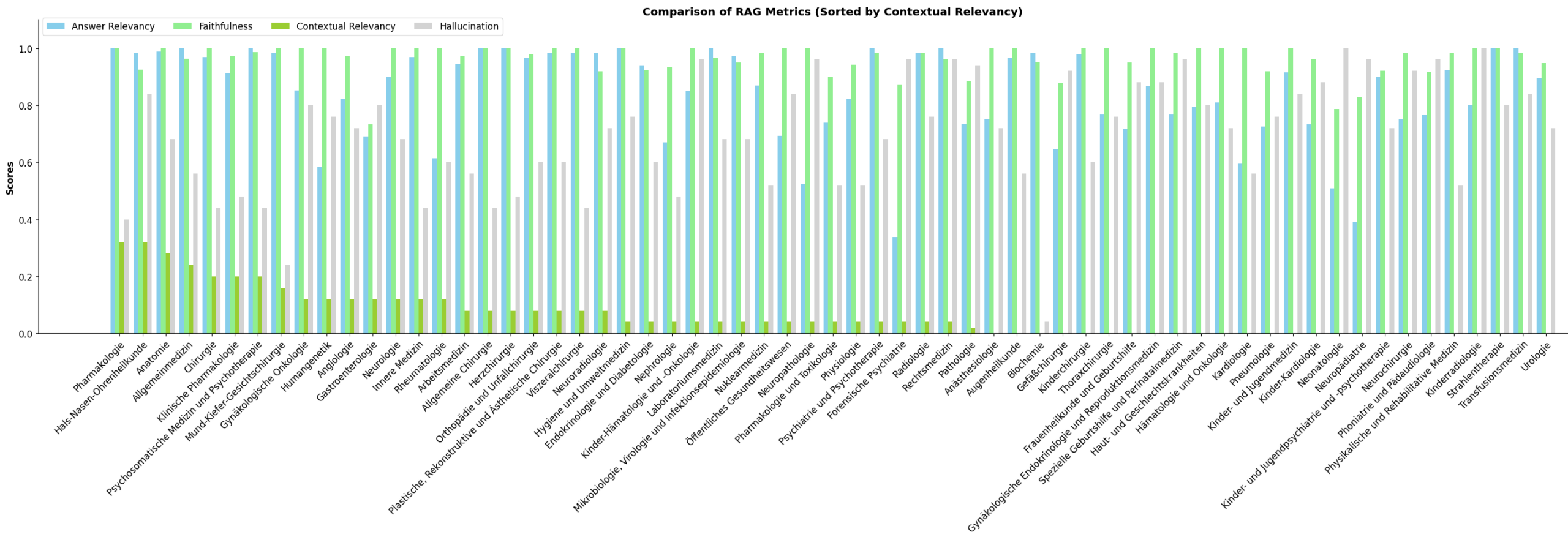
RAG Evaluation

1. Generate 5 Questions for each medical field using LLM
2. Generate answers using RAG system
3. Use „DeepEval“ to compute the following metrics from question, answer and context
   - Answer Relevancy
   - Faithfulness
   - Contextual Relevancy
   - Hallucination

Video-RAG Metrics (sorted by Contextual Relevancy)



Comment-RAG Metrics (sorted by Contextual Relevancy)



Comment Data Potentials

- Observation: RAG retrieves comments containing similar questions to user question
  - Mitigation: Reformat comments to be question/reply pairs

  $\Rightarrow$ no better results

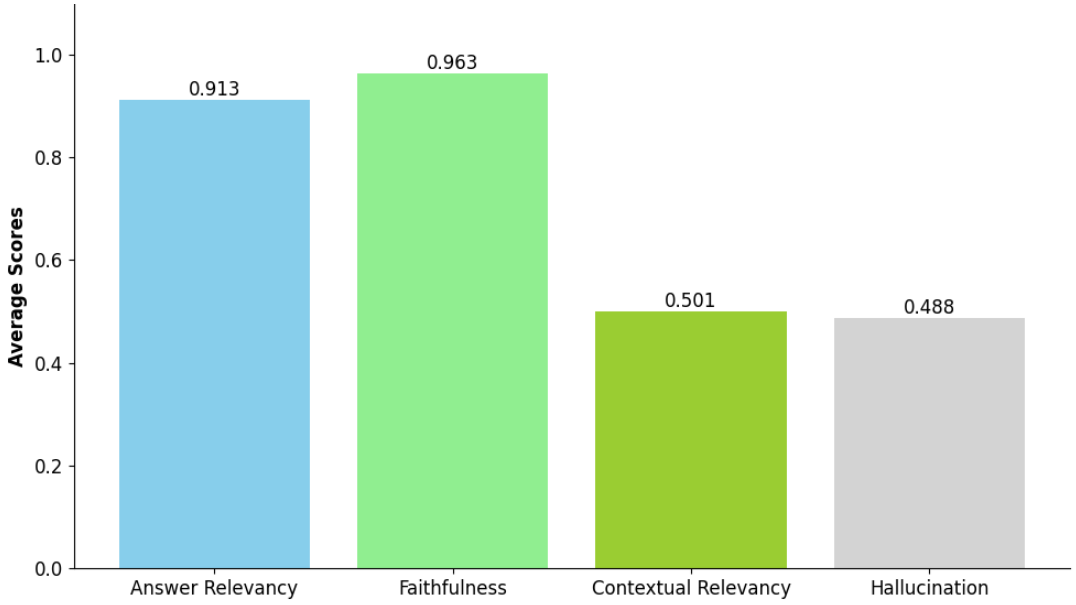
  $\Rightarrow$ comments do not contain required information

  $\Rightarrow$ comments not suitable for answering questions

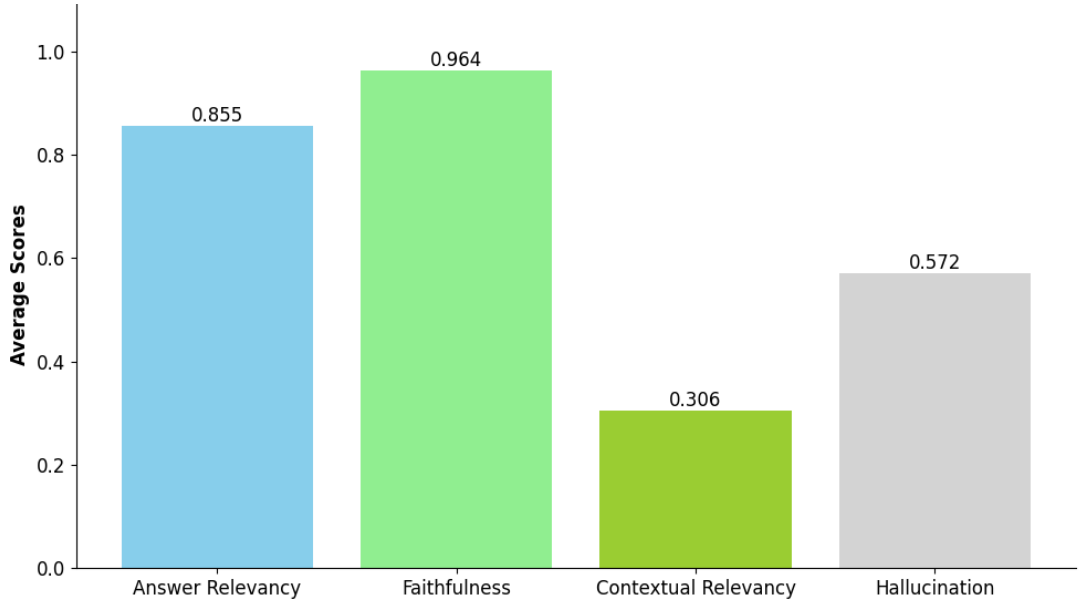
- However Comments are suitable for generating questions!
  1. Parse Comments to retrieve Questions
  2. Instruct LLM to generate high quality question using comment + video description
     - Necessary to generate self-contained questions
  3. Instruct LLM to rate resulting questions from 1 to 10
  4. Use questions rated at least 8

Evaluating RAG on Comment Questions

LLM Questions



Comment Questions



Conclusion

- YouTube Data can be scraped in large quantities
- Data contains diverse range of topics
- Video Transcriptions are suitable for Question Answering
- Video Comments are not suitable for Question Answering
- Video Comments can be used to generate high-quality Questions

Thank You!

Additional Information (Optional)

Examples „Dankbarkeit“

Dankbarkeit (c0)

- Danke
- Danke
- Danke
- …

Dankbarkeit (c91)

- Danke ❤
- Danke ❤
- Danke ❤
- …

Dankbarkeit (c3)

- Thank You
- Thank You
- Thank You
- …

Dankbarkeit (c110)

- Vielen Dank :)
- Super! 😊👍
- Vielen Dank :)
- …

Dankbarkeit (c5)

- Vielen Dank!
- Vielen Dank!
- Vielen Dank!
- …

Dankbarkeit (c531)

- Danke für das gute Feedback. Viele Grüße JB
- Vielen Dank für dein Feedback und gute Besserung. Liebe Grüße JB
- Vielen Dank für dein tolles Feedback! Toll 😊👍 dann weiterhin viel Erfolg :)…
- …

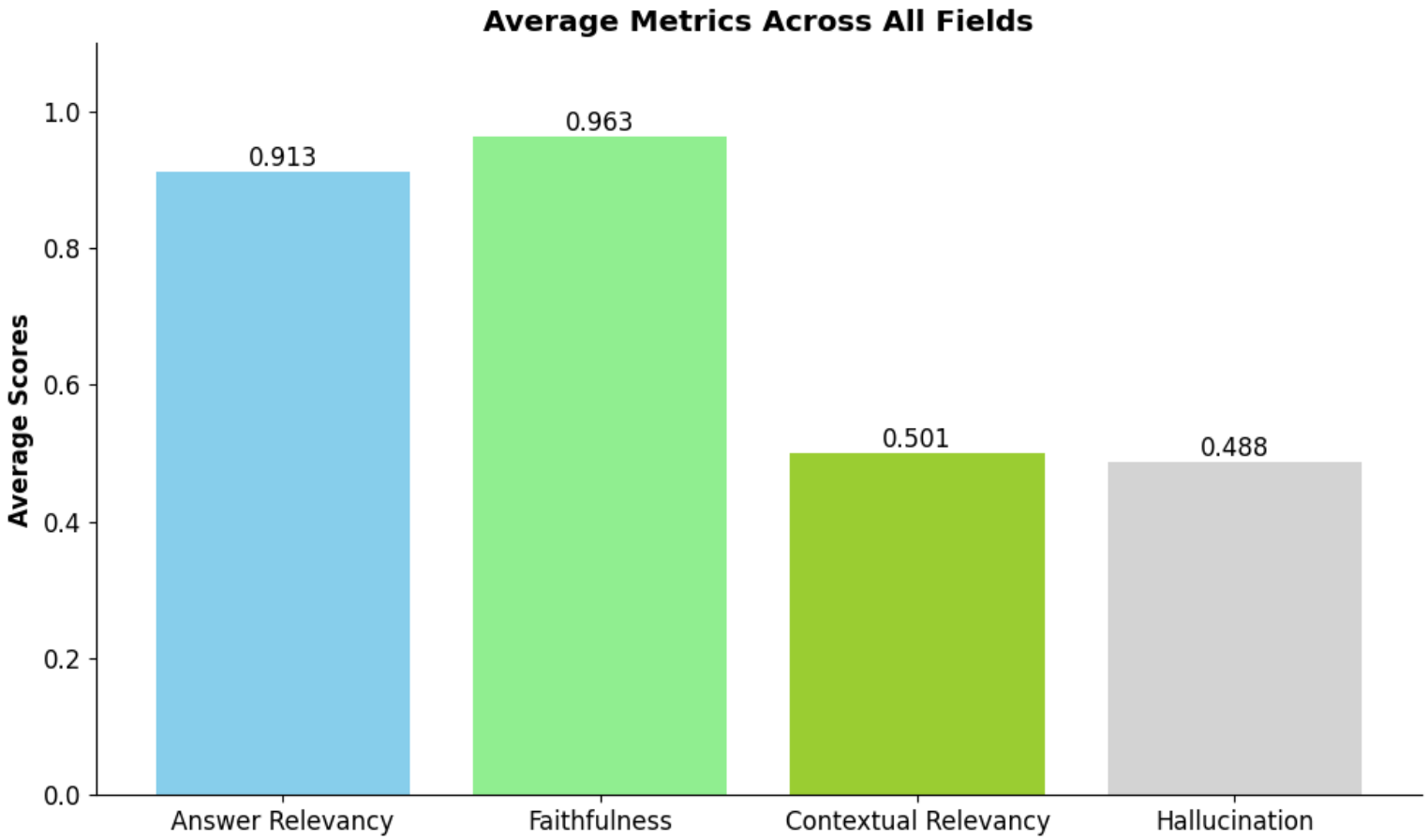
Examples „Unwissenheit“

- Immer wieder schockierend, obwohl man vieles davon schon wusste. Einige Infos kannte ich jedoch nicht.
- Hier bekommt man nie eine Antwort
- @michaela stehst bestimmt voll im nebel 🤣 nie gesehen und nie gehört
- I am suffering from the same…but no one can understand 😢
- Ich bekomme es nicht hin. Ich weiß nicht wie ich den beckenboden ansteuer. Bitte erkläre es noch einmal
- Diese Frage können wir aus der Ferne nicht beantworten und es gibt auch keine pauschale Antwort darauf. Tut uns leid.

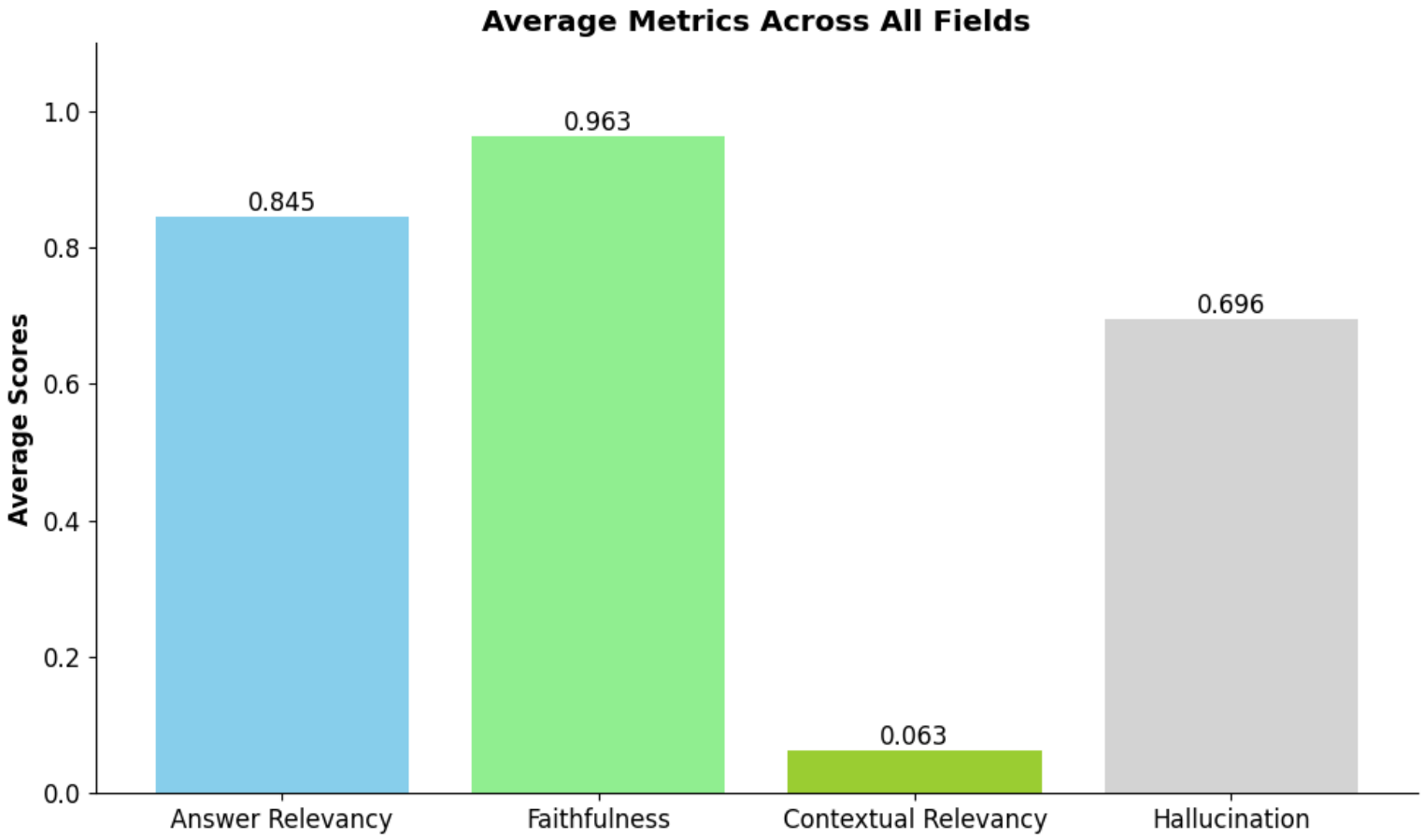
Examples „Meinung“

- … Ganz fern bleiben. Sie werden sich nicht ändern und sie manipulieren weiter und die Beziehung aufrechtzuerhalten ist schädlich.
- … Genau das sehe ich auch, immer mehr Menschen ganz gelb im Gesicht, zum Teil schon junge Menschen …
- These are the kind of people we should look up to. Not someone who puts a ball into a basket, hits a ball with a stick or plays make believe.
- … I don’t care what naysayers think about this behaviour. It works like a charm for me …

Video-RAG Metrics Averaged



Comment-RAG Metrics Averaged



Future Work

- Data preprocessing
  - Missing punctuations in many transcripts + errors in transcripts
  - Unicode + smileys in comments
  - Many duplicate phrases in comments
- Interesting Comment Cluster: Opinion
  - Further investigate this cluster
  - Map out the types of opinions formulated
- Can more data improve RAG performance?
- What are other approaches to improve context relevancy?
- Add a Question Answering system to the RAG system
- Analyze the complexity of the language (layman vs. professional)
- Use more questions per medical field to get a statistically more reliable result

Project Description

How can medical data from YouTube be used for ML applications?

1. Data Retrieval
2. Data Analysis
3. Retrieval Augmented Generation (RAG)
4. RAG Evaluation