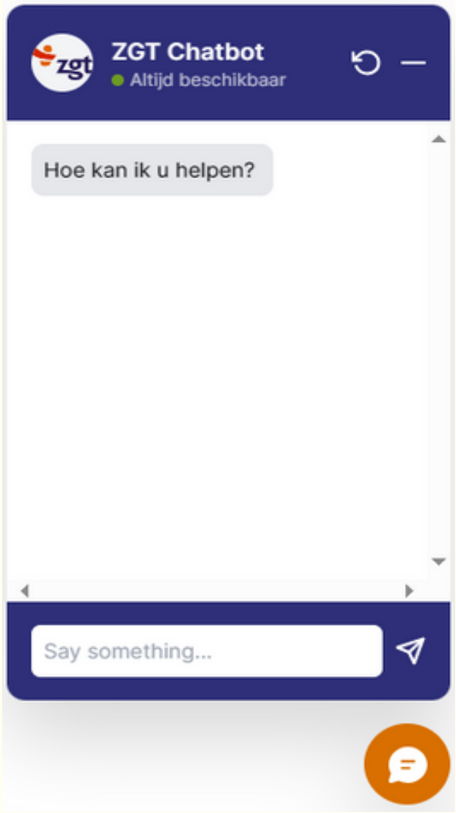
Topzorg voor uw levenskwaliteit

zgt

Automated Question-Answer Generation for Evaluating RAG-based Chatbots

ZGT Chatbot



ZGT chat for patients

MISSION

To answer general questions from the patients.

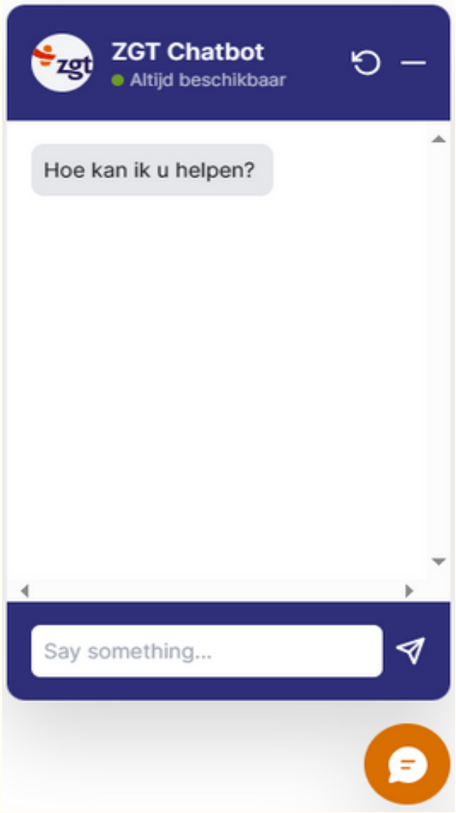
Examples:

- My insulin pump indicates low blood sugar, what now?
- What specialties does ZGT offer?

EVALUATION

Set of 19 questions

ZGT Chatbot



ZGT chat for patients

MISSION

To answer general questions from the patients.

Examples:

- My insulin pump indicates low blood sugar, what now?
- What specialties does ZGT offer?

EVALUATION ?

Set of 19 questions

Retrieval-augmented generation

ZGT chat for patients

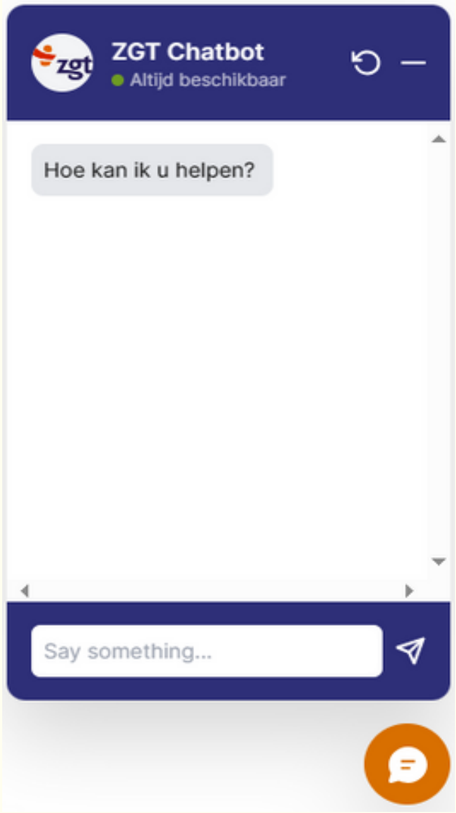
GPT-3.5-turbo

vector DB

~1300 flyers

question

ZGT Chatbot



ZGT chat for patients

MISSION

To answer general questions from the patients.

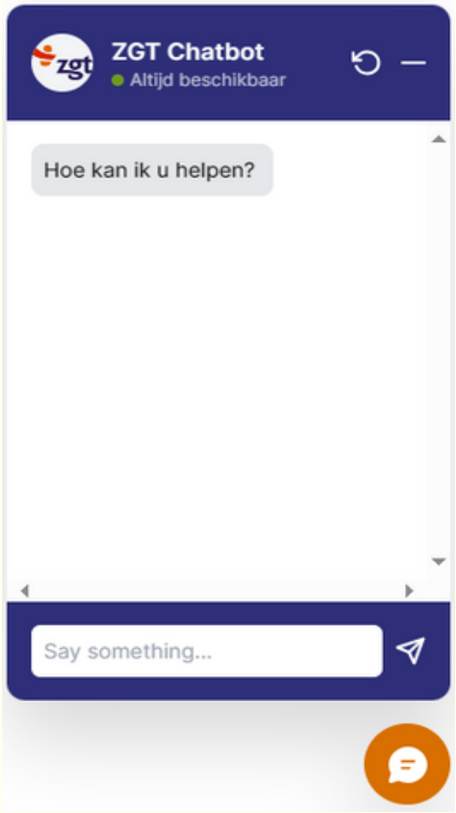
Examples:

- My insulin pump indicates low blood sugar, what now?
- What specialties does ZGT offer?

EVALUATION

Set of 19 questions

ZGT Chatbot



ZGT chat for patients

EVALUATION

Set of 19 questions

BUT...

- How does the chatbot deal with questions whose answer it does not know?
- Is the performance similar on factoid and long-answer questions?
- Is the large variety of topics on the documents covered in the questions set?

Problem statement

Currently, the system is evaluated with only 19 QA pairs

1. There’s a need to replace the LLM with an open source one
2. A statistically significant comparison of LLMs requires a large test set
3. The manual creation of a comprehensive QA set takes a lot of time

Research question

How to create an automated framework for testing RAG-based LLMs starting from a set of documents?

Proposed framework

Controles bij uw baby
Omdat er bij langdurig gebroken vliezen een verhoogde kans op infectie bestaat, wordt u samen met uw baby na de bevalling minimaal twaalf uur opgenomen op de afdeling moeder en kind. Ook komt de kinderarts langs om uw baby na te kijken. In overleg met de kinderarts wordt ontslag afgesproken.

Als uw baby mee naar huis mag, moet u contact opnemen met uw verloskundige als uw baby:

- Slecht drinkt;
- Regelmatig kreunt;
- Een te snelle of te langzame ademhaling heeft;
- Slap is;
- Een nare kleur heeft (grauw);
- Geen constante temperatuur tussen de 36.5º C en 37.5º C heeft, maar een sterk wisselende temperatuur.

Meer informatie vind u in de folder "Vroege infectie herkennen bij uw baby".

Extract sections from documents

Cluster sections into topics

Create folds for validation and test dataset

Select sections using stratified sampling

Generate and filter questions and answers

Gather human feedback on the data produced

Perform LLM cross-validation for n models

Get raking of LLMs

Topic Generation

Document

Filgotinib (Jyseleca®) bij IBD

Uw behandelend medisch specialist of verpleegkundig specialist heeft met u gesproken over de behandeling met filgotinib (Jyseleca®). In deze folder krijgt u informatie over de werking en het gebruik van dit medicijn en hoe u moet handelen in geval van bijwerkingen. Het is echter géén vervanging van de bijsluiter. Hebt u na het lezen van deze folder nog vragen, dan kunt u daarmee bij uw behandelend medisch specialist of verpleegkundig specialist terecht. 1

Algemeen
Tot nu toe is de oorzaak van Colitis Ulcerosa niet bekend. Elke behandeling met medicijnen is gericht op het onderdrukken van ontstekingsreacties. Dit leidt tot vermindering van klachten en verkleint de kans op complicaties die zich bij deze ziekte voor kunnen doen. De ziekte geneest niet door de behandeling met medicijnen. Na het afbouwen van of stoppen met medicijnen kunnen de klachten weer terugkomen.

U gaat met filgotinib starten omdat andere medicijnen niet of onvoldoende voor u hebben geholpen. Filgotinib behoort tot de groep van de JanusKinase remmers (JAK-remmers). JAK's zijn signaal-eiwitten binnenin de cel die betrokken zijn bij het ontstaan en onderhouden van ontstekingen. JAK-remmers zijn kleine moleculen die de activiteit van de Januskinasen verminderen. Op die manier doorbreken ze de vicieuze cirkel van het in stand houden van de ontstekingen in de darm bij Colitis Ulcerosa. 2

Effect van filgotinib
Het effect is binnen een aantal weken tot maximaal 16 weken te verwachten. 3

Voorzorgsmaatregelen
Informeer uw behandelend medisch specialist altijd indien er sprake is van één of meer van onderstaande omstandigheden.
- U heeft een infectie op één plaats op uw lichaam (zoals een zweer op uw been).
- U heeft een infectie in uw hele lichaam (zoals griep).

4

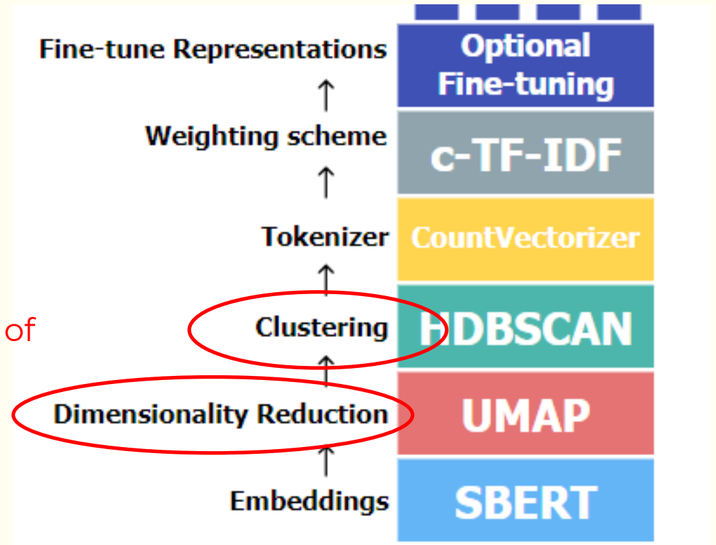
- U heeft een verhoogde vatbaarheid op infecties (bijvoorbeeld door diabetes mellitus, chronische longziekten).
- U heeft een langdurige of steeds terugkerende infectie (bijvoorbeeld koortsblaasjes, genitale herpes of gordelroos).
- U heeft een hoog cholesterol waarde.
- U had in het verleden een besmetting met tuberculose.
- U heeft in de afgelopen vier weken een levend vaccin gekregen.
- U heeft een ernstige leveraandoening.
- U heeft een galactose-intolerantie, lactasedeficiëntie of een glucose-galactosemalabsorptie.
- U heeft in het verleden een vorm van kanker gehad.
- U heeft in het verleden een diep veneuze trombose of een longembolie gehad.
- U gebruikt andere medicatie. Neem altijd een lijstje mee naar uw behandelend medisch specialist of verpleegkundig specialist met daarop de medicijnen die u gebruikt.

4

Voorbereiding start filgotinib
Omdat filgotinib het afweersysteem onderdrukt, kunnen bepaalde ziekten plotseling actief worden, bijvoorbeeld tuberculose (TBC) en hepatitis (besmettelijke leverontsteking). Je kunt tuberculose en hepatitis meedragen zonder dat je hiervan klachten hebt. Bij alle patiënten wordt voor de start met filgotinib gekeken of er een vroegere besmetting met tuberculose heeft plaats gevonden en een eventuele besmetting wordt uitgesloten. Dit wordt getest via een bloedafname. 5

Topic and questions generation

- 13,216 sections
- 11,292 unique sections
- BERTopic



Number of clusters

Embeddings dimensionality

Topics generation

	How many dimensions?	How many clusters?
Too many?		
Too few?		

Topics generation

Hyperparameters fine tuning

Geometric evaluation

Silhouette coefficient

Calinski-Harabasz coefficient

Davies-Bouldin coefficient

Robustness evaluation

Adjusted rand score

Adjusted mutual information score

Normalized mutual information score

Document-cluster evaluation

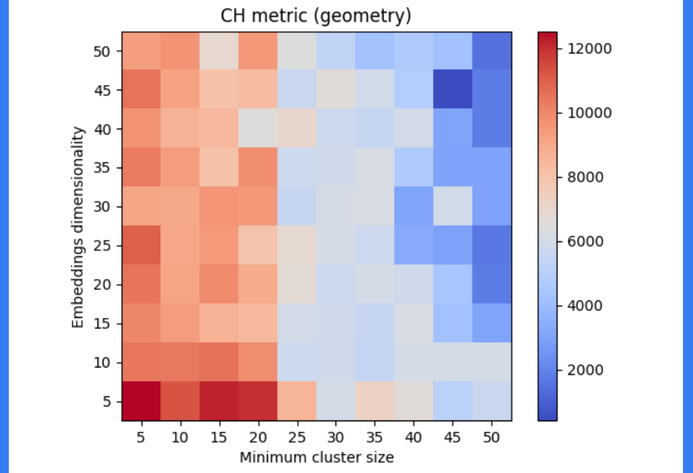
Hypothesis: the majority of sections from one given document should be clustered together

Topics generation

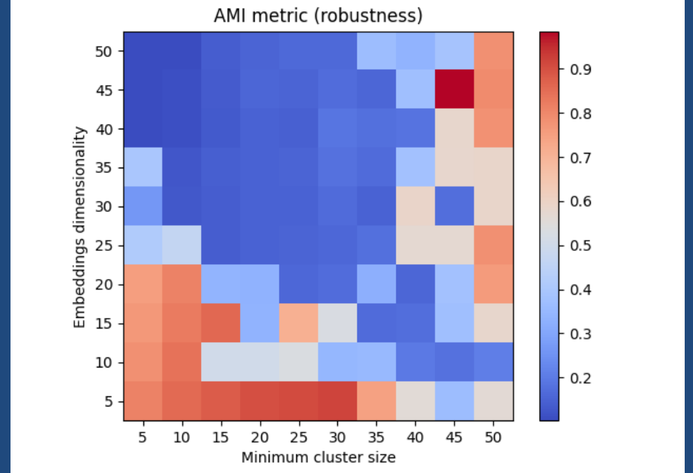
Hyperperameter choice:

- Minimum cluster size: 30
- Embedding dimension: 15

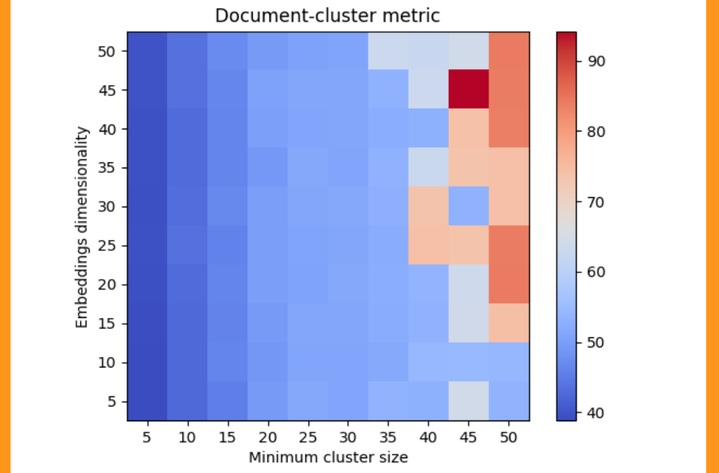
Geometric evaluation (Calinski-Harabasz index)



Robustness evaluation (Adjusted Mutual Info)



Document-cluster evaluation



Topic analysis

Topics

Two sets of topics created using BERTopic configured with different types of embeddings

Characteristics

- compact clusters of sections
- sections represented as embeddings with 15 dimensions

statistic	DBMC-v1	PMM-L12-v2
min topic size	32	30
median topic size	66	79
max topic size	634	1063
topics no	71	62
outliers %	40.55%	32.17%

Question-answer types

Domain

General (working hours, prices, etc.)
Medical (symptoms, medical interventions, etc.)

Answerable

Yes
No (flyers that contain the info are not provided)

Answer

Factoid
Long form

+

Metrics

Quantitative evaluation

Hallucination:
% of questions answered without available info

Factoid answers:
% of correct answers

Long-form answers:
BLEU, ROUGE, BLEURT

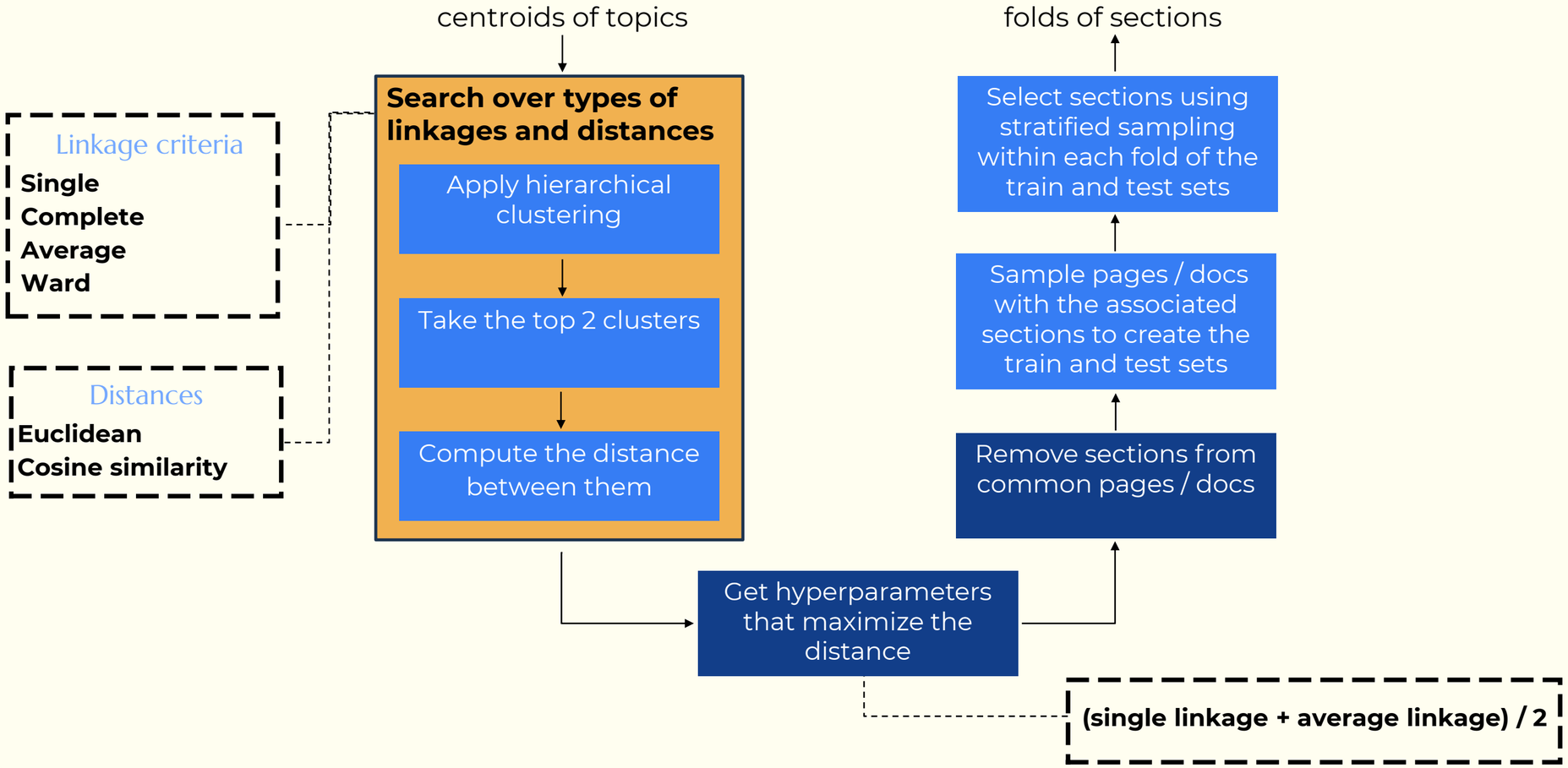
Qualitative evaluation

Human judgment
Performed by professionals who work for ZGT

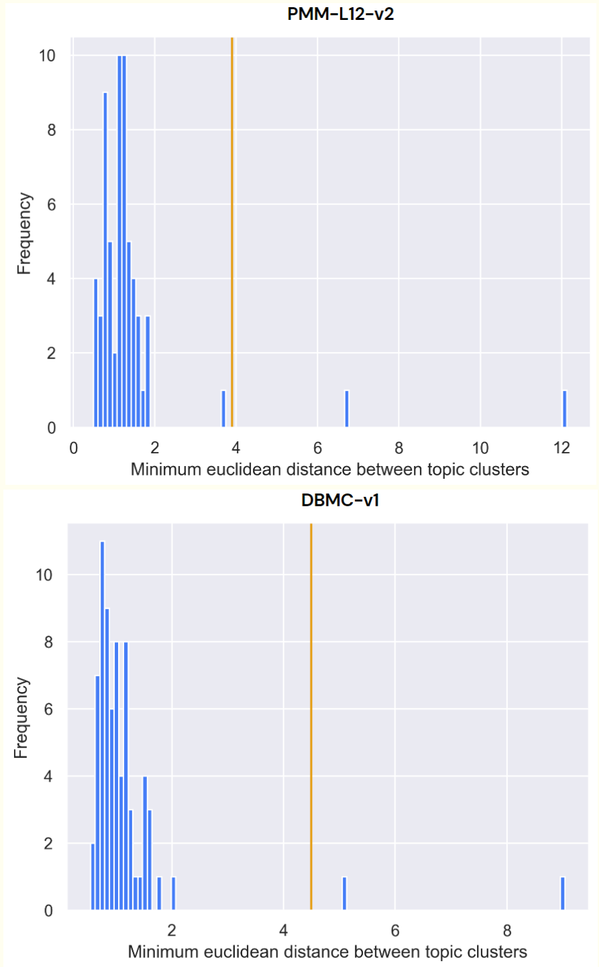
Folds creation and sections sampling +

Goals: create two validation folds and two test folds

minimize the probability of having overlapping info in any pair of folds



Folds creation and sections sampling: results



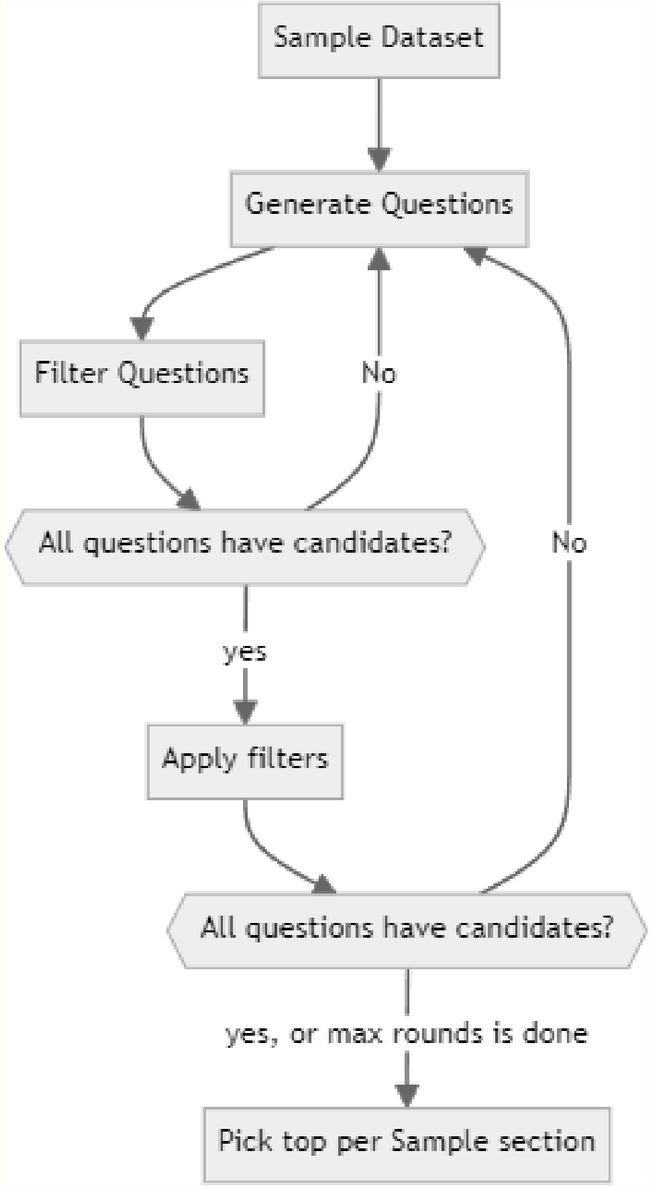
Minimum distance to any other cluster computed with average linkage

set of topics	DBMC-v1	DBMC-v1	DBMC-v1	PMM-L12-v2
linkage criterion	ward	complete	complete	ward
distance type	euclidean	cosine	euclidean	euclidean
min dist topics	0.2783	0.2783	0.2783	0.2481
min dist folds	2.2882	1.923	1.8662	1.9548
small fold ratio	0.1175	0.3741	0.1574	0.4727
avg folds per doc	1.2104	1.3285	1.1208	1.3399
valid sections %	72.12%	55.84%	84.22%	55.28%

Results of the hierarchical clustering

Validation – test folds split: 80% – 20%
Sampled sections: val fold 1 – 200 + val fold 2- 200
test fold 1 - 50 + test fold 2 - 50

Generate and filter questions



Filter Sections

Similarity between Questions

- Vector Similarity
- Should be different

Similarity between source and answer

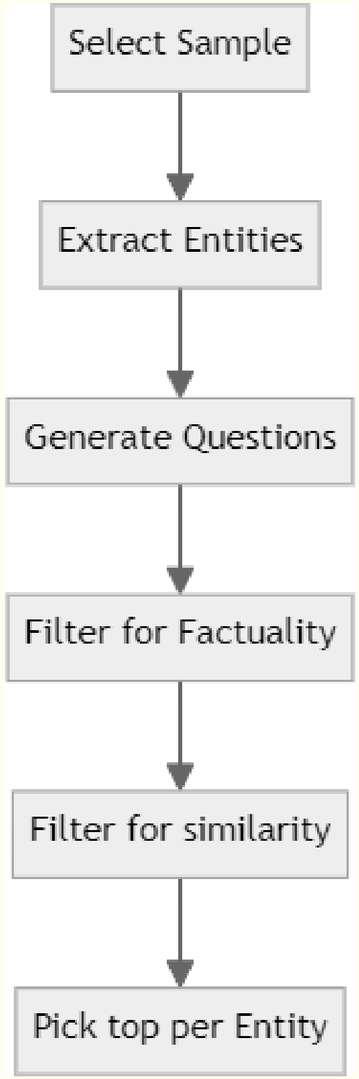
- Should be similar
- ROUGE score

LLM Scoring

Remove “Fact” questions

- Answer short, only containing “entity”

Fact questions



To check hallucination of the chatbot

Extracting Phone numbers, emails

- Verify exact match

Create targeted question

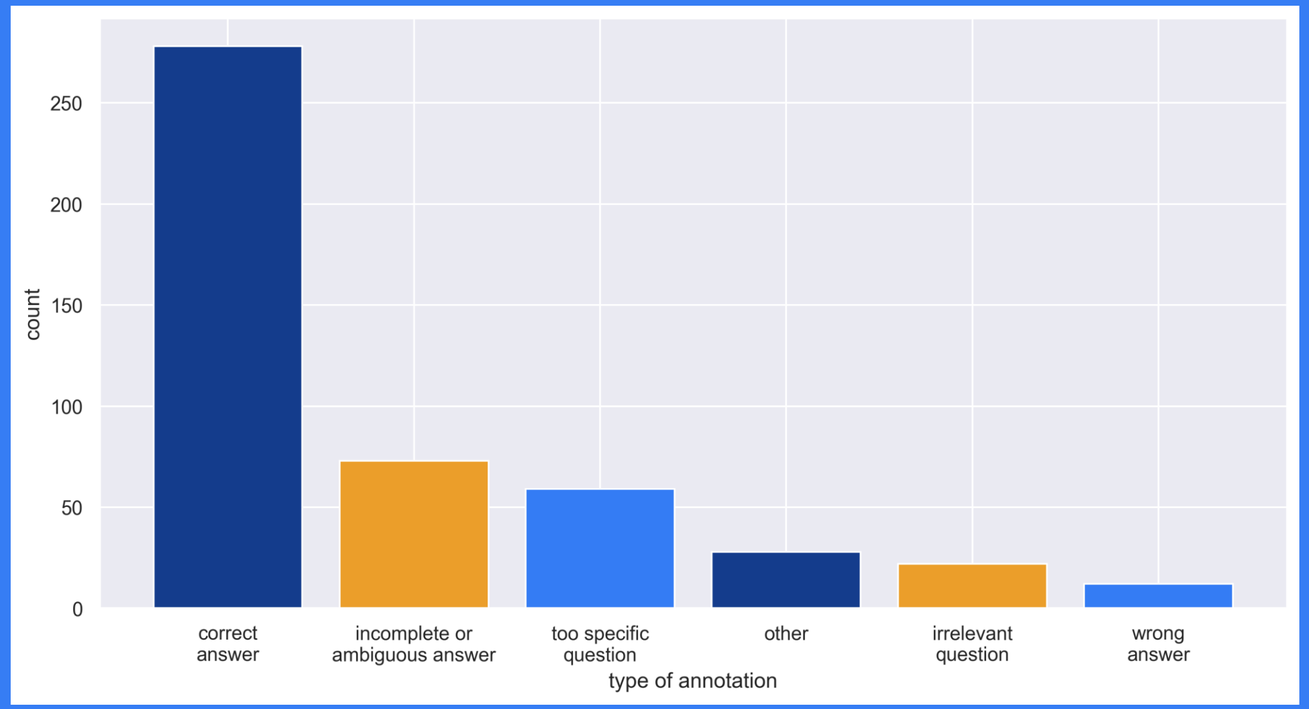
Human evaluation

Number of answers

1 question without any answer
404 questions with only one answer
85 questions with two answers

Agreement between annotators

Full agreement in 64.71% of cases
Partial agreement in 15.29% of cases



Distribution of answers

How can I apply for insulin pump therapy (Patch pump)?

Where should I report when I arrive at the hospital on the day of admission?

What are the preparations at home that should be done for the schisis, lip closure?

How much liquid is there in 10 glasses or 12 cups?

- Too specific

How can I prevent adverse effects on mother and child?

- Not specific enough

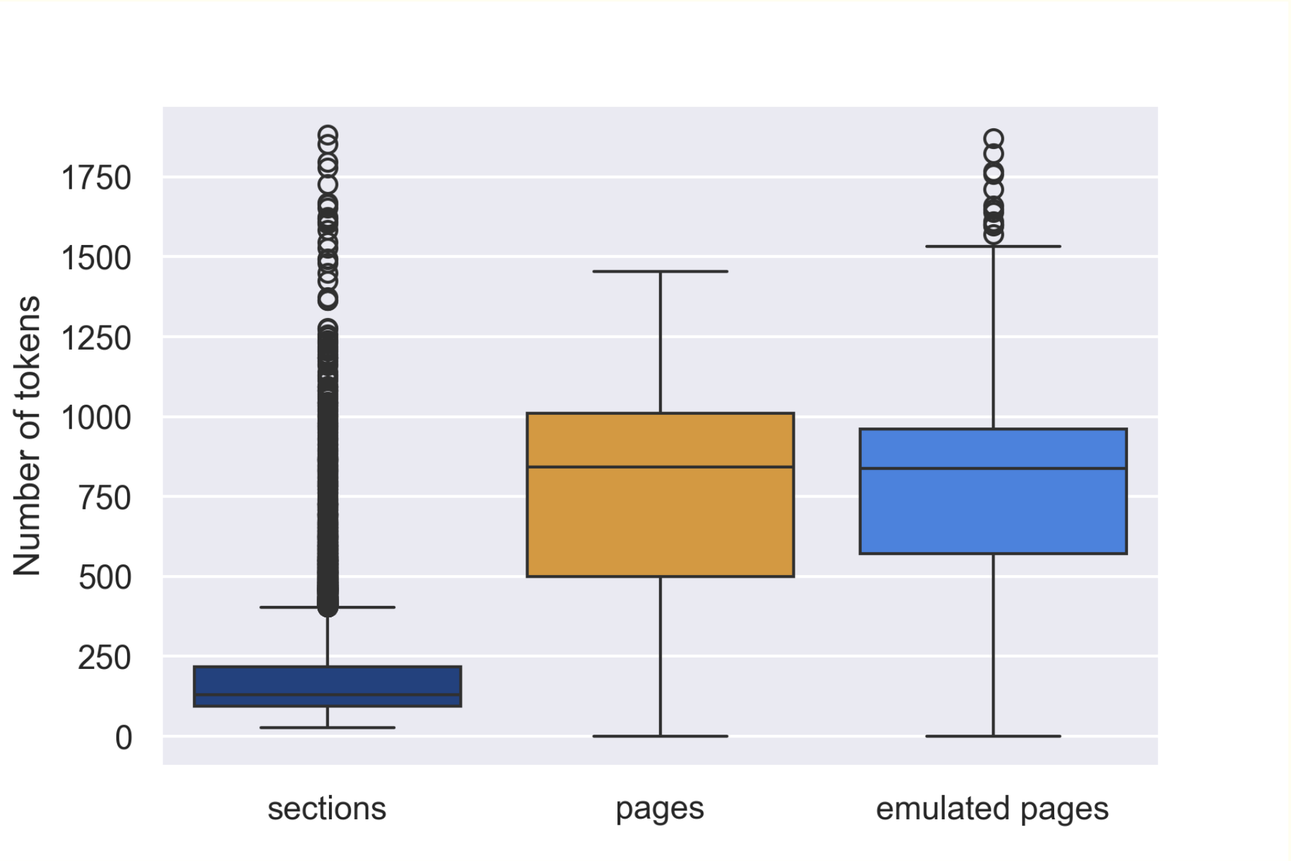
Can the Methotrexate_Metoject be used with another syringe or pen? --- No, the Methotrexate_Metoject must be used with the specially filled syringe or pen.

- Answer contains assumption

Sections and RAG pages

- Short sections for Question generation
- Chatbot works on "full pages" of the document
  - May split sections between pages

- Emulating pages from sections
  - Appending sections while under page length limit
  - This can be used for RAG in the evaluation
  - Making sure it does not violate the fold generation



Cross-validation

do this for each fold:

Include in RAG only flyers associated with answerable questions

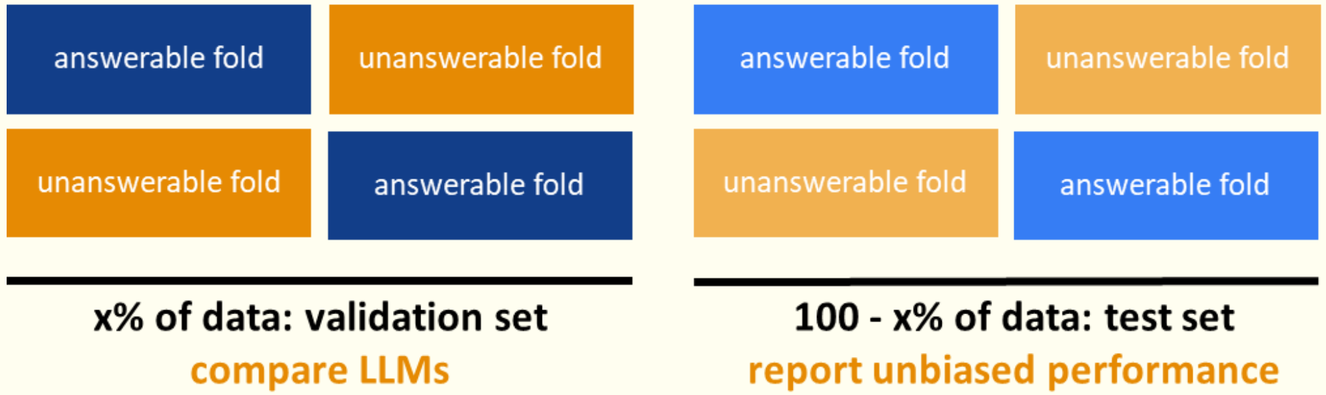
Ask answerable and unanswerable questions w.r.t. current folds

Measure metrics

after CV:

Compute average metrics

folds are sets of flyers that are perfectly separated from the topic's perspective



+

Metrics

Quantitative evaluation

Hallucination:

% of questions answered without available info

Factoid answers:

% of correct answers

Long-form answers:

BLEU, ROUGE, BLEURT

Qualitative evaluation

Human judgment

Performed by professionals who work for ZGT

Conclusion

- Created a framework for creating dataset for comparing LLMs Based on a set of documents.
- 90%+ Relevant questions
- 60%+ Correct answers

Thank you for your attention!

Questions?

Outliers and number of clusters