No more searches, just verifiable answers!

Verif.ai: Towards an Open-Source Scientific Generative Question-Answering System with Referenced and Verifiable Answers

CHALLENGES WE ARE ADDRESSONG

- Generative language models (like GPT) transform the way we use information
- These models mimic human-like understanding
- LLMs still tend to hallucinate - undermining trust in AI
- The specifics and sensitive fields, like biomedicine, require factual information
- Reliability and accuracy in scientific application are of great importance
- Our system will produce scientifically accurate, reference-backed answers

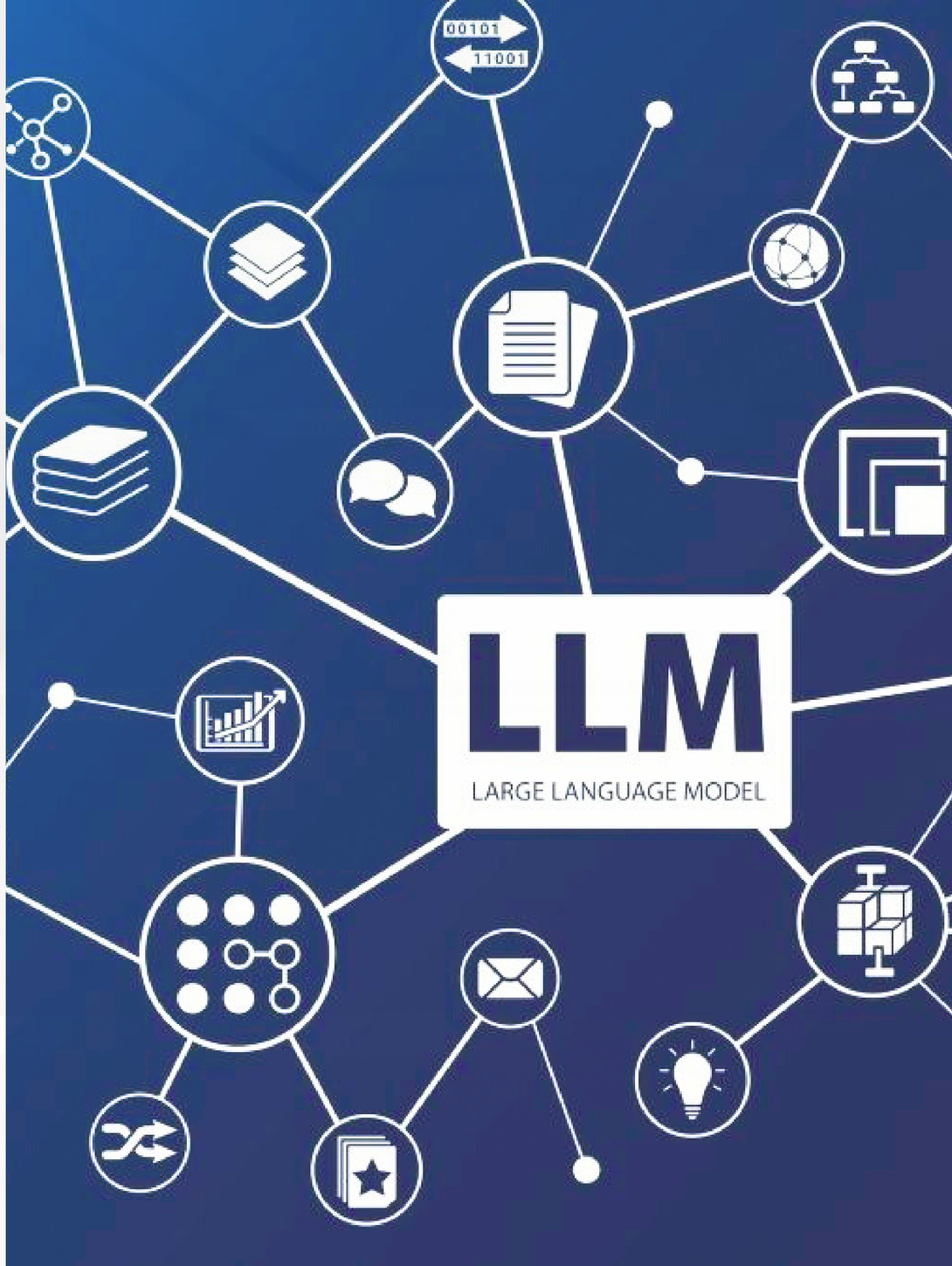
WHAT ARE WE BUILDING

Generative question-answering engine:

- focused on BioMedicine
- open-source
- leveraging RAG and fine-tuned models for accurate, referenced and verifiable answers

The aim of this project is:

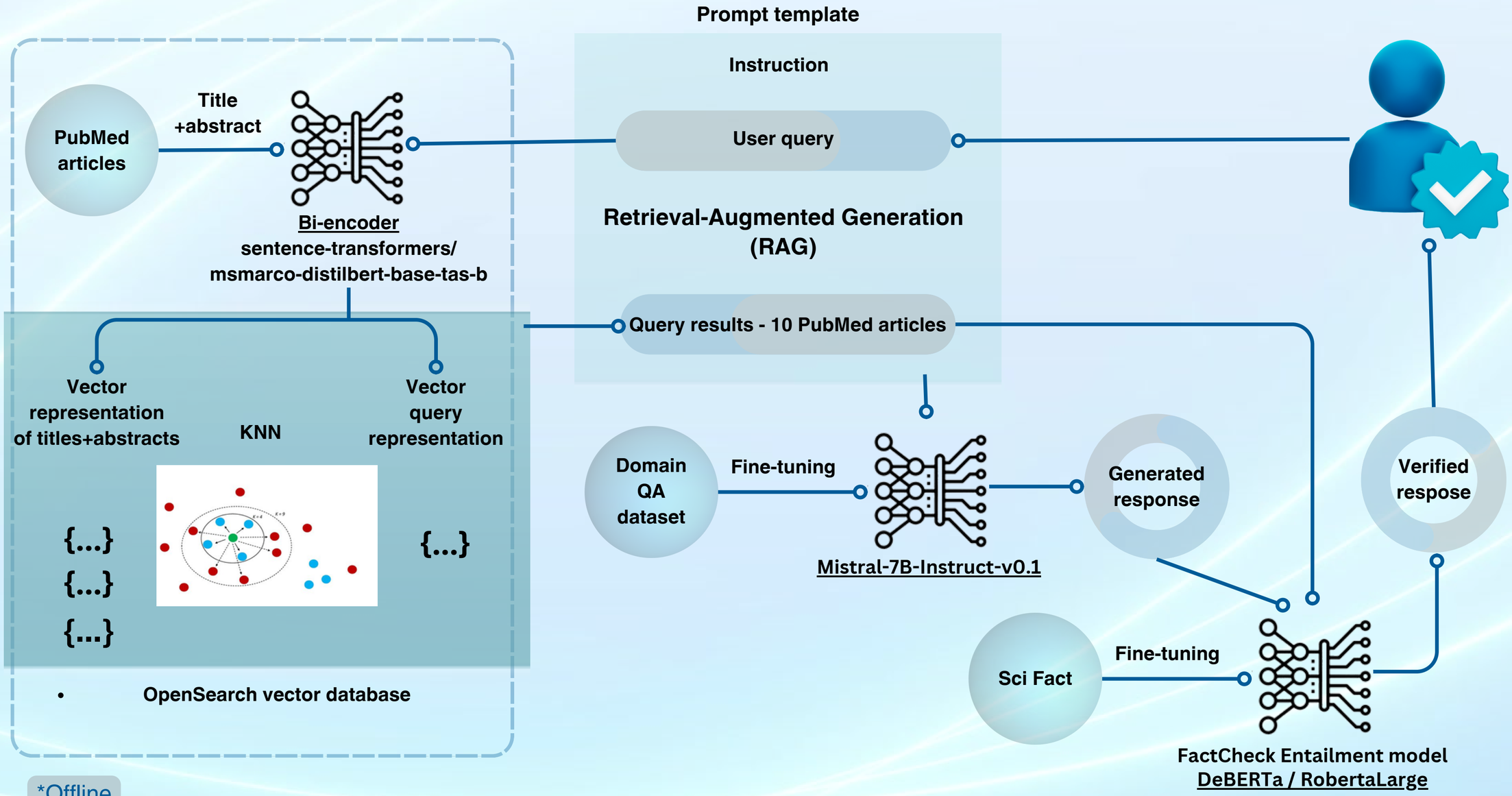
- enhance trust in generative LLM by
- reduce hallucination in generative model (using RAG)
- provide verifiable information (referencing and claim checking)

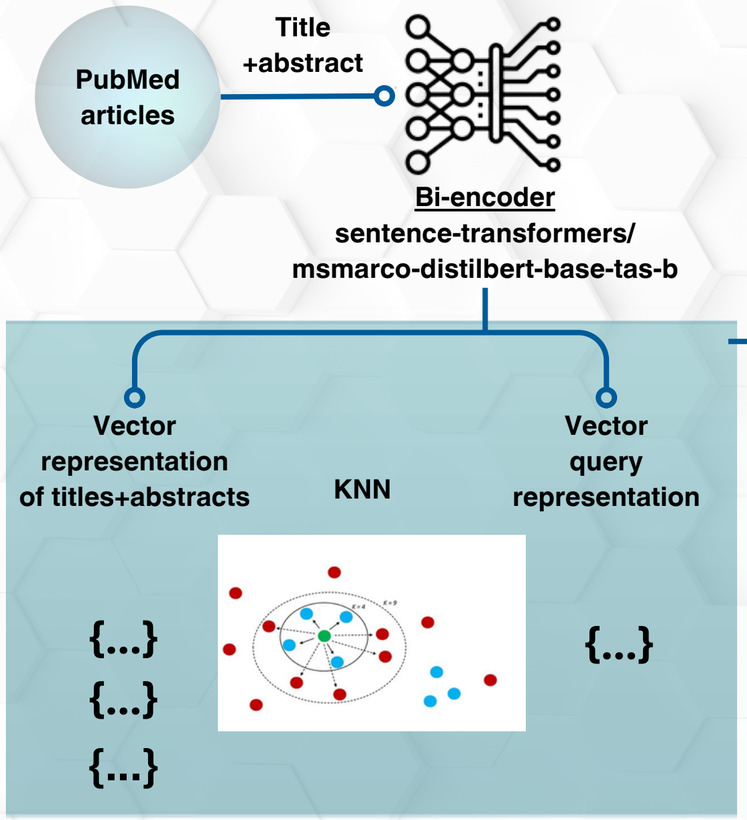


PROJECT COMPONENTS

- **Information retrieval system** combining semantic and lexical search techniques over scientific papers (PubMed)
- **RAG using fine-tuned generative model (Mistral 7B)** taking top answers from the information retrieval system and generating answers with references to the papers from which the claim was derived
- **Verification engine** that cross-checks the generated claim and the abstract or paper from which the claim was derived, verifying whether there may have been any hallucinations in generating the claim

SYSTEM ARHITECTURE





INFORMATION RETRIEVAL - SEMANTIC/HYBRID SEARCH

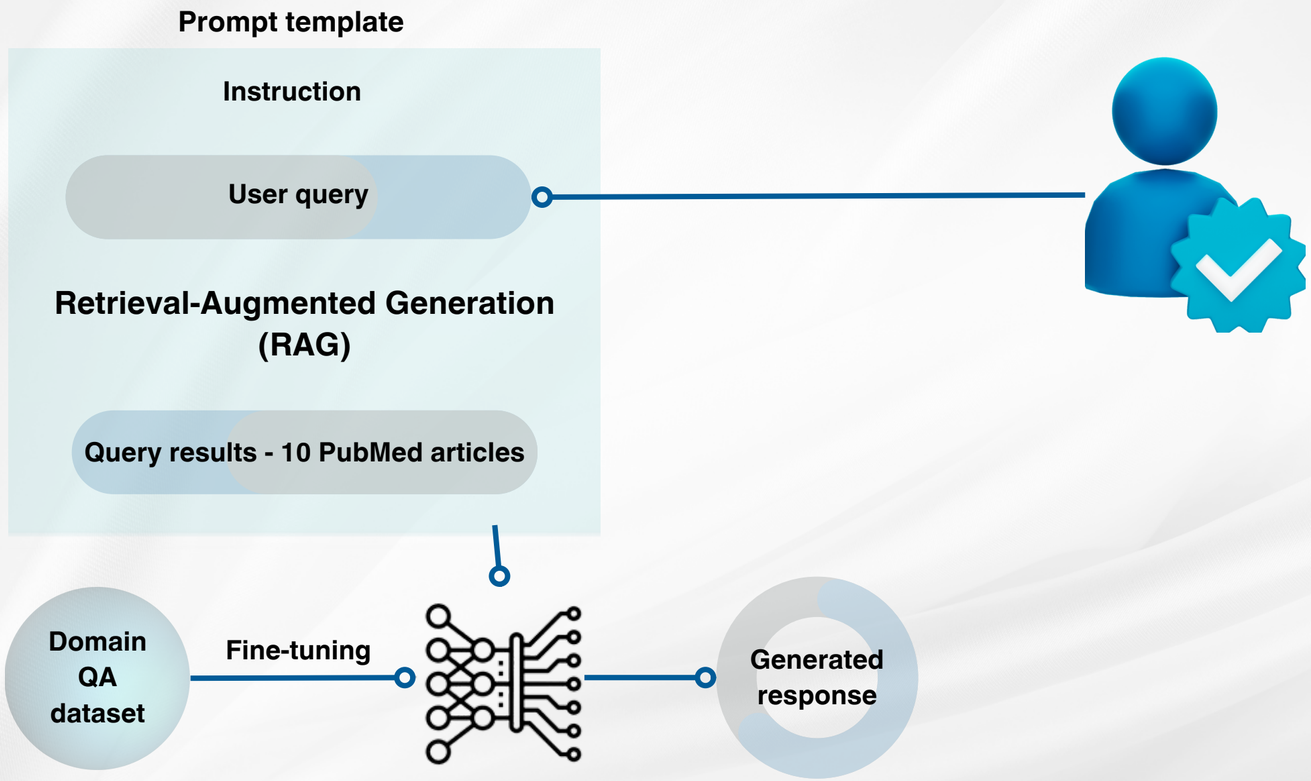
OpenSearch vector database

- Vector representations of tile+abstracts
- Sentence transformer model for embeddings
- knn_vector
- Lexical and semantic search
- FAISS, HNSW, ineerproduct
- One node in the cluster, 8 shards

The output of this component is used for:

- input to the generative model template (RAG technique)
- fact-checking using the textual entailment model

RAG COMPONENT



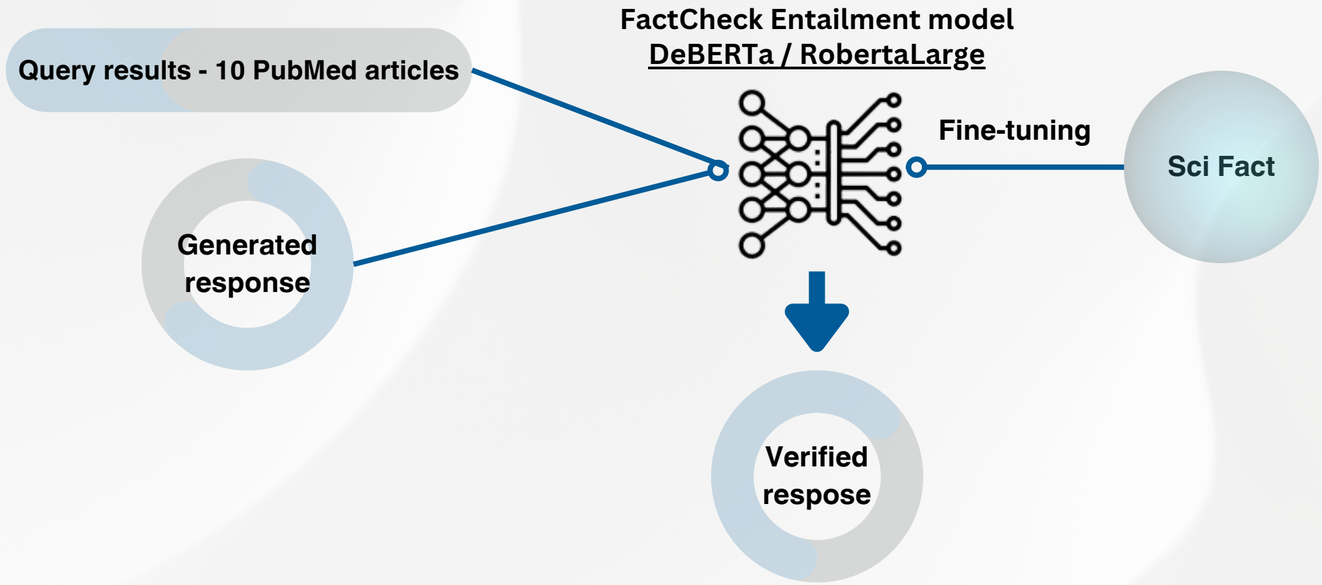
[INST] Answer the question using the given abstracts. Reference claims with the relevant abstract id in brackets (e.g. (PUBMED:123456) at the end of the sentence). Answer may contain references to many abstracts. Be as factual as possible and always use references in brackets. Use exclusively provided abstracts and their ids. Make answer look similar to the following: Several genes play role in breast cancer. For example BRAC1, BRAC2 are well studied targets (PUBMED:554433). The other targets involve IRAK4, CAS2 and HMPA (PUBMED:665544).

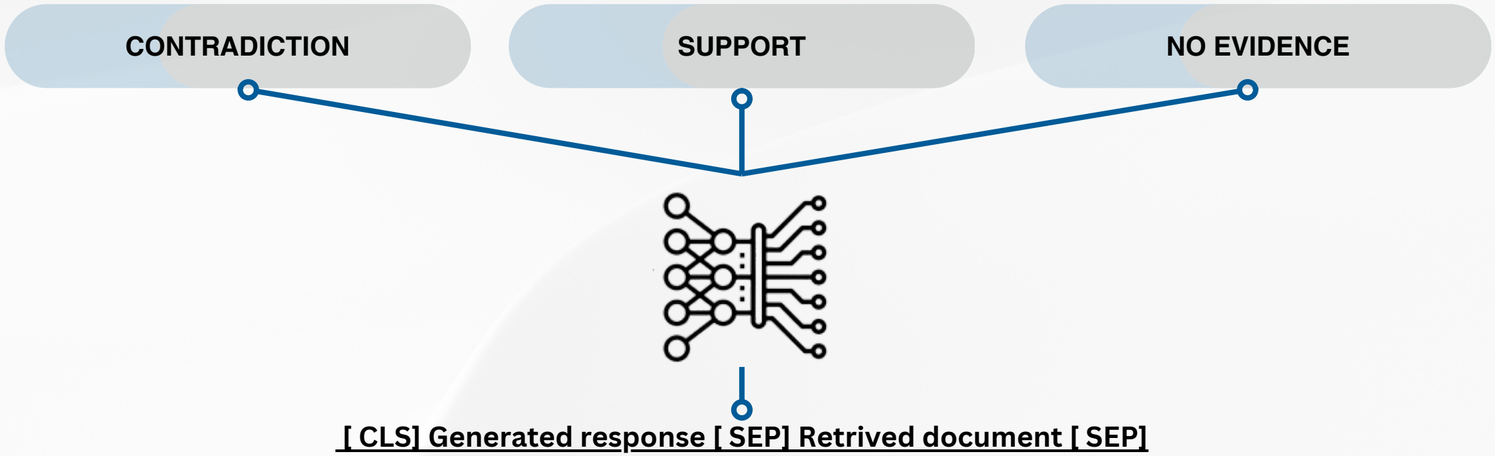
User query: Central corneal thickness: will one measurement suffice?
Semantic query result:
[0]
Corrected Intraocular Pressure Variability with Central Corneal Thickness Measurement
Purpose: To evaluate variability in measured intraocular pressure (IOP) values...
[1]
...
[2]
...
[3]
Effects of diabetic keratopathy on corneal optical density, central corneal thickness, and corneal endothelial cell counts.
Diabetic keratopathy is an ocular complication that occurs with diabetes...

,,,

Mistral response: The provided papers suggest that measuring central corneal thickness (CCT) using either ultrasonic or optical tools is sufficient for evaluating intraocular pressure (IOP) [0]. However, … Additionally, diabetic keratopathy can affect CCT, and medial and intimal corneal optical density and central corneal thickness are sensitive indicators for early diabetic keratopathy [3]. Overall, while one measurement of CCT may suffice for evaluating IOP, the choice of method should be carefully considered based on the specific application and the desired level of accuracy.

VERIFICATION - TEXTUAL ENTAILMENT MODEL





EVALUATION OF RETRIVAL SYSTEM

- Lexical search is most effective for direct, exact term matches in documents.
- Semantic search is adept at identifying documents with paraphrased text or synonymous terms, enhancing the search's flexibility.
- Hybrid search leverages the strengths of both lexical and semantic searches, effectively finding documents with both exact matches and similar meanings. It uniquely prioritizes documents with exact term matches at the top of search results, offering a balanced approach.

EVALUATION OF RAG COMPONENT

We compared the results of our model and GPT-3.5 and GPT-4 models on a test set of 50 questions and extracted abstracts.

Conclusions are:

- Our model is **comparable to those of much larger GPT-3.5 and GPT-4 models** for the referenced question-answering task.
- N**o model showed a clear advantage over the others.**
- The quality, referenced abstracts, and length of the answers varied within each model and among the models.
- Most of the time all three models **referenced the same abstracts** as relevant

EVALUATION OF VERIFICATION COMPONENET

- Bert-based models (XLM-Roberta-Large and DeBERTa) fine-tuned on SciFact dataset
- Fine-tuned on natural language inference task
- 10% of the dataset used for evaluation (80% training, 10% validation)
- We also evaluated the SciFact label prediction task using the GPT-4 model, resulting in a precision of 0.81, recall of 0.80, and an F-1 score of 0.79.

TABLE I
THE EVALUATION OF THE ENTAILMENT MODEL FINE-TUNED FROM XLM-ROBERTA-LARGE AND DEBERTA-LARGE MODEL USING SCIFACT DATASET

	XLM-RoBERTa			DeBERTa		
	Precision	Recall	F1-score	Precision	Recall	F1-score
NO_EVIDENCE	0.91	0.96	0.95	0.88	0.86	0.87
SUPPORT	0.91	0.75	0.82	0.87	0.92	0.90
CONTRADICT	0.59	0.81	0.68	0.88	0.81	0.85
Weighted Avg	0.87	0.85	0.85	0.88	0.88	0.88

SOTA, surpassing the reported scores in

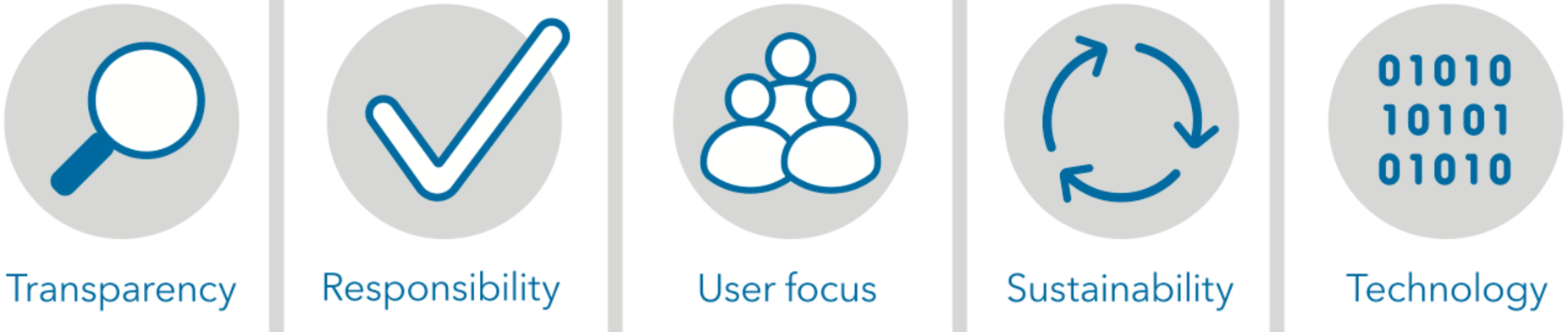
[Fact or Fiction: Verifying Scientific Claims] (https://aclanthology.org/2020.emnlp-main.609) (Wadden et al., EMNLP 2020)

WE HOPE TO

- Provide methodology for better generative search
- Provide verifiability score/metric for each statement/source document
- Increase trust in LLMs
- Reduce misinterpretation of the answer
- Limit the spread of misinformation
- Initially scope is life science domain

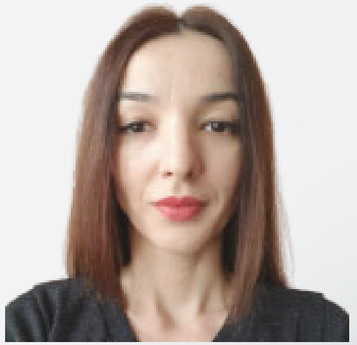
FUTURE PLANS

- Broaden Verif.ai's mission to enhance trust in generative AI across multiple scientific areas, ensuring reliable, verifiable information
- Expand the project's reach to explore new domains beyond biomedicine
- Incorporate scientific community feedback to continuously improve our systems, addressing evolving research requirements and information accuracy



OUR TEAM

TEAM LEAD Dr Nikola Milošević



Dr Adela Ljajić

Dr Bojana Bašaragin

Lorenzo Cassano
Intern

Miloš Košprdić

Darija Medvecki

BAYER

+

THE INSTITUTE FOR ARTIFICIAL INTELLIGENCE
RESEARCH AND DEVELOPMENT OF SERBIA

Angela Pupovac
Intern

www.verifai-project.com

verif.ai.project@gmail.com

LinkedIn
Verif.ai Project

Instagram
verif.ai_project

X / Twitter
Verif.ai Project

Facebook
Verif.ai Project

TikTok
verif.ai

THANK YOU FOR YOUR ATTENTION