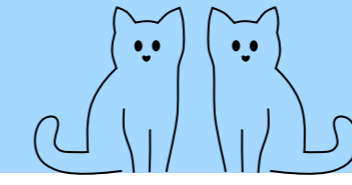


# Document Retrieval with Fine-grained Relevance Cues

Autor: Bc. Antonín Jarolím Supervisor: Ing. Martin Fajčík Ph.D.



## Ever disappointed by Google's highlighting?

**Google** moon phases reason

The Planetary Society  
https://www.planetary.org › Articles

**The phases of the Moon explained**

25 Apr 2023 — The phases of the Moon are caused by the changing positions of the Moon, Earth, and the Sun. As the Moon goes around the Earth, different parts

BBC  
https://www.bbc.co.uk › bitesize › articles

**Phases of the Moon - BBC Bitesize**

Learn about what causes the phases of the Moon it orbits the Earth and what a lunar month is with this KS3 Physics guide for BBC Bitesize.

**treatment of varicose veins in legs**

ca yen ne pepper . ca yen ne pepper is considered a miracle treatment for varicose veins . being a very rich source of vitamin e and bio flavonoids , it increases blood circulation and eases the pain of congested , swollen veins . add one tea spoon of ca yen ne pepper powder to a cup of hot water and stir it well .

**what is priority pass**

priority pass . priority pass is an independent airport lounge access program membership provides you with access to their network of over 700 lounge s these lounge s are managed by a variety of airlines and companies , meaning that significant variation among them a but great international coverage .

### Fine-Grained Cues Motivation

- highlight => get information faster
- lowering hallucination in RAG
- token-cues without calling LLM

with thresholding

**what is priority pass**

priority pass . priority pass is an independent airport lounge access program membership provides you with access to their network of over 700 lounge s these lounge s are managed by a variety of airlines and companies , meaning that significant variation among them a but great international coverage .

Model	Precision	Recall	F1 Score
human annotations (reference)	1.0000	1.0000	1
select-all (baseline)	0.3249	1.0000	0.4905
tF-idf (baseline)	0.5264	0.1716	0.2588
gemma2:27b-instruct-fp16	0.7635	0.7139	<b>0.7379</b>
gemma2:27b-instruct-q8	0.7592	0.7093	0.7334
gemma2:9b-instruct-fp16	0.6717	0.4037	0.5043
gemma2:9b-instruct-q8	0.6883	0.3654	0.4774
gpt-4o-2024-08-06	0.7667	0.6197	0.6854
gpt-4o-mini-2024-07-18	0.6823	0.6959	0.689
llama3.1:70b-instruct-q8	0.7495	0.6587	0.7012
llama3.1:70b-instruct-q4	0.7587	0.6371	0.6926
llama3.1:8b-instruct-fp16	0.5977	0.6042	0.6009

Table 1: Performace of LLMs on fine-grained extraction task.

Dataset	1	2	3	4	5	6	7	>8
Train	497,922	250,182	33,030	7,561	2,238	816	335	364
Dev	4,472	2,101	267	61	25	5	1	6

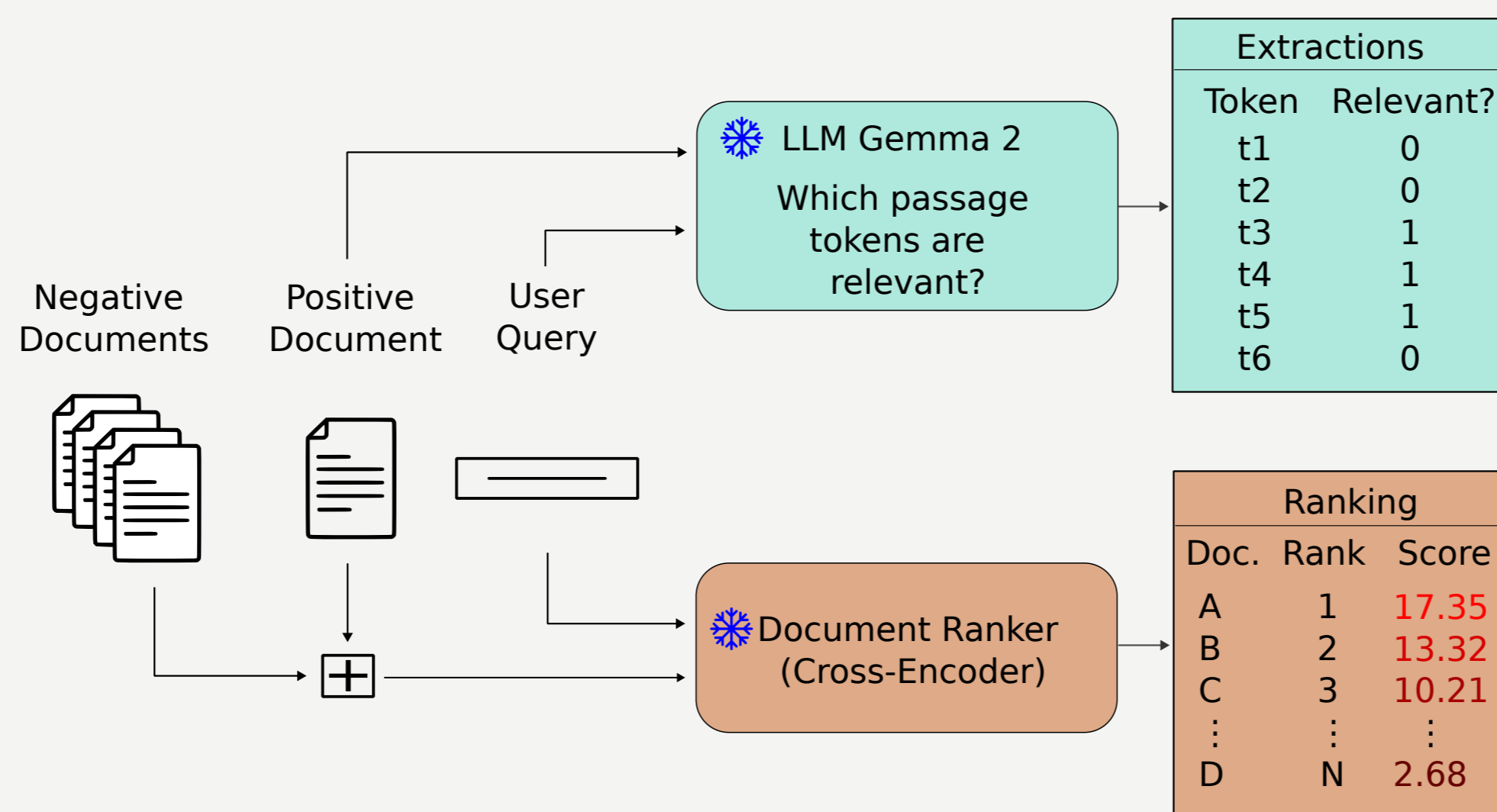
Table 2: Number of extracted spans in a train and dev. sets (24 is max).

1. Let Human Annotate Small Dataset

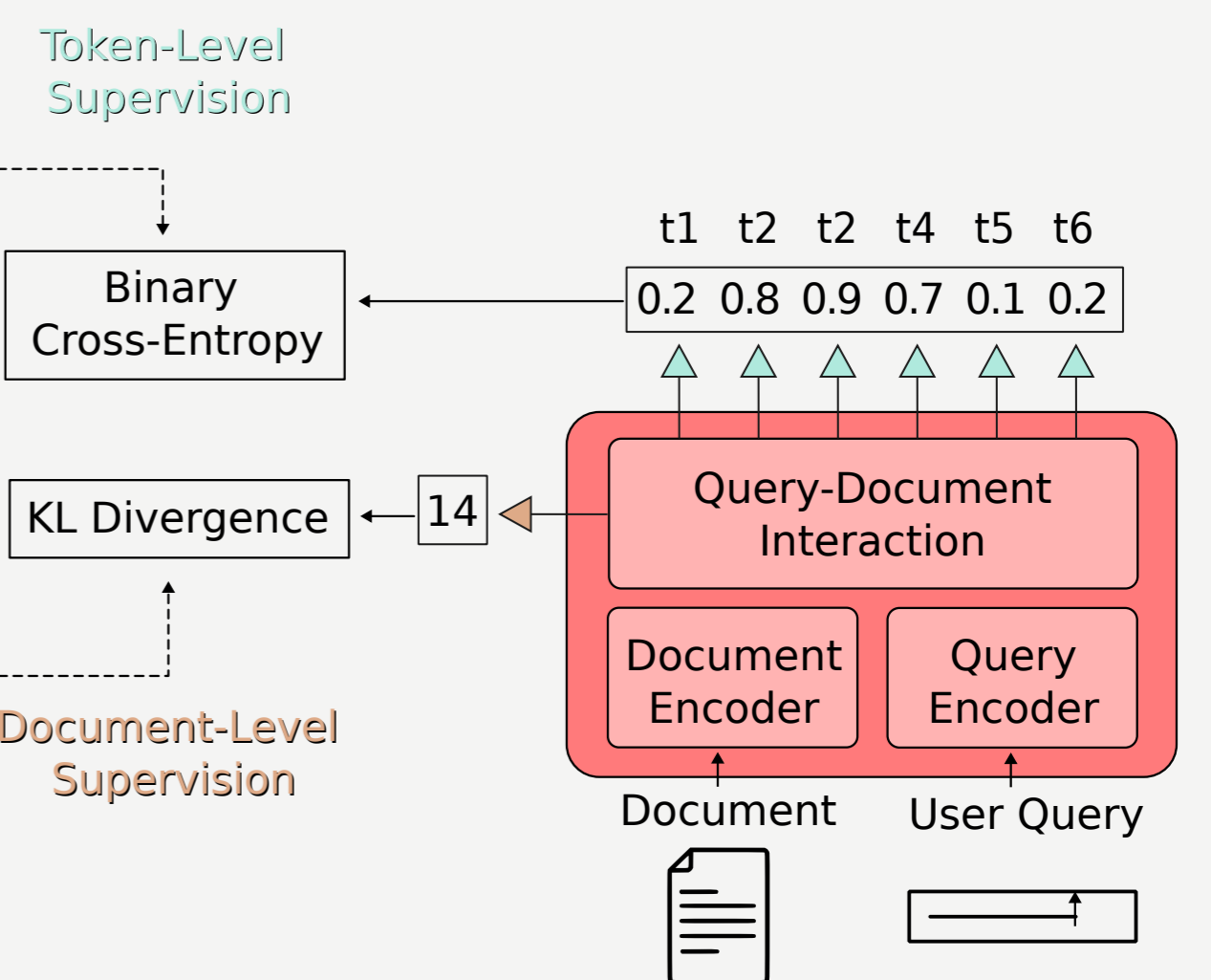
2. Evaluate LLM on Fine-Grained Extraction Task

3. Use Winner LLM To Create Large-Scale Training Dataset

### Generate Data For Distillation (Offline)

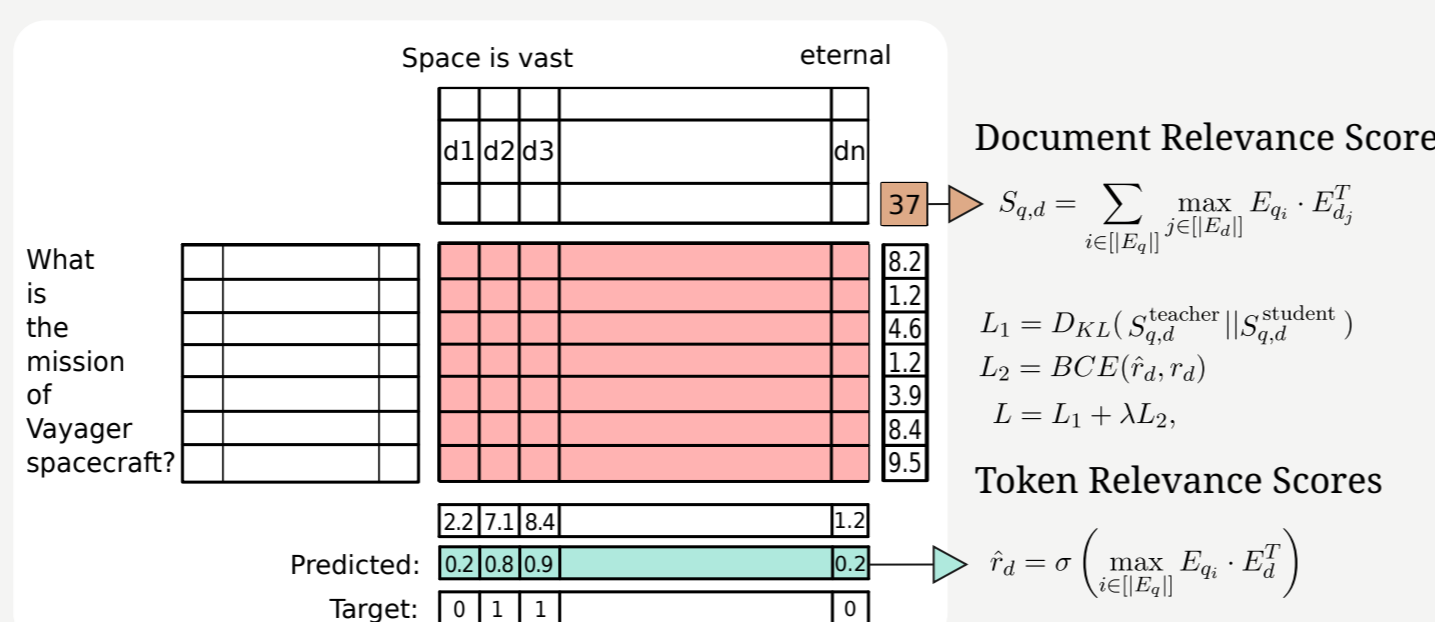


### Joint Training of FGR-CoBERT



4. Modify Retrieval Model To Enable Token Extraction

### Query-Document Interaction



### Encoding Architectures

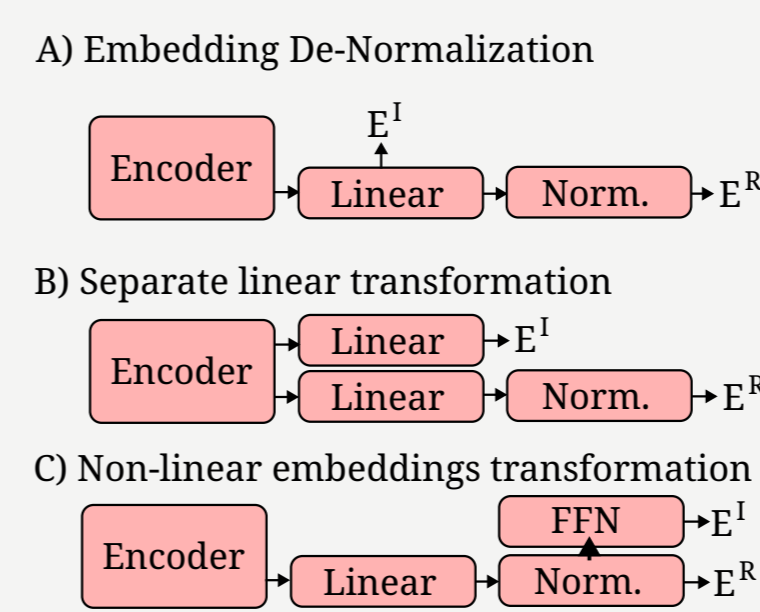
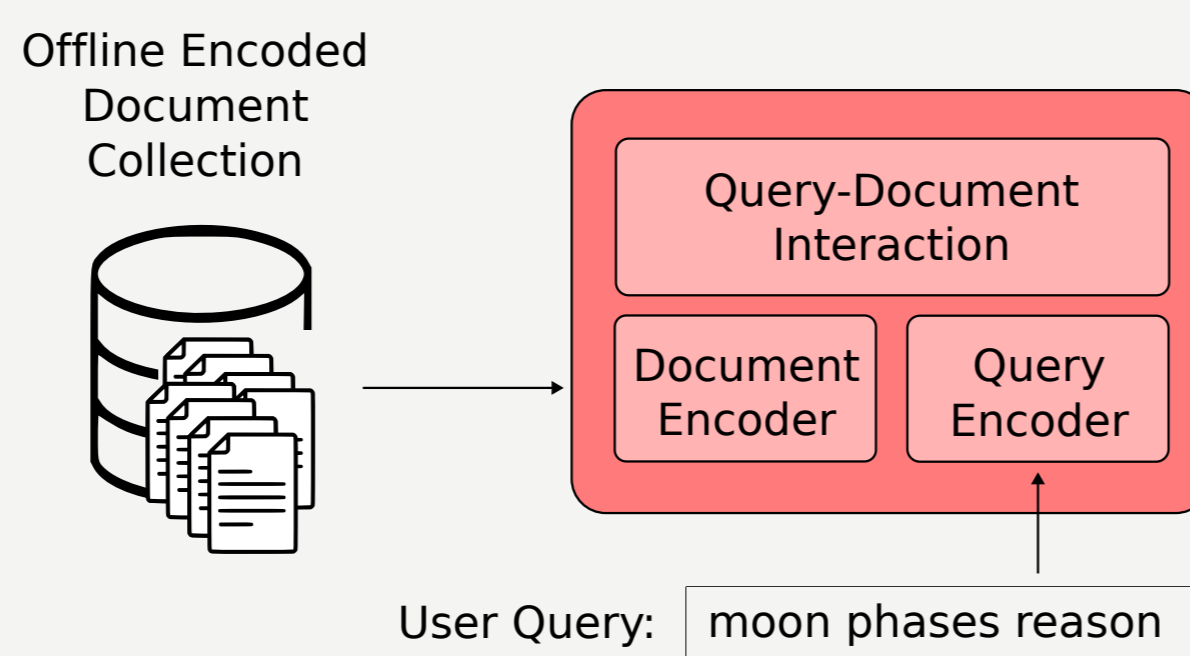


Figure 1: Extraction performance training progress depicted on PR curve.

5. Get LLM Quality Token Relevance Cues During Retrieval

### FGR-CoBERT Inference Overview



### TOP ranked Documents With Highlights

The Moon looks different because of how sunlight hits it. Phases happen **due to the changing angles between the Earth, Moon, and Sun.** A full cycle takes about 29 days.

The Moon orbits Earth and reflects sunlight. We see **different parts of its lit side, causing the phases.** A full cycle takes about 29 days.

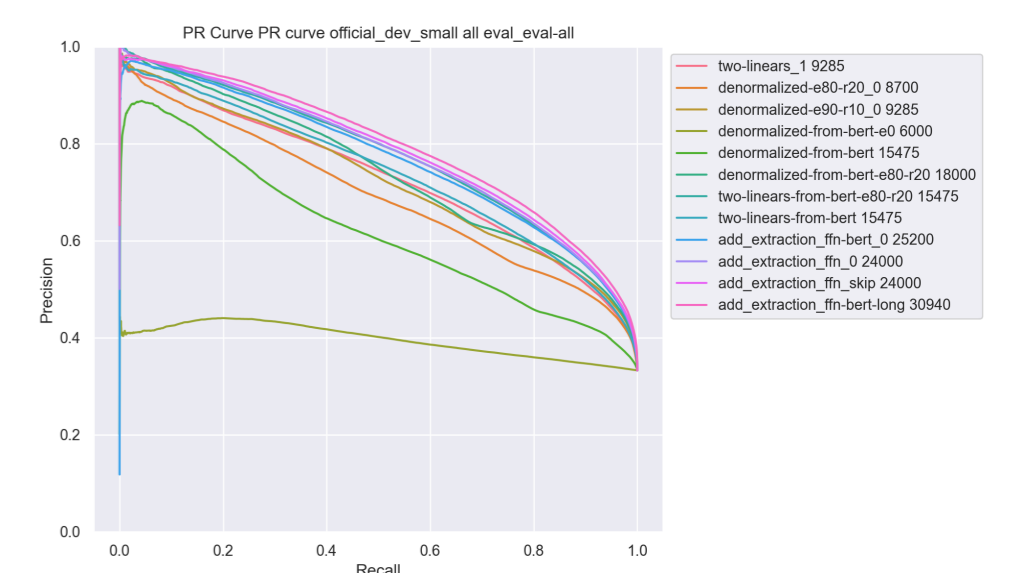


Figure 2: Best PR curves (by F1-score) for different runs and architectures.

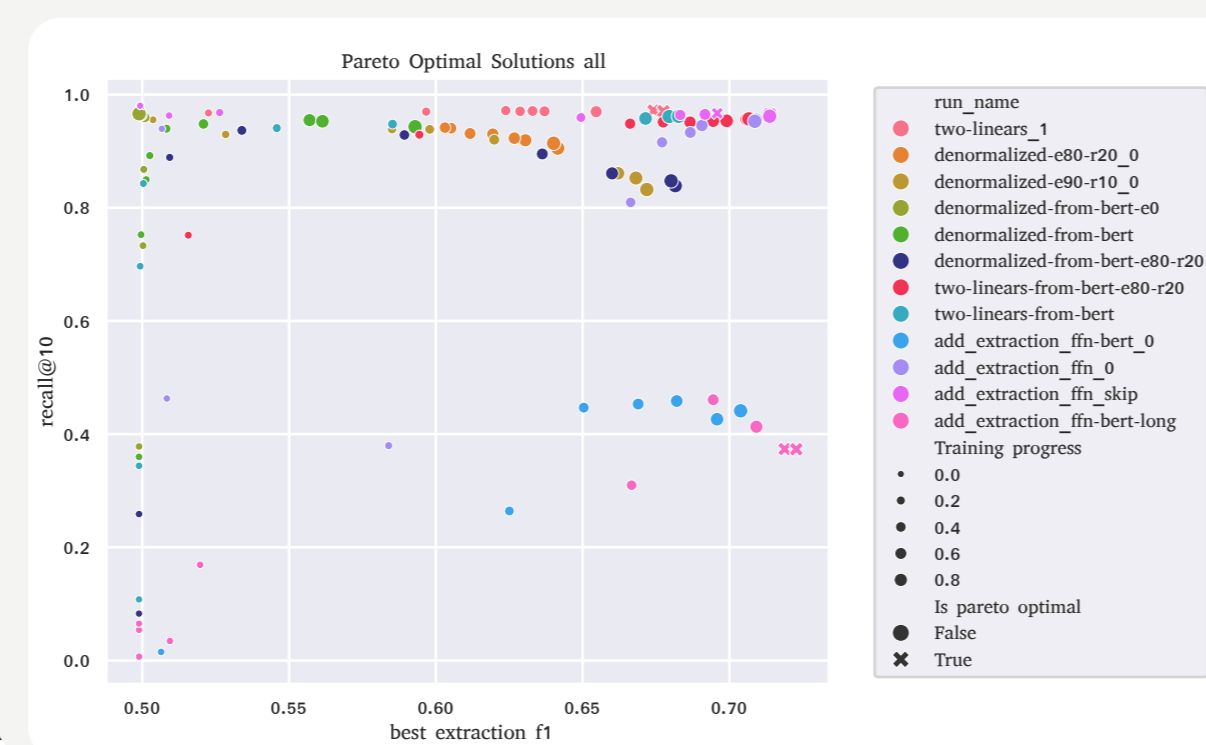
### Conclusions

modified retrieval approach providing built-in relevance explainability

cues obtained from our retrieval COLBERT (120 M) model even matched Gemma-2 (27 B)

three different approaches offering deployment flexibility

### Training & Evaluation



Model Architecture	Initialization	F1-score
Non-linear embeddings transformation	BERT	0.7038
	CoBERT	0.7093
Separate linear transformation	BERT	0.7067
	CoBERT	0.6775
Embedding de-normalization	BERT	0.6816
	CoBERT	0.6719
Retrieval Only	BERT	0.5008

Table 3: Human Match vs Doc Retrieval Performance

### Quick Facts and Notes

Gemma size = 27B  
Colbert size = 120M

Generation Errors:  
792k / 808k total - (470k MS-MARCO + 317k Top-1 Retrieved)  
98% generated correctly

Span Not Found (8400) + Nothing Selected (7800)

Huristically Fixed 4600 Samples

Another Dataset of 160 unique samples  
3 annotators inter agreement fleiss kappa 0.445

Previous evaluation on MD2D (Grad-SAM, AttCAT, Attention-Rollout)