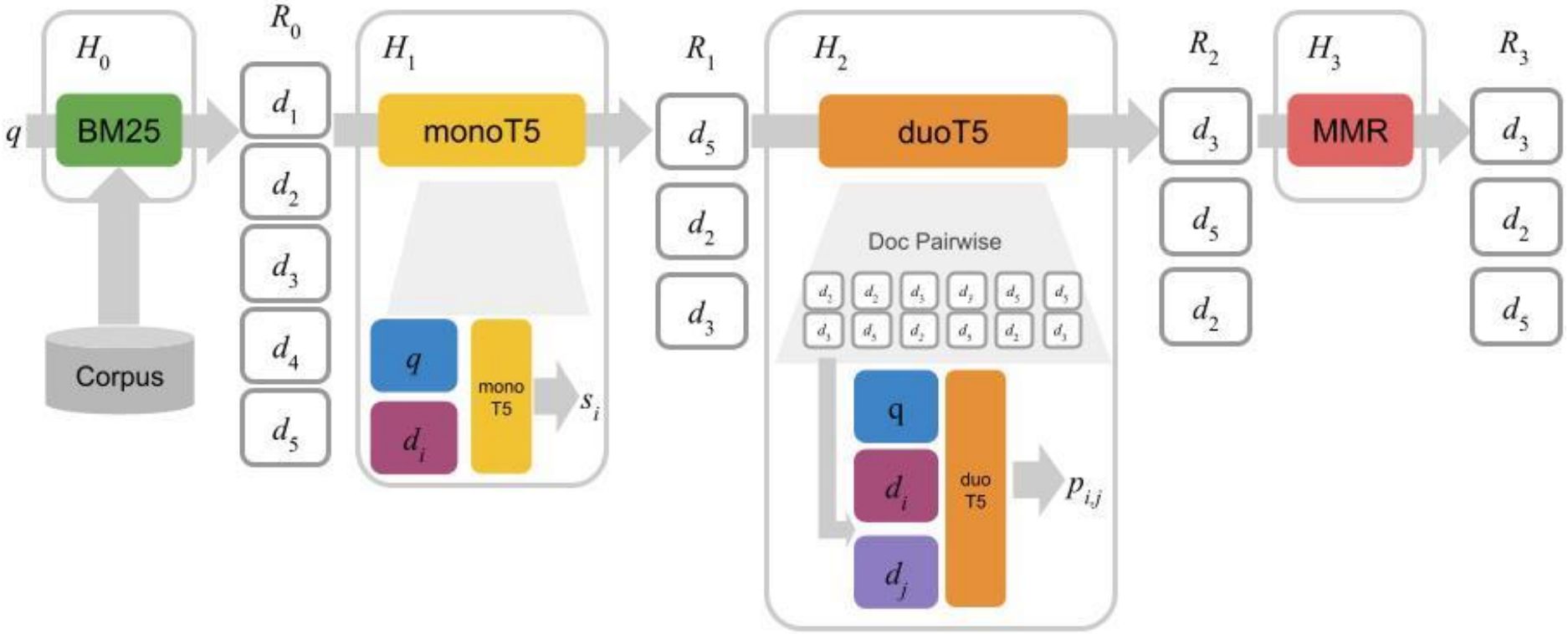
Epidemic QA with $h_2$oloo

Justin Borromeo, Ronak Pradeep, and Jimmy Lin [University of Waterloo]
TAC 2020
February 23, 2021

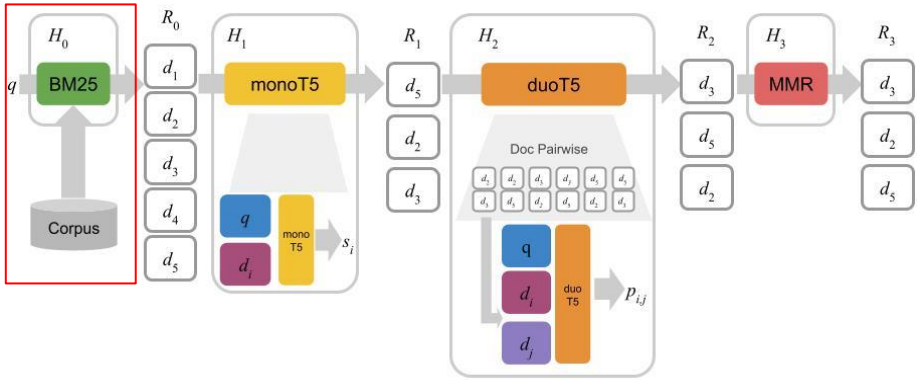
Our Multi-Stage Ranking Pipeline



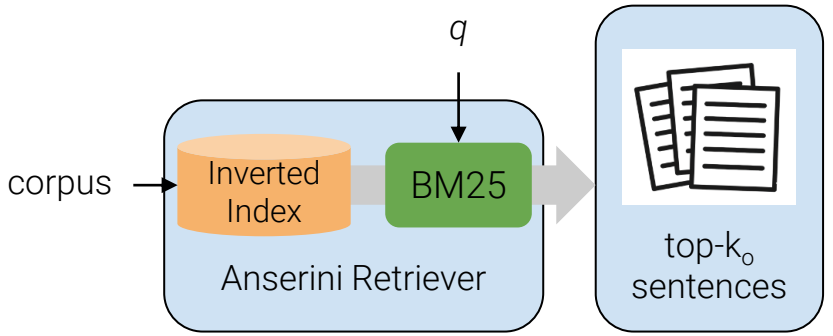
Definitions
Segment: A section of text we're interested in ranking
H_n: Pipeline stage number
R_n: The ranked set of segments outputted by H_n
k_n: The number of segments in R_n

R. Pradeep, R. Nogueira, and J. Lin. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models, 2021.

H_0: Anserini BM25 Retrieval

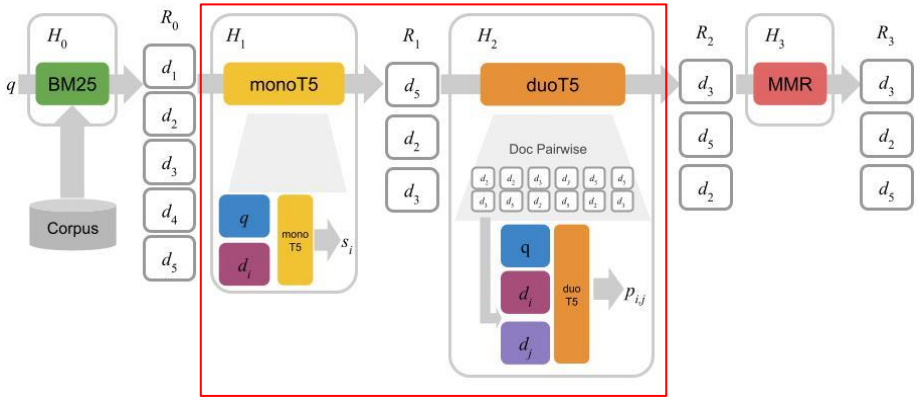


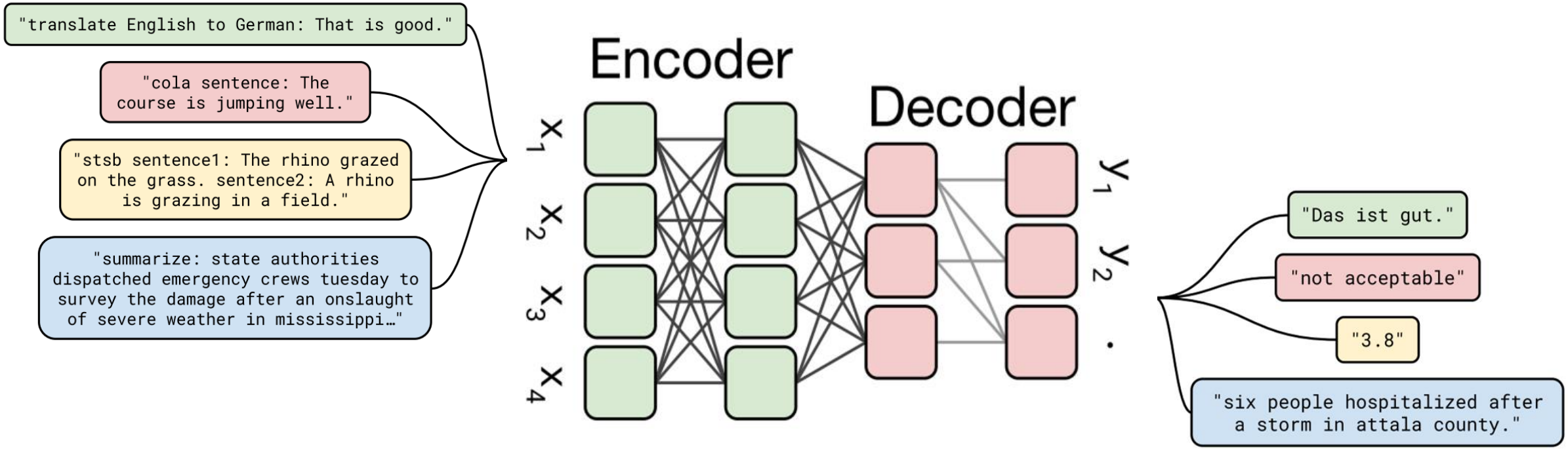
- BM25 retrieval function provides a first-stage ranking of relevant sentences for each query
- Bag-of-words approach
- k_0= 10000
- Performed with the Anserini toolkit, which is built on top of Lucene



P. Yang, H. Fang, and J. Lin. Anserini: Enabling the Use of Lucene for Information Retrieval Research. SIGIR 2017

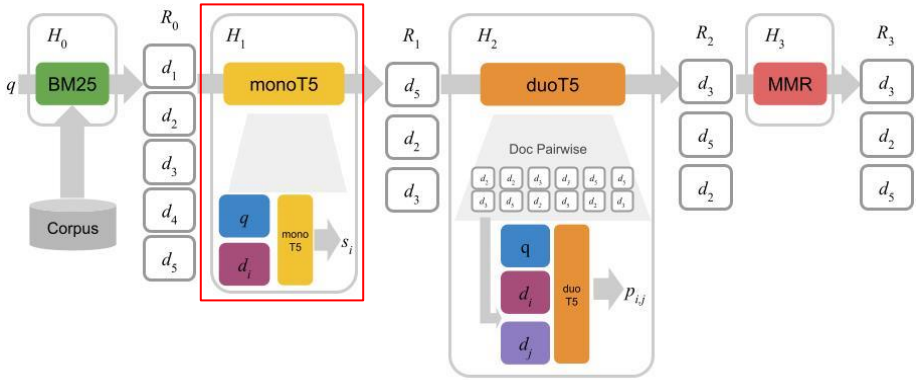
T5





Raffel, Colin, et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *JMLR 2020*.

H_1: monoT5 Re-ranking



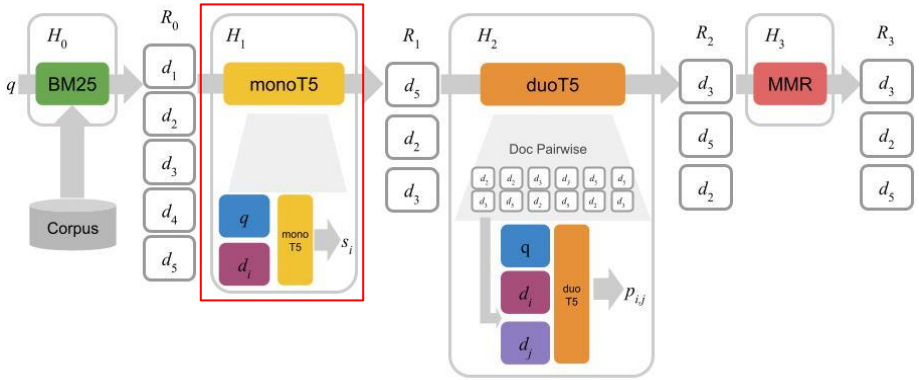
- Sentences from R_0 are augmented with 3 sentences before and 2 sentences after for context
- Augmented segments are re-ranked using monoT5
- Input (d is the segment):

Query: q Document: d Relevant:

- Output: probability that the segment is relevant to the query (i.e. probability that the model produces "true" or "false")

R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin. Document Ranking with Pretrained Sequence-to-Sequence Model. In Findings of EMNLP, 2020.

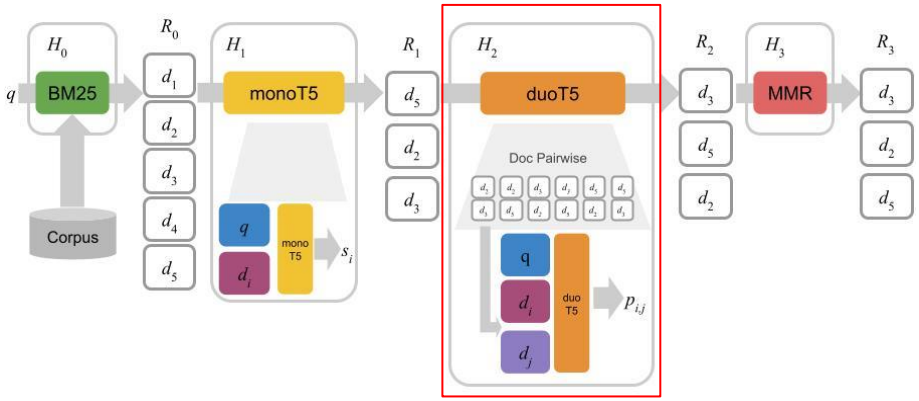
H_1: monoT5 Training



- Trained on the MS MARCO passage dataset, which contains pairs of queries and relevant passages
- Fine-tuned on Med-MARCO, a medically-focused subset of MS MARCO

S. MacAvaney, A. Cohan, and N. Goharian. SLEDGE: A Simple Yet Effective Baseline for COVID-19 Scientific Knowledge Search. arXiv:2005.02365, 2020.

H_2: duoT5 Re-ranking

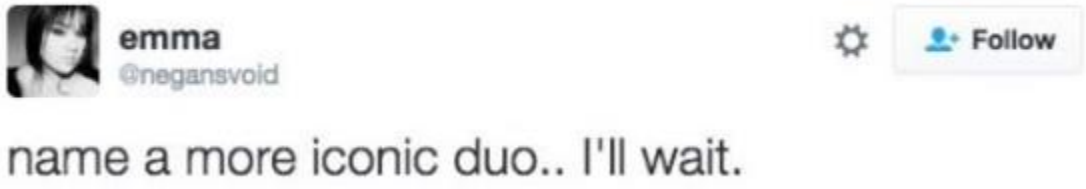


duoT5 model:

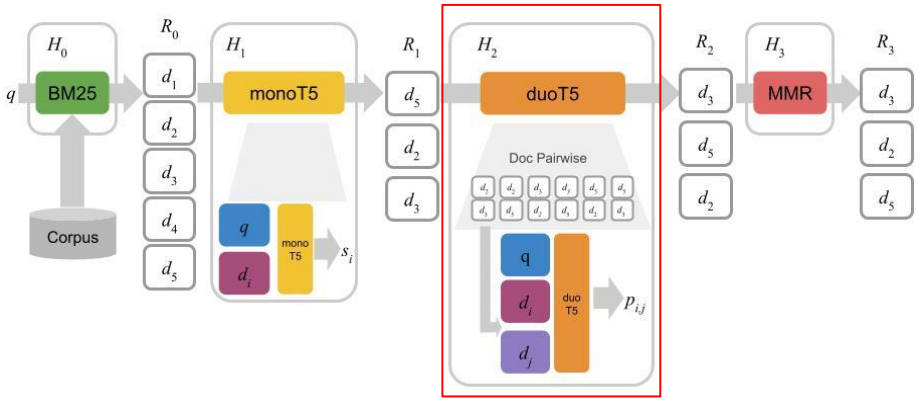
- Input:

Query: q Document0: d_0 Document1: d_1 Relevant:

- Output: probability that d_0 is more relevant than d_1 (i.e. probability that the model produces "true" or "false" token)
- Trained in the same manner as monoT5



H_2: duoT5 Re-ranking

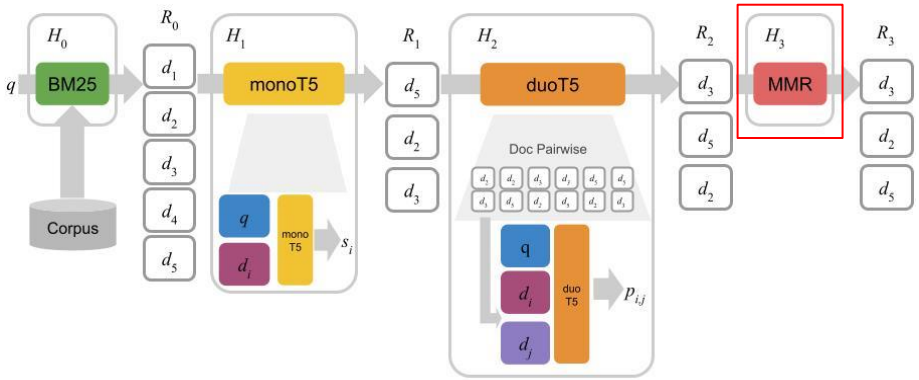


- duoT5 is used for every combination of documents in the top-k_2 documents of R_1
- duoT5 score aggregation:

$$\text{SYM-SUM} : s_i = \sum_{j \in J_i} (p_{i,j} + (1 - p_{j,i}))$$

- Top-k_2 documents are re-ranked according to their aggregated scores

H_3: Maximal Marginal Relevance Re-ranking



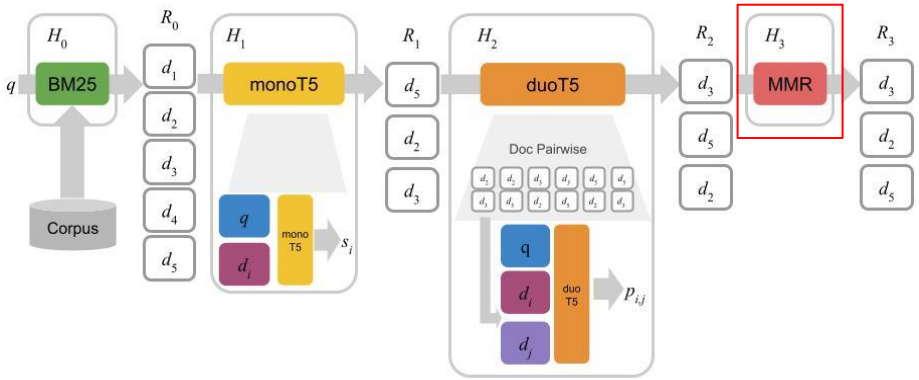
- Motivation: in NDNS, repeated nuggets don't count...diversity matters!
- We use MMR to incrementally build R_3 and improve diversity of R_2

$$MMR = \arg \max_{d_i \in R_2 \setminus S} \left[\lambda Sim_1(d_i, q) - (1 - \lambda) \max_{d_j \in S} Sim_2(d_i, d_j) \right]$$

- Sim_1 is the sym-sum aggregated duoT5 score from the previous stage
- Sim_2 is measured as cosine similarity of the sentences' BM25 vectors
- This stage returns a ranked set of singular sentences (without context sentences)

J. Carbonell and J. Goldstein-Stewart. "The use of MMR, diversity-based reranking for reordering documents and producing summaries." *SIGIR 1998*.

H_3: Maximal Marginal Relevance Re-ranking



- λ was tuned using preliminary round judgments
- Consumer task runs 1, 2, and 3 use λs of 0.75, 0.7, and 1, respectively.
- Expert task runs 1, 2, and 3 use λs of 0.375, 0.42, and 1, respectively.

$$MMR = \arg \max_{d_i \in R_2 \setminus S} \left[\lambda Sim_1(d_i, q) - (1 - \lambda) \max_{d_j \in S} Sim_2(d_i, d_j) \right]$$

Track A (Expert QA) Primary Round Results

	Run	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
	(1) Median	0.3377	0.3387	0.3802
	(2) Max	0.3700	0.3709	0.4207
λ=0.375	(3) Run 1	0.3381	0.3390	0.3880
λ=0.42	(4) Run 2	0.3404	0.3412	0.3901
λ=1	(5) Run 3	0.3284	0.3292	0.3755

Table 1: Mean (across questions) NDNS for EPIC-QA Task A (Expert) Primary Round

Track B (Consumer QA) Results

	Run	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
	(1) Median	0.3142	0.2858	0.2845
	(2) Max	0.3662	0.3675	0.4143
λ=0.75	(3) Run 1	0.3593	0.3607	0.4065
λ=0.7	(4) Run 2	0.3662	0.3675	0.4143
λ=1	(5) Run 3	0.3382	0.3395	0.3825

Table 2: Mean (across questions) NDNS for EPIC-QA Task B (Consumer) Primary Round

Discussion of Results

- MMR (Runs 1 and 2 for both tasks) shows improvement over mono-duo baseline (Run 3)
- Performance compared to median is much worse for expert track. Potential reasons include:
  - Suboptimal MMR tuning
  - Lack of fine-tuning on preliminary round judgments
  - Ineffectiveness of Med-MARCO fine-tuning (ablation study required)

Run	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
(1) Median	0.3377	0.3387	0.3802
(2) Max	0.3700	0.3709	0.4207
(3) Run 1	0.3381	0.3390	0.3880
(4) Run 2	0.3404	0.3412	0.3901
(5) Run 3	0.3284	0.3292	0.3755

Table 1: Mean (across questions) NDNS for EPIC-QA Task A (Expert) Primary Round

Run	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
(1) Median	0.3142	0.2858	0.2845
(2) Max	0.3662	0.3675	0.4143
(3) Run 1	0.3593	0.3607	0.4065
(4) Run 2	0.3662	0.3675	0.4143
(5) Run 3	0.3382	0.3395	0.3825

Table 2: Mean (across questions) NDNS for EPIC-QA Task B (Consumer) Primary Round

PyGaggle

- You too can replicate our results!
- Gaggle of Deep Neural Architectures for Text Ranking and Question Answering.
- Epidemic QA systems to be shared soon!
- Find us at pygaggle.ai!

Thank you!

EPIC Q&A