Ozel Yilmazel

Exploring and Improving Cross- and Multi-Domain Personalized QA via a Dual Retrieval Augmentation

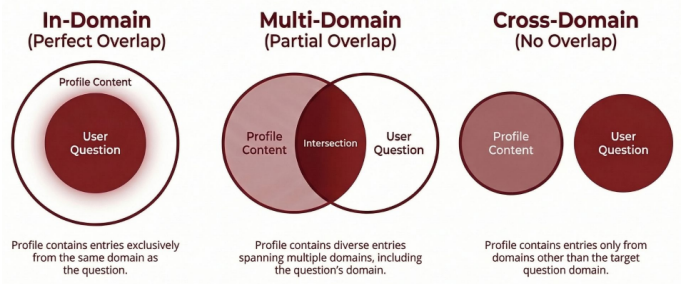
UMassAmherst
Manning College of Information & Computer Sciences
Center for Intelligent Information Retrieval

Problem

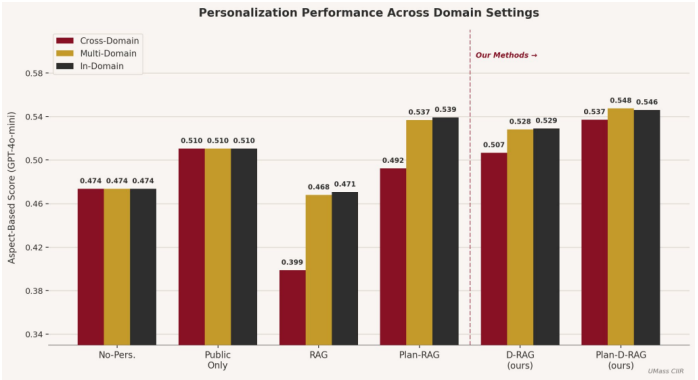
How does a user profile's domain composition affect personalization?

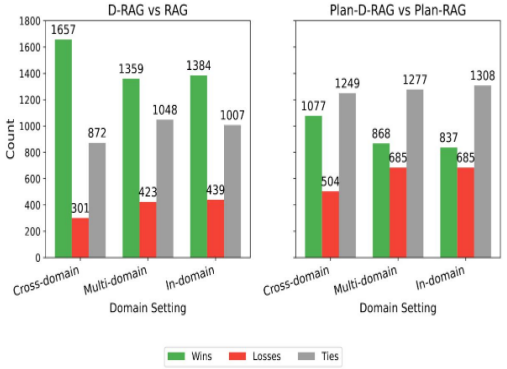
Personalization backfires when profiles don't match the question domain. Misaligned retrieval misleads the model, worse than no personalization at all.

Experiments

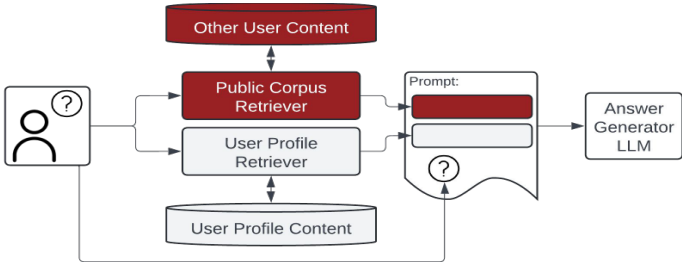


Results





Solution: D-RAG



Ozel Yilmazel

Self-Distillation Experiments for DPR

Problem - Motivation

Knowledge distillation typically requires a large, task-trained teacher model. This is expensive, training the teacher is itself a full training run, and then a second training run is needed for distillation.

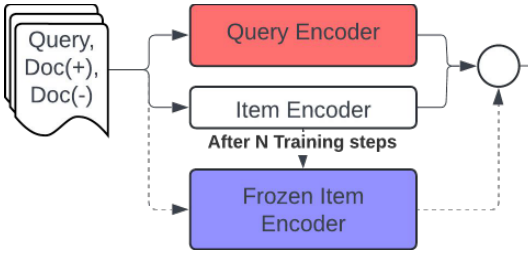
For tasks with relevance judgements, **can a model distil itself using only a self-supervised signal?**

Can this method be used to **improve InfoNCE?**

ANTIQUE, Rel<2 used as hard-negatives, BGE Base EN used to fine tune.
https://wandb.ai/oyilmazel-umass-amherst/self-dist-retrieval?nw=nwuseroyilmazel

Proposed Solution - Combine Encoder's Existing Similarity

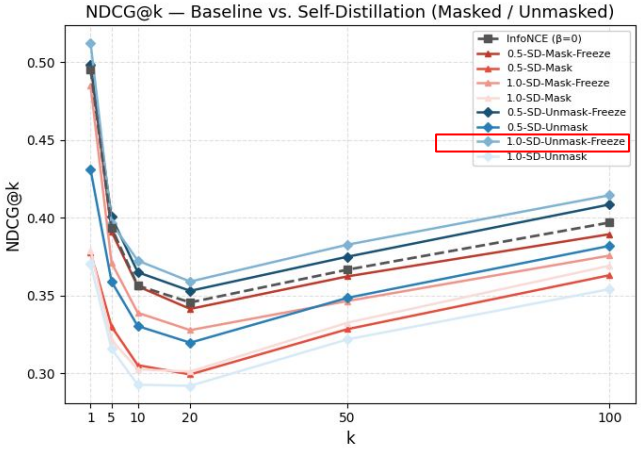
$$L = L_{InfoNCE} + \lambda L_{distill}$$

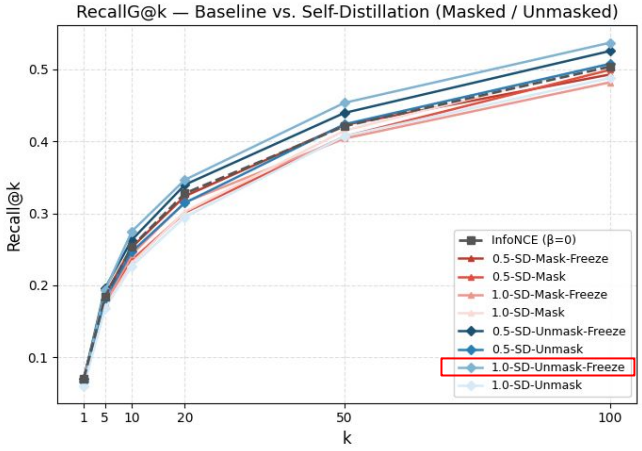


Distill-Loss

	D1	D2	D3	D4
Q2	0.2	0.9	0.1	0.1
		KL-Div		
D2 (+)	0.2	1.0	0.8	0.1

Results - Unmasking HN and Freezing Helps





Ozel Yilmazel

CPU-native Adaptive Vectors for Local Retrieval.

CAVaLRy Embeddings

UMassAmherst
Manning College of Information & Computer Sciences
Center for Intelligent Information Retrieval

Motivation

Built-in OS search suffers from poor accuracy, aggressive re-indexing, and no support for scoped or ad-hoc queries.

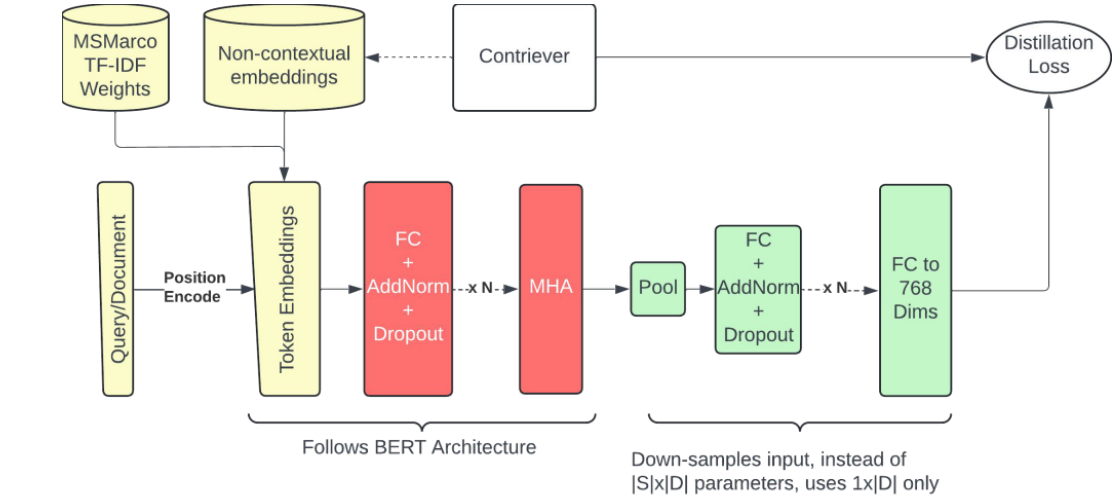
Grep and Lexical retrieval models offer a practical baseline, but fall short of contextual embeddings.

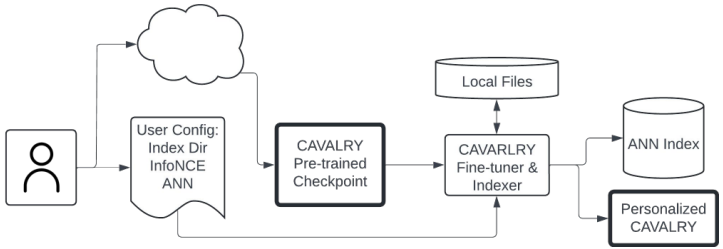
Can we have contextual embeddings while staying within the constraints of common CPU hardware?
Reduce the expensive layers in Transformers, operate with less parameters.

How to reduce reliance on GPU?
Pre-train on GPU, released model can fine-tune and index on CPU overnight.

How to adapt a pre-trained cavalry encoder to your local corpus?
Build a retrieval dataset from local files by sampling n-grams, train contrastively.

Proposed Solution





Unlike small transformers that maintain token-level representations through every layer, Cavalry collapses the sequence into a single vector early, making inference cost nearly independent of document length.

Funnel Transformer for Retrieval