# Retrieval-Augmented Generation: Is Dense Passage Retrieval Retrieving?

Benjamin Reichman and Larry Heck
AI Virtual Assistant (AVA) Lab, Georgia Institute of Technology

## Introduction

- LLMs, widely used but hallucinate often → mislead people and erode trust in LLMs
- RAG addresses hallucinations by adding information to query.
- Important for retrieval to have both high recall and precision.
- To improve retrieval performance we analyze retrieval models from multiple perspectives.

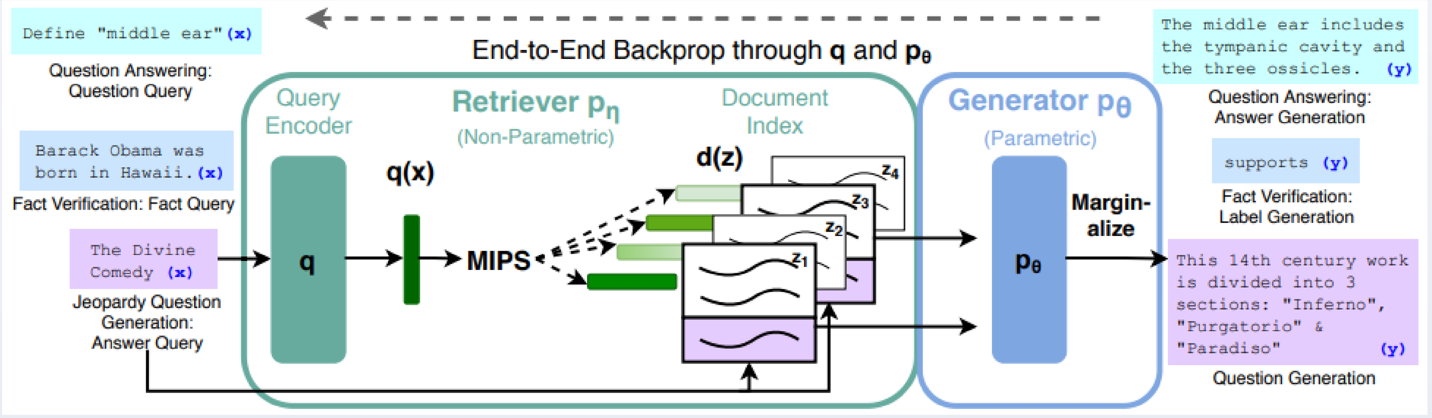


Figure 1: Example Retrieval-Augmented Model Architecture [1]

- Why does DPR training improve on BERT?

| | R@1 | R@5 | R@10 | R@20 | R@50 | R@80 | R@100 |
|---|---|---|---|---|---|---|---|
| Pre-trained BERT | 0.03 | 0.10 | 0.14 | 0.2 | 0.28 | 0.33 | 0.36 |
| DPR BERT | 0.46 | 0.68 | 0.75 | 0.80 | 0.84 | 0.86 | 0.86 |

Table 1: Performance of pre-trained BERT and DPR BERT on retrieval.

- In this poster we:

1. Probing model to determine if pre-trained BERT features are as discriminative as DPR-BERT in matching a query to correct passage amongst hard-negative passages.
2. Compare relative strength and number of activations of the feedforward layers throughout the original pretrained and DPR-trained models
3. Add and remove knowledge from network → investigate how knowledge interacts with DPR training.

## Knowledge Consistency

- Linear probing reveals mutual information shared between model's primary task and probing task [2].
- Probe trained for each BERT block to discriminate between true positive and hard negative passages.
- Performance disparity between probes for pretrained and DPR-trained BERT relatively minor.

| Task | Model | Layer 0 | Layer 1 | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 | Layer 7 | Layer 8 | Layer 9 | Layer 10 | Layer 11 | Layer 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2-Passage Probing | Pre-trained BERT – Untrained Probe | 0.50 | 0.50 | 0.51 | 0.48 | 0.50 | 0.52 | 0.51 | 0.51 | 0.50 | 0.49 | 0.50 | 0.54 | 0.50 |
| | Pre-trained BERT | 0.51 | 0.69 | 0.74 | 0.74 | 0.77 | 0.79 | 0.81 | 0.81 | 0.81 | 0.82 | 0.83 | 0.84 | 0.84 |
| | DPR-BERT Query Model | 0.51 | 0.68 | 0.74 | 0.77 | 0.79 | 0.80 | 0.81 | 0.83 | 0.82 | 0.83 | 0.83 | 0.82 | 0.82 |
| | DPR-BERT Context Model | 0.51 | 0.68 | 0.74 | 0.77 | 0.79 | 0.80 | 0.81 | 0.83 | 0.82 | 0.83 | 0.83 | 0.82 | 0.82 |
| 3-Passage Probing | Pre-trained BERT | 0.34 | 0.53 | 0.59 | 0.59 | 0.65 | 0.64 | 0.67 | 0.67 | 0.68 | 0.69 | 0.69 | 0.73 | 0.73 |
| | DPR-BERT | 0.34 | 0.54 | 0.60 | 0.63 | 0.66 | 0.66 | 0.66 | 0.70 | 0.71 | 0.69 | 0.73 | 0.72 | 0.71 |
| 4-Passage Probing | Pre-trained BERT | 0.26 | 0.43 | 0.47 | 0.49 | 0.53 | 0.57 | 0.61 | 0.60 | 0.56 | 0.62 | 0.64 | 0.66 | 0.66 |
| | DPR-BERT | 0.26 | 0.46 | 0.51 | 0.54 | 0.57 | 0.58 | 0.60 | 0.63 | 0.64 | 0.63 | 0.65 | 0.63 | 0.63 |
| 5-Passage Probing | Pre-trained BERT | 0.21 | 0.35 | 0.42 | 0.43 | 0.43 | 0.50 | 0.53 | 0.53 | 0.54 | 0.56 | 0.57 | 0.60 | 0.61 |
| | DPR-BERT | 0.21 | 0.36 | 0.42 | 0.48 | 0.49 | 0.51 | 0.54 | 0.56 | 0.58 | 0.58 | 0.60 | 0.56 | 0.56 |

Table 2: Per layer probing performance on 2-5 passage matching task.

- Findings suggest capabilities to discern relevant from irrelevant passages already present in BERT.

## Knowledge Decentralization

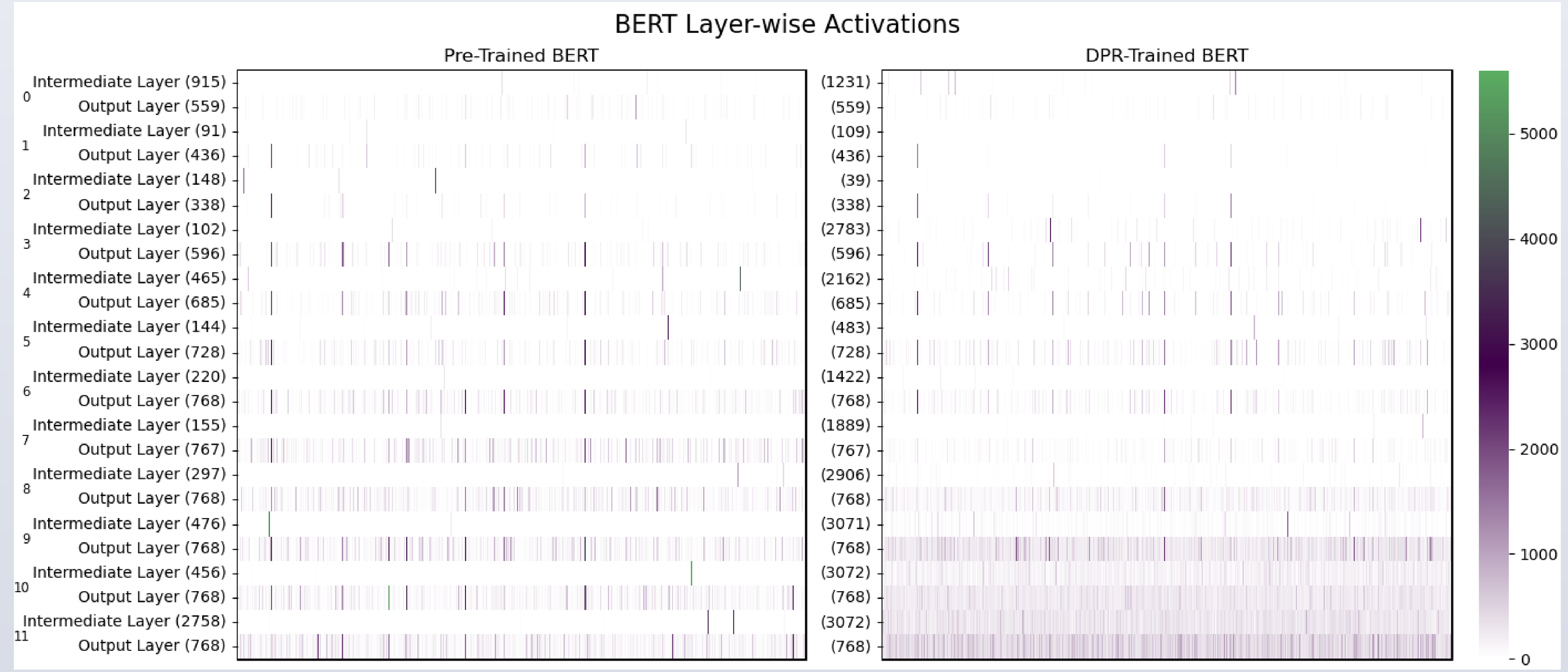


Figure 4: Per layer attribution scores in pre-trained and DPR-trained BERT

- Examined neuron activation patterns for pre-trained and DPR-trained models
- Knowledge attribution method from [3] used
- Following [3], a threshold of $0.1 * \max(\text{Attr})$ was applied to identify coarse set of knowledge neurons.

$$\text{Attr}^{(l)}(w_i) = w_i^{(l)} \int_{\alpha=0}^{1} \frac{\partial P_x(\alpha w_i^{(l)})}{\partial w_i^{(l)}} d\alpha$$

- DPR expands "keys" available to access a given volume of semantic knowledge.
- Decentralization strategy for semantic knowledge.
- Decreases accessible volume of syntactic knowledge.

| Query | Answer in Top-1? Pre-trained BERT | Answer in Top-1? DPR-BERT | # Strongly Activated Neurons Pre-trained BERT | # Strongly Activated Neurons DPR-BERT | Title of Top-5 Retrieval Pre-trained BERT | Title of Top-5 Retrieval DPR BERT |
|---|---|---|---|---|---|---|
| where is the most distortion on a robinson projection | ✗ | ✗ | 220 | 1323 | Circle of latitude, Scale-invariant feature transform, Line moiré, Theil–Sen estimator, Pole splitting | Robinson projection, Robinson projection, Arthur H. Robinson, Robinson projection, Arthur H. Robinson |
| are pure metals made of atoms or ions | ✓ | ✗ | 69 | 1268 | Alloy, Common attributes, Metal, Resonance ionization, Alloy | Properties of metals, metalloids and nonmetals, Properties of metals, metalloids and nonmetals, Solid, Metal, Metal |
| who is the bad guy in lord of the rings | ✗ | ✓ | 100 | 533 | Millennium Earl, The Sword of Shannara, Eye of Ra, The Enchanted Apples of Oz, Ys I & II | Saruman, Saruman, Sauron, Morgoth, Legolas |
| when did mozart compose his first piece of music | ✓ | ✓ | 74 | 364 | Wolfgang Amadeus Mozart, Der Messias, Life of Franz Liszt, Die Entführung aus dem Serail, Quattro versioni originali della Ritirata notturna di Madrid | Wolfgang Amadeus Mozart, Wolfgang Amadeus Mozart, Leopold Mozart, Wolfgang Amadeus Mozart, Wolfgang Amadeus Mozart |

Table 2: Example queries with counts of strongly activated neurons. DPR BERT has more strongly activated neurons and more focused retrievals.

## Adding and Removing Knowledge

| 284 Facts Added | Probing Added | Off-Target Flips - Probing | DPR Added | Off-Target Edits - DPR |
|---|---|---|---|---|
| Transformer-Patch | 0.54 | 581 | 0.44 | 222 |
| MalMen | 0.57 | 592 | 0.37 | 236 |
| Mend | 0.57 | 592 | 0.38 | 229 |

| 284 Facts Removed | Probing Removed | Off-Target Flips - Probing | DPR Removed | Off-Target Edits - DPR |
|---|---|---|---|---|
| Transformer-Patch | 0.16 | 689 | 0.87 | 183 |
| MalMen | 0.11 | 721 | 0.81 | 261 |
| Mend | 0.11 | 722 | 1.00 | 252 |

Table 3: Results of adding and removing facts from BERT and then DPR-training BERT

- Do facts that pre-trained BERT knows reappear in DPR-BERT?
- Both knowledge addition and removal experiments show DPR training refines how pre-existing knowledge within BERT rendered more "retrievable".
- Added facts became retrievable, removed ceased to be retrievable.

## Conclusions

- DPR does not add knowledge to networks
- It decentralizes knowledge representation
  - Allows for more pathways to trigger same information.
- Retrieval limited by knowledge present in network after pre-training

## References

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2021. Retrieval-augmented generation for knowledge intensive nlp tasks.
[3] Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1):207–219
[4] Damai Dai, Li Dong, et al. 2022. Knowledge neurons in pretrained transformers. In Proceedings of ACL (Volume 1: Long Papers), pages 8493–8502. ACL.

## Acknowledgments

This work was supported in part by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.