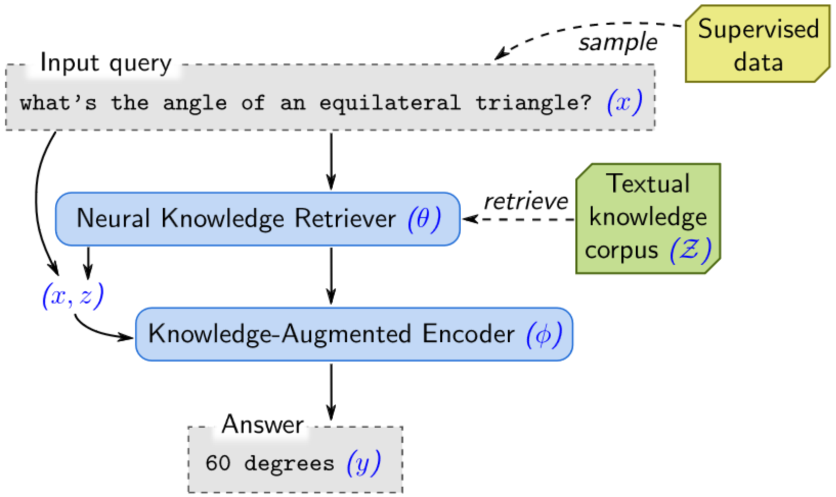
Variational Open-Domain Question Answering

Valentin Liévin [1 2] **Andreas Geert Motzfeldt** [1] **Ida Riis Jensen** [1] **Ole Winther** [1 2 3 4]

Open-Domain Question Answering (ODQA)

- LLMs are limited by the implicit knowledge they possess
  - Incomplete, flawed, out of date etc.
- Retrieval-augmented models:
  - Augmenting LMs with external knowledge bases indexed with a retrieval mechanism
  - Popular for ODQA (e.g. REALM (Guu et al. 2020))
- This paper:
  - Proposes a probabilistic framework for retrieval-augmented tasks with end-to-end learning
  - Based on Rényi divergence variational inference



REALM: Retrieval-Augmented Language Model Pre-Training. Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. (2020)

Reader-Retrieval Model

For a question and answer $\mathbf{q}, \mathbf{a} \in \Omega$ (Ω is the space of sequences of tokens)

and a corpus of N documents $\mathbb{D} := \{\mathbf{d}_1, \ldots, \mathbf{d}_N\} \in \Omega^N$

the reader-retrieval model is given by:

$$p_\theta(\mathbf{a}|\mathbf{q}) := \sum_{\mathbf{d} \in \mathbb{D}} \underbrace{p_\theta(\mathbf{a}|\mathbf{d}, \mathbf{q})}_{\text{reader}} \underbrace{p_\theta(\mathbf{d}|\mathbf{q})}_{\text{retriever}}$$

(where both reader and retriever are BERT models)

Reader-Retrieval Model with Traditional VI

$$p_\theta(\mathbf{a}|\mathbf{q}) := \sum_{\mathbf{d} \in \mathbb{D}} \underbrace{p_\theta(\mathbf{a}|\mathbf{d}, \mathbf{q})}_{\text{reader}} \underbrace{p_\theta(\mathbf{d}|\mathbf{q})}_{\text{retriever}}$$

Traditional variational inference:

- Estimate $p_\theta(\mathbf{a}|\mathbf{q})$ by drawing samples from an approximate posterior (a "static retriever")

$$r_\phi(\mathbf{d}|\mathbf{a}, \mathbf{q})$$

- and evaluating the ELBO

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\mathbf{a}, \mathbf{q}) &:= \log p_\theta(\mathbf{a}, \mathbf{q}) - \mathcal{D}_{\text{KL}}(r_\phi(\mathbf{d}|\mathbf{a}, \mathbf{q}) || p_\theta(\mathbf{d}|\mathbf{a}, \mathbf{q})) \\ &\leq \log p_\theta(\mathbf{a}, \mathbf{q})\end{aligned}$$

But the authors suggest that Rényi divergence VI may be better...

Variational Rényi Bound (RVB)

- A generalisation of the ELBO with a parameter $\alpha \in [0, 1)$

$$\mathcal{L}_\alpha(\mathbf{a}, \mathbf{q}) := \frac{1}{1-\alpha} \log \mathbb{E}_{r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})} \left[w_{\theta,\phi}^{1-\alpha}(\mathbf{a}, \mathbf{q}, \mathbf{d}) \right]$$

$$w_{\theta,\phi}(\mathbf{a}, \mathbf{q}, \mathbf{d}) := \frac{p_\theta(\mathbf{a}, \mathbf{d}|\mathbf{q})}{r_\phi(\mathbf{d}|\mathbf{a}, \mathbf{q})}$$

- Can be extended for $\alpha = 1$

$$\mathcal{L}_{\alpha=1}(\mathbf{a}, \mathbf{q}) := \lim_{\alpha \to 1} \mathcal{L}_\alpha(\mathbf{a}, \mathbf{q}) = \mathcal{L}_{\mathrm{ELBO}}(\mathbf{a}, \mathbf{q})$$

- Gives a lower bound on the marginal task log-likelihood

$$\mathcal{L}_{\alpha=0}(\mathbf{a}, \mathbf{q}) = \log p_\theta(\mathbf{a}|\mathbf{q})$$

$$\mathcal{L}_{\alpha \geq 0}(\mathbf{a}, \mathbf{q}) \leq \log p_\theta(\mathbf{a}|\mathbf{q})$$

Importance-Weighted RVB (IW-RVB)

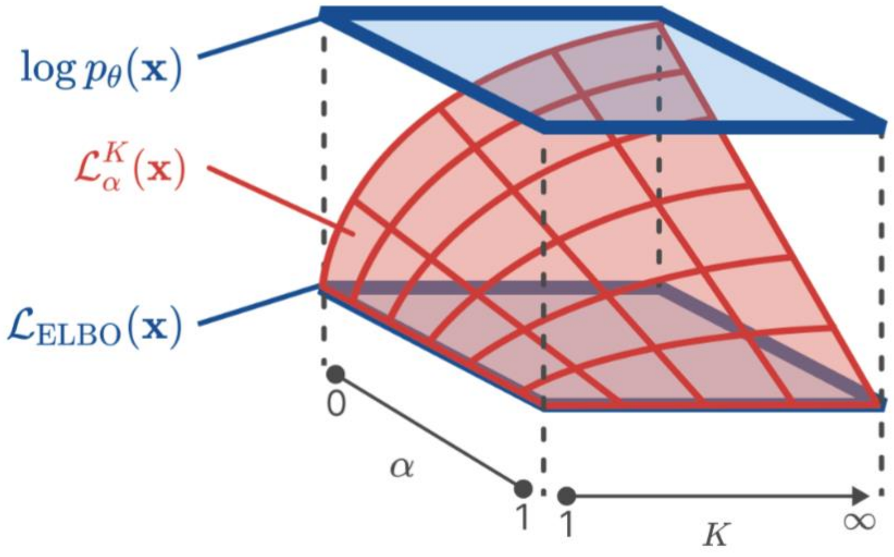
- RVB is usually intractable

$$\mathcal{L}_\alpha(\mathbf{a},\mathbf{q}) := \frac{1}{1-\alpha}\log \mathbb{E}_{r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})}\left[w_{\theta,\phi}^{1-\alpha}(\mathbf{a},\mathbf{q},\mathbf{d})\right] \quad \text{(RVB)}$$

- So we often use the importance-weighted RVB:

$$\hat{\mathcal{L}}_\alpha^K(\mathbf{a},\mathbf{q}) := \frac{1}{1-\alpha}\log \frac{1}{K}\sum_{i=1}^{K} w_{\theta,\phi}^{1-\alpha}(\mathbf{a},\mathbf{q},\mathbf{d}_i) \quad \text{(IW-RVB)}$$

$$\mathbf{d}_1,\ldots,\mathbf{d}_K \overset{\text{iid}}{\sim} r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})$$



- For which we still have: $\mathcal{L}_{\text{ELBO}}(\mathbf{a},\mathbf{q}) \le \hat{\mathcal{L}}_\alpha^K(\mathbf{a},\mathbf{q}) \le \log p_\theta(\mathbf{a},\mathbf{q})$

The RVB & IW-RVB

- The idea is that we can control α to improve our model's training

- Evaluating the IW-RVB (and its gradient w.r.t. θ) has $\mathcal{O}(N)$ complexity
  - Requires importance weights for every document in the corpus

$$w_{\theta,\phi}(\mathbf{a},\mathbf{q},\mathbf{d}) := \frac{p_\theta(\mathbf{a},\mathbf{d}|\mathbf{q})}{r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})}$$

- This is problematic in ODQA applications, where N is often very large

- Solution: use priority sampling instead of regular importance sampling

Priority Sampling

- **Main point:** priority sampling allows us to sample K documents (and corresponding importance weights s_i) without computing sums over the whole corpus

$$(\mathbf{d}_1, s_1), \ldots, (\mathbf{d}_K, s_K) \overset{\text{priority}}{\sim} r_\phi(\mathbf{d}|\mathbf{a}, \mathbf{q})$$

- Such that for a function $h(\mathbf{d})$ we have a consistent estimator $\sum_{i=1}^{K} s_i h(\mathbf{d}_i) \approx \mathbb{E}_{r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})}[h(\mathbf{d})]$

- We have N documents, each weighted by $x_i := r_\phi(\mathbf{d}_i|\mathbf{a}, \mathbf{q})$

1. Sample random weights $u_1, \ldots, u_N \overset{\text{iid}}{\sim} \text{Uniform}(0, 1]$
2. Generate "priorities" x_i/u_i
3. Let τ be the $(K + 1)$ th largest priority
4. Select the K items with the largest priorities, return these alongside their corresponding importance weights $s_i := \max(x_i, \tau)$

Variational Open-Domain (VOD) Objective

- So instead of IW-RVB ($\mathcal{O}(N)$ due to importance sampling):

$$\hat{\mathcal{L}}_\alpha^K(\mathbf{a},\mathbf{q}) := \frac{1}{1-\alpha}\log\frac{1}{K}\sum_{i=1}^{K} w_{\theta,\phi}^{1-\alpha}(\mathbf{a},\mathbf{q},\mathbf{d}_i)$$

$$\mathbf{d}_1,\ldots,\mathbf{d}_K \overset{\text{iid}}{\sim} r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})$$

- We use the VOD objective ($\mathcal{O}(K)$ because of priority sampling):

$$\hat{L}_\alpha^K(\mathbf{a},\mathbf{q}) := \frac{1}{1-\alpha}\log\sum_{i=1}^{K} s_i\hat{v}_{\theta,\phi}^{1-\alpha}(\mathbf{a},\mathbf{q},\mathbf{d}_i)$$

$$(\mathbf{d}_1,s_1),\ldots,(\mathbf{d}_K,s_K) \overset{\text{priority}}{\sim} r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})$$

$$\hat{v}_{\theta,\phi} \approx w_{\theta,\phi}$$

Technical details:

$$\hat{v}_{\theta,\phi} := p_\theta(\mathbf{a}|\mathbf{q},\mathbf{d}_i)\zeta(\mathbf{d}_i)\left(\sum_{j=1}^{K} s_j\zeta(\mathbf{d}_j)\right)^{-1} \approx w_{\theta,\phi}$$

1 reader evaluation

$$\zeta(\mathbf{d}) \propto \frac{p_\theta(\mathbf{d}|\mathbf{q})}{r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})}$$

1 retriever and approx. posterior evaluation per document

Q: How are we actually modelling the retriever(s)?

- Via score functions, $f_\theta : \Omega^2 \rightarrow \mathbb{R}$ and $f_\phi : \Omega^3 \rightarrow \mathbb{R}$ for retriever and approx. posterior respectively

$$f_\theta(\mathbf{d}, \mathbf{q}) \;=\; \mathrm{BERT}_\theta(\mathbf{d})^T \mathrm{BERT}_\theta(\mathbf{q})$$

$$f_\phi(\mathbf{a}, \mathbf{q}, \mathbf{d}) := f_\phi^{\mathrm{ckpt}}(\mathbf{d}, [\mathbf{q}; \mathbf{a}]) + \tau^{-1}(\mathrm{BM25}(\mathbf{q}, \mathbf{d}) + \beta \cdot \mathrm{BM25}(\mathbf{a}, \mathbf{d}))$$

  - BERT (language model)
  - BM25 (bag-of-words document ranking algorithm)

(where f_ϕ^{ckpt} is a saved version of $f_\theta(\mathbf{d}, \mathbf{q})$ at a certain point in training and initially $f_\phi^{\mathrm{ckpt}} = 0$)

Q: How are we actually modelling the retriever(s)?

- Then softmax over the scores to get distributions

$$p_\theta(\mathbf{d}|\mathbf{q}) := \frac{\mathbb{1}[\mathbf{d} \in \mathcal{T}_\phi] \exp f_\theta(\mathbf{d}, \mathbf{q})}{\sum_{\mathbf{d}' \in \mathcal{T}_\phi} \exp f_\theta(\mathbf{d}', \mathbf{q})} \qquad r_\phi(\mathbf{d}|\mathbf{a}, \mathbf{q}) := \frac{\mathbb{1}[\mathbf{d} \in \mathcal{T}_\phi] \exp f_\phi(\mathbf{a}, \mathbf{q}, \mathbf{d})}{\sum_{\mathbf{d}' \in \mathcal{T}_\phi} \exp f_\phi(\mathbf{a}, \mathbf{q}, \mathbf{d}')}$$

- $\mathcal{T}_\phi$ is the set of P documents with highest $f_\phi(\mathbf{a}, \mathbf{q}, \mathbf{d})$ scores, which we can put in cache:
  - So in priority sampling, we only sample K from P documents ($K < P \ll N$)
  - Good for memory management
  - Also serves as an "exploration/exploitation threshold"

- The reader is modelled with BERT in a similar way, with softmax over a scoring function $g_\theta : \Omega^2 \to \mathbb{R}$

$$g_\theta(\mathbf{d}, \mathbf{q}_j) = \text{Linear}_\theta(\text{BERT}_\theta([\mathbf{d}; \mathbf{q}_j])) \qquad p_\theta(\mathbf{a}_\star|\mathbf{D}, \mathbf{Q}) := \frac{\exp g_\theta(\mathbf{d}_\star, \mathbf{q}_\star)}{\sum_{j=1}^{M} \exp g_\theta(\mathbf{d}_j, \mathbf{q}_j)}$$

Training Intuition from the RVB

- Ultimately, using $\alpha = 0$ should give us a tighter bound on the marginal log-likelihood, since

$$\mathcal{L}_{\alpha=0}(\mathbf{a}, \mathbf{q}) = \log p_\theta(\mathbf{a}|\mathbf{q})$$

- But, a looser bound may be better at the start of training, e.g. the ELBO/RVB$_{\alpha = 1}$:
  - Encourages knowledge transfer from approx. posterior to retriever

$$\nabla_{\theta_{\text{RETR.}}} \mathcal{L}_{\alpha=1}(\mathbf{a}, \mathbf{q}) = -\nabla_\theta D_{\text{KL}} \left(r_\phi(\mathbf{d}|\mathbf{a}, \mathbf{q}) \| p_\theta(\mathbf{d}|\mathbf{q}) \right)$$

  This is useful if we set the initial approx. posterior to a domain specific baseline:

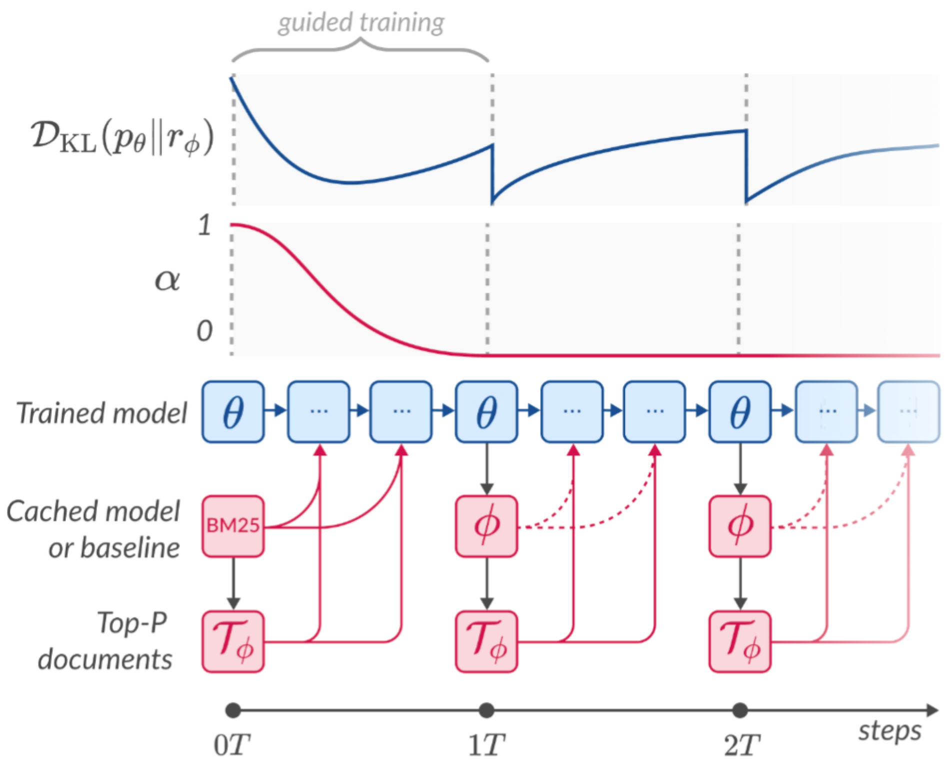
$$f_\phi^{\text{ckpt}} = 0 \implies f_\phi(\mathbf{a}, \mathbf{q}, \mathbf{d}) = \tau^{-1}(\text{BM25}(\mathbf{q}, \mathbf{d}) + \beta \cdot \text{BM25}(\mathbf{a}, \mathbf{d}))$$

  - Also maximises the answer likelihood independently of the retriever (in expectation over the approx. posterior)

$$\nabla_{\theta_{\text{READ.}}} \mathcal{L}_{\alpha=1}(\mathbf{a}, \mathbf{q}) = \mathbb{E}_{r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})} \left[\nabla_\theta \log p_\theta(\mathbf{a}|\mathbf{d}, \mathbf{q}) \right]$$

Training Procedure

- So, for the first T training steps we move from $\alpha = 1$ to $\alpha = 0$
  - Allows for initial knowledge distillation from BM25 to the retriever $p_\theta(\mathbf{d}|\mathbf{q})$
  - Then able to train on tighter bound
- At each iteration we use $r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})$ (fixed) to sample K documents, then we evaluate the VOD objective & gradients to update θ
- Every T steps we update $r_\phi(\mathbf{d}|\mathbf{a},\mathbf{q})$ by setting $f_\phi^{\text{ckpt}} = f_\theta$



Evaluation on Multiple-Choice Q&A

- Given a question $\mathbf{q}$ and each potential answer $\mathbf{a}$, evaluate 10 Monte Carlo estimates of the VOD and choose most likely answer.

$$\hat{L}_{\alpha}^{K}(\mathbf{a},\mathbf{q}) := \frac{1}{1-\alpha}\log\sum_{i=1}^{K} s_i \hat{v}_{\theta,\phi}^{1-\alpha}(\mathbf{a},\mathbf{q},\mathbf{d}_i)$$

$$(\mathbf{d}_1, s_1),\ldots,(\mathbf{d}_K, s_K) \overset{\text{priority}}{\sim} r_{\phi}(\mathbf{d}|\mathbf{a},\mathbf{q})$$

Experiments

Table 3. Open-domain question answering accuracy.

Method	Params.	Finetuning	MedMCQA Valid.	MedMCQA Test	USMLE Valid.	USMLE Test
VOD BioLinkBERT+BM25	110M	MedMCQA	51.6	55.3	–	–
VOD BioLinkBERT+BM25	110M	USMLE	–	–	41.0	40.4
VOD 2×BioLinkBERT	220M	MedMCQA	58.3	**62.9**	47.2	46.8
VOD 2×BioLinkBERT	220M	USMLE	–	–	45.8	44.7
VOD 2×BioLinkBERT	220M	MedMCQA→USMLE*	–	–	53.6	55.0
Disjoint PubMedBERT+DPR[1]	220M	MedMCQA	43.0	47.0	–	–
Disjoint PubMedBERT+BM25[2]	110M	USMLE	–	–	–	38.1
Disjoint BioLinkBERT+BM25[3]	110M	USMLE	–	–	–	40.0
Disjoint BioLinkBERT-L+BM25[3]	340M	USMLE			–	44.6
Reader only PubMedGPT[4]	2.7B	MedMCQA+USMLE	–	50.3	–	–
Reader only Galactica[5]	120B	MedMCQA	52.9	–	–	44.4
Reader only Codex 5-shot CoT[6]	175B	∅	59.7	62.7	–	60.2
Reader only FLAN-PaLM[7]	540B	∅	–	56.5	–	60.3
Reader only Med-PaLM[7]	540B	MedMCQA+USMLE	–	57.6	–	**67.6**
Random Uniform			25.0	25.0	25.0	25.0
Human Passing score[6]			50.0	50.0	60.0	60.0
Human Merit candidate[6]			90.0	90.0	87.0	87.0

Figure 1. Parameter efficiency. Answering accuracy of baseline methods and of VOD (BioLinkBERT backbone) on MedMCQA.

- Better results on MedMCQA (entry-level med-student knowledge) than USMLE (trained medical professional knowledge)
  - "BERT-sized model may not be sufficient for handling reasoning-intensive questions"

Experiments

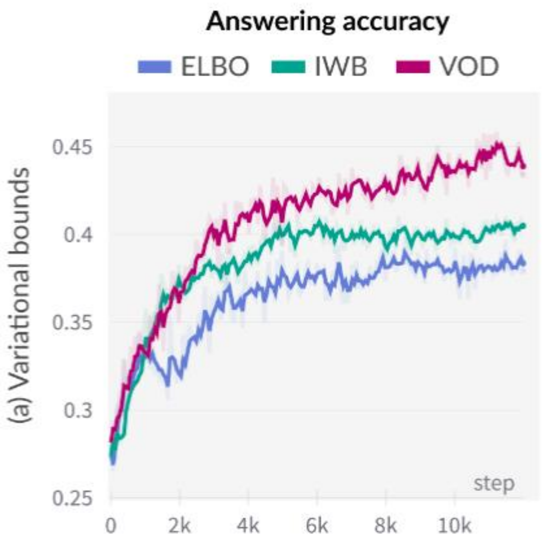
Table 4. Zero-shot accuracy on MMLU (%).

Task	**Subcategory**	**Unified QA**	**GPT-3**	**VOD**
medical_genetics	health	40.0	40.0	**76.0**
high_school_psychology	psychology	**70.0**	61.0	60.6
college_biology	biology	40.0	45.0	**59.7**
anatomy	health	43.0	46.0	**58.5**
clinical_knowledge	health	57.0	50.0	**58.5**
professional_medicine	health	43.0	38.0	**57.4**
nutrition	health	48.0	50.0	**56.5**
high_school_biology	biology	53.0	48.0	**55.2**
college_medicine	health	43.0	**47.0**	46.8
human_aging	health	**55.0**	50.0	44.4
virology	health	43.0	**44.0**	42.2
professional_psychology	psychology	**49.0**	45.0	42.2
Average	-	48.7	47.0	**54.8**

- Trained on MedMCQA only and evaluated on MMLU

Ablation Study

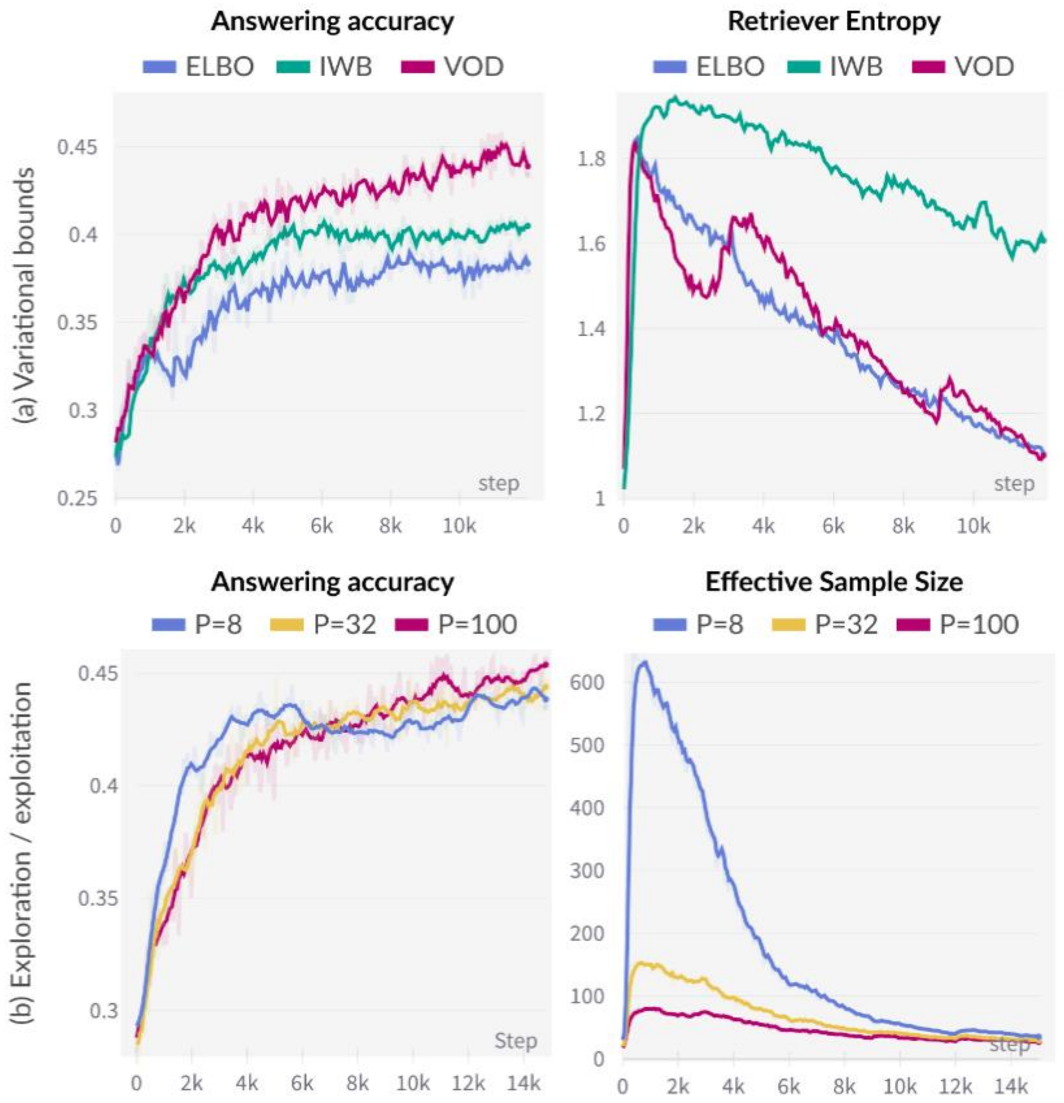
- Better performance when trained on VOD (interpolating α from 1 to 0) than with IWB ($\alpha = 0$) and ELBO ($\alpha = 1$)



Ablation Study

- Better performance when trained on VOD (interpolating α from 1 to 0) than with IWB ($\alpha = 0$) and ELBO ($\alpha = 1$)

- Larger P leads to:
  - Smaller effective sample size (measure of importance sampling quality, higher is generally better)
  - Slower learning
  - Better end performances



Re-purposing MCQA Retrievers for Sematic Search

- Using the trained VOD retriever to teach a query-only retriever "student" model by minimising:

$$L_{\text{DISTILL.}} = D_{\text{KL}}\big(\underbrace{r_\phi(\mathbf{d} \mid [\mathbf{q}; \mathbf{a}_\star])}_{\substack{\text{MCQA Teacher} \\ \text{(question+answer)}}} \,\|\, \underbrace{p_\theta(\mathbf{d} \mid \mathbf{q})}_{\substack{\text{Student} \\ \text{(question only)}}}\big)$$

- MRR = 100 * mean reciprocal rank (of first document with correct "disease concept" label)
- Hit@20 = fraction of queries for which correct document is in top-20 returned articles

Table 5. Retrieval performances on the FindZebra benchmark for a BioLinkBERT retriever trained using VOD on MedMCQA and one trained using task-specific distillation, with and without coupling with a BM25 score during evaluation.

Method	**Distillation**	**MRR**	**Hit@20**
VOD	✗	27.8	56.9
VOD	✓	31.7	58.1
VOD + BM25	✓	**38.9**	**64.1**
BM25	–	26.4	48.4
FINDZEBRA API	–	30.1	59.3

Conclusion

- Good results, authors hope more attention will be given to Rényi divergence VI in NLP
- VOD could be used in areas other than the ODQA presented here:
  - Generative and extractive ODQA outside of multiple-choice ODQA
  - Retrieval-augmented language modelling: retrieve one document per input token
  - Fusion-in-Decode (FiD): using a reader model that takes in multiple documents
- Authors want more training (on larger datasets) and to use larger models than BERT
- "Additional theoretical analysis is required" to investigate bias induced in VOD objective via self-normalised priority sampling.

$$s_i := \bar{s}_i / \sum_{j \in \mathbb{S}} \bar{s}_j$$