HINDSIGHT: POSTERIOR-GUIDED TRAINING OF RETRIEVERS FOR IMPROVED OPEN-ENDED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many text generation systems benefit from using a retriever to retrieve passages from a textual knowledge corpus (e.g., Wikipedia) and providing these passages as additional context to the generator. For open-ended generation tasks (like generating informative utterances in conversations) many varied passages may be equally relevant and we find that existing methods that jointly train the retriever and generator underperform: the retriever may not find relevant passages even amongst the top-10 and the generator may hence not learn a preference to ground its generated output in them. We propose using an additional guide retriever that is allowed to use the target output and "in hindsight" retrieve relevant passages during training. We model the guide retriever after the posterior distribution Q of passages given the input and the target output and train it jointly with the standard retriever and the generator by maximizing the evidence lower bound (ELBo) in expectation over Q. For informative conversations from the Wizard of Wikipedia dataset, with posterior-guided training, the retriever finds passages with higher relevance in the top-10 (23% relative improvement), the generator's responses are more grounded in the retrieved passage (19% relative improvement) and the end-to-end system produces better overall output (6.4% relative improvement).

1 INTRODUCTION

In knowledge-intensive NLP tasks, models must use open-domain knowledge to answer questions (Kwiatkowski et al., 2019; Joshi et al., 2017), fact-check claims (Thorne et al., 2018), engage in informative conversations (Dinan et al., 2019), or otherwise facilitate information access through natural language (Nguyen et al., 2016). To leverage the information present in human-readable corpora (e.g., Wikipedia), many recent models for open-domain question answering (Lee et al., 2019; Khattab et al., 2021) are *retrieval-augmented*: they extract passages from the corpus using a learned *retriever* and process it with a task-specific *reader*. If the relevant passage is known, the retriever is supervised to find it; otherwise, it is weakly supervised with passages that contain the answer string (e.g., a name or a date). In recent work (e.g., Lewis et al., 2020), for end-to-end tasks with arbitrary text output, where relevant passages are not known, a jointly-trained generative sequence-to-sequence model provides the relevance signal.

Current methods work well for short-answer QA-like problems, where the query itself has high overlap with the relevant passage and there is a unique answer, for example factoid Open-domain QA such as Natural Questions (Kwiatkowski et al., 2019), HotPotQA (Yang et al., 2018), fact-checking (Thorne et al., 2018), and slot-filling (Levy et al., 2017). However, for *open-ended* generation tasks like free-form QA (MS-Marco NLGen), informative dialogue (Dinan et al., 2019), or Wikipedia abstract generation (Liu* et al., 2018), many responses can be equally acceptable and yet the data source (e.g., human-written answers, real-world conversations, Wikipedia abstracts) supplies only one plausible output from a space that includes many. Figure 1 illustrates this with a conversational context x and multiple relevant passages z: while only one of them z_{gold} is useful for generating the gold-standard answer y, other passages could have led to other coherent responses. This setting with one input and many possible unobserved outputs (**one-to-many**) is challenging, as it creates a distinction between the numerous passages that are relevant to the input (which we dub **context-relevant** passages) and the few passages that pertain to the observed target output (which we dub **label-relevant** passages). Had we known z_{gold} corresponding to the target output, we could have supervised the retriever with z_{gold} and trained the generator conditioned on z_{gold}, but we don't!

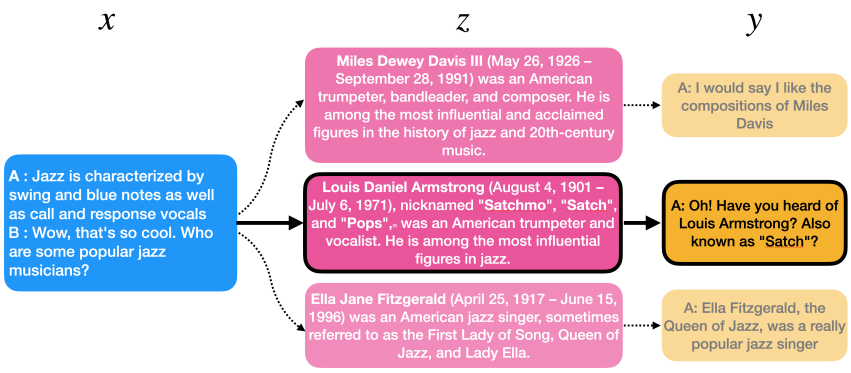

Figure 1: A conversational turn with multiple plausible responses. The input (blue) can be answered based on 3 equally relevant passages but only one possible response (yellow) is observed in the training set (shown outlined in black) based on only one of the pink relevant passages (black outline).

Current methods (Lewis et al., 2020) attempt to use the generator's probability distribution $P_\theta(y|x, z)$ as a proxy for label relevance and train the retriever $P_\eta(z|x)$ by marginalizing $p(y|x)$ over retrieved documents z: $P(y|x) = \sum_{z \in \text{top-k}(P_\eta(.|x))} P_\eta(z|x)P_\theta(y|x, z)$. We find empirically that for one-to-many tasks, this objective is suboptimal in three ways: the generator is less grounded in the retrieved passages (Figure 3), the retriever performance saturates at low recall (Figure 3), and at training time the top-k retrieved passages miss out on many label-relevant passages weakening the supervision (Table 1).

We propose explicitly estimating the label-posterior distribution $Q(z|x, y)$ in the form of a **guide-retriever** that has access to the target output, and with this "hindsight" can capture label-relevance precisely. We jointly optimize the retriever, posterior-guide, and generator using the evidence lower bound (ELBo): $\mathbb{E}_{z_i \sim Q(.|x,y)}[\log P_\theta(y|x, z)] - D_{\text{KL}}(Q|P_\eta)$. In theory, posterior-guided training is better for joint-training on open-ended tasks on all three counts: (1) The generator gets conditioned on passages weighted directly by their label-relevance from the label-posterior distribution, reducing the tendency of the generator to ignore the retrieved passage; (2) The retriever is trained with reverse-KL divergence, whose mode-seeking nature encourages the retriever's probability distribution to match some modes with the guide (label-relevant passages), without explicitly penalizing other modes (other context-relevant passages); (3) As the label-posterior distribution $Q(z|x, y)$ is modeled by a full-fledged retriever, it can retrieve label-relevant passages from the entire collection, which is a generalization of weak supervision approaches that use span overlap for short-answer QA.

Our main contribution in this paper is the HINDSIGHT training system that: (1) uses a guide-retriever to provide a stronger learning signal for both the generator and the retriever, (2) is amenable to index-updates with iterative closed-set training (Section 3). We thoroughly evaluate the retrieval-augmented systems, going beyond end-to-end performance with a granular evaluation of the individual models (retriever and generator) at varying passage depths, important for evaluating one-to-many open-ended generation tasks (Section 4). Using HINDSIGHT on the Wizard of Wikipedia dataset of informative conversations our retriever achieves a 23% relative improvement in success@10 (its ability to have the label-relevant passage among the top-10 retrieved passages), our generator is more grounded with 19% improvement in Novel-F1 overlap with the top-1 retrieved passage (i.e., its overlap with rare words from the retrieved passage that were not already in the input) and 6.4% relative improvement in Novel-F1@1 overlap with the gold utterance (the best matching generation when considering top-1 retrieved passage). We also validate that HINDSIGHT improves performance on the one-to-one generation task of free-form QA using the MS-MARCO NLGen dataset.

2 BACKGROUND

Open-domain Question Answering While Question-Answering datasets originally focused on extracting the answer from a given passage (a.k.a. reading comprehension), recent work has made the task more challenging by not supplying a gold passage but instead expecting the model to con-

duct open-domain QA directly over a large document collection. The first neural system for answering open-domain factoid questions was Dr. QA (Chen et al., 2017), which used an off-the-shelf retriever (e.g., TF-IDF, BM25) to retrieve relevant passages and a separately trained reader to extract the answer span. The advent of efficient nearest-neighbour search algorithms (Jegou et al., 2010; Johnson et al., 2017) opened up the possibility of encoding the query and the passage collection into a vector space and finding the relevant passage as nearest neighbors to the query embedding. Subsequent work trained neural retrievers in various ways such as: pretraining with the inverse cloze task then weakly supervising using span matches (Lee et al., 2019), using gold passages with in-batch negatives (Karpukhin et al., 2020), and retrieval-guided supervision with span-based positives (Khattab et al., 2021).

Open-ended Generation Natural language generation tasks involve generating a sequence of tokens (maybe, word-pieces), often contextualized on some input (another sequence of tokens for sequence-to-sequence tasks, an image for an image captioning task). However, they differ in how open-ended they can be. Some tasks such as factoid question-answering having a single correct short answer are less open-ended than free-form long answers. For example, in informative dialogue the speakers can lead the conversation in many different directions (Dinan et al., 2019), also referred to as one-to-many generation. Thus it is more open ended, or, equivalently, it has a higher entropy than say machine translation Bojar et al. (2014) which has few correct translations that are very similar to each other. Many more generation tasks such as summarization (Narayan et al., 2018) and story generation (Mostafazadeh et al., 2016) lie on this spectrum.

Retrieval for Language Modeling To improve the perplexity of a pre-trained language model, Khandelwal et al. (2020) retrieve similar contexts from the training set at each time-step and increase the likelihood of tokens that were predicted in similar contexts. Guu et al. (2020) instead pre-train a retrieval-augmented masked language model using salient-span masking and fine-tune it on downstream QA tasks. Under the paradigm of retrieval-augmented generation, for input x and output y, Lewis et al. (2020) retrieve top-k passages (z) from a corpus with a retriever and jointly train the generator (P_θ) and the retriever (P_η) by maximizing the likelihood of the output marginalized over the top-k documents, which we refer to as the MARGINALIZEDLOSS:

$$(y|x) = \sum_{z \in \text{top-k}(P_\eta(.|x))} P_\eta(z|x) P_\theta(y|x,z) \tag{1}$$

Here $P_\theta(y|x,z)$ has two roles: supervision, i.e. providing label-relevance by scoring label-relevant passages higher than other passages and grounding, i.e. maximizing the probability of the target output given context-relevant passages. For one-to-many datasets, where few context-relevant passages are label-relevant, the two roles are at odds with each other. By increasing label-relevance signal the generator reduces the ability to ground in passages that are context-relevant but not label-relevant. By increasing grounding in all context-relevant passages it reduces the label-relevance signal. In the next section, we describe our method we separates these two concerns, leading to improved performance.

3 TRAINING WITH HINDSIGHT

To precisely identify label-relevant passages, we propose explicitly modeling the posterior distribution: $Q(z|x,y)$ with a learned neural model. Unlike the retriever, the label-posterior model has access to the target output and in hindsight can differentiate label-relevant from other context-relevant passages. We learn the label-posterior jointly with the retriever and the generator by maximizing the evidence lower bound, ELBOLOSS, as given by the formula:

$$\log P(y|x) \geq \mathbb{E}_{z \sim Q(.|x,y)}[\log P_\theta(y|x,z)] - D_{\text{KL}}(Q|P_\eta) \tag{2}$$

For an intuitive understanding we look at the two terms separately. The first term is an expectation of the generator's log-likelihood P_θ over the label-posterior Q. This ensures that the generator need attend only to the label-relevant passages and therefore learns to "trust" the retrieved passages better. The second term is the KL divergence from the retriever to the label-posterior, also referred to as reverse KL divergence:

$$D_{\text{KL}}\big[Q(z|x,y) \mid P_\eta(z|x)\big] = \sum_{z \sim Q(.|x,y)} Q(z|x,y)\big(\log Q(z|x,y) - P_\eta(z|x)\big)$$

This term is again weighted by $Q(z|x, y)$, which is the probabilistic equivalent of implication: high $Q(z|x, y)$ implies high $P(z|x)$, i.e., label-relevance implies context-relevance but not vice-versa.

Posterior as a retriever While we can certainly model the label-posterior $Q(z|x, y)$ as a re-scorer that takes in documents as retrieved by the retriever P_η, we instead model it as a guide retriever that can retrieve label-relevant passages from the entire corpus. This is helpful because we can sample passages directly from the label-posterior distribution, and estimate the ELBOLOSS accurately than using passages from $P_\eta(z|x)$. The guide retriever generalizes weak supervision approaches (Lee et al., 2019; Guu et al., 2020) and relevance-guided supervision (Khattab et al., 2021), to **posterior-guided supervision** with a learned posterior retriever rather than brittle heuristics based on word-overlap.

Iterative closed-set training Prior work (Guu et al., 2020; Khattab et al., 2021) has shown the utility of intermittently updating the passage index during training. To allow for such a workflow, we organize our training into rounds (see Figure 2). At the beginning of each round, in the outer loop we encode the passages and the queries with various retrievers and find the highest scoring r passages that we dub the closed-set. In the inner loop that runs for many epochs, we sample passages from the closed-set ($r = 100$ for our experiments). This is fast because we are no longer retrieving from the entire corpus in the inner loop and also sufficient because the closed-set has a high recall. We update the retrievers (both document and query encoders) during the inner loop and use the latest model parameters for computing the loss functions. A round results in trained models that are then used for the next round. We find that 2 rounds are often sufficient, with decreasing marginal utility from the third round onward.

Distributional repositioning before inference According to ELBOLOSS, the expectations are computed over $z \sim Q(.|x, y)$. We cannot compute the expectations exactly because of the large size of the passage corpus, so we approximate it by sampling k passages from the closed-set $Q_{\text{top-r}}(Q(.|x, y))$. This leads to faster model training, especially at the beginning because passages from $Q(.|x, y)$ provide better supervision. However, due to this sampling scheme, the models only ever get exposed to passages from the $Q(.|x, y)$ distribution, which limits their ability to generalize over passages from $P_\eta(.|x)$ during inference. To remedy this, we sample passages from an α-mixture of the two distributions: $z \sim \begin{cases} P_\eta(.|x) & \text{with prob. } \alpha \\ Q(.|x, y) & \text{with prob. } 1-\alpha \end{cases}$. In the initial rounds we set low values of α and increase it toward the end to reposition the passage distribution and better match with $P_\eta(.|x)$ at test time.

Training individual models to convergence A practical issue with joint training is that the retriever and generator train at different rates, reaching convergence at different times. The single term in MARGINALIZEDLOSS (Eq. 1) does not indicate convergence of individual models leading to situations where one model starts to overfit while the other model still hasn't converged. With ELBOLOSS there are two terms in Eq. 2: the first connecting $Q(z|x, y)$ and $P_\theta(y|x, z)$, the second connecting $Q(z|x, y)$ and $P_\eta(z|x)$. This allows us to fix the guide and train the retriever and generator independently for a differing number of gradient steps until each model reaches convergence.

4 EXPERIMENTAL EVALUATION

We evaluate on two open-ended knowledge-intensive tasks: informative conversations and free-form question answering. We ask the following three research questions:

RQ$_1$ Relevance: Are the retrieved passages more relevant? (Section 4.3)
RQ$_2$ Groundedness: Does the generator make better use of the retrieved passages? (Section 4.4)
RQ$_3$ Generation Quality: Does this lead to better end-to-end performance? (Section 4.5)

4.1 MODELS

Retriever Models We use ColBERT (Khattab & Zaharia, 2020) to model the retriever $P_\eta(z|x)$ and the guide-retriever Q. The query tokens q_i and the document tokens d_j are independently en-

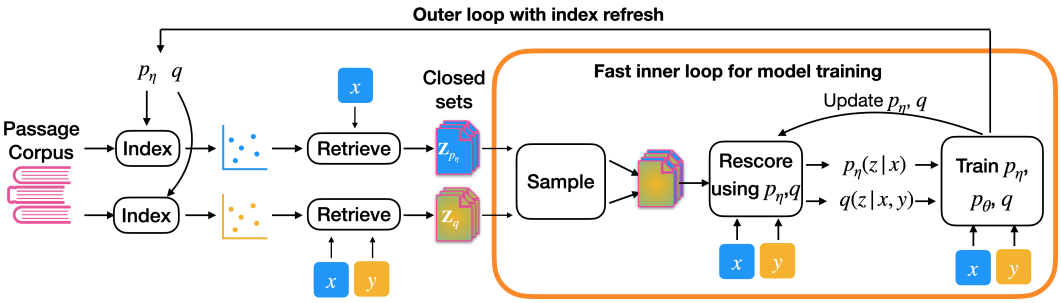

Figure 2: An overview of iterative closed-set training: We iterate through the outer-loop and call each execution a round. At the beginning of the round we re-index the passage corpus using the latest retriever p_η and guide-retriever q to create a high-recall closed-set of top-r passages for each retriever and query. Then, in the fast inner loop, we train the models for multiple epochs by sampling passages from the fixed closed-set and recomputing the probability distributions. The trained models are then used in the next round.

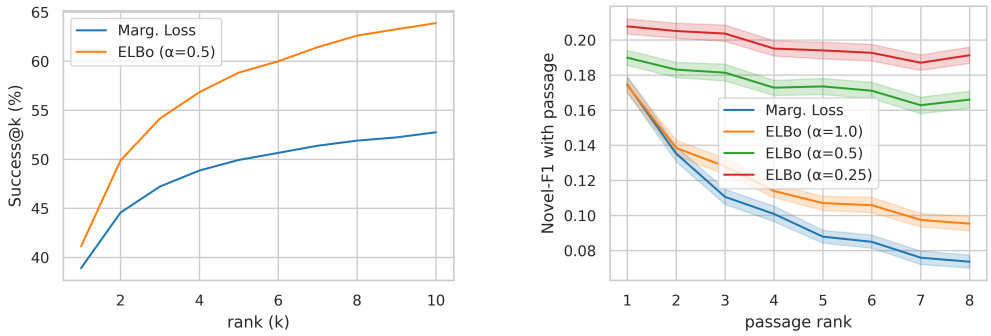

Figure 3: Relevance and Groundedness of models trained on the Wizard of Wikipedia dataset: (**left**) success@k of retrieved passages w.r.t. rank and (**right**) Novel-F1 between decoded output and retrieved passage w.r.t retrieved passage rank. The ELBoLoss retriever is more effective at retrieving the gold passage than MarginalizedLoss retriever, especially when we consider the top-10 passages for this one-to-many task. The ELBoLoss generators have higher overlap with top-k retrieved passages and the overlap increases as α decreases

coded using BERT and normalized to produce unit-vectors E_{q_i} and E_{d_j}. The similarity between the query and the document is defined as $S_{q,d} = \sum_i \max_j E_{q_i} E_{d_j}^T$. The late-interaction paradigm of ColBERT is more expressive than using just the [CLS] token, as in DPR (Karpukhin et al., 2020), because it allows the query and document tokens to retain their identities and provide finer-grained term-wise similarity. Recently, Khattab et al. (2021) has shown that ColBERT establishes state of the art retrieval results on open-domain QA benchmarks. To calculate the probability distributions during training, we sample a small set of documents $\mathcal{R}$ using a suitable sampling strategy (as explained in Section 3) and compute the softmax of the scores over $\mathcal{R}$. For the posterior-retriever, we concatenate the input and the output to create the query, $q = [x\ y]$. For the Wizard of Wikipedia task, we use a ColBERT model that has been pre-trained on the MS-MARCO passage ranking dataset. However, for the MS-MARCO NLGen task, we use another ColBERT model that has been pre-trained using the Natural Questions dataset instead.

Generation Model Following Lewis et al. (2020) we use pre-trained BART model and fine-tune it for the respective tasks during training. It is conditioned on both the context and the document and trained to produce the target. At test time, we decode using beam-search with 4 beams.

Table 1: Relevance evaluation: Our method (ELBoLoss Retriever, $\alpha = 1$) strongly improves over the baseline (MarginalizedLoss Retriever) for the one-to-many Wizard of Wikipedia dataset, in particular for $k = 5, 10$. Our method shows smaller but consistent improvements on the one-to-one MS MARCO NLGen dataset. The ELBo posterior finds z_{gold} with high success providing better supervision during training. (MRR = Mean Reciprocal Rank, Success@k both in percentages)

	Wizard of Wikipedia				MS MARCO NLGen			
Method	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10
Marg. Retriever	43.8	38.9	49.9	52.8	30.4	19.4	43.4	53.2
ELBo Retriever	**49.0**	**41.1**	**58.8**	**63.9**	**32.1**	**21.2**	**45.3**	**54.4**
ELBo Posterior	78.5	72.4	86.0	88.4	67.8	56.7	81.9	86.2

4.2 Tasks

Informative conversations We first evaluate on generating utterances in informative conversations as a one-to-many generation task that is open-ended and knowledge-intensive. Informative conversations are one-to-many because they are open-ended and people have the agency to drive the conversation in a different direction at every turn. We use the Wizard of Wikipedia (WoW) dataset (Dinan et al., 2019), where an "apprentice" chats (via text) with a "wizard", being curious about different topics, and the "wizard" grounds their response in a sentence from Wikipedia. The input for this task is the conversational history x and the output is the wizard's utterance y with the models provided access to individual passages (z) from all of Wikipedia ($\approx$26 million passages). We use the version of this dataset provided in the KILT benchmark (Petroni et al., 2021). As the test set outputs are not available to us, we show granular evaluation numbers on the dev set and report leaderboard performance on the test set.

Free-form Question Answering We also validate the efficacy of our method on on-to-one generation task: free-form open-domain QA. We use the MS-MARCO NLGen dataset (Nguyen et al., 2016) where the task is to generate natural-sounding answers to questions, which is more challenging than other *extractive* open-domain QA datasets. This is a subset of MS-MARCO questions whose answers were reviewed by a separate editor and rewritten if there was a high overlap between the answer and one of the provided passages (indicating that the original editor may have copied the passage directly). These "well-formed answers" are meant to be complete sentences (such as can be read out by a conversational assistant) and have a median length of 11 words. The input for this task is a query x, the output is a well-formed answer y, and the models are required to retrieve from the entire collection provided with the MS-MARCO QA consisting of 8.8 million passages from the web. As the public benchmark is no longer available we could not evaluate on the benchmark test set. Instead we split the public validation set into a validation and test set and show results on the test set.

While both datasets annotate the passages referred to by the person who wrote the target output (gold passages), we only use them for evaluation and validation and not for training nor for validation.

4.3 Relevance Evaluation

We begin by investigating the quality of the retrieved passages ($\mathbf{RQ}_1$), leveraging the gold passage labels supplied by each dataset. To evaluate relevance, we report Success@k (S@k for short), the percentage of inputs for which the gold provenance passage is retrieved within the the top-k passages by each system. While multiple passages may lead to the desired output, our datasets mark only a single passage as the gold provenance for each input, and so we report success at retrieval depths $k = \{1, 5, 10\}$. We also report Mean Reciprocal Rank (MRR), an evaluation metric commonly used to evaluate IR systems.

Our results are shown in Table 1. Starting off with Wizard of Wikipedia, we observe that our ELBoLoss retriever markedly outperforms MarginalizedLoss at finding the label-relevant passage. While both systems easily handle 38.9–41.1% of the examples (evident in relatively high success@1), only the ELBoLoss retriever continues to find many more relevant passages at larger

Table 2: Groundedness evaluation: Our method ELBOLOSS ($\alpha = 0.25$) shows more overlap between generated output and the retrieved passage than MARGINALIZEDLOSS and for the Wizard of Wikipedia dataset the gap increases as we consider the maximum over top-5 passages. (Novel-F1: discounts commonly occurring words and context words (x))

Dataset	Method	Top-1 F1	Top-1 Nov-F1	Max. of Top-5 F1	Max. of Top-5 Nov-F1
WoW	Marg. Gen.	18.63	17.46	26.19	25.39
	ELBo Gen. ($\alpha = 0.25$)	**21.34**	**20.78**	**34.16**	**34.24**
MSM	Marg. Gen.	33.12	25.39	45.76	39.45
	ELBo Gen.($\alpha = 0.5$)	**34.52**	**26.49**	**46.91**	**40.47**

retrieval depths k, reaching success@10 of 63.9%. This considerably exceeds MARGINALIZED-LOSS's success@10 (at 52.8%). Interestingly, the success@5 rate of ELBOLOSS is, in fact, higher than MARGINALIZEDLOSS's success@100 (not reported in the table), where the latter saturates at 55.8% despite containing 20× more passages. In contrast, ELBOLOSS's success@100 reaches 69.3%.

To explain the much stronger success@k scores for $k > 1$ under ELBOLOSS, we hypothesize that besides straightforward examples where both systems find a unique "best" passage, the label-relevant passages are often less apparent and, in those cases, the ELBOLOSS retriever's better-refined training signal allows us to find provenance for more challenging examples. This is confirmed by our informal inspection of results, where we found that the sets of passages retrieved by the two methods on Wizard of Wikipedia are qualitatively different. While not captured by word-overlap measures, we found that the MARGINALIZEDLOSS retriever would select many similar "safe" passages while the ELBOLOSS retriever would select a more diverse and rewarding set of passages (see examples in Appendix A.1).

Shifting our attention to MS MARCO NLGen, we notice that ELBoLoss still outperforms MarginalizedLoss by 1–2 points across our metrics, reflecting smaller—but nonetheless consistent—gains when compared with the one-to-many generation task. We also observe that success@1 is lower for both methods, when compared with Wizard of Wikipedia, a fact we largely attribute—based on manual inspection—to the presence of many false negatives, i.e., passages that contain the answer but aren't marked as gold passages, which is consistent with the findings from related studies Arabzadeh et al. (2021). Overall, we find that ELBOLOSS improves relevance of retrieved passages over MARGINALIZEDLOSS for two qualitatively different tasks, with larger gains for the one-to-many generation task.

4.4 GROUNDEDNESS EVALUATION

We now examine $\mathbf{RQ}_2$, studying the degree to which the generator relies on the retrieved passages for producing its output. To quantify this *groundedness*, we compute F1-overlap between a *retrieved passage* (not necessarily the gold passage) and the produced text when generation is conditioned on that passage.

As an analogue of Success@k, we propose *Max. F1@k*, the largest F1-overlap exhibited by any generated output with the corresponding retrieved passage fed to the generator. We also propose Novel-F1, a new metric that discounts words the occur frequently and words that already appear in the context x, since otherwise these tokens dominate raw F1 in practice (up to 80%, see Appendix A.2). We restrict this to the novel words, since it is easy for the generator to "copy" common words without that indicating grounding in the content of the passage.

Our results are shown in Table 2 and Figure 3. For the Wizard of Wikipedia dataset, we observe that our ELBOLOSS generator outperforms MARGINALIZEDLOSS by 2.7 F1 (14.5% r.i.) and 3.3 Novel-F1 (19% r.i.) when given the top retrieved passage. In Figure 3 (right), we can see that beyond the top passage, MARGINALIZEDLOSS generator's overlap decays rapidly whereas ELBOLOSS ($\alpha = 0.25$) generator's overlap declines gradually. This shows that the ELBOLOSS generator stays grounded beyond just the top passage, a desirable property in one-to-many generation systems. The

Table 3: Wizard of Wikipedia KILT leaderboard evaluation: ELBOLOSS achieves SoTA on generation metrics (F1, ROUGE-L, KILT-F1, KILT-ROUGE-L indicated with †) as of Oct 2021 and improves relevance over MARGINALIZEDLOSS

	R-Prec	Recall@5	F1	ROUGE-L	KILT-F1	KILT-ROUGE-L
Marg.	53.94	68.12	18.11	16.21	11.78	10.47
ELBo	**56.08**	**74.26**	**19.19**†	**17.06**†	**13.39**†	**11.92**†

above plot could hide the scenario where MARGINALIZEDLOSS is grounded in just one out of the top-k passages (say the gold passage), reducing the average overlap while still being grounded. Therefore we consider Max. F1@k and Max Novel-F1@k and observe that ELBOLOSS generator outperforms MARGINALIZEDLOSS by 8 F1@5 and 8.8 Novel-F1@5, which shows that even when allowed to pick the most suitable passage, MARGINALIZEDLOSS heavily underperforms. For the MS MARCO NLGen dataset, we observe smaller but consistent gains in groundedness (1–2 F1, Novel-F1) with MARGINALIZEDLOSS compared to ELBOLOSS. This is expected because for a QA task $P_\eta(z|x)$ is a good proxy for label-relevance and ELBOLOSS gains little by optimizing the expectation of the log-likelihood over $Q(z|x, y)$. Overall, we see that ELBOLOSS is more grounded than MARGINALIZEDLOSS, with a bigger margin for one-to-many tasks.

4.5 END-TO-END EVALUATION

To evaluate the end-to-end quality of our systems, we calculate F1 and Novel-F1 (defined in Section 4.4) of the decoded output with the *human-written gold output*. To allow for the possibility of the generator using any part of the long passages (150 words) for the Wizard Of Wikipedia task, we use Knowledge-F1 (defined by Shuster et al. (2021)): F1 between the sampled generation and the *gold passage*. Since it is reasonable to expect the gold passage to be in top-k for k>1 for one-to-many tasks (as in Section 4.4), we also compute the Max. over top-k retrieved passages.

The results are summarized in Table 4. For Wizard of Wikipedia, we see that using only the top retrieved passage ELBOLOSS only slightly improves performance over MARGINALIZEDLOSS. This is expected in the one-to-many setting: the label-relevant passage is an arbitrary choice from amongst the context-relevant passages, and the 6.7% relative improvement in Novel-F1@1 is consistent with the relatively small improvements in relevance and grounding for the rank 1 passage. We see larger improvements for ELBOLOSS with Max. overlap over the top-5 passages, namely 1 F1, 2 Novel-F1 ($\sim 15\%$ r.i.), and 2 K-F1 ($> 10\%$ r.i.). This shows the retriever and generator improvements for $k > 1$ contribute to a better end-to-end system, with one of the generated outputs having a higher overlap with the target output because it was better grounded in a label-relevant passage in the top-k retrieved passages. For MS Marco NLGen, we see a small but consistent increase due to ELBOLOSS over MARGINALIZEDLOSS: 1.5 F1 and 2 Novel-F1 across passage depths.

We also submit models trained using ELBOLOSS and MARGINALIZEDLOSS on Wizard of Wikipedia to the KILT leaderboard. The results are reported in Table 3. As shown, ELBOLOSS consistently outperforms the baseline MARGINALIZEDLOSS across all metrics. The table also reports Recall@5, which evaluates retrieval at a coarser granularity, namely at the *full Wikipedia page* level, though so far we have investigated it directly at the passage level. Consistent with the results in Table 1, our method also outperforms MARGINALIZEDLOSS in retrieval metrics. In fact, our ELBOLOSS model achieves state-of-the-art performance across all the generation metrics (F1, ROUGE-L, KILT-F1, KILT-ROUGE-L) on the leaderboard, though it is not the strongest on Recall@5.[1]

To conclude, we have evaluated the ELBOLOSS and MARGINALIZEDLOSS using a one-to-one free-form QA dataset and a one-to-many dataset of informative conversations. Our results show that our method **ELBoLoss** trains a better retriever, a more grounded generator and improves end-to-end performance, especially in the one-to-many setting.

[1]Earlier results on the KILT leaderboard for Wizard of Wikipedia should be interpreted with caution, as the KILT authors recently updates the train/dev splits due to anomalies in the preprocessing script. We have used the updated version for our model and baseline.

Table 4: End-to-end evaluation: Our method ELBOLOSS improves over MARGINALIZEDLOSS when considering Max. overlap of generated output with target output over top-5 passages for the Wizard of Wikipedia dataset and also for top-1 with MS Marco NLGen dataset. (Novel-F1: discounts commonly occurring words and context words (x), Knowledge-F1: overlap of generated output with gold passage)

		Top-1			Max. of Top-5		
Dataset	Method	F1	N-F1	K-F1	F1	N-F1	K-F1
WoW	Marg.	18.79	10.45	12.61	26.52	16.42	12.45
	ELBo	**18.86**	**11.12**	**13.08**	**27.56**	**18.67**	**14.62**
MSM	Marg.	60.18	37.19	–	72.22	56.06	–
	ELBo	**61.46**	**39.65**	–	**73.18**	**58.19**	–

5 DISCUSSION

Hallucination, grounding and correctness Shuster et al. (2021) show that providing retrieved passages to a generator has been shown to reduce hallucination. Our work increases grounding in the retrieved passage, promising to further reduce hallucination. Even though the generator is now more likely to use content from the provided passage (rather than hallucinating from parametric memory), that does not guarantee correctness.This is not captured perfectly by our token-level overlap metrics that evaluate grounding. There is scope for future work to address this gap with better generator training methods that not only produce grounded but also correct outputs.

Passage selection for controllable generation Our results show that the HINDSIGHT generator stays grounded beyond the top retrieved passage. As a consequence, we can exert considerable control over the generated content by controlling the passage provided to the generator, potentially complementing controllable language generation using special tokens or attributes (Dathathri et al., 2020). This can have significant impact when such systems are deployed in real-life settings (e.g., in open-domain socialbots) where external business logic with concrete objectives can help select an appropriate passage from the top-k retrieved passages and control the generator's output.

Modulating trust For tasks like QA, we want a "conservative" generator: it should abstain from using a passage that doesn't contain the answer. For more open-ended tasks like informative conversations, we would like the generator make use of diverse passages. In the first case, we want the generator to "trust" the retrieved passages less and in the second case, more . In our work, we show that by changing the distribution of the passage from $P_\eta(.|x)$ to $Q(.|x, y)$, the generator increasingly trusts the retrieved passages (Figure 3). Based on the nature of the task, system designers can use the α-mixture to modulate the degree of trust placed by the generator in the retrieved passages.

Comparison with Fusion-in-Decoder Izacard & Grave (2021a) provide multiple passages to the generator simultaneously and in Izacard & Grave (2021b) they use the decoder's attention weights over each passage for relevance supervision. While this is a valid mechanism, perhaps conditioning on individual passages like we do is more precise for relevance supervision. Indeed, recent work (Sachan et al., 2021) illustrates by using Fusion-in-Decoder during inference but foregoing the decoder's attention weights and using an equivalent version of the MARGINALIZEDLOSS for training the retriever. Furthermore, Fusion-in-Decoder is uniquely useful for QA style tasks, where it has to select the correct answer from many passages. But for one-to-many tasks, our system can condition on each passage separately and generate diverse outputs whereas with Fusion-in-Decoder that cannot happen.

In this paper, we propose HINDSIGHT, a system that introduces a guide-retriever to improve supervision for both the retriever and the generator for retrieval-augmented, open-ended generation. During training, the guide retriever is allowed to use the target output of each example—alongside the input context—in order to find relevant passages, improving the coverage of useful passages during training and, in turn, leading to better retrieval and more grounded generation. The resulting system achieves considerable empirical improvements over existing work, improving retrieval quality by up to 23%, grounding by up to 19%, and end-to-end output quality by up to 6.4%.

REFERENCES

Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles LA Clarke. Shallow pooling for sparse labels. *arXiv preprint arXiv:2109.00062*, 2021.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3302. URL https://aclanthology.org/W14-3302.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL https://aclanthology.org/P17-1171.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1l73iRqKm.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL http://proceedings.mlr.press/v119/guu20a.html.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, April 2021a. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.74.

Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=NTEz-6wysdb.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*, 2017.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://www.aclweb.org/anthology/2020.emnlp-main.550.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pp. 39–48, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Omar Khattab, Christopher Potts, and Matei Zaharia. Relevance-guided Supervision for OpenQA with ColBERT. *Transactions of the Association for Computational Linguistics*, 9:929–944, 09 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00405. URL https://doi.org/10.1162/tacl_a_00405.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL https://aclanthology.org/P19-1612.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. *CoRR*, abs/1706.04115, 2017. URL http://arxiv.org/abs/1706.04115.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Peter J. Liu*, Mohammad Saleh*, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hyg0vbWC-.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1098. URL https://aclanthology.org/N16-1098.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, 2016. URL http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. End-to-end training of neural retrievers for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6648–6662, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.519. URL https://aclanthology.org/2021.acl-long.519.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. URL https://nlp.stanford.edu/pubs/yang2018hotpotqa.pdf.

rank	text
1.0	Bucatini > Abstract | Bucatini , also known as perciatelli , is a thick spaghetti-like pasta with a hole running through the center. The name comes from , meaning "hole", while "bucato" or its Nea...
2.0	Pasta con le sarde > Ingredients. | The principal ingredients are olive oil, onions, pasta and a finely chopped mixture of sardines and anchovy. Various types of pasta are used for the dish, but b...
3.0	Bucatini > Preparation. | Standard pasta machines will roll out sheets of flat pasta which are then cut into ribbons to make flat, ribbon-style pasta like fettuccine, tagliatelle, or pappardelle. ...
4.0	Bocconcini > Abstract | This cheese is described by its Italian name, which means "small mouthfuls". It is made in the "pasta filata" manner by dipping curds into hot whey, and kneading, pulling, ...
5.0	Carbonara > Abstract | Carbonara () is an Italian pasta dish from Rome made with egg, hard cheese, guanciale (or pancetta), and black pepper. The dish arrived at its modern form, with its current ...

Table 5: Passages retrieved by ELBoLoss retriever while talking about Italian Cuisine. They include passages about various ingredients (rank=2), cheeses (rank=4), dishes (rank=5) alongside more information about Bucatini Pasta (rank=1,3).

rank	text
1.0	Bucatini > Abstract | Bucatini , also known as perciatelli , is a thick spaghetti-like pasta with a hole running through the center. The name comes from , meaning "hole", while "bucato" or its Nea...
2.0	Bucatini > Preparation. | Standard pasta machines will roll out sheets of flat pasta which are then cut into ribbons to make flat, ribbon-style pasta like fettuccine, tagliatelle, or pappardelle. ...
3.0	Rotini > Abstract | Rotini is a type of helix- or corkscrew-shaped pasta. The name comes from a 17th-century Italian word meaning "small wheels". Rotini is related to fusilli, but has a tighter he...
4.0	Vermicelli > History.:The Americas. | The "fideo" is a type of noodle, produced in Europe ever since the Roman times, best known as "fideus" or "fidelis", and then spread to Mexican and Latin Amer...
5.0	Rollatini > Abstract | Rollatini (sometimes also spelled rolatini or rolletini) is an Italian-style dish (called "rollatini di melanzane" in faux Italian) that is usually made with thin slices of ...

Table 6: Passages retrieved by MarginalizedLoss retriever while talking about Italian Cuisine. All passages talk about Pastas.

A Appendix

A.1 Examples of retrieved passages

A.1.1 Conversation 1: Italian cuisine

Other Ooh I like that! Stick some nice spicy arrabbiata sauce with it, ahhhh! Have you ever had bucatini before?

Self Oh yeah! I love that spicy garlic and tomato sauce. No I have not had bucatini. Is that a type of cheese?

Other Now you're speakin' my language. No no, it's a style of noodle, like a really long straw. Bucatini amatraciana is insanely good.

A.1.2 CONVERSATION 2: ROCK AND ROLL

Self Do you mean Elvis Aaron Presley, the American singer and actor?

Other That's the one. I think his nickname was the king of rock 'n roll.

Self I had just heard of him being "The King". There probably would not have been a Sun Records if not for Elvis and Sam Phillips.

Other He was revolutionary for his time. Many older people thought he was straight from the devil.

rank	text
1.0	Sam Phillips > Abstract | Samuel Cornelius Phillips (January 5, 1923 – July 30, 2003) was an American record producer who played an important role in the development of rock and roll during the 19...
2.0	Cultural impact of Elvis Presley > Abstract | Since the beginning of his career, Elvis Presley has had an extensive cultural impact. According to "Rolling Stone", "it was Elvis who made rock 'n' r...
3.0	Freddie King > Abstract | Freddie King (September 3, 1934 – December 28, 1976) was an American blues guitarist and singer. He recorded several hits for Federal Records in the early 1960s. His soul...
4.0	Elvis Presley > Abstract | With a series of successful network television appearances and chart-topping records, he became the leading figure of the newly popular sound of rock and roll. His energ...
5.0	Elvis Presley > Abstract | Elvis Aaron Presley (January 8, 1935 – August 16, 1977), also known mononymously as Elvis, was an American singer, musician, and actor. Regarded as one of the most signi...

Table 7: Passages retrieved by ELBOLOSS while talking about Rock and Roll. Relevant passages about cultural impact of Elvis Presley (rank=2) and details about his career (rank=4) alongside introductary paragraphs of other musicians

rank	text
1.0	Elvis Presley > Abstract | Elvis Aaron Presley (January 8, 1935 – August 16, 1977), also known mononymously as Elvis, was an American singer, musician, and actor. Regarded as one of the most signi...
2.0	Sam Phillips > Abstract | Samuel Cornelius Phillips (January 5, 1923 – July 30, 2003) was an American record producer who played an important role in the development of rock and roll during the 19...
3.0	Johnny Otis > Abstract | Johnny Otis (born Ioannis Alexandres Veliotes; December 28, 1921 – January 17, 2012) was an American singer, musician, composer, arranger, bandleader, talent scout, disc j...
4.0	Carl Perkins > Abstract | Called "the King of Rockabilly", he was inducted into the Rock and Roll Hall of Fame, the Rockabilly Hall of Fame, the Memphis Music Hall of Fame, and the Nashville Songw...
5.0	Chubby Checker > Abstract | Chubby Checker (born Ernest Evans; October 3, 1941) is an American rock 'n roll singer and dancer. He is widely known for popularising many dance styles including the t...

Table 8: Passages retrieved by MARGINALIZEDLOSS while talking about Rock and Roll. All passages are the introductory paragraphs from various related artists

A.2 NOVEL-F1

Rationale We conducted a small experiment with the generated output on Wizard of Wikipedia dataset using top-8 retrieved passages. We removed the gold passage and computed overlap of the generated output with the target output. We consistently found (across models and passage ranks) the F1 overlap to be close to 15. This meant that by conditioning on arbitrary passages the generator

(likely by ignoring them altogether) is able to achieve around 80% of the F1-overlap of the best performing models ($\sim$ 19 F1). This can be a confounding factor for selecting models based on high F1 overlap. A model that simply copies content from the input x can achieve high F1-overlap but fail to using the retrieved passage to generate the output. Removing commonly occurring words reduces it to 8 F1, but removing words from input context reduces it further down to 4 F1. Thus we find Novel-F1 to be the cleanest measure of overlap as it discounts two confounding factors and only looks at "Novel" words, words that are rare and were not in the input text x.

We construct the list of common words based on their frequency in the training corpus. We sort words by frequency and take the most frequent words that contribute: 50% of the probability mass toward Wizard of Wikipedia utterances (amounting to 121 words) following Shuster et al. (2021). However, we found that using the same heuristic for MS-MARCO NLGen answers included numbers and rarer tokens that could potentially be in the answer span. So we instead use only 33% of the probability mass (amounting to 55 words). We also ran evaluation using 50% of the probability mass but found the trends to be consistent.

MS Marco NLGen list of common words (sorted by frequency)
is, of, in, to, and, for, or, are, that, on, from, as, by, you, with, it, county, can, at, per, was, your, average, cost, be, between, which, used, one, united, states, there, years, located, name, not, new, have, takes, number, has, means, days, when, blood, system, year, should, no, most, first, hours, up, minutes, 1

Wizard of Wikipedia list of common words (sorted by frequency)
is, of, in, to, and, for, or, are, that, on, from, as, by, you, with, it, county, can, at, per, was, your, average, cost, be, between, which, used, one, united, states, there, years, located, name, not, new, have, takes, number, has, means, days, when, blood, system, year, should, no, most, first, hours, up, minutes, 1 i, and, of, in, is, to, it, that, are, you, they, have, was, but, for, as, its, like, with, on, so, be, or, not, yes, do, can, from, there, by, well, also, one, my, know, has, some, he, their, love, most, people, think, really, all, about, just, too, them, im, which, sure, more, been, at, would, many, were, good, very, dont, when, thats, no, yeah, what, other, great, if, because, used, actually, first, since, lot, me, even, your, how, we, time, different, world, use, get, called, only, out, much, over, had, though, music, around, popular, his, am, made, than, such, back, up, us, make, usually, who, favorite, new, food, oh, long, she, now, did, pretty, any, where, years, this, way, go

B INTUITION BEHIND IMPROVEMENTS DUE TO ELBOLOSS

To understand the intuition behind suboptimality of MARGINALIZEDLOSS for open-ended generation tasks consider the following: We would want a good retriever to assign similar but high probabilities to all context-relevant passages because they are similarly relevant but a good generator to only assign high probabilities when using label-relevant passages because only label-relevant passages are pertinent to the target output. But the training signal to a model (partial derivative w.r.t the model and a passage) is modulated by the probability of the other model:

$$\frac{\partial P(y|x)}{\partial P_\eta(z_i|x)} = P_\theta(y|x, z_i) \qquad \frac{\partial P(y|x)}{\partial P_\theta(y|x, z_i)} = P_\eta(z_i|x)$$

Since context-relevant passages have similar $P(z_i|x)$ the gradient encourages the generator to assign equal probabilities to the target output using all context-relevant passages. We see this issue play out empirically when using MARGINALIZEDLOSS for two different tasks: Open-Domain QA (Natural Questions by Kwiatkowski et al. (2019)) and informative dialogue (Wizard of Wikipedia by Dinan et al. (2019)) (Figure 4). We see that on the Natural Questions dataset, where there is typically one correct answer, the generator produces distribution with a sharp peak that can potentially serve as an accurate proxy for label-relevance and in turn train a good retriever. But on the Wizard of Wikipedia dataset, the generator produces a flatter distribution which is a bad proxy for label-relevance. This provides weaker supervision for the retriever which learns a flatter probability distribution as well and is less able to differentiate context-relevant from irrelevant passages.

We see in Figure 5 that for the Wizard of Wikipedia dataset with ELBOLOSS we obtain a sharp distribution for $Q(z|x, y)$ (nearly as good as $P_\theta(y|x, z)$ on NQ from Figure 4) and that the $P_\eta(z|x)$ and $P_\theta(y|x, z)$ are now sharper than MARGINALIZEDLOSS. While a sharper distribution does not

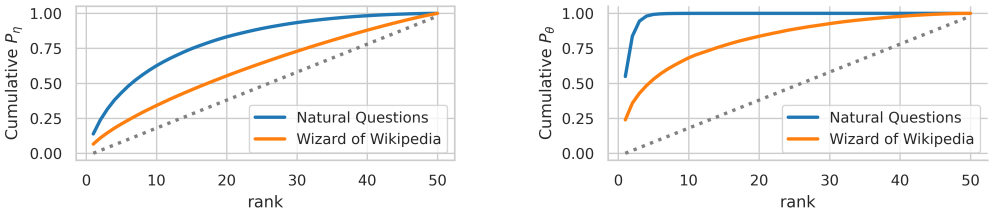

Figure 4: With MARGINALIZEDLOSS, the generator $P_\theta(y|x,z)$ learns a sharp distribution for Natural Questions (NQ) dataset (**right**) but learns a flatter distribution for a one-to-many open-ended generation task using the Wizard of Wikipedia dataset (WoW). The flatter distribution in the case of WoW Generator shows that it has not learned label-relevance as well. Consequently, for WoW we see a weaker retriever (**left**) that has a flatter distribution than NQ. (**Left**) Cumulative probability $P_\eta(z|x)$ w.r.t. rank for passages. (**Right**) Assuming a uniform prior $P(z|x)$, the cumulative probability $P_\theta(y|x,z)$ w.r.t. rank for passages, plotted as $P(z|x,y) \propto P(y|x,z)P(z|x)$. The gray dotted line shows a hypothetical model that assigns equal probabilities to all passages.

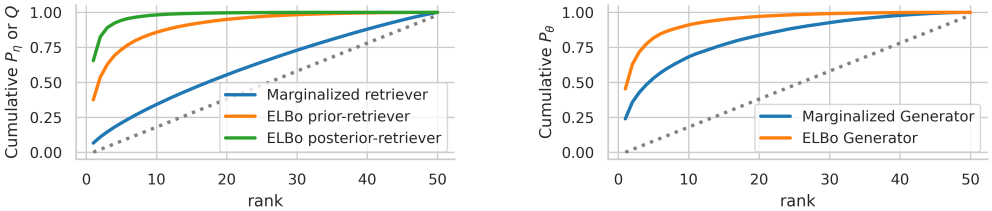

Figure 5: Analogous plots to Figure 4 but with ELBOLOSS on the one-to-many Wizard of Wikipedia (WoW) dataset. Training with ELBOLOSS produces a sharp distribution for $Q(z|x,y)$ and subsequently sharper $P_\eta(z|x)$ and $P_\theta(y|x,z)$ than MARGINALIZEDLOSS.

imply a better retriever and generator (they may still assign high probability to the wrong passage), a flatter distribution limits their potential. As we will see in Section 4, ELBOLOSS indeed utilizes the potential and trains a better retriever and more a grounded generator.