Zemi: Learning Zero-Shot Semi-Parametric Language Models from Multiple Tasks

Zhenhailong Wang* UIUC wangz3@illinois.edu
Xiaoman Pan Tencent AI Lab xiaomanpan@global.tencent.com
Dian Yu Tencent AI Lab yudian@global.tencent.com
Dong Yu Tencent AI Lab dyu@global.tencent.com
Jianshu Chen Tencent AI Lab jianshuchen@global.tencent.com
Heng Ji UIUC hengji@illinois.edu

Abstract

Although large language models have exhibited impressive zero-shot ability, the huge model size generally incurs high cost. Recently, semi-parametric language models, which augment a smaller language model with retrieved related background knowledge, alleviate the need for storing everything into the model parameters. Although existing semi-parametric language models have demonstrated promising *language modeling* capabilities, it remains unclear whether they can exhibit competitive *zero-shot* abilities as their fully-parametric counterparts. In this work, we introduce **Zemi**, a semi-parametric language model for zero-shot task generalization. To our best knowledge, this is **the first semi-parametric language model that can demonstrate strong zero-shot performance on a wide range of held-out unseen tasks.** We train Zemi with semi-parametric multitask training, which shows significant improvement compared with the parametric multitask training as proposed by T0 (Sanh et al., 2021). Specifically, during both training and inference, Zemi is equipped with a retrieval system based on the unlabeled pretraining corpus of our backbone model. To address the unique challenges from large-scale retrieval, we further propose a novel **retrieval-augmentation fusion** module that can effectively incorporate noisy retrieved documents. Finally, we show detailed analysis and ablation studies on the key ingredients towards building effective zero-shot semi-parametric language models. Notably, our proposed $\text{Zemi}_{\text{LARGE}}$ model outperforms T0-3B by 16% across seven diverse evaluation tasks while being 3.8x smaller in scale.[1]

1 Introduction

Achieving strong generalization ability on unseen tasks while maintaining a reasonably small parameter size is a long-lasting challenge for natural language processing (NLP) models. Although large language models (Brown et al., 2020; Lieber et al., 2021; Rae et al., 2021; Smith et al., 2022; Hoffmann et al., 2022; Zhang et al., 2022; Ouyang et al., 2022; Chowdhery et al., 2022) have shown impressive zero-shot ability on various NLP tasks, the huge model size generally incurs high cost. Alternatively, instead of stuffing everything in the model parameters, recent work on semi-parametric language models (Grave et al., 2016; Khandelwal et al., 2019; Yogatama et al., 2021; Borgeaud et al., 2021; Zhong et al., 2022) demonstrated competitive *language modeling* performance compared with much larger fully-parametric language models. The intuition is to use a relatively small language model as a reasoning module and augment it with a retriever to retrieve related background knowledge, which effectively alleviates the need for increasing the model capacity to align with the growing data size.

However, what really makes large language models the focus of attention in the past two years is their strong zero-shot in-context learning abilities. Unfortunately, it is still unclear whether semi-parametric language models can exhibit similar *zero-shot* ability on unseen tasks as their fully-parametric counterparts such as T0 (Sanh et al., 2021) and GPT-3 (Brown et al., 2020). Moreover, improvements in language modeling metrics such as perplexity may not guarantee better performance on downstream tasks especially in low-shot settings (Wei et al., 2022). Thus, in this work, we aim to investigate this unexplored research question, *can semi-parametric language models exhibit strong zero-shot generalization abilities on various downstream tasks?*

To this end, we introduce **Zemi**, a zero-shot semi-parametric language model. To the best of our knowledge, this is the first semi-parametric language model that shows strong zero-shot perfor-

*Work was done when interning at Tencent AI Lab.
[1]Code and data are available for research purpose at https://github.com/MikeWangWZHL/Zemi.

Findings of the Association for Computational Linguistics: ACL 2023, pages 3978–4004
July 9-14, 2023

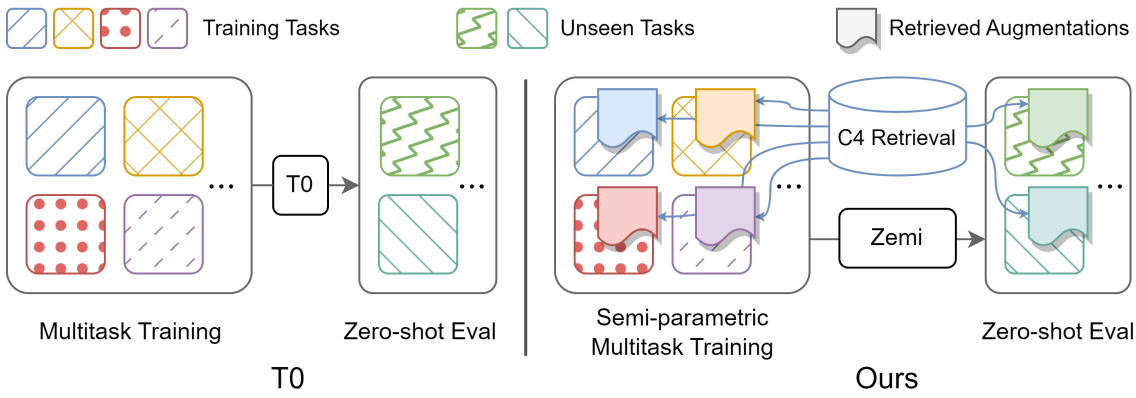

Figure 1: Overview of the semi-parametric multitask prompted training. Each training and evaluation instance is formatted with unified text-to-text prompt templates (Sanh et al., 2021; Bach et al., 2022). In this work, we further augment the prompted instances with retrieved passages from a large-scale task-agnostic corpus, C4 (Sanh et al., 2021), which is the same unlabeled pretraining corpus used in T5 (Raffel et al., 2020) and T0 (Sanh et al., 2021). An example of the prompted input and the retrieved documents can be found in Figure 2.

mance on a wide range of downstream tasks. In order to effectively train Zemi, we propose to extend the multitask prompted training (Sanh et al., 2021) into semi-parametric settings (Section 2.1). Specifically, during both the training and the inference stage, we augment the prompted instances with retrieved plain text documents. To cover a wider range of unseen tasks, instead of retrieving from specific corpora for certain tasks, such as exploiting Wikipedia for open-domain question answering (Lee et al., 2019; Karpukhin et al., 2020; Izacard and Grave, 2020), we retrieve documents from a large-scale task-agnostic corpus, C4 (Raffel et al., 2020) (Section 2.2). Notably, C4 is the unlabeled pre-training corpus of our backbone model (Raffel et al., 2020), which means that every document is seen by the model and we do not require any annotated or curated resources. This guarantees fair comparison with the parametric counterpart T0 (Sanh et al., 2021).

In our preliminary experiments, we find that existing methods (Izacard and Grave, 2020; Brown et al., 2020) for incorporating retrieved text cannot effectively handle the noise inevitably introduced by retrieving from large-scale corpora. To address this challenge, we propose a novel **retrieval-augmentation fusion** module that can selectively ignore noisy retrieved text. Specifically, we introduce a light-weight *perceiver resampler* and a *gated cross-attention* layer (Alayrac et al., 2022) to enforce the model to attend to salient information of each augmentation and gate out noisy ones (Section 2.3).

We train Zemi on eight multiple-choice question answering (QA) tasks (4.5x fewer than T0) and evaluate on a diverse set of seven unseen tasks from five categories (Section 3.1). In order to investigate the impact of the retrieval-augmentation, we favor knowledge-intensive tasks over extractive tasks.

Experimental results show that Zemi outperforms both parametric and semi-parametric baselines. Notably, $\text{Zemi}_{\text{LARGE}}$ outperforms T0-3B by 16% across seven evaluation tasks while being 3.8x smaller in scale (Section 3.2). We further conduct extensive analysis on *why Zemi works*. We show that the source of the improvements comes from the interplay of our proposed retrieval-augmentation fusion architecture along with the semi-parametric multitask training paradigm. Finally, we perform in-depth ablation studies on all aspects of our model design including the gated mechanism.

To sum up, the main contributions of this paper are threefold:

- We introduce Zemi, which is to our knowledge the first semi-parametric model that demonstrates strong zero-shot task generalization ability.
- We propose a novel retrieval-augmentation fusion module which can effectively handle multiple potentially noisy retrieved documents and is essential towards the effectiveness of semi-parametric multitask training.
- We show detailed analysis and ablation studies on *why Zemi works* which shed light on future work for developing large-scale universal semi-parametric language models with strong zero-shot ability.

2 Method

2.1 Semi-parametric multitask training

In this section, we introduce how we extend the multitask training paradigm to semi-parametric language models. We follow the overall text-to-text framework proposed by the previous parametric multitask prompted training (Sanh et al., 2021) where each input-output pair of a certain task is converted into a prompted text input and a generated text output via human-written templates (Bach et al., 2022).[2] For Zemi, as illustrated in Figure 1, during both training and inference, we augment Zemi with a retrieval system. Instead of using specific corpora for different tasks, such as Wikipedia for open-domain question answering (Chen et al., 2017; Karpukhin et al., 2020; Izacard and Grave, 2020) and textbooks for science question answering (Mihaylov et al., 2018), we retrieve texts from a large-scale task-agnostic corpus, **C4** (Raffel et al., 2020) (Section 2.2). Retrieving from a larger corpus brings wider coverage but also more noisy augmentations. To address this problem we further propose a novel semi-parametric architecture for Zemi that specializes in handling a large number of potentially noisy augmentations (Section 2.3). After semi-parametric multitask training, we perform zero-shot evaluation on seven diverse held-out unseen tasks (Section 3).

2.2 C4 retrieval

To build a universal semi-parametric language model that can generalize to various types of NLP tasks, we retrieve documents from Colossal Clean Crawled Corpus (C4) (Raffel et al., 2020). Notably, C4 is the unlabeled pre-training corpus of our backbone model T5 (Raffel et al., 2020), which guarantees fair comparison with non-retrieval methods in our zero-shot evaluation settings. The C4 corpus (750GB in size) contains more than 364 million documents. Performing dense retrieval (Karpukhin et al., 2020) on such a wide-coverage corpus is very expensive. Thus, for efficiency consideration, we perform document-level indexing and retrieval based on BM25 (Robertson et al., 1995) with ElasticSearch (ElasticSearch) and Huggingface Datasets (Lhoest et al., 2021). Despite its simplicity, recent work (Wang et al., 2022a) has demonstrated the effectiveness of using BM25 for retrieving clean training data as augmentations. To further improve the retrieval efficiency, we use 5% of the entire C4 corpus, which is still 3x larger than the Wikipedia corpus (Foundation), as our retrieval corpus in our experiments. For each query, we truncate the query length at 20 tokens and truncate each retrieved document at 256 tokens. See details on the query fields for each dataset in Appendix D.

[2] https://github.com/bigscience-workshop/promptsource.

2.3 Zemi model architecture

One major challenge of retrieving from a large-scale task-agnostic corpus is that the retrieved augmentations (documents) can be noisy and inaccurately ranked. Examples of good and noisy retrieved documents can be found in Appendix A. To address this problem, intuitively, we want the model to have the following two properties: (1) be able to simultaneously pay attention to multiple retrieved augmentations instead of only the top-1 document. (2) be able to identify salient information from the retrieved augmentations and selectively ignore uninformative ones.

To this end, we propose the Zemi architecture, a semi-parametric language model capable of selectively incorporating multiple potentially noisy retrieved augmentations. The main idea is to jointly train a light-weight **retrieval-augmentation fusion** module between the encoder and decoder, which contains two major components, a *perceiver resampler* and a *gated cross-attention*, which are inspired by recent work on vision-language fusion (Alayrac et al., 2022).

Figure 2 shows an illustration of the Zemi model architecture. We consider a prompted text input I and a few retrieved textual augmentations $A_1, A_2, ..., A_k$. Let l_I, l_A^i be the length of the prompted input and the ith augmentation. Let d be the hidden dimension of our backbone model. We first independently encode I and $A_1, A_2, ..., A_k$ with a shared T5 (Raffel et al., 2020) encoder Enc. We then feed the latent representation of the augmentations $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_k$ through the perceiver resampler.

$$\mathcal{I} = \text{Enc}(I) \quad (1)$$

$$\mathcal{A}_i = \text{Enc}(A_i) \quad (2)$$

$$\mathcal{A}'_i = \text{PerceiverResampler}(\mathcal{A}_i, \mathcal{Q}) \quad (3)$$

where $\forall\, i \in \{1, .., k\}$, $\mathcal{I} \in R^{l_I \times d}$, $\mathcal{A}_i \in \mathcal{R}^{l_A^i \times d}$ and $\mathcal{A}'_i \in \mathcal{R}^{l_Q \times d}$.

As shown on the bottom right of Figure 2, the perceiver resampler is a variant of Perceiver IO (Jaegle et al., 2021), where a cross-attention is

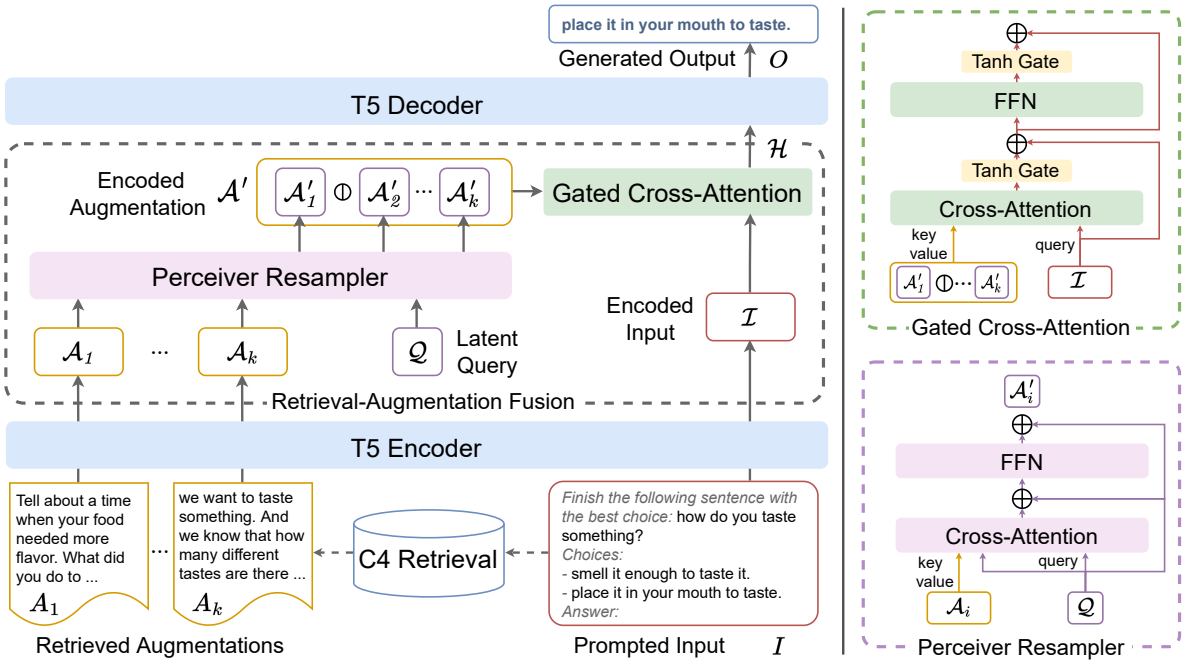

Figure 2: Zemi model architecture with an example of a prompted input and a generated output from the Piqa (Bisk et al., 2020) task. The *italic text* in the prompted input I indicates the prompt template. A_1 and A_k shows two examples of the corresponding retrieved augmentations (documents) from the C4 corpus. To incorporate the potentially noisy retrieved augmentations, we introduce a light-weight retrieval-augmentation fusion module that contains two major components, a single layer perceiver resampler and a single layer gated cross-attention (detailed on the right).

performed between the variable-length latent representation of an augmentation $\mathcal{A}_i$ and a fixed-length learnable latent query vector $\mathcal{Q}$. Let l_Q be the predefined length of the latent query, which is typically smaller than the original length of an augmentation l^i_A. The output of the perceiver resampler is a compressed fixed-length latent representation of each augmentation. This resampling mechanism not only allows the model to include *longer and a larger number of augmentations* but also encourages the model to select *salient information* from the original augmentations. After the resampler, we concatenate the encoded augmentations $\mathcal{A}' = [\mathcal{A}'_i, .., \mathcal{A}'_k]$ and perform gated cross-attention with the encoded prompted input $\mathcal{I}$. As shown in the top right of Figure 2, the gated cross-attention layer contains two Tanh gates controlling the information flow from the cross-attention layer and the feed-forward layer before the addition with the skip connections, i.e., the original encoded input $\mathcal{I}$. Finally, the hidden states from the gated cross-attention module $\mathcal{H}$ is fed into the T5 decoder Dec to generate the output sequence O.

$$\mathcal{H} = \text{GatedCrossAttn}([\mathcal{A}'_i, .., \mathcal{A}'_k], \mathcal{I}) \quad (4)$$

$$O = \text{Dec}(\mathcal{H}) \quad (5)$$

where $[\mathcal{A}'_i, .., \mathcal{A}'_k] \in R^{(k \times l_Q) \times d}$ and $\mathcal{H} \in \mathcal{R}^{l_I \times d}$.

Following (Alayrac et al., 2022), we initialize the parameter of the Tanh gate to be 0, allowing the forward pass of the prompted input through the pre-trained T5 encoder-decoder to be intact at the beginning of the training process. With the gated mechanism, the model can learn to *gate out noisy augmentations* and rely more on the skip connections during semi-parametric multitask training.

3 Experiments

3.1 Experimental setup

Following (Sanh et al., 2021) we partition various types of NLP tasks into two groups, training tasks and held-out unseen tasks. In this work, we are particularly interested in investigating the impact of the retrieval augmentation Thus, when choosing the training and evaluation tasks, we favor knowledge-intensive tasks over extractive tasks such as summarization, where most knowledge for solving the problem is already self-contained in

Method	semi-param	# train tasks	# param	Tasks							Avg_5	Avg_7
				OBQA	Piqa	RT	CB	COPA	WiC	HSwag		
BART0	No	36	0.4B	34.4	36.1	-	39.6	-	46.7	39.4	39.3	-
T0-3B	No	36	3B	42.8	59.3	73.6*	45.5	75.9	50.0	27.3	45.0	53.5
T0-11B	No	36	11B	59.1	72.5	81.8*	70.1	91.5	55.2	33.5	58.1	66.3
ReCross	Yes	36	0.4B	39.6	41.4	-	44.8	-	50.6	47.3	44.7	-
$Zemi_{BASE}$ (ours)	Yes	8	0.2B	35.6	59.2	68.6	50.1	63.6	49.6	29.7	44.8	50.9
$Zemi_{LARGE}$ (ours)	Yes	8	0.8B	51.5	67.9	84.1	62.1	84.5	50.4	35.8	53.5	62.3
GPT-3	No	-	175B	57.6	81.0*	59.7	46.4	91.0	49.4†	78.9	62.7	66.3

Table 1: Comparison to both parametric (*BART0, T0, GPT-3*) and semi-parametric (*ReCross*) state-of-the-art. $Zemi_{LARGE}$ significantly outperforms T0-3B while being 3.8x smaller in scale. $Zemi_{BASE}$ slightly outperforms *ReCross* while being 1.7x smaller. Note that Avg_7 indicates averaged performance across all seven tasks. Avg_5 indicates averaged performance on five tasks excluding *RT* and *COPA* due to unreported baseline results. * indicates the task is seen during training. † indicates few-shot results with 32 examples.

the input. Furthermore, we avoid including large datasets, such as DBPedia (Lehmann et al., 2015) (630K instances) and QQP (Shankar Iyer, 2017) (400K instances), due to limited computational resources.

Training Tasks We use a subset of T0's (Sanh et al., 2021) training mixture for our semi-parametric multitask prompted training. Specifically, our training mixture contains *eight* multiple-choice QA datasets, including, **CommonsenseQA** (Rajani et al., 2019), **CosmosQA** (Huang et al., 2019), **DREAM** (Sun et al., 2019), **QASC** (Khot et al., 2020), **QUARTZ** (Tafjord et al., 2019), **SciQ** (Johannes Welbl, 2017), **Social IQa** (Sap et al., 2019), and **WIQA** (Tandon et al., 2019). We choose the subset in multiple-choice QA tasks because they are diverse in domains and overall task formats. Ablation studies on including more types of training tasks can be found in Section 3.4.

Evaluation Tasks For evaluation tasks, we consider *seven* datasets from *five* diverse categories following the task taxonomy of T0, including, two sentence completion tasks, **COPA** (Roemmele et al., 2011) and **HellaSwag** (HSwag) (Zellers et al., 2019), two multiple-choice QA tasks, **OpenbookQA** (OBQA) (Mihaylov et al., 2018) and **Piqa** (Bisk et al., 2020), one word sense disambiguation task, **WiC** (Pilehvar and Camacho-Collados, 2018), one sentiment task, **Rotten Tomatoes** (RT) (Pang and Lee, 2005), and one natural language inference task, **CB** (De Marneffe et al., 2019). All scores are reported on the validation set of each dataset. The detailed prompt templates used for training and evaluation can be found in Appendix C.

Prompts We use *PromptSource* (Bach et al., 2022) with *Huggingface Datasets* (Lhoest et al., 2021) to construct prompted inputs for each training and evaluation instance. During training, we randomly select two templates for each dataset. During evaluation, we follow the exact evaluation procedure as in T0 (Sanh et al., 2021) and report the mean accuracy across all available templates. All scores are reported on the validation set of each dataset. The detailed templates used for training and evaluation can be found in Appendix C.

Model We consider two variants of Zemi with a different pre-trained backbone, i.e., T5-base and T5-large (Raffel et al., 2020). Following T0 (Sanh et al., 2021), we use the language modeling adapted[3] checkpoint, which is trained for an additional 100k steps on a language modeling objective. By default, we use five retrieved passages as augmentations for each instance. More implementation details can be found in Appendix B

3.2 Main results

We aim to explore whether Zemi can exhibit competitive zero-shot performance against larger state-of-the-art language models. We compare Zemi with both parametric (*T0* (Sanh et al., 2021), *BART0* (Lin et al., 2022), *GPT-3* (Brown et al., 2020)) and semi-parametric (*ReCross* (Lin et al., 2022)) models on seven zero-shot tasks. Table 1 shows the mean zero-shot accuracy across all templates for each task. The last two columns of Table 1 show the averaged performance across different sets of tasks, where Avg_7 is averaged across all seven tasks, and Avg_5 considers five tasks excluding RT and COPA due to their unavailable baseline

[3] https://huggingface.co/google/t5-base-lm-adapt.

results.

For *BART0*, *ReCross*, and *GPT-3*, we copy the reported scores directly from their original papers. For the missing score of RT on *GPT-3*, we run the original text completion API [4] to get the generated outputs which is then mapped to the most similar answer choice using SentenceBert (Reimers and Gurevych, 2019). For *T0* models, there are some tasks such as OBQA and Piqa that are not evaluated in the original paper (Sanh et al., 2021), and some tasks such as CB and WiC are evaluated with slightly different templates. Thus, for fair comparison, we re-evaluate all seven tasks on T0-3B and T0-11B using the official implementation and checkpoints[5] with the exact same set of templates as our model. See details on the templates used for each task in Appendix C.

Result Table 1 shows that $\text{Zemi}_{\text{BASE}}$ outperforms previous retrieval-based method, ReCross, on the average of five tasks (Avg_5) while being 2x smaller in scale. Notably, **$\text{Zemi}_{\text{LARGE}}$, significantly outperforms T0-3B** on seven evaluation tasks (Avg_7) by 16% with **3.8x fewer parameters**. This shows that Zemi scales up well with larger backbone models. We also observe that although trained with 4.5x fewer training tasks (8 v.s. 36), Zemi effectively achieves state-of-the-art zero-shot performance. In Section 3.4, we show that adding more tasks into multitask training does not necessarily improve the performance. And the training mixture with multiple-choice QA tasks seems to be highly effective in generalizing to various kinds of unseen tasks.

3.3 Analysis: semi-parametric v.s. parametric

In order to further analyse the source of the strong performance of $\textit{Zemi}_{\textit{LARGE}}$, we compare ***$\textit{Zemi}_{\textit{LARGE}}$*** with a baseline (***No Aug***) trained with parametric multitask training on the same set of training tasks and with the same backbone model, T5-Large (Raffel et al., 2020). To show the impact of our newly proposed retrieval-augmentation fusion module, we further compare $\textit{Zemi}_{\textit{LARGE}}$ against two semi-parametric baselines with a different fusion method for incorporating the retrieved augmentations (***Concat*** and ***FiD***). In Table 2, we show that the source of benefit comes from **the interplay of the proposed retrieval-augmentation fusion and the semi-parametric multitask training**.

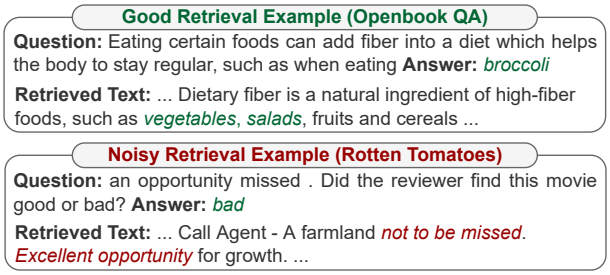

Figure 3: Example of good and noisy retrieved augmentations. See Appendix A for more examples.

Specifically, for *Concat*, we directly concatenate all retrieved augmentations with the prompted input text. The concatenated input is then truncated to the maximum acceptable length of 1024 tokens and fed to our backbone model. For *FiD*, we implement the model following (Izacard and Grave, 2020) where we first independently encode each pair of retrieved augmentation and the prompted input text. Then we concatenate the encoder outputs and feed them to the decoder. Note that we keep everything else identical except the retrieval-augmentation fusion module for $\textit{Zemi}_{\textit{LARGE}}$, *Concat* and *FiD*.

Zemi architecture improves zero-shot task generalization. In Table 2, we first notice that the semi-parametric setting in itself does not necessarily bring consistent positive gains compared with the *No Aug* baseline, as shown in the results of *Concat* and *FiD*. This can be explained by the fact that the retrieved documents are not always highly correlates with the task of interest, as shown in the examples in Figure 3. The fact that *FiD* performs better than *Concat* further verifies this hypothesis, since *FiD* preserves more input text information in the encoding step and only do fusion with all the retrieved augmentations in the decoder, whereas *Concat* perform unified self-attention on all augmentations concatenated directly to the input.

On the other hand, with the proposed retrieval-augmentation fusion module that contains the explicit resampling and gating mechanism, $\textit{Zemi}_{\textit{LARGE}}$ was able to achieve the best performance on six out of seven tasks, and brings a overall gain of **+5%** against the *No Aug* baseline. This result shows that the retrieval-augmentation fusion module in *Zemi* can effectively enable the model to leverage potentially noisy retrieved augmentations during semi-parametric multitask training, which

[4] For consistency with other results, we report the RT result from the original "davinci" model.

[5] https://github.com/bigscience-workshop/t-zero.

Method	# Param	Tasks							Avg
		OBQA	Piqa	RT	CB	COPA	WiC	HSwag	
No Aug	0.8B	50.5	65.5	82.2	52.4	80.0	50.2	34.1	59.3
Concat	0.8B	48.8	65.9	74.9	44.6	82.7	50.0	30.5	56.8
FiD	0.8B	51.0	66.7	67.1	60.7	**86.3**	50.2	32.9	59.3
Zemi$_{\text{LARGE}}$ (Ours)	0.8B	**51.5**	**67.9**	**84.1**	**62.1**	84.5	**50.4**	**35.8**	**62.3**

Table 2: Comparison to parametric multitask trained baseline (No Aug) and alternative augmentation fusion methods (Concat, FiD) with an identical backbone model, T5-large. # Param indicates the model size.

brings significant improvement in zero-shot task generalization. In abaltion study 3.4, we further verify that the gated cross-attention is an important factor contributing to the effectiveness of the Zemi architecture.

3.4 Analysis: ablation studies

In this section, we continue investigating *why Zemi works* by conducting comprehensive ablation studies on different aspects of the model design. As shown in Table 3, we consider the following five categories of ablated settings on *Zemi$_{BASE}$* [6]:

(i) Tanh gate. We replace the gated cross-attention module with vanilla cross-attention in the ablated version. Specifically, we remove the two Tanh gates as shown in Figure 2. We find that **removing Tanh gate hurts the zero-shot performance**. Note that the Tanh gate is also the main difference between *Zemi* and *FiD* (Izacard and Grave, 2020).

(ii) Number of augmentations. We ablate on the number of augmentations. Note that for settings with 20 and 30 augmentations, in order to reduce the computation complexity, we propose another variant of *Zemi$_{BASE}$* where we encode augmentations with a separate frozen augmentation encoder. We find that increasing the number of augmentations from single to multiple (five) improves the performance. However, further increasing the number to 10 starts to hurt the performance, which again indicates that the noise starts to overwhelm the useful signals introduced by the retrieval. We also observe that the performance with 30 augmentations outperforms 20 augmentations, we hypothesis that this is due to inaccurate retrieval ranking that leads to some more informative documents being ranked lower. We show an example of this case in Figure 11. Nevertheless, the fact that we are able to achieve positive gain with as many as 30 augmentations shows the **robustness of our model to very noisy augmentations**.

(iii) Perceiver resampler latent size. We ablate on the size of the latent query vector in the perceiver resampler. Note that here the latent size is different from the hidden state size of the backbone model. The trade-off of the size of the latent query vector is that, a larger latent size preserves more information from the original augmentation but also includes more noise. A larger latent size can also increase the computational complexity. We find that Zemi is **relatively robust to the change of the latent size** and achieves the best performance with a latent size of 64.

(iv) Per augmentation length. We investigate the impact of different ways of constructing augmentations from the retrieved documents. Specifically, we increase the maximum length of each augmentation from 256 to 512 and fit two retrieved documents into one augmentation. We keep the number of augmentations the same as default, i.e., 5. We then compare this ablated setting with the 10-augmentation variant in (ii). We find that with the same set of retrieved documents, **augmenting the model with longer but fewer augmentations generally outperforms using a larger number of shorter augmentations**.

(v) Training mixture. We investigate the impact of adding new types of training tasks to the original training mixture. We dub the models trained with this new training mixture as ***No Aug+*** and ***Zemi+***. Specifically, apart from the eight multiple-choice QA tasks, we further include four more tasks: one closed-book QA task **WikiQA** (Yang et al., 2015), one topic classification task **TREC** (Li and Roth, 2002), one sentence completion task **COPA** (Roemmele et al., 2011), and one sentiment task **Rotten**

[6] We ablate on Zemi$_{\text{BASE}}$ instead of Zemi$_{\text{LARGE}}$ mainly to reduce the computation overheads of a large amount of experiments.

	Ablated setting	Zemi value	Changed value	Tasks OBQA	Piqa	RT	CB	COPA	WiC	HSwag	Avg
	No Augmentation (No Aug$_{BASE}$)			36.6	60.2	64.1	41.5	68.5	49.9	28.0	49.8
	Zemi$_{BASE}$			35.6	59.2	68.6	50.1	63.6	49.6	29.7	**50.9**
(i)	Tanh Gate	✓	✗	35.0	57.8	55.8	49.9	71.5	51.6	27.9	49.9
(ii)	Num of Augs	5	1	35.6	58.9	67.1	47.6	65.2	49.5	30.1	50.6
			10	35.3	59.4	62.1	46.3	64.6	51.4	29.4	49.8
			20*	34.7	58.7	60.3	46.5	61.6	50.1	28.4	48.6
			30*	35.1	60.5	58.7	48.2	67.2	50.7	28.5	49.8
(iii)	Latent size	64	32	34.9	58.8	64.3	44.5	67.9	51.2	28.6	50.0
			128	34.7	57.6	63.1	47.8	69.8	50.4	28.1	50.2
(iv)	Aug length	256	512	35.3	58.8	58.6	52.5	68.9	50.3	28.8	50.5
(v)	Training mixture	Zemi	No Aug+	37.6	58.4	-	43.3	-	50.7	28.0	-
			Zemi+	34.5	58.7	-	42.8	-	50.1	29.3	-

Table 3: Ablation study. Each ablated setting should be compared with the first two rows, i.e., the original *No Augmentation (No Aug$_{BASE}$)* setting and *Zemi$_{BASE}$*. The superscripted "⋆" in ablated setting *(ii)* indicates using the model variant with a frozen augmentation encoder. See descriptions of each setting in Section 3.4.

Tomatoes (RT) (Pang and Lee, 2005) [7]. We find that adding new types of tasks does not necessarily increase the performance. Although trained with only 8 tasks (v.s. 36 tasks) we are able to achieve state-of-the-art performance (Section 3.2), which shows that the **multiple-choice QA mixture is highly effective for generalizing to a wide range of held-out unseen tasks**.

3.5 Analysis: computation overheads

There are two main computation overheads compared with the fully-parametric counterpart, i.e., the *No Aug* baseline. First, retrieving from a large-scale corpus can be time-consuming. As mentioned in Section 2.2, we apply document-level retrieval with BM25 and truncation on the query to reduce the retrieval time. We also perform the retrieval offline to avoid repeated time commitment. As a result, indexing 5% of the C4 corpus takes 1 hour. Offline retrieval for the entire training and evaluation mixture takes 11 hours, which is approximately 0.28 seconds per instance. Furthermore, we measure the computation overhead on inference which is caused by the additional retrieved inputs as well as a small amount of newly introduced parameters (+4.6%). The average computation overhead across all evaluation datasets during inference is around 4x compared with the *No Aug* baseline. Notably, Table 2 shows that Zemi$_{BASE}$ achieves competitive performance with T0-3B while being 15x smaller in scale, indicating that the benefit of the retrieval augmentation overwhelms the computation overhead.

[7] We follow T0 to move tasks that are originally in the evaluation split, i.e. COPA and RT, into the training split in this ablated setting.

4 Related Work

4.1 Semi-parametric models

Semi-parametric models (Sun et al., 2021; Verga et al., 2021; Chen et al., 2017; Lee et al., 2019; Guu et al., 2020; Wang et al., 2019; Karpukhin et al., 2020; Yang et al., 2019; Lewis et al., 2020; Izacard and Grave, 2020), which augmenting a parametric neural network with external knowledge bases or text corpora, have been widely applied to knowledge-intensive NLP tasks such as open-domain question answering. Recent advancements in semi-parametric *language models* (Khandelwal et al., 2019; Yogatama et al., 2021; Borgeaud et al., 2021; Zhong et al., 2022) have demonstrated improved language modeling performance with a relatively small language model and a retrieval system based on a large-scale corpus. Although the aforementioned semi-parametric language models have shown competitive performance on language modeling, compared with fully-parametric counterparts such as GPT-3 (Brown et al., 2020), it is unclear whether the superiority generally holds on down-

stream tasks. While concurrent work (Izacard et al., 2022) showed initial success in few-shot settings relying on Fusion-in-Decoder (FiD) (Izacard and Grave, 2020) framework, this work focus on the more challenging **zero-shot** settings (Sanh et al., 2021; Zhou et al., 2022; Gu et al., 2022). Furthermore, instead of reusing FiD framework as in (Izacard et al., 2022), we show that our newly proposed fusion module is more effective than FiD due to the **gated mechanism**, which is inspired by Highway Networks (Srivastava et al., 2015; Chai et al., 2020), Gated Convolution (Dauphin et al., 2017) and Vision-Language Fusion(Alayrac et al., 2022).

4.2 Massive multitask prompted training

Based on the assumption that the reasonable zero-shot ability of large language models may come from implicit multitask learning during pretraining, recent studies (Sanh et al., 2021; Wei et al., 2021; Ye et al., 2021; Wang et al., 2022b) have demonstrated that explicitly training a language model on a mixture of diverse tasks can effectively improve its zero-shot performance on unseen tasks. In this work, we extend T0's multitask prompted training to a **semi-parametric** setting, where we further augment the training and evaluation instances with retrieved documents. Notably, our work is distinguished from previous work ReCross (Lin et al., 2022), which uses upstream training data for augmentation, in twofold. First, we retrieve documents from a much larger task-agnostic corpus instead of clean upstream training instances. Second, in addition to directly concatenating the augmentation with the input just as FiD (Izacard and Grave, 2020), we further propose a novel retrieval-augmentation fusion module to handle retrieval noise.

4.3 Fusion of retrieved augmentations

In this work, the main challenge of designing the semi-parametric language model architecture is how to effectively leverage potentially noisy retrieved documents. Existing methods on incorporating external texts fall in two categories, *direct concatenation* (Lin et al., 2022; Brown et al., 2020; Liu et al., 2021; Lewis et al., 2020; Wang et al., 2022a) and *cross-attention* (Izacard and Grave, 2020; Prabhumoye et al., 2021; Borgeaud et al., 2021). However, we find that prior work lacks an explicit design for preventing the model from attending to noisy augmentations. Inspired by recent visual language models (Alayrac et al., 2022; Yu et al., 2022; Li et al., 2022; Jiang et al., 2022), we find that we can actually borrow ideas from vision-language fusion for text-text fusion. We identify two key differences from Flamingo architecture: first, we use a much smaller encoder-decoder model that is jointly trained with the newly initialized layers instead of frozen layers. Second, instead of inserting the gated cross-attention module into a large frozen language model (Hoffmann et al., 2022), we add only one layer of gated cross-attention on top of the encoder to alleviate the need for extensive pre-training.

5 Conclusion

In this work, for the first time, we show that semi-parametric language models have the potential to exhibit strong zero-shot task generalization ability by introducing Zemi. Through extensive analysis and ablation study, we further demonstrate that the interplay of the proposed retrieval-augmentation fusion and the semi-parametric multitask training is essential towards Zemi's empirical success. Notably, our proposed $\text{Zemi}_{\text{LARGE}}$ model outperforms T0-3B by 16% across seven diverse evaluation tasks while being 3.8x smaller in scale.

6 Limitation

In Section 3.2, we show that our training mixture with multiple-choice QA tasks, although small, is highly effective for multitask training. However, it is still unclear why multiple-choice QA tasks are particularly effective. Identifying the key factors towards positive or negative transfer from different tasks in the multitask training mixture would greatly help improve zero-shot task generalization. Future work includes investigating why certain mixtures are more effective than others and expanding the evaluation set to a wider range of tasks. Computation overhead is another noticeable limitation of semi-parametric models which is discussed in detail in Section 3.5. Moreover, future work on developing more efficient and accurate retrieval methods for retrieving from large-scale task-agnostic corpus can definitely improve semi-parametric language models.

Acknowledgements

We would like to express our gratitude to the anonymous reviewers for their insightful comments and

suggestions. We would also like to thank our colleagues and fellow interns at Tencent AI Lab for their valuable internal discussions and feedback, as well as the students from Blender Lab at the University of Illinois Urbana-Champaign for their insightful feedback.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2021. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yekun Chai, Shuo Jin, and Xinwen Hou. 2020. Highway transformer: Self-gating enhanced self-attentive networks. *arXiv preprint arXiv:2004.08178*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

ElasticSearch. Elasticsearch.

Wikimedia Foundation. Wikimedia downloads.

Edouard Grave, Armand Joulin, and Nicolas Usunier. 2016. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*.

Yuxian Gu, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Learning instructions with unlabeled data for zero-shot cross-task generalization. *arXiv preprint arXiv:2210.09175*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *arXiv:1909.00277v2*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2021. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2022. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*.

Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, and Xiang Ren. 2022. Unsupervised cross-task generalization via retrieval augmentation. *ArXiv*, abs/2204.07937.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2018. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. *arXiv preprint arXiv:2104.12714*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Kornél Csernai Shankar Iyer, Nikhil Dandekar. 2017. First quora dataset release: Question pairs.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

Haitian Sun, Patrick Verga, Bhuwan Dhingra, Ruslan Salakhutdinov, and William W Cohen. 2021. Reasoning over virtual knowledge bases with open predicate relations. In *International Conference on Machine Learning*, pages 9966–9977. PMLR.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. Quartz: An open-domain dataset of qualitative relationship questions. *arXiv preprint arXiv:1909.03553*.

Niket Tandon, Bhavana Dalvi Mishra, Keisuke Sakaguchi, Antoine Bosselut, and Peter Clark. 2019. Wiqa: A dataset for" what if..." reasoning over procedural text. *arXiv preprint arXiv:1909.04739*.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William Cohen. 2021. Adaptable and interpretable neural memoryover symbolic knowledge. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 3678–3691.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022a. Training data is more valuable than you think: A simple and effective method by retrieving from training data. *arXiv preprint arXiv:2203.08773*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. Adaptive semiparametric language models. *Transactions of the Association for Computational Linguistics*, 9:362–373.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. *arXiv preprint arXiv:2205.12674*.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Prompt consistency for zero-shot task generalization. *arXiv preprint arXiv:2205.00049*.

A Qualitative analysis of the retrieved documents

Here we visualize one good and one noisy example of the retrieved documents for each evaluation task. A full list of examples for each training and evaluation task can be found in the supplementary material under the "visualization" folder. As shown in Figure 4, 5, 6, 7, 9, 8, and 10, the retrieved augmentations can contain highly correlated information that can be directly helpful for solving a certain task, however, they can also be very noisy. As mentioned in Section 2.2, the retrieved documents can also be inaccurately ranked, for example in Figure 11, we show that the 21th ranked retrieval result can contain more correlated information than the top ranked ones. Furthermore, as shown in the noisy example of Figure 7, for some tasks such as sentiment analysis, even though the retrieved document is highly correlated with the input text, i.e., with a high BM25 score, the content can steer the prediction into a wrong direction. These observations motivate us to propose the augmentation fusion module with a gated mechanism.

B Implementation details

We use T5-base and T5-large as backbone model for $\text{Zemi}_{\text{BASE}}$ and $\text{Zemi}_{\text{LARGE}}$, respectively. We follow (Alayrac et al., 2022) to implement the perceiver resampler and the gated cross-attention. For both $\text{Zemi}_{\text{BASE}}$ and $\text{Zemi}_{\text{LARGE}}$, unless otherwise specified, we use one layer of gated cross-attention and one layer of perceiver resampler with a latent size of 64. A comprehensive ablation study on the impact of different aspects of our model design such as the Tanh Gate can be found in Section 3.4. All models are trained on the same training mixture as mentioned in Section 3.1 for ten epochs with a batch size of 32 and a learning rate of 1e-4. We report results from the checkpoint that achieved the best overall performance across all tasks. All experiments are done on eight NVIDIA-V100 32GB GPUs.

C Full list of tasks and templates

Following T0 (Sanh et al., 2021), we use tasks from Hugginface Datasets (Lhoest et al., 2021) and templates from PromptSource (Bach et al., 2022) marked as "original task" and with "choices_in_prompt". Specifically, for tasks in the training mixture, we randomly sample two

templates per task for semi-parametric multitask prompted training. For tasks in the held-out evaluation mixture, we use all available templates. Table 4, and 5 shows the full list of templates we used for each task during multitask training and zero-shot evaluation.

D Retrieval query key for each task

In order to retrieve most relevant documents for each instance, we specify a certain field for each dataset which will be served as the query to the retrieval system. For example, for most multiple-choice QA tasks, we use the "question" string as our query. Table 6 shows a full list of field names we use as retrieval query keys for each dataset. Note that the field name shown in the table is what appears to be in the corresponding Huggingface Dataset (Lhoest et al., 2021).

E Broader impact

One major benefit of developing powerful semi-parametric language models is that we can reduce the negative environmental impact from training huge parametric models. However, since the backbone language model is pretrained on massive internet-scale text data, there might be unexpected output that can have potential negative impact on the society, such as bias against people of a certain gender, race or sexuality. We are fully aware of the risks of potential misuses and will actively work with the community to improve the responsibility of large NLP models.

Mixture	Task	Template Name
Semi-T0 Training	cos_e/v1.11	question_option_description_text description_question_option_id
	cosmos_qa	context_description_question_answer_id description_context_question_answer_text
	dream	baseline read_the_following_conversation_and_answer_the_question
	qasc	qa_with_separated_facts_1 qa_with_separated_facts_4
	quartz	answer_question_below read_passage_below_choose
	sciq	Multiple Choice Multiple Choice Question First
	social_i_qa	Show choices and generate answer Show choices and generate index
	wiqa	effect_with_string_answer effect_with_label_answer
Semi-T0+ Training	wiki_qa	Decide_good_answer found_on_google
	trec	what_category_best_describe trec1
	super_glue/copa	more likely best_option
	rotten_tomatoes	Sentiment with choices Reviewer Opinion bad good choices

Table 4: PromptSource template names used for each task (Part1).

Mixture	Task	Template Name
Semi-T0 Evaluation	openbookqa/main	choose_an_answer_with_options which_correct pick_using_id choices only_options which_correct_inverse pick_answer_with_options
	piqa	what_is_the_correct_ending pick_correct_choice_with_choice_given_before_goal pick_correct_choice_index finish_sentence_with_correct_choice choose the most appropriate solution
	rotten_tomatoes	Reviewer Opinion bad good choices Sentiment with choices
	super_glue/cb	can we infer based on the previous passage claim true/false/inconclusive does it follow that justified in saying always/sometimes/never GPT-3 style consider always/sometimes/never guaranteed true must be true guaranteed/possible/impossible does this imply MNLI crowdsource should assume take the following as truth
	super_glue/copa	exercise . . . What could happen next, C1 or C2? i_am_hesitating plausible_alternatives C1 or C2? premise, so/because. . . . . . As a result, C1 or C2? best_option . . . which may be caused by more likely cause_effect . . . why? C1 or C2 choose
	super_glue/wic	question-context-meaning-with-label grammar_homework affirmation_true_or_false same_sense GPT-3-prompt-with-label polysemous
	hellaswag	complete_first_then Randomized prompts template Predict ending with hint if_begins_how_continues

Table 5: PromptSource template names used for each task (Part2).

Task	Query Key
cos_e/v1.11	question
cosmos_qa	question
dream	question
qasc	question
quartz	question
sciq	question
social_i_qa	context
wiqa	question_stem
openbookqa/main	question_stem
piqa	goal
rotten_tomatoes	text
super_glue/cb	hypothesis
super_glue/copa	premise
super_glue/wic	sentence1
hellaswag	ctx
wiki_qa	question
trec	text

Table 6: Retrieval query key used for each task.

Good Retrieval Example

======= Instance Index 6063 =========
Input Text: If a description of a situation begins like this: [header] How to set macgo mac blu ray player as default player [title] Download mac blu-ray menu player and install it at once. [step] There will be watermark on your screen if you play blu-ray with the trial version. Only 39.95 dollars for the full version of mac blu-ray menu player for now, please buy mac blu-ray player with discount.... Then how
does it continue?

Ending 1: [title] Click " check file associations " under " tools ". [title] Click and macgo mac blu-ray player will be your default player.

Ending 2: [title] Choose your video size and port size from the dropdown menu at the top of mac blu-ray menu. [step] Once you have downloaded the blu-ray menu player and installed it, you have to choose your video size and port size.

Ending 3: [title] Run the make app and then the itunes installer. [title] Determine the output type for each file in your mac blu-ray player.

Ending 4: [title] Uncheck the sidebar at the bottom of " applications ". [step] These are the files that are currently currently on your mac blu-ray player.

Target Text: Ending 1

Retrieved Documents
Rank: 0
Score: 136.6041
Retrieved Text: Macgo Mac Blu-ray Player has added itself Auto Play function, which means when you insert a disc into your Blu-ray drive, the player will automatically start and play. In order to make this whole process smoother, you'd better set Mac Blu-ray Player as default player on your Mac. Now I'll tell you how to do it.
After installing Mac Blu-ray Player, you can go to "Launchpad" and click on its icon to launch the program. The simplified main interface will reduce certain misoperations. You can see a menu at the top of the interface. Click "Check File Associations" under "Tools".
Then it will come up with a pop up window. You can choose some media formats which you want to play with Macgo Blu-ray player, then click "Make Mac Blu-ray player my default player".
Click "OK" to continue. Then Macgo Mac Blu-ray Player will be your default player.
After you set Macgo Mac Blu-ray Player as your default player, you also need to enable Auto Play function to freely enjoy Blu-ray this player.
Open "Preferences" under "Mac Blu-ray Player".Open "Playback" and tick under "Auto play when you insert a disc", and then click "OK".
Insert a Blu-ray disc into the drive and wait for the program automatically start and display the Blu-ray Menu. You can make some adjustments there or directly click "Play Movie" to enjoy some Blu-ray time.

Noisy Retrieval Example

======= Instance Index 122 =========
Input Text: If a description of a situation begins like this: A group of people are in a house. a man... Then how
does it continue?

Ending 1: is holding cored soap in his hand as he washes with a bottle.

Ending 2: is mopping the floor with a mop.

Ending 3: is shown wearing skis as he talks about areas he will like to ski on.

Ending 4: uses a paintball gun on his child.

Target Text: Ending 2

Retrieved Documents
Rank: 0
Score: 17.592176
Retrieved Text: #52704883 - Red lanterns, oriental charm, the Spring Festival atmosphere.
#35618548 - Silhouette of a man Happy successful raising arms to the sky..
#93113276 - Backlighting portrait of a joyful mother raising her baby outdoors..
#108745447 - Backlighting portrait of a joyful mother raising her baby outdoors..
#37541508 - Group of cheerleaders performing outdoors - Concept of cheerleading..
#108747918 - Back view of young backlit man looking into the distance on illuminated..
#38536931 - Working man walking near airplane wing at the terminal gate of..
#73301104 - Group of urban friends walking in city skate park with backlighting..
#76682984 - People silhouettes putting puzzle pieces together on city background..
#77013901 - People silhouettes putting puzzle pieces together on abstract..
#104666805 - Stylish light gray kitchen interior with modern cabinets with..
#108748125 - Back view of young backlit man looking into the distance on illuminated..
#86815910 - Best friends taking selfie outdoors with backlighting - Happy..
#86815909 - Best friends taking selfie outdoors with backlighting - Happy..
#117963685 - Back view backlighting silhouette of a man alone on a swing looking..
#86815911 - Best friends taking selfie outdoors with backlighting - Happy..
#118172724 - Back view backlighting silhouette of a man sitting on swing alone..
#96363446...

Figure 4: Example of retrieved documents on HellaSwag.

Good Retrieval Example

======= Instance Index 285 =========
Input Text: Decomposition occurs when a decomposer recycles nutrients from dead organisms back to the soil by eating them; what is an example of this?

Which is the correct answer?
- flies laying eggs on a body
- worms devouring a corpse
- wet leaves denigrating in a pile
- slugs digging through mulch

Target Text: worms devouring a corpse

Retrieved Documents
Rank: 0
Score: 77.088646
Retrieved Text: In most terrestrial ecosystems the bulk of nutrient cycling occurs in the topmost layers of soil. The main sources of the nutrient inputs to these soil layers comes from weathering, rainfall, fertilizers, atmospheric fallout, and organisms. Organism add nutrient matter via excreted wastes, shed tissues, and from the decomposition of their tissues when they die. Under most conditions, plants are the greatest single source of nutrients to soils. Plants not only supply nutrients released by organic decomposition of shed tissues and dead body parts, but also substances carried in from the plant leaves when water flows over them (foliar leaching). Losses or outputs of nutrients within ecosystems are by leaching, erosion, gaseous loss (like denitrification), and plant root uptake for growth purposes. Within the soil, nutrients are found attached to the surface of soil particles by chemical bonds, stored within the chemical structure of dead organic matter, or in chemical compounds.
Organic matter decomposition is the main process that recycles nutrients back into the soil. Decomposition of organic matter begins with large soil organisms like earthworms, arthropods (ants, beetles, and termites), and gastropods (slugs and snails). These organisms breakdown the organic matter into smaller pieces which can be decomposed by smaller organisms like fungi and heterotrophic bacteria (Figure 9q-1).

Noisy Retrieval Example

======= Instance Index 361 =========
Input Text: A scale can

Which is the correct answer?
- give an estimate of a dog's age
- measure how long a dog is
- let you know if the dog needs to lose a few pounds
- make an educated guess about a dog's breed

Target Text: let you know if the dog needs to lose a few pounds

Retrieved Documents
Rank: 0
Score: 9.894971
Retrieved Text: What is the abbreviation for Vertical Scale Measurement?
A: What does Y-SCALE stand for?
Y-SCALE stands for "Vertical Scale Measurement".
A: How to abbreviate "Vertical Scale Measurement"?
"Vertical Scale Measurement" can be abbreviated as Y-SCALE.
A: What is the meaning of Y-SCALE abbreviation?
The meaning of Y-SCALE abbreviation is "Vertical Scale Measurement".
A: What is Y-SCALE abbreviation?
One of the definitions of Y-SCALE is "Vertical Scale Measurement".
A: What does Y-SCALE mean?
Y-SCALE as abbreviation means "Vertical Scale Measurement".
A: What is shorthand of Vertical Scale Measurement?
The most common shorthand of "Vertical Scale Measurement" is Y-SCALE.
You can also look at abbreviations and acronyms with word Y-SCALE in term.

Figure 5: Example of retrieved documents on OpenbookQA.

Good Retrieval Example

======= Instance Index 653 =========
Input Text: Sentence: To make a graham cracker crust, to turn graham crackers to crumbs, you can

Choice 1: Run the graham crackers through a food processor

Choice 2: Run the graham crackers through a cheese grater

What is the index of the correct choice for ending for the sentence?

Answer:

Target Text: 1

Retrieved Documents
Rank: 0
Score: 82.00445
Retrieved Text: A graham cracker crust recipe for baked pies and no bake pies! We could also title this post The Anatomy of a Graham Cracker Crust. In other words, we're making our own graham cracker crust from scratch today and it'll be the best graham cracker crust you've ever had!
You can use it for no-bake pies or you can bake it first. That's a summer #win if you ask me!
Graham cracker crust is one of my favorite pie crusts. I don't think I can choose an absolute favorite, because I love all of them too much. But a good graham cracker crust is a must have in your baking arsenal. So many pies can be made to pair with the graham cracker flavor because it's so versatile. You can fill it with creamy s'mores chocolate pudding or even an easy blueberry-lemon dessert filling.
I think everyone needs a from-scratch graham cracker crust recipe in their arsenal. What if you want a pie right now and can't get to the store? And, let's face it. As good as those store-bought crusts are, they sorta taste like the aluminum foil pie tin, right? Or is it just me?
So today, I'm showing you my favorite from-scratch homemade graham cracker pie crust recipe. And this is even more perfect because you can use it for recipes that call for baking the crust OR you can use it no-bake.
Because when it's 106° like it has been this week in Sacramento, the last thing you want to do is turn on your oven. A

Noisy Retrieval Example

======= Instance Index 1111 =========
Input Text: Sentence: water

Choice 1: can drown a man

Choice 2: can drown a fish

What is the index of the correct choice for ending for the sentence?

Answer:

Target Text: 1

Retrieved Documents
Rank: 0
Score: 5.867839
Retrieved Text: best water pitcher filter lead water filter pitcher water filter lead best water filter for lead removal core pitcher lead reduction water pitcher filter 3 pk water filter aquagear water filter pitche.
best water pitcher filter water pitchers that remove lead water filter pitcher that removes fluoride fluoride water filter pitcher plus water pitchers water pitcher filter fluoride.
best water pitcher filter water filter pitcher water pitcher best water filter pitchers marina water filter pitcher zero water pur water filter pitcher target.
best water pitcher filter water filtration pitcher reviews water filtration pitchers comparison carafe water filters target water filter pitcher reviews.
best water pitcher filter water filter pitcher reviews best water pitcher filter best water filter pitchers water filter pitcher water filter pitcher reviews consumer reports.
best water pitcher filter our three picks for best water filter pitcher water pitcher filter cartridge.
best water pitcher filter the best water filter pitcher water filter pitchers best water pitchers best water filter pitchers best water filtration pitcher zero water pitcher filter replacement instruc.
best water pitcher filter filter pitcher target best water pitchers does zero water pitcher filter fluoride.
best water pitcher filter water filter best water pitcher filter water
...

Figure 6: Example of retrieved documents on Piqa.

Good Retrieval Example

======= Instance Index 575 =========
Input Text: every joke is repeated at least four times . every joke is repeated at least four times . every joke is repeated at least--annoying , isn't it ? Did the reviewer find this movie good or bad?

Target Text: bad

Retrieved Documents
Rank: 2
Score: 52.421543
Retrieved Text: Just very poor riddles in bad English and repeated 10 times each!
Several misspelled words (pretty unprofessional for a published "app book"). Also, some of the riddles are a bit morbid & makes me wonder what's going on in the mind of the one who came up with them..?! Not very challenging or logical for my taste.
This book has SOME useful riddles, but most of them repeat and don't make sense. Almost every riddle is misspelled and poorly written. Don't read this, find another book because this obviously looked like a 5th grader typed it from a cellphone.
Why did you put multiple of the exact same joke like a million times?!
The title says "8000+ riddles" but it doesn't say that all the riddles were DIFFERENT. It repeats the same riddles for pages after pages. Also, some riddles don't even make sense! And so much misspelling! Please update this and correct some misspelling and include more riddles so I'll rate 4 stars.

Noisy Retrieval Example

======= Instance Index 703 =========
Input Text: paul bettany is good at being the ultra-violent gangster wannabe , but the movie is certainly not number 1 . Did the reviewer find this movie good or bad?

Target Text: bad

Retrieved Documents
Rank: 0
Score: 58.04047
Retrieved Text: If you like retro crime movies this is a good one, its ultra-violence and unrelentingly crude language not withstanding. Much of the credit goes to Paul McGuigan's stylish direction which is so good that it makes you wonder why there are so many pedestrian films made. A good of credit should also go to Johnny Ferguson's amped-up screenplay and the fine performances by the three leads, Malcolm McDowell, David Thewlis and Paul Bettany. Although McDowell gets top billing this is really Paul Bettany's film whilst David Thewlis gives a solid and unusually restrained performance that counterbalances the familiarly thuggish ambiance.
The film opens potently with a Reservoir Dogs-like round table discussion amongst a troupe of aging East End crims recalling past times. The subject of Freddy Mays (Thewlis) comes up and this sets Malcolm McDowell's character referred to in the credits as Gangster 55 to recalling his rise in Mays' Kray-era gang. We then go into flash back and follow his story with Paul Bettany playing the McDowell character. Quite a few people will have difficultly accepting the casting of the handsome and refined looking Bettany playing a hard man, let alone McDowell's younger self, but he burns with the icily ambitious and sociopathic energy that the character requires. Set in the mid-60s, the production design is a treat, McGuigan's direction dynamic and the use of incidental music excellent.
The last act returns us to the starting point and now we understand why the name of Freddie Mays has derailed Gangster 55. The film looses some of its

Figure 7: Example of retrieved documents on Rotten Tomatoes.

Good Retrieval Example

======= Instance Index 10 =========
Input Text: The bowling ball knocked over the bowling pins.

What's the best option?
- The man rolled the bowling ball down the alley.
- The man dropped the bowling ball on his foot.

We are looking for a cause

Target Text: The man rolled the bowling ball down the alley.

Retrieved Documents
Rank: 0
Score: 54.388268
Retrieved Text: There are different types Chesterfield Bowling Clubs in Derbyshire.
Ten pin bowling is the most fashionable form of bowling. In ten pin bowling, matches consist of each player bowling a game. Each game is divided into ten frames. A frame allows a bowler 2 chances to bang down all 10 pins. The number of pins knocked over in each frame is recorded, a running total is made beneath the specific frame score as each frame goes on, and the player with the highest score in his/her game wins the match. Scores can be greater than the actual number of pins knocked over if strikes or spares are bowled. A strike is scored when a player knocks down all pins on the first roll in the frame. Rather than a score of just 10 for the frame, the player's score will be 10 plus the total pins knocked down on the next two rolls in the next frame(s). A spare is scored when all pins are knocked down using the second roll in the frame. The player's score for that frame will be 10 plus the number of pins knocked down on the first roll in the next frame. A player who rolls a spare or strike in the last frame is given one (if it was a spare in the previous frame) or two more rolls (if it was a strike in the previous frame) to score additional points.
As standard in most sports there are colloquialisms for various occurences in a game. Two consecutive strikes is acknowledged

Noisy Retrieval Example

======= Instance Index 2 =========
Input Text: The woman retired.

What's the best option?
- She received her pension.
- She paid off her mortgage.

We are looking for an effect

Target Text: She received her pension.

Retrieved Documents
Rank: 0
Score: 15.455591
Retrieved Text: What a fun and unique Valentine's gift!!!
Categories: Retirement, Woman, Man, Book.
Categories: Funny Gift, Retirement, Woman, Man, Book.
Our Name is Mud "Retired" Cuppa Doodle Porcelain Mug, 16 oz.
Categories: Funny Gift, Retirement, Woman, Decorative Items.
Our Name is Mud "Retirement Plan" Stoneware Mug, 16 oz.

Figure 8: Example of retrieved documents on COPA.

Good Retrieval Example

======= Instance Index 8 =========
Input Text: Given that A: And I haven't quite figured that out, if they figure they have got it won or if there's no real hurry because the first three quarters or, uh, uh, if something happens that that adrenalin starts flowing. They say, hey, we got to do something now. And then start playing the game the way the game should be played toward the last few minutes. B: Yeah. A: So, I don't know I'm looking for a good year. I guess we're always looking for a good year. B: So, obviously though, do you think they're going to do anything in the playoffs to make it to the Super Bowl this year Therefore, it must be true that "they're going to do anything in the playoffs to make it to the Super Bowl this year"? Yes, no, or maybe?

Target Text: Maybe

Retrieved Documents
Rank: 0
Score: 41.988216
Retrieved Text: Two-time super bowl champion and CNN Sport contributor Hines Ward shares his Week 9 takeaways with CNN's Jill Martin.
We're at the halfway point, and you start to see teams separate the contenders from the pretenders. You really see what teams are made of. This is a crucial month for a lot of teams in the NFL.
Let's start with the NFC South, where the Panthers and Saints need our attention.
The Carolina Panthers -- I don't think anyone expected them to have the year that they're having.
Cam Newton is looking like he's back to his MVP form, from back in 2015. What they're doing with running back Christian McCaffrey I just think is amazing. It's showing his versatility both running and catching the ball.
They're only one game behind the New Orleans Saints, and they have key matchups at the end of the year. In the last three weeks of the season, they play each other twice. Right now, it looks like it should be for the division.
Meanwhile, the Saints just knocked off the Rams. What, if anything does that performance show you?
Well, it's a tough place to play. I think, right now, it's really a two-team race to try to get that home field advantage for the playoffs.
I've played in New Orleans. I've been there. I know what their fans are like. It's one of the toughest places to play. It's loud. They get rowdy, and they love their Saints.
Definitely having Drew Brees playing at home in the playoffs helps the Saints' chances of making it to the

Noisy Retrieval Example

======= Instance Index 7 =========
Input Text: Given that It grew bigger with incredible speed, she was whizzing towards it. She must slow down or she 'd miss it. She took her foot off the accelerator and put it on the brake and as the car slowed she could see now that it was a child a toddler with a red woolly hat on. Therefore, it must be true that "it was a child"? Yes, no, or maybe?

Target Text: Yes

Retrieved Documents
Rank: 0
Score: 10.499066
Retrieved Text: The Plan Has Been Executed – My Grace Is..
This is my love letter to you son. Forever you will remain a child dear to me my daughter. I know you have read and heard that I have a plan to prosper you, to give you a hope and a future.
Child oh my child, yes I had a plan for you back then, back, back then. It was all true. But here is a thing today for you grab hold of my child. To master and rejoice in. The plan has been executed. The plan is sealed and delivered. My child, yes I had a plan for you, a plan for you to live a happy life. To live a joyous life. My plan was great for you my child. My plan was great for you my precious child.
Like every other parent, I had a plan for you my child. The plan was drawn down. Well designed, well traced and well set out. Just like a cartoonist would first draw before he brings the characters he has drawn to motion, I too, did that. I too my son had a plan in mind for you. I could not put you on earth and not have a plan at all. I did it and set it up my child. Worry not my son, the plan is executed. For long you heard the words that I had a plan for you, my precious child, please know this, the plan has been executed. The plan has come to life.
My plan

Figure 9: Example of retrieved documents on CB.

Good Retrieval Example

======= Instance Index 517 =========
Input Text: The word "knuckleball" has multiple meanings. Does it have the same meaning in sentences 1 and 2? Yes or no?

Sentence 1: Even the pitcher doesn't know where his knuckleball is going.
Sentence 2: Boston Red Sox pitcher Tim Wakefield is best known for his use of the knuckleball.

Target Text: Yes

Retrieved Documents
Rank: 0
Score: 70.62252
Retrieved Text: Tonight at approximately 5PM, Tim Wakefield will announce his retirement from baseball at the age of 45. "Wake" will finish his 19 year career with 200 wins, a feat he reached this past September.
His career accomplishments also include 2 World Series rings, an All-Star berth in 2009, 1995 AL Comeback Player of the Year, and 2010 Roberto Clemente Award winner, an honor he was nominated for eight times.
To Sox fans however, the knuckleballer will be remembered for being a world class team player who's sacrifices as a pitcher and an athlete in general are unparalleled. He was constantly asked to change his roles from front line starter, to middle reliever, and even a successful stint as a closer. This was something that most fans thought was easy since his style allowed it, but Tim has come forward recently as saying it was extremely difficult and uncomfortable.
In my mind, all you need to know about Wake happened in 2007. After finishing as one of the more reliable starters for Boston with a 17-12 record that season, he volunteered his roster spot in the World Series for a healthier rookie, Jon Lester, who won the clinching game against the Rockies. Name the players who have done that in the history of professional sports and you will undoubtedly come up with a very short list.
After being drafted as a first baseman by the Pirates in 1988, a scout told Wake that he would never make it above the AA level as a position player. Doing "anything he could to

Noisy Retrieval Example

======= Instance Index 406 =========
Input Text: The word "state" has multiple meanings. Does it have the same meaning in sentences 1 and 2? Yes or no?

Sentence 1: State your name.
Sentence 2: State your opinion.

Target Text: Yes

Retrieved Documents
Rank: 0
Score: 13.911013
Retrieved Text: The Washington attorney general issues formal published opinions in response to requests by the heads of state agencies, state legislators, and county prosecuting attorneys. When it appears that individuals outside the attorney general's office have information or expertise that will assist in the preparation of a particular opinion, a summary of that opinion request will be published in the state register. If you are interested in commenting on a request listed in this volume of the register, you should notify the attorney general's office of your interest by January 22, 2014. This is not the due date by which comments must be received. However, if you do not notify the attorney general's office of your interest in commenting on an opinion request by this date, the opinion may be issued before your comments have been received. You may notify the attorney general's office of your intention to comment by writing to the Office of the Attorney General, Solicitor General Division, Attention Jeffrey T. Even, Deputy Solicitor General, P.O. Box 40100, Olympia, WA 98504-0100, or by e-mail jeff.even@atg.wa.gov. When you notify the office of your intention to comment, you may be provided with a copy of the opinion request in which you are interested; information about the attorney general's opinion process; information on how to submit your comments; and a due date by which your comments must be received to ensure that they are fully considered.
1. Is an individual who has been convicted of aggravated assault, or other serious offenses, in a foreign country prohibited from possessing

Figure 10: Example of retrieved documents on WiC.

Example of Inaccurate Ranking

======= Instance Index 0 =========
Input Text: Sentence: How do I ready a guinea pig cage for it's new occupants?

Choice 1: Provide the guinea pig with a cage full of a few inches of bedding made of ripped paper strips, you will also need to supply it with a water bottle and a food dish.

Choice 2: Provide the guinea pig with a cage full of a few inches of bedding made of ripped jeans material, you will also need to supply it with a water bottle and a food dish.

What is the index of the correct choice for ending for the sentence?

Answer:

Target Text: 1

Retrieved Documents
Rank: 1
Score: 46.520477
Retrieved Text: how do I find neat names?
How to go about finding a vet?
guinea pig dali apparently on mend, again?
hamster cage Hammock pattern. . .
hamster cage Secure your cage doors!
ear infection, ear infections, inner ear infection degu sick am having rant!!!!
Do males hump each other?...
...
########
Rank: 10
Score: 41.217316
Retrieved Text: Contact Alittlebitiffy Animal Sanctuary at Alittlebitiffy Animal Rescue to express your interest.
Another successful adoption - amazing work Alittlebitiffy Animal Rescue!
More successful adoptions - amazing work Alittlebitiffy Animal Rescue! ...
...
########
Rank: 21
Score: 39.48997
Retrieved Text: Keeping your little furry friend healthy and happy should be a priority for any owner and, along with providing the right food for guinea pigs, finding an appropriate cage for them should be at the top of your priority list. Although, as you can see in this post here, there are numerous options on the market when it comes to commercially available guinea pig cages, some owners have opted towards a more do-it-yourself approach.
Many guinea pig parents complain that the regular pet store-sized cages are nothing but 'glorified litter boxes' and therefore are looking to improve the well-being of their cavies by making them a healthy and large-enough living enclosure, rather than buying one.
If you are one of those owners, this article will guide you through what you need to know before you start making a DIY cage for your guinea pig and what options you have when it comes to materials, design and features....

Figure 11: Example of the inaccurate ranking of the retrieval. Here we show the ranked retrieved documents for instance 0 in Piqa. We can see that the 21th ranked document is more correlated than many of the higher ranked ones, such as rank 1 and rank 10.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Appendix F Broader Impact

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Footnote 1. See supplementary material

☑ B1. Did you cite the creators of artifacts you used?
Reference. All datasets and models used in this paper are cited.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Reference and footnotes. All the datasets and models used and created by this work are publicly avaliable for research purpose.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3.1

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 3.1

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
See README of the supplementary code.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.1

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3, Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3, Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3, Appendix B, C, D

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.