Pick the Better and Leave the Rest: Leveraging Multiple Retrieved Results to Guide Response Generation

Bowen Wu, Yunhan Deng, Donghang Su, Jianyu Xiang, Chao Yang, Zongsheng Wang, Ying Li, Junhong Huang and Baoxun Wang

***Abstract*—Neural Response Generation (NRG) has achieved remarkable progress recently, though they suffer from the safe response problem. Some researchers proposed leveraging the retrieval-based chatbots' results to enhance the NRG models to generate diverse and informative responses. However, picking helpful information inside the multiple retrieved references and avoiding errors and noise brought from retrieval systems is still challenging. This paper proposes a variational neural response generating framework in which the validity of retrieved results is measured and those useful ones are taken as guidance explicitly. Moreover, if all the retrieved results fail to provide sufficient information, the framework also can let the model regress to a regular query-based NRG automatically. According to the thorough experimental comparisons with other retrieval-guided models, our proposed model can better utilize the useful information of retrieved results to generate appropriate and diverse responses.**

***Index Terms*—Response generation, Retrieval Guided, Conversational agents, Deep Learning**

I. Introduction

BASICALLY, models for building non-task oriented conversational agents (a.k.a, Chatbots) can be categorized into two architectures: the retrieval based models [1], [2] and the End-to-End generation based ones [3], [4]. Nowadays, generative Chatbots have drawn much attention due to their notable academic potential. However, the "safe response" problem [5] still remains a great challenge to be addressed. By contrast, the retrieval-based Chatbots can generally provide informative responses with better diversity [6], [7]. Thus, it is natural and reasonable to investigate methodologies for leveraging the capacities of retrieval based models to improve the diversity and informativeness of responses from generative models.

The typical methodology of the retrieval-based system [1] is to take the results as a condition (i.e. guidance) and perform response generation. Specifically, given a query q_i and the ground truth response r_i, the retrieval-based system can produce n replies $R_i = \{r_i^{(1)}, r_i^{(2)}, \cdots, r_i^{(n)}\}$ noted as references. Then, the retrieval-guided neural response generator is trained to maximize the objective written as:

$$\mathcal{L} = \sum_{i=1}^{N} P(r_i|q_i, R_i) \tag{1}$$

These methods conduct attention from the query to retrieved query-response pairs to ensure generators utilize those replies most related to the query [8], [9], [10], [11]. Some studies [8] also utilized the copy mechanism to exploit such references explicitly.

However, as the example shown in Table I, there exist noisy references (e.g. those marked as red or yellow) which have little help on learning to generate the ground-truth reply. To address this issue, a straightforward solution is to pick relevant and helpful references as guidance to generate the expected response, i.e. r_i in the training procedure and a reference given by the retrieval-based system noted as ***anchor*** a_i. Accordingly, various studies introduce to employ a $r_i^{(j)} \in R_i$ that lexically overlapped or semantically similar to the anchor as the reference [2], [12]. Specifically, in early studies [13], [14], [15], the selected reference is encoded into a real-valued vector to influence the generation process implicitly. Further studies take the reference explicitly as editing basis [2] or extracted skeleton [12], [16] for generation guiding.

Nevertheless, the hazard of such methodology lies in that, when irrelevant responses such as $r^{(3)}$ or $r^{(5)}$ in Table I are used during inference, models tend to be trapped in noise. Besides, valid information contained in multiple references was omitted since only one reference was supported in the

TABLE I: An example of the materials for training a retrieval guided generation model. The retrieved set of responses $R = \{r^{(1)}, r^{(2)}, ..., r^{(5)}\}$ can be classified with properties of: appropriate to q and helpful for generating r; appropriate to q but has little help for generating r; irrelevant to q (marked as green $\surd$, yellow $\square$ and red $\times$ respectively). Both the references marked with $\square$ and $\times$ are noise when a response generator aims to learn to generate r.

q	Does puppy's personality really match with the owner?	
r	That's true, puppy's personality indeed matches with the owner.	
$r^{(1)}$	It depends on character: "introverted or extroverted"	$\square$
$r^{(2)}$	I do find that puppy's personality is influenced by the owner	$\surd$
$r^{(3)}$	Gosh, are you talking about me?	$\times$
$r^{(4)}$	Absolutely, its quite similar with the own's personality"	$\surd$
$r^{(5)}$	Not yet.	$\times$

B. Wu and Y. Li are with School of Software & Microelectronics, Peking University, Beijing, China (email: jason_wbw@pku.edu.cn; li.ying@pku.edu.cn).

Y. Deng, D. Su, J. Xiang, C. Yang, Z. Wang, J. Huang and B. Wang are with Platform and Content Group, Tencent, China (email: yunhandeng, ashersu, jianyuxiang, adacyang, jasoawang, vincenthuang, asulewang@tencent.com).

[1]In this article, the retrieval-based system represents the retrieval-based Chatbot other than the search engine.

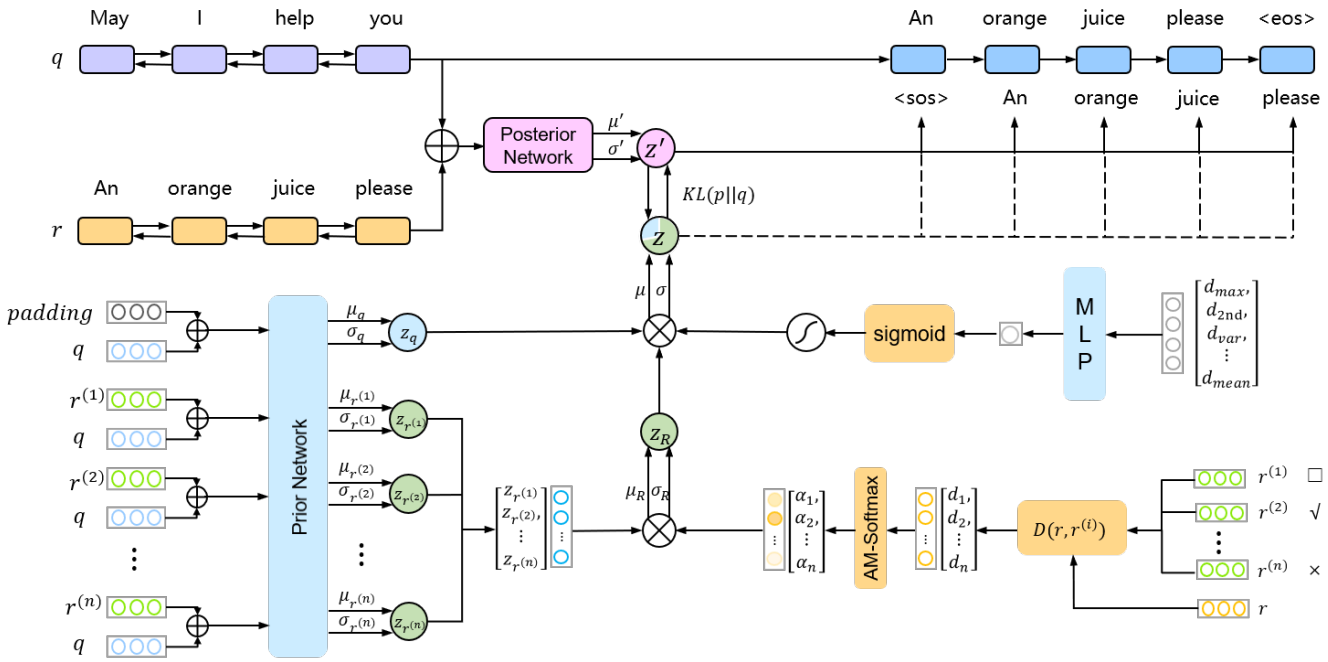

Fig. 1: The architecture of the Retrieval-Enhanced Response Generator (REGenerator) described in this paper. $\oplus$ represents the concatenation of inputs and $\otimes$ denotes the weighted sum of latent variables.

model design. Consequently, to better utilize the references to help model generating anchor-like replies, it is necessary to drive the response generators to sense and integrate useful information in multiple references. Meanwhile, it is better to maintain the original quality of the generation when the retrieval system only provides useless information.

To achieve these objectives, in this paper, we introduce a methodology to improve the multiple retrieved responses selection and utilization for better optimization and response generation. On one hand, a mechanism is proposed to represent the extent of preference for taking a retrieved response as guidance. It measures the references' semantically semantic relevance to the anchor and guides the generator to pay the most attention to those relevant ones. On the other hand, a gate module is proposed to endow the response generator with a capacity of annealing to a non-retrieval-guided model when only noisy references are provided. Moreover, to better integrate diverse semantic information of references and guarantee the generation quality, the entire model is conducted using the Conditional Variational AutoEncoders (CVAE) architecture [17]. Experimental results on the Retrieval Generation Chat dataset [16] reveal that the proposed model can notably improve the performance of retrieval-guided response generation. Especially, due to the annealing capacity, our model better maintains the appropriateness of generated results in the face of different quality references.

II. Approach

Utilizing latent variables in response generation has become a widely accepted methodology in NRG due to their Bayesian essence [18], [17]. It helps model the one-to-many relationships between post and response and deal with external information such as retrieved references efficiently. Hence, our proposed approach is built based on the variational generator. As shown in Figure 1, it consists of two main components: an anchor-guided selector and a retrieval-enhanced variational generator. The selector is designed to pick out anchor-related responses from multiple retrieved candidates and provide the retrieval-augmented prior distribution. Both the original and augmented prior distributions are coordinated by the generator in simulating the posterior distribution, as well as determining when and how much to take retrieved responses as references.

A. Anchor-Guided Selector

Let q, r, a stand for query, reply and the given anchor respectively (during the training phase, $a = r$). Meanwhile, $R = \{r^{(1)}, r^{(2)}, \cdots, r^{(n)}\}$ is a set of responses from any retrieval based Chat-bot, where n is the collection size. Firstly, each candidate $r^{(i)}$ is modeled by the prior network separately. In order to better representing the semantics of the retrieved references, a frozen pre-trained BERT [19] is utilized to transform q, r, R into their sentence vectors, written as v_q, v_r, $V = \{v^{(1)}, v^{(2)}, \cdots, v^{(n)}\}$. Then, given the i-th candidate, the prior network takes $v^{(i)}$ and v_q as inputs to generate the distribution $q_\theta(z_{r^{(i)}}|q, r^{(i)})$ of latent variable $z_{r^{(i)}}$. Since we assume $z_{r^{(i)}}$ follows isotropic Gaussian distribution, the distribution can be formulated as:

$$q_\theta(z_{r^{(i)}}|q, r^{(i)}) \sim \mathcal{N}(\mu_{r^{(i)}}, \sigma^2_{r^{(i)}}\mathbf{I}) \tag{2}$$

where the corresponding means $\mu_{r^{(i)}}$ and variances $\sigma^2_{r^{(i)}}$ are computed as:

$$\begin{bmatrix} \mu_{r^{(i)}} \\ \log(\sigma^2_{r^{(i)}}) \end{bmatrix} = W_{r'} \begin{bmatrix} v_q \\ v^{(i)} \end{bmatrix} + b_{r'} \tag{3}$$

Here $W_{r'}$ and $b_{r'}$ are the trainable parameters. Reparameterization[20] is taken for all the $(z_{r^{(1)}}, z_{r^{(2)}}, \cdots, z_{r^{(n)}})$ to obtain their corresponding samples.

Among the obtained latent variables, those corresponding to the references similar to the anchor, e.g. $r^{(2)}$ and $r^{(4)}$ in Table I, should be selected and integrated to generate responses. Firstly, we define the importance d_i of the i-th retrieved response as the cosine similarity between v_r and $v^{(i)}$. After that, based on it, a weighted average upon all latent variables is employed to integrate information.

$$\begin{aligned} z_R &= \sum_{i=1}^{n} \alpha_i z_{r^{(i)}}, \\ \alpha_i &= \frac{exp(\tau d_i)}{\sum_{k=1}^{n} exp(\tau d_k)} \end{aligned} \tag{4}$$

where α_i represents the impact factor of the i-th reference and the scaling factor τ is a hyperparameter for performing an AM-Softmax [21] without the margin. Compared to the standard normalization, such method assigns higher weights to the expected references rather than produces relatively-balanced weights. Accordingly, the model can explicitly extract useful information and avoid being misled by undesirable ones.

In addition, the sampling process of z_R can also be calculated using reparameterization as follows:

$$\begin{aligned} z_R &= \sum_{i=1}^{n} \alpha_i (\mu_{r^{(i)}} + \sigma_{r^{(i)}} \epsilon_{r^{(i)}}) \\ &= (\sum_{i=1}^{n} \alpha_i \mu_{r^{(i)}}) + (\sum_{i=1}^{n} \alpha_i \sigma_{r^{(i)}} \epsilon_{r^{(i)}}) \\ &= (\sum_{i=1}^{n} \alpha_i \mu_{r^{(i)}}) + (\sqrt{\sum_{i=1}^{n} {\alpha_i}^2 \sigma^2_{r^{(i)}}}) \epsilon \end{aligned} \tag{5}$$

where $\epsilon_{r^{(i)}}$ and ϵ are sampled from $\mathcal{N}(0, \mathbf{I})$. According to the fact that $\epsilon_{r^{(1)}}, \epsilon_{r^{(2)}}, \cdots, \epsilon_{r^{(n)}}$ are independent to each others, it can be derived that $\sum_{i=1}^{n} \alpha_i \sigma_{r^{(i)}} \epsilon_{r^{(i)}} \sim \mathcal{N}(0, (\sum_{i=1}^{n} {\alpha_i}^2 \sigma^2_{r^{(i)}}) \mathbf{I})$. Then followed by reparametrization trick again, z_R can be re-written as the last formula in Equation 5. In practice, rather than sampling n times, we construct the distribution $q_\theta(z_R) \sim \mathcal{N}(\sum_{i=1}^{n} \alpha_i \mu_{r^{(i)}}, (\sum_{i=1}^{n} {\alpha_i}^2 \sigma^2_{r^{(i)}}) \mathbf{I})$ to directly get z_R's sample.

In conclusion, the Anchor-guided Selector is designed to select anchor-relevant responses and generate samples (also noted as z_R in Figure 1) from their collectively affecting distribution.

B. Retrieval-enhanced Generator

Although the selector is capable of choosing valuable responses, if all retrieved results are of limited help in generating the expected results, the weighted average combination still fails to give an accurate representation. Thus, based on the importance of the retrieved responses, we introduce a gate mechanism to control the volume of adopted information dynamically. Firstly, as shown in Table II, ten features are constructed to reflect whether the retrieved results providing useful information. The features are then fed into a multi-layer perception to get $\beta \in [0, 1]$. We denote β and $1 - \beta$ as *decay* gate and *enhance* gate respectively. When β is close to 1, the model ignores the retrieved information and degenerates into a standard VAE model that mainly focuses on the query. Besides, the prior network is shared to compute the distribution of latent variable z_q of query q. Since the network's inputs being v_q instead of v_q and $v_{r^{(i)}}$, a trainable and random initialized variable $padding$ is applied to pad the position of $v_{r^{(i)}}$. Finally, the prior latent variable z is computed as follows:

TABLE II: The constructed features. d_{max}, d_{2nd}, d_{3rd}, d_{mean} and d_{var} indicate maximum, the second largest value, the third largest value, mean and variance of $\{d_1, d_2, \cdots, d_n\}$, respectively. The subscripts of difference features represent difference between two statistical values, e.g., $d_{max-mean}$ represents $d_{max} - d_{mean}$.

Feature class	**Features**
statistical features	$d_{max}, d_{2nd}, d_{3rd}$ d_{mean}, d_{var}
difference features	$d_{max-mean}, d_{2nd-mean}$ $d_{3rd-mean}, d_{max-2nd}$ $d_{max-3rd}$

$$z = \beta z_q + (1 - \beta) z_R \tag{6}$$

Again, the reparameterization trick is utilized to perform the sampling of z.

In addition, we inherit the encoder-decoder architecture and the posterior network from the standard VAE model. A bidirectional LSTM is first employed to encode the query and response into fixed size vectors h_q and h_r. After that, the posterior network (parameterized by ϕ) takes h_q, h_r as inputs to generate the distribution $p_\phi(z|q, r)$ of latent variables. Besides, we use the bag-of-word (BOW) loss [17] to tackle the latent variable vanishing problem and an extra $KL(p_\phi(z|q, r)||q_\theta(z_q|q))$ to regular the padding variable. On this basis, the whole network is trained by maximizing the variational lower-bound [18] of the objective:

$$\begin{aligned} \mathcal{L}(\theta, \phi; q, r, R) = & \mathrm{E}_{p_\phi(z|q,r)}[\log P(r|z, q, R)] \\ & - KL(p_\phi(z|q, r) \| q_\theta(z|q, R)) \\ & - KL(p_\phi(z|q, r) \| q_\theta(z|q)) \\ & + \mathrm{E}_{p_\phi(z|q,r)}[\log P(r_{bow}|z, q)] \end{aligned} \tag{7}$$

C. Inference

Given that the objective of the training is to maximize the likelihood of the ground truth response, it is reasonable to use it as the anchor. However, during inference, the response is unknown. Therefore, it is crucial to identify an appropriate mechanism to select an expected response from the retrieved responses to provide guidance.

On the one hand, a useful anchor should help select the responses conversationally related to the given query. Thus, we utilize the matching model proposed by [16] to calculate the relevance score γ_i between the query and the i-th candidate response. On the other hand, by design, our retrieval-enhanced generator is more likely to generate a desired result if more

related references are provided. Therefore, another standard for anchor selection is to choose a cohesive anchor which can be helped by other replies. To achieve it, when inference, we use each retrieved response as the anchor and fed it into the selector to get its corresponding *enhance* score. A larger *enhance* score indicates the model can use more information as well as generate higher-quality responses. Besides, the candidate anchor itself is removed from the referenced responses. Finally, the i-th candidate reference's score is computed as follows, and the top-k results are selected to stand for the anchor.

$$\delta_i(q, r^{(i)}) = \gamma_i + (1 - \beta_i) \tag{8}$$

III. Experiments

A. Dataset

We conduct experiments on the Retrieval Generation Chat dataset [16], which contains about five million query-response pairs and provides 3 to 10 retrieved references for each query. Note that there exist samples where the retrieved query is exactly the user's utterance, and such retrieved candidates are taken as the ground truth and removed from the rest candidates. After that, we also remove those queries with more than ten responses since each reference is probably irrelevant to most replies. Moreover, following the setting of previous studies [16], [22], only the samples with at least 20% of corresponding R satisfying $\mathit{Jaccard}(r, r^{(i)}) > 0.3$ are leveraged for training, where *Jaccard* stands for the Jaccard distance. Finally, since each query corresponds to multiple replies, we split the filtered corpus into training (1,179,374), validation (21,462), and test (20,896) sets based on the query. The modified dataset has been published[2], and the samples are given in the **Supplementary Material**.

B. Model Variations

Retrieval-Sys The underlying retrieval system. A dual-LSTM model [23] is trained to re-rank the retrieved results by measuring their matching degrees to the current context.

S2S The basic Seq2Seq model with the attention mechanism [4].

CVAE The conditional variational auto-encoder for response generation [17].

GPT2 Generative Pre-trained Transformer 2 [24] is one of the most well-known neural language model instances, and it has shown surprisingly realistic text generation results including dialogue generation [25].

Edit-Vec The model proposed by [22]. An edit vector representing the lexical difference between the current context and the retrieved one is introduced to enhance the decoder.

Skeleton-Match The best method presented in [16]. We refer to it as Skeleton-Match because its skeleton extractor is trained by the matching model that reveals token-level matched pairs between a query and its response.

Multi-S2S & Multi-GPT2 The multi-seq2seq model [8] (Multi-S2s) which encodes N-best response candidates using N encoders and utilizes the attention and copy mechanism to refer to these semantics. Based on this framework, Multi-GPT2 is the model that uses GPT2 as the backend. The integration of information of references follows the scheme proposed in [26]. In detail, each reference is encoded by BERT, and then the last state as the key and value vectors to perform the Pseudo Self-Attention in decoder layers.

REGenerator & REGPT2 REGenerator is the proposed model in this paper. Since our model is a typical variational framework, the obtained latent variable z can be used to feed into a GPT2-based VAE model [26], noted as REGPT2.

ChatGPT & ChatGPTRetrieve & REChatGPT Due to ChatGPT's [27] exceptional performance in various text generation tasks, we utilized it as a backend model to confirm the efficacy of retrieval-guided models. To enhance ChatGPT's capability to accommodate daily conversation scenarios, we utilize the prompt "Please provide a suitable response for {query}." [3] as input. To facilitate generating replies based on all the retrieved references, we also incorporate "Given that {query1} can be answered with {reply1}, {query2} can be answered with {reply2}, ... Please provide a suitable response for {query}.", noted as ChatGPTRetrieve. Due to our inability to fine-tune ChatGPT with the VAE component, we are relying solely on anchor and REGenerator-identified effective question-reply pairs as references to perform our methods upon ChatGPT. Specifically, when β computed by REGenerator is greater than 0.9, we only use the query as input for ChatGPT. In all other cases, we select the minimum number of retrieved responses to ensure the sum of their corresponding αs is greater than 0.5, in addition to the selected anchor, and use them together as references for input into ChatGPT.

C. Implementation Details

For a fair comparison, all models share the same configuration. A 256-dimensional word embedding is randomly initialized, and the sentence vector is extracted from the last transformer layer of BERT. We employ a BiLSTM of hidden size = 256 for encoding, and a LSTM of hidden size = 512 for decoding. The dimensions of all the latent variables are set to 512. We employ Adam [28] for optimization with batch size of 128 and a learning rate of 0.0002. The size of vocabulary and beam search is set to 30,000 and 15, respectively. For the Edit-Vec, Skeleton-Match, and our model, the top three scored retrieved results are taken as anchors to guide the generation, and five responses are randomly selected for each. For all the GPT2-based models, the number of transformer layers, hidden size and heads of attention is set as 12, 768, 12, respectively. Besides, the code is provided in the **Supplementary Material**.

D. Automatic Evaluation Metrics

Since the generated responses tend to be more specific when retrieved responses provided, the one-to-many characteristics of the dataset can help us better evaluate a retrieval-enhanced model. As discussed in previous studies [29], given multiple ground truth responses, it is more effective to measure whether

[2]https://drive.google.com/drive/folders/1W0aEZ0oj123Hj2uOZ-m55OVGwCWglynV?usp=sharing

[3]All the prompts are translations of the Chinese version used in practice.

TABLE III: Evaluation results on the Retrieval Generation Chat dataset.

Methods	Max-BLEU	Max-Embedding			Diversity		
		Average	Extrema	Greedy	Distinct	Ent-1	Ent-2
Retrieval-Sys	0.036	0.787	0.608	0.575	0.018	9.95	15.12
S2S	**0.121**	0.814	0.723	0.721	0.004	6.67	9.46
CVAE	0.092	0.831	0.729	0.727	0.007	7.25	10.53
Edit-Vec	0.079	0.834	0.720	0.727	**0.014**	7.74	12.16
Skeleton-Match	0.093	0.841	0.742	0.740	0.010	6.51	12.16
Multi-S2S	0.104	0.840	0.747	0.743	0.010	7.93	11.94
REGenerator	0.113	**0.846**	**0.751**	**0.749**	0.011	**8.08**	**12.27**
GPT2	**0.134**	0.817	0.721	0.713	0.007	8.55	12.35
Multi-GPT2	0.122	0.852	0.757	0.758	0.010	9.12	12.89
REGPT2	0.125	**0.864**	**0.761**	**0.767**	**0.015**	**9.45**	**13.49**
ChatGPT	0.155	**0.853**	0.671	0.658	0.015	**8.91**	13.28
ChatGPTRetrieve	0.124	0.775	0.624	0.606	0.014	8.86	12.96
REChatGPT	**0.159**	0.848	**0.677**	**0.664**	**0.017**	8.88	**13.29**

TABLE IV: Human labeled results upon the generated responses.

Methods	Appropriateness				Informativeness			
	0	1	2	Average	0	1	2	Average
Retrieval-Sys	77.3%	12.7%	10.0%	0.327	6.7%	47.3%	46.0%	1.393
S2S	26.7%	46.7%	26.6%	0.999	18.7%	65.0%	16.3%	0.976
CVAE	46.0%	42.3%	11.7%	0.657	10.0%	80.3%	9.7%	0.997
Edit-Vec	56.0%	26.0%	18.0%	0.620	13.3%	77.0%	9.7%	0.964
Skeleton-Match	30.3%	45.7%	24.0%	0.937	16.0%	69.3%	14.7%	0.987
Multi-S2S	23.0%	53.7%	23.3%	1.003	20.0%	62.0%	18.0%	0.980
REGenerator	***15.0%***	53.0%	**32.0%**	**1.170**	***5.0%***	76.3%	**18.7%**	**1.137**
GPT2	24.1%	49.3%	26.6%	1.025	6.3%	84.0%	9.7%	1.034
Multi-GPT2	20.0%	52.7%	27.3%	1.073	4.9%	79.2%	15.9%	1.110
REGPT2	***13.2%***	49.3%	**35.5%**	**1.203**	***4.4%***	74.4%	**21.2%**	**1.168**
ChatGPT	13.6%	4.5%	81.9%	1.684	***0.1%***	11.7%	88.2%	1.883
ChatGPTRetrieve	11.3%	12.0%	76.7%	1.653	1.4%	11.3%	87.3%	1.860
REChatGPT	***8.1%***	6.0%	**85.9%**	**1.779**	0.7%	9.4%	**89.9%**	**1.893**

the generated result is close to any reference. Inspired by above conclusion, we adopt the following automatic evaluation metrics:

Max-BLEU introduced by [29], which evaluates a generated response based on multiple ground truths. Unigram BLEU (BLEU-1) is utilized to compute the amount of the word overlaps.

Diversity We use Distinct [5] to measure the diversity by counting the ratio of generated unique unigram. Besides, Entropy (Ent-n) [30] is employed to reflect how evenly the empirical generated n-gram distribution is for a response.

Max-Embedding Embedding-based metrics [31] compute the cosine similarity between the sentence embeddings of a given response and a generated result. We also calculate the result's similarities to multiple ground truths and select the highest one as the score. Besides, stopwords are removed during evaluation.

E. The Human Evaluation Criterion

We recruit 12 annotators from the crowd-sourcing labeling resources of our organization to manually evaluate the quality of generated responses. For each model, 300 generated samples are judged by 12 annotators, and each query-response pair is cross-evaluated by three annotators. Besides, human evaluation is performed from two aspects: appropriateness and informativeness. The levels of **Appropriateness** which evaluate whether the response is appropriate and relevant to the query are defined as 0 (irrelevant), 1 (acceptable), and 2 (great).

Moreover, the retrieval-involved models often use the skeleton or copy mechanism which tends to directly choose phrases from retrieved responses. Such characteristics might improve sentence diversity superficially but also raise the hazard of grammar incoherence. So, we evaluate the **Informativeness** based on grammatical correctness. Three levels are conducted following the criteria: **0**: the response has obvious grammar errors, even if it incorporates enough information; **1**: the

response is not informative but grammatically acceptable; **2**: the response is informative.

IV. Results and Analysis

A. Experimental Results

The results of automatic evaluating metrics are illustrated in Table III. Firstly, we compare those non-GTP2-based models. It can be observed that the Max-BLEU scores of various models are relatively low, and the Seq2Seq model is higher than the others. As discussed in previous studies [32], higher BLEU scores usually come from frequent words or keywords covered by the ground truth. Therefore the score of S2S is the highest due to more safe responses generated. Moreover, the last three retrieval-guided models outperform the rest models. We attribute this to the fact that these models are effectively referred to the retrieved responses. Furthermore, our model obtains the highest scores on all three embedding-based metrics, suggesting that our model's generated responses are more semantically relevant to the ground truths.

According to Diversity results, the retrieval system produces most diverse responses with no doubt. Correspondingly, all the models enhanced by these retrieved results outperform other baselines significantly. Especially by performing explicit edits, Edit-Vec achieves the highest Distinct score. However, its Entropy scores are relatively low, which reflects the lack of diversity among the non-copied words. By contrast, despite not generating most diverse words, REGenerator uses the words most impartial. We attribute this to the fact that, by consulting multiple retrieved references, our model better understands different anchors' guidance and generates corresponding replies.

As shown in Table IV, Edit-Vec and Skeleton-Match generate even more irrelevant results than the S2S model. Especially, we find that the responses generated by the Edit-Vec performs inadequately due to its strong dependency on anchor responses even if the anchors themselves are inappropriate. By contrast, the skeleton selection is more flexible as it can better ignore the useless information. Moreover, Multi-S2S obviously generates fewer irrelevant responses and improves the replies' informativeness, which indicates referring on multiple retrieved samples other than a specific one can guarantee better diversity and accuracy. Furthermore, our proposed model further outperforms all baselines by utilizing multiple references more effectively. The low percents of irrelevant and not fluent responses demonstrate that our specially designed architecture balances the consulting form retrieved responses and the pure variational generator.

The experimental results of GPT2-based models align with those of basic models, indicating the efficacy of large-scale transformer-based models. Notably, utilizing GPT2 alone can yield comparable or even superior results to most models, with a lower ratio of non-fluent responses marked as Informativeness level 0. However, Table IV shows that GPT2 tends to generate only acceptable and non-informative responses due to insufficient information. Retrieval-guided models, built on the foundation of powerful GPT2, can better explore effective information in retrieval system results, leading to improved performance. Our proposed framework, the GPT2-based REGenerator, outperforms other models, including GPT2-based ones, further validating the framework's ability to effectively utilize retrieval results.

TABLE V: Ablation studies of retrieval-enhanced generator. *ave.*, *ext.* and *gre.* are the abbreviations for Average, Extrema, and Greedy respectively. $\beta = 1$ represents the model abandoning the retrieved information, i.e., the standard VAE part inside our model. $\beta = 0$ stands for the situation that the model has to leverage message from R. Besides, Dynamic β represents the whole proposed model.

Methods	**Max-Embedding**			**Ent-2**
	ave.	ext.	gre.	
$\beta = 1$	0.820	0.717	0.713	10.77
$\beta = 0$	0.839	0.747	0.748	12.09
Dynamic β	**0.846**	**0.751**	**0.749**	**12.27**

Although we cannot use training data for finetuning ChatGPT, it is evident through manual evaluation that its generated responses are of much higher quality than other basic models, with an average appropriateness score improvement of at least 0.48 and nearly 90% of responses being informative. By further providing retrieval results that are deemed helpful for the query by REGenerator as references for in-context learning, REChatGPT further reduces the proportion of inappropriate responses by 5.5% and increases the proportion of high-quality responses by 4%. However, as a comparison, providing all retrieval results to ChatGPT as references would compromise the quality and informativeness of generated responses, which validates the effectiveness of our proposed reference response selection method. Furthermore, REChatGPT exhibits increased diversity compared to ChatGPT and other baseline models. However, due to the fact that the information content of generated results based on ChatGPT is much higher than that of the ground truth, the embedding-based evaluation metrics are not as good as those of other basic models.

B. Ablation Study

To gain a better understanding of the effectiveness of each component in the Retrieval-enhanced generator, we conducted ablation studies to analyze their contributions. As presented in Table V, the last two forms of our model, which leverage retrieved references, achieved remarkable improvements compared to $\beta = 1$. Moreover, when $\beta = 0$, the generator is compelled to rely entirely on the retrieval-guided latent semantic. We observed a significant increase in both the embedding and entropy scores after switching to dynamic fusion from static. This finding demonstrates that the gate structure can effectively limit redundant information derived from retrieved results, given that a specific part of queries contains only a small amount of useful retrieved responses.

Additionally, when $\beta = 1$, the model is essentially a CVAE model with the exception of a fixed sentence vector obtained from BERT being used to derive the latent variables. However, as demonstrated in Table III, utilizing pretrained BERT without finetuning does not enhance the quality of generated responses

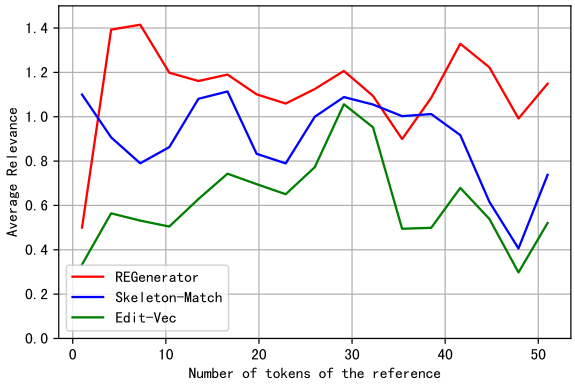

Fig. 2: Correlation curve between averaged relevance scores and the length of the anchor. Because most candidate anchors have less than 50 words, the cases with anchor longer than 50 are summarized in the last node.

compared to the standard CVAE model. Therefore, it can be concluded that the proposed model's improvement is attributed to the effective utilization of multiple retrieved references.

C. Influence of the Anchor

During inference, Edit-Vec, Skeleton-Match and our model select an anchor from the retrieved results to select useful references from R. Especially, since either Edit-Vec or Skeleton-Match is designed to have only one retrieved response, they utilize the anchor itself as guidance. In this section, we discuss the influence of different types of selected anchors based on the performance of the above three models.

1) Complexity: Generally, a longer reference contains more information as well as noise for replying, so it is more challenging to extract its "nourishment". To evaluate the model's robustness on anchors with various lengths, we demonstrate the average relevance scores labeled by annotators in Figure 2. Guided by the long anchors especially those containing more than 35 words, the quality of generated responses from Edit-Vec and Skeleton-Match drops significantly. By contrast, the relevance scores of REGenerator are relatively stable. This phenomenon indicates that, compared to directly using the anchor's literal information, leveraging multiple related retrieved responses can better process the useful information to help the future generation.

2) Appropriateness: Except for the complexity, whether an anchor itself is able to reply to the query may also influence the generation. Both Edit-Vec and Skeleton-Match actually use the retrieved results that are more relevant to the query as anchors. Thus, annotators are also asked to label all selected anchors' relevance for analyzing the influence of anchors' quality on generated results. Consequently, 33% of selected retrieval results are irrelevant to users' utterances.

As shown in Table VI, the last column shows the Pearson correlation coefficient between the anchors' appropriateness scores and the ones given to the corresponding generated replies. Obviously, Edit-Vec highly relies on the quality of the anchor, reflected by the correlation score of 0.29 and the measure of relevance. When guided by anchors great appropriate to the query, the average score of Edit-Vec reaches 0.83, but the model struggles to generate acceptable responses if the quality of selected anchor is low. By contrast, the matching based skeleton extraction of Skeleton-Match can produce an empty skeleton for an irrelevant anchor. Thus, replacing great anchors with irrelevant ones, its corresponding performance only decreases by 0.2 on average. However, explicitly consulting an utterance regardless of its degree of relevance still brings in noises inevitably. As we propose to extract useful information implicitly with the help of diverse retrieved responses, consequently, nearing zero correlation demonstrates our proposed model's performance is almost not influenced by irrelevant anchors.

TABLE VI: The appropriateness results vary with the quality of the selected anchor. AR-k indicates the group of results is drawn from the cases with the anchor labeled as k level appropriate to the query. S-Match and REGen in the table are the abbreviations for Skeleton-Match and REGenerator, respectively.

Methods	Average Appropriateness			Pearson
	AR-0	AR-1	AR-2	
Edit-Vec	0.30	0.55	0.83	0.29
S-Match	0.81	0.97	1.02	0.11
REGen	**1.12**	**1.15**	**1.22**	***0.06***

TABLE VII: An example of generated responses from different models with the same anchor.

Query	Let's talk about your reasons for break up.
S2S	I also want to break up.
CVAE	So am I.
Multi-S2S	Impossible.
<Anchor>	Is it necessary to have a reason for break up?
Edit-Vec	Is it necessary to have a reason for break up.
S-Match	Is there any reason.
REGen	All the reasons for break up are excuses.

D. Case Study

Some generated results from different methods are also presented in Table VII. It can be observed that both S2S and CVAE fail to discuss the topic "reasons for break up". By contrast, Multi-S2S gives an indirect but reasonable answer to the query. Furthermore, since the chosen anchor itself is a good response, all three anchor-guided models give more informative results about break up reasons. However, the reply generated by Edit-Vec is relatively meaningless as it is basically a copy of anchor (result of the retrieval-based system), while the S-Match rephrases the anchor. Overall, our REGen produce the most relevant and informative response. Beside the case shown in Table VII, more diverse cases are given and discussed in the **Appendix**.

V. Related Work

A. Neural Response Generation

As the widely applied mainstream architectures of conversational agents, both retrieval-based chat-bots [1], [2] and end-to-end generative conversation models [3], [4] have been developing rapidly during recent years. Especially, the neural generative models are considered as the typical representation of linguistic intelligence. For generative models, the major challenge is to improve the diversity and informativeness of generated results with relevance maintained, and thus, various variational response generators are introduced to address this problem since they can naturally promote diversity by involving sampling in the generate stage [17], [18], [33], [34], [35]. Moreover, consistent with other NLP tasks, the advent of large-scale pre-trained language models further improved the quality of generated responses [36], [37], [38].

Besides the model-driven approaches, the fusion of knowledge is also an effective way to improve the quality and informativeness of responses, so as to generate diverse results [39], [40], [41], [42], [43]. The knowledge-enhanced response generation methods can be categorized according to the organization of knowledge, which includes Knowledge base (KB), Knowledge graph (KG), and Knowledge grounded text [44]. Correspondingly, various studies focus on retrieving helpful facts from KB [39], [40], reasoning the context basis on the KG [45], [46], conducting effective queries and summarizing the unstructured text [47], [48], and then fusing all the processed knowledge into generation.

B. Retrieval-Guided Response Generation

Retrieval-based chat-bots have the natural advantage in feeding highly-diverse responses. Thus, it is of great necessity to leverage retrieved responses to improve the diversity of results given by neural response generation models. There are three perspectives to classify the retrieval-guided response generation methods: (a) implicate or explicate utilization of references; (b) single or multiple references; (c) anchor-guided references selection or not.

The implicitly guided methods transform one [13], [49], [14] or multiple [9], [11] retrieved responses into vectors to influence the generation. And the attention mechanisms including multi-level attention are widely used in these works [8], [10]. Besides, a re-weight function of the loss is designed in [15] by considering the similarities between retrieved queries and the input query to control the contributions of different retrieved responses. By contrast, some studies attempt to explicitly extract key information from retrieved results to guide the response generation. Copy mechanism can help the model directly generate the words provided in the references [8]. An extra encoder for the lexical differences between retrieved context and original context is introduced to represent the available information in an anchor selected reference [22]. Moreover, some researchers propose to construct skeletons or frames for response generation by editing one selected retrieved response [12], [16], [50].

However, although various methods have demonstrated that both multiple references [8], [11] and anchor-based reference selection [22], [51] can guide the model to generate better results, current models those selecting references by anchor only utilize one reference to perform guidance. The motivation of such design lies in that these methods aim to explicitly modify a specific reference to obtain results, so as can guarantee better diversity by employing different references directly [12], [50]. From our perspective, the architecture proposed in this paper is the first anchor-involved multiple retrieved results guided response generation.

In addition, there is another kind of response generation involving a retrieval system that belongs to the knowledge-enhanced method [52], [53], [54]. These methods first generate a query according to the context and then retrieve a pre-built search engine or directly use the internet to gain knowledge grounded text. And then, knowledge-enhanced mechanisms are employed to utilize the information. Thus, they do not belong to the retrieval-guided methods discussed in this paper that generate responses referencing other replies that are judged as suitable replies for the current query by another retrieval-based dialogue system.

VI. Conclusions

In this paper, we proposed a neural generation model named REGenerator which leverages multiple retrieved responses to produce informative and relevant responses. From experimental results on multiple related baselines, the following conclusions can be drawn: 1) our model captures the advantages of both generative and retrieval-based methods, reflected by evaluating metrics; 2) the ablation study reveals the capability of our dynamic-gate structure on both response selection and auto-regression; 3) compared to other anchor-related models, REGenerator can successfully unitize the comprehensive semantics from multiple retrieved responses and generates more complex, relevant, and coherent sentences.

Appendix A
Influence of the Anchor Selection

TABLE VIII: Average relevance and informativeness scores of the generated responses.

Criteria	Relevance	Informativeness
By relevance	**1.18**	1.09
By cohesion	1.16	**1.21**

As we conclude that REGenerator is less sensitive to the anchor's relevance, it is reasonable to seek other criteria that can lead the model to generate high-quality responses. Precisely, different from previous studies, REGenerator not only selects the retrieved results relevant to the query as anchors, but also considers each candidate's cohesion. In this section, the anchors selected by δ are firstly categorized into two groups: (a) those can also be picked out by γ; (b) those cannot be picked out by γ. Then, we compare the performance of the REGenerator on this two categories of anchors. As shown in Table VIII, the average relevance scores of results on anchor groups are 1.18 and 1.16, while the corresponding informative scores are 1.09 and 1.21 respectively. Both two types of anchors can

lead the model to generate highly relevant responses. However, guided by the anchors that are cohesive to other references, the generated responses are more informative. We attribute it to the fact that REGenerator utilizes adequate information from multiple retrieved results to better represent the guidance hidden in the anchor. Thus, employing the anchors both relevant to the query and cohesive with other references can leverage more information to carry forward the conversation as well as maintain the relevance.

APPENDIX B
CASE STUDIES

In this section, we illustrate the cases corresponding to the experimental results discussed in Section IV-A, Section IV-B and Section IV-C.

A. Cases of Various Models

Three groups of cases are shown in Figure 3. Taking cases in the first group for example, obviously, with the help of retrieved results, Multi-S2S generates more informative and relevant results than S2S and CVAE. By contrast, Edit-vec and S-Match fail to produce better results, given the additional information. Different from the case discussed in Section IV-D, we attribute it to the fact that even though the anchor provides valid information "Admission Ticket Number", it is conversationally irrelevant to the query. Edit-vec and S-Match pay over-attention to "know ... want to" and "Admission Ticket Number ... forgot ... it" respectively, so that their responses deviate from the query. Overall, our REGen produces a relevant response and maintains the additionally provided topic.

B. Influence of the Anchor

Corresponding to the objective analysis in Section IV-C, Figure 4 and Figure 5 show cases when the anchor is irrelevant to the query and complex respectively. These settings indicate that the anchor is relatively difficult to use. The results of the subjective comparison are consistent with the ones given in Section IV-C.

C. Ablation Study

The effectiveness of each component in the Retrieval-enhanced generator is discussed in Section IV-B. In this subsection, we utilize cases with different β scores from REGenerator to demonstrate how dynamic β helps the model deal with various situations. Figure 6 shows the generated responses from different methods with the queries obtained high decay scores. Obviously, given the design of dynamic β, a high β indicates a limited relation between anchor and other retrieved responses. Therefore anchors in Figure 6 are usually inappropriate responses. Given that, those methods which explicitly utilize the anchor usually struggle to give desirable replies. By contrast, the influence of these retrieved responses is suppressed by large in REGenerator, thus our model generates proper results which are almost irrelevant with those noise-like anchor and retrieved responses.

Oppositely, Figure 7 illustrates the results that the model gives low decay scores to the queries. Correspondingly, these anchors can be considered as high-quality responses, and all models generate informative and somewhat relevant results. However, due to retrieved responses from different perspectives, Multi-S2S faces trouble in information choosing, reflected by its result in the second case. Compared with the results from Edit-Vec and S-Match, our REGenerator picks the first three retrieved responses according to the anchor and reorganizes the information. Based on such adequate information, it generally produce better replies.

REFERENCES

[1] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., pp. 2042–2050.

[2] Y. Wu, W. Wu, C. Xing, C. Xu, Z. Li, and M. Zhou, "A sequential matching framework for multi-turn response selection in retrieval-based chatbots," *Computational Linguistics*, vol. 45, no. 1, pp. 163–197, Mar. 2019.

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014, pp. 3104–3112.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015 timestamp = Wed, 17 Jul 2019 10:40:54 +0200, biburl = https://dblp.org/rec/journals/corr/BahdanauCB14.bib, bibsource = dblp computer science bibliography, https://dblp.org.

[5] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 110–119.

[6] J. Feng, C. Tao, W. Wu, Y. Feng, D. Zhao, and R. Yan, "Learning a matching model with co-teaching for multi-turn response selection in retrieval-based dialogue systems," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3805–3815.

[7] J. Lu, X. Ren, Y. Ren, A. Liu, and Z. Xu, *Improving Contextual Language Models for Response Retrieval in Multi-Turn Conversation*. Association for Computing Machinery, 2020, p. 1805–1808.

[8] Y. Song, C.-T. Li, J.-Y. Nie, M. Zhang, D. Zhao, and R. Yan, "An ensemble of retrieval-based and generation-based human-computer conversation systems," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 4382–4388.

[9] Q. Zhu, L. Cui, W.-N. Zhang, F. Wei, and T. Liu, "Retrieval-enhanced adversarial training for neural response generation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3763–3773.

[10] Y. Zhu, Z. Dou, J.-Y. Nie, and J.-R. Wen, "Reboost: a retrieval-boosted sequence-to-sequence model for neural response generation," *Information Retrieval Journal*, vol. 23, no. 1, pp. 27–48, 2020.

[11] N. Majumder, D. Ghosal, D. Hazarika, A. Gelbukh, R. Mihalcea, and S. Poria, "Exemplars-guided empathetic response generation controlled by the elements of human communication," *IEEE Access*, vol. 10, pp. 77 176–77 190, 2022.

[12] D. Cai, Y. Wang, W. Bi, Z. Tu, X. Liu, W. Lam, and S. Shi, "Skeleton-to-response: Dialogue generation guided by retrieval memory," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1219–1228.

Query	大家晒一晒四六级成绩啊	Let's show our scores of CET-4/6.
Retrieval	上班这么久,英语都快忘完了	I have graduated for such a long time, almost forget my English.
S2S	裸考的飘过~	I took the exam without review~
CVAE	我也准备考试了	I'm also preparing for the exam
Multi_S2S	我六级 424.	My score of CET-6 is 424 .
<Anchor>	可不可以知道你的准考证号,我的忘了我想推一下	May I know your Admission Ticket Number? I forgot my own and I want to speculate it.
Edit-vec	我也想知道,你是什么	I also want to know what you are.
S-Match	我也是准考证号的,忘了	I am also Admission Ticket Number, I forgot it.
REGen	只有准考证号,我还没考过	I only have the Admission Ticket Number but have not taken the exam yet.
Query	喜欢是乍见之欢,爱是久处不厌	Like is a glance of joy, love is an everlasting companionship
Retrieval	嗯，吃完饭主动去洗碗的女生也很有唉	Actually, many girls will do the dishes after eating off their own bat.
S2S	不喜欢,不喜欢就不喜欢	Don't like, don't like, absolutely don't like.
CVAE	是的,真的是爱的人	Yeah, this person is really the one you love.
Multi_S2S	喜欢的人是不一样的~	The people you like are different~
<Anchor>	说的简单。但是有些事是不能用文字表达的	It's simple to say, but many things are beyond words.
Edit-vec	喜欢,但是不能说的	Like it, but I can't speak it out.
S-Match	说久了,就会有文字了	Words appeared after a long time saying.
REGen	喜欢这句话,爱是深深的喜欢	I like this sentence, "love is deep like".
Query	说说分手的理由吧	Let's talk about your reasons for break up.
Retrieval	一起失恋~	Let's fall out of love together ~
S2S	我也想分手	I also want to break up.
CVAE	我也是这样的	So am I.
Multi_S2S	不可能的	Impossible.
<Anchor>	分手需要理由吗?	Is it necessary to have a reason for break up?
Edit-Vec	分手需要理由吗	Is it necessary to have a reason for break up.
S-Match	有理由吗	Is there any reason.
REGen	分手的理由都是借口	All the reasons for break up are excuses.

Fig. 3: Comparisons of cases generated by baselines and REGenerator.

Query	关于绝命毒师	About Breaking Bad
<Anchor>	绝命方是八宅风水学中一个凶位	Breaking-direction is a bad position in Fengshui theory.
Edit-Vec	我喜欢一个人中~	I like middle myself ~
S-Match	看了中一季	Watched half first season
REGen	我觉得第一季的结局还是不错的	I think the ending of the first season is great.
Query	你喜欢直发女孩还是卷发女孩	Do you like girls with straight hair or curly hair.
<Anchor>	大粗腿，其他很好	Thick legs, the others are good.
Edit-Vec	大粗腿很好	The thick leg is good.
S-Match	我喜欢直发的好	I like straight hair’s good
REGen	自然卷的最好看	Naturally curly hair looks best

Fig. 4: Comparisons for the case when the anchor is irrelevant to the query.

Query	生日送什么	What gift for the birthday?
<Anchor>	送她你亲手绣的十字绣,时间短你就绣个简单的,不要觉得男生绣十字绣很丢人,真的会让她感动	Send her your hand-made cross stitch. If there's no enough time, do a simple stitch. Don't feel embarrassing for boys to do cross stitch, it's really touching.
Edit-Vec	你是个很好的女生	You're a nice girl.
S-Match	送你你亲手的礼物吧,如果是时间的话,那就绣个的,不要给他绣十字绣,感动	Send you your hand-made gift. If is time, stitch a one. Don't do cross stitch for him， touching.
REGen	送你自己的心意,最好的礼物	Send a gift of your own gratitude, this is the best gift.
Query	狗狗不吃饭怎么办	What if puppy doesn't eat?
<Anchor>	这种状况有几天了?除了不吃饭以外,精神状态,排便的形状怎样?狗多大了?是否打过疫苗?还有前几天吃过什么?	How many days has it been? Besides not eating, what about its mental state and the shape of defecation? How old is the puppy? Have it ever been vaccinated? What did it eat previous day?
Edit-Vec	这种状况不吃饭,除了吃多了	This kind of sympton not eating, except eat too much
S-Match	这种狗吃饭以外的精神状态和狗,打了疫苗就好了	Besides mental state of this dog eating and puppy, will be fine after been vaccinated
REGen	去宠物医院看看吧,里面的狗狗都是这样的	Let's go to pet hospital, many puppies there have sympton like this.

Fig. 5: Comparisons for the case when the anchor is complex.

Query	下雨天适合做什么 $\beta = 0.90$	What matches with a rainy day?
<Anchor>	没有可想的人	There's no one to miss
Retrieved Responses	感觉不适合	It doesn't feel right
	鋤禾!	Hoe!
	下雨最适合我陪你	It's best for me to stay with you on a rainy day.
	来来来	Come on
	有些情绪要说给懂的人听	Some emotions should be told to someone who can understand
	天天都喜欢	Like it every day.
	雨天就想趴在窗台上看下雨	I want to lie on the windowsill and watch the raindrop.
Multi-S2S	适合打伞	Suitable for umbrella
Edit-Vec	没有人做的	No one does it
S-Match	下雨	Rainy
REGen	下雨天,适合睡觉	Rainy day matches with sleeping
Query	你对相亲怎么看? $\beta = 0.91$	What do you think of blind dates ?
<Anchor>	你怎么不和他相亲	Why don't you go on a blind date with him.
Retrieved Responses	他家邻居介绍的	His neighbor introduced him
	那个年代...相亲应该是比较普遍的方式吧	At that time... Blind date should be common
	没怎么，起码家长的关过了	Nothing special, at least passed the parents' test.
	再相个	Blind date another one
	想知道你跟男友怎么认识的	I want to know how you met your boyfriend
	看不上我的人太多了…	Many people just don't give me a thought...
	就是家里人介绍的丫。这是她相亲的第二个对象。目测身高 180+虽然没有她前任帅但是对她超级好,各种浪漫。而且很有钱	Her family introduced him.This was her second blind date. The guy seems over 180. Although not as handsome as her ex, he is super nice to her and romantic. Rich as well btw.
Multi-S2S	我妈说,我也想相亲了	My mom said, I want to go on a blind date too
Edit-Vec	不知道怎么和你看	Don't know how to think of you
S-Match	同相亲	Same blind date
REGen	相亲的对象不靠谱	The partner of blind date is not reliable

Fig. 6: Comparisons for the case when REGenerator give higher β scores.

Query	你一生受制于什么星座 $\beta = 0.28$	Which constellation are you controlled by for life
<Anchor>	天秤座:一生受制于巨蟹座	Libra: controlled by Cancer for life
Retrieved Responses	0.32 金牛座:一生受制于天秤座	0.32 Taurus: controlled by Libra for life
	0.30 天蝎座:一生受制于狮子座	0.30 Scorpio: controlled by Leo for life.
	0.26 白羊座:一生受制于天蝎座	0.26 Aries: controlled by Scorpio for life.
	0.10 水瓶座:一生受制于巨蟹座。巨蟹座能把水瓶座的人完全玩弄于鼓掌之中	0.10 Aquarius: controlled by Cancer for life. Cancer can twist Aquarius around his/her little finger.
	0.02 白羊座:一生受制于天蝎座。只有天蝎座的人才能使白羊座的人爱得死心塌地。	0.02 Aries: controlled by Scorpio for life.Only Scorpio can receive Aries's desperate love.
	0.00 我一生会受制于金牛的	0.00 I am controlled by Taurus for life.
Multi-S2S	天秤座:一生受制于天蝎座	Libra: controlled by Scorpio for life
Edit-Vec	我双鱼一生受制于	I'm a pisces is controlled by.
S-Match	水瓶一生受制于天蝎	Aquarius is controlled by Scorpio for life.
REGen	我天秤,一生受制于巨蟹座	I'm a libra, controlled by cancer for life.
Query	荷兰 vs 西班牙 $\beta = 0.19$	Holland VS Spain
<Anchor>	西班牙 2:1 荷兰	Spain 2: 1 Holland
Retrieved Responses	0.96 西班牙 1:0 荷兰	0.96 Spain 1: 0 Holland
	0.03 荷兰 2:1 胜西班牙	0.03 Holland 2: 1 win Spain
	0.01 1:0 西班牙胜	0.01 1：0 Spain win
	0.00 进球的那个	0.00 The one who scores
	0.00 西班牙	0.00 Spain
	0.00 西班牙没攻击力啊,有几个球是前锋进的啊	0.00 Spain has no attacking power. How many goals are scored by strikers
	0.00 荷兰	0.00 Holland
	0.00 乌拉圭很可能把荷兰灭了乌拉圭可不弱啊	0.00 Uruguay is likely to beat Holland. Uruguay isn't weak
Multi-S2S	西班牙 2:1:1:4:1:4:1:1:4:1:1:4:1:1:1:4	Spain 2 : 1 : 1 : 4 : 1 : 4 : 1 : 1 : 4 : 1 : 1 : 4 : 1 : 1 : 1 : 4
Edit-Vec	2:1 西班牙	2：1 Spain
S-Match	西班牙,荷兰	Spain, Holland
REGen	西班牙必胜!	Spain is invincible !

Fig. 7: Comparisons for the case when REGenerator give lower β scores. The scores shown in front of retrieved responses are α scores.

[13] Y. Song, R. Yan, X. Li, D. Zhao, and M. Zhang, "Two are better than one: An ensemble of retrieval- and generation-based dialog systems," *CoRR*, vol. abs/1610.07149, 2016 archivePrefix = arXiv, eprint = 1610.07149, timestamp = Mon, 15 Jun 2020 12:39:01 +0200, biburl = https://dblp.org/rec/journals/corr/SongYLZZ16.bib, bibsource = dblp computer science bibliography, https://dblp.org.

[14] J. Weston, E. Dinan, and A. Miller, "Retrieve and refine: Improved sequence generation models for dialogue," in *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 87–92.

[15] G. Pandey, D. Contractor, V. Kumar, and S. Joshi, "Exemplar encoder-decoder for neural conversation generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1329–1338.

[16] D. Cai, Y. Wang, W. Bi, Z. Tu, X. Liu, and S. Shi, "Retrieval-guided dialogue response generation via a matching-to-generation framework," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1866–1875.

[17] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 654–664.

[18] I. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014 timestamp = Thu, 04 Apr 2019 13:20:07 +0200, biburl = https://dblp.org/rec/journals/corr/KingmaW13.bib, bibsource = dblp computer science bibliography, https://dblp.org.

[21] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[22] Y. Wu, F. Wei, S. Huang, Y. Wang, Z. Li, and M. Zhou, "Response generation by context-aware prototype editing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 7281–7288, Jul. 2019.

[23] R. Lowe, N. Pow, I. V. Serban, and J. Pineau, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," in *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, p. 285.

[24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners."

[25] C. Li, X. Gao, Y. Li, B. Peng, X. Li, Y. Zhang, and J. Gao, "Optimus: Organizing sentences via pre-trained modeling of a latent space," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4678–4699.

[26] L. Fang, T. Zeng, C. Liu, L. Bo, W. Dong, and C. Chen, "Transformer-based conditional variational autoencoder for controllable story generation," *arXiv preprint arXiv:2101.00828*, 2021.

[27] ChatGPT, "Chatgpt: Optimizing language models for dialogue," OpenAI, 2023, accessed: 4.1.2023. [Online]. Available: https://openai.com/blog/chatgpt/

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015 timestamp = Thu, 25 Jul 2019 14:25:37 +0200, biburl = https://dblp.org/rec/journals/corr/KingmaB14.bib, bibsource = dblp computer science bibliography, https://dblp.org.

[29] Z. Xu, N. Jiang, B. Liu, W. Rong, B. Wu, B. Wang, Z. Wang, and X. Wang, "Lsdscc: a large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2070–2080.

[30] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan, "Generating informative and diverse conversational responses via adversarial information maximization," *Advances in Neural Information Processing Systems*, vol. 31, pp. 1810–1820, 2018.

[31] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2122–2132.

[32] B. Wu, N. Jiang, Z. Gao, S. Li, W. Rong, and B. Wang, "Why do neural response generation models prefer universal replies?" *CoRR*, vol. abs/1808.09187, 2018 archivePrefix = arXiv, eprint = 1808.09187, timestamp = Mon, 18 Jan 2021 13:55:35 +0100, biburl = https://dblp.org/rec/journals/corr/abs-1808-09187.bib, bibsource = dblp computer science bibliography, https://dblp.org.

[33] X. Shen, H. Su, Y. Li, W. Li, S. Niu, Y. Zhao, A. Aizawa, and G. Long, "A conditional variational framework for dialog generation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 504–509.

[34] X. Shen, H. Su, S. Niu, and V. Demberg, "Improving variational encoder-decoders in dialogue generation," in *AAAI*, 2018, pp. 5456–5463.

[35] J. Gao, W. Bi, X. Liu, J. Li, G. Zhou, and S. Shi, "A discrete CVAE for response generation on short-text conversation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1898–1908.

[36] Y. Cao, W. Bi, M. Fang, and D. Tao, "Pretrained language models for dialogue generation with multiple input sources," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 909–917.

[37] J. Li, T. Tang, W. X. Zhao, and J. rong Wen, "Pretrained language models for text generation: A survey," in *IJCAI*, 2021.

[38] Y. Zhao, W. Wu, and C. Xu, "Are pre-trained language models knowledgeable to ground open domain dialogues?" *arXiv preprint arXiv:2011.09708*, 2020.

[39] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, "Knowledge-grounded dialogue generation with pre-trained language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 3377–3390.

[40] J. Bai, Z. Yang, X. Liang, W. Wang, and Z. Li, "Learning to copy coherent knowledge for response generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, 2021, pp. 12 535–12 543.

[41] Y. Xu, E. Ishii, S. Cahyawijaya, Z. Liu, G. I. Winata, A. Madotto, D. Su, and P. Fung, "Retrieval-free knowledge-grounded dialogue response generation with adapters," in *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, 2022, pp. 93–107.

[42] Z. Tian, W. Bi, X. Li, and N. L. Zhang, "Learning to abstract for memory-augmented conversational response generation," in *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 2019, pp. 3816–3825.

[43] Y. Ling, F. Cai, X. Hu, J. Liu, W. Chen, and H. Chen, "Context-controlled topic-aware neural response generation for open-domain dialog systems," *Information Processing & Management*, vol. 58, no. 1, p. 102392, 2021.

[44] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, and M. Jiang, "A survey of knowledge-enhanced text generation," *ACM Computing Surveys (CSUR)*.

[45] S. Moon, P. Shah, A. Kumar, and R. Subba, "Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 845–854.

[46] H. Zhang, Z. Liu, C. Xiong, and Z. Liu, "Grounded conversation generation as guided traverses in commonsense knowledge graphs," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2031–2043.

[47] L. Yang, J. Hu, M. Qiu, C. Qu, J. Gao, W. B. Croft, X. Liu, Y. Shen, and J. Liu, "A hybrid retrieval-generation neural conversation model," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1341–1350.

[48] M. Komeili, K. Shuster, and J. Weston, "Internet-augmented dialogue generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8460–8478.

[49] L. Zhang, Y. Yang, J. Zhou, C. Chen, and L. He, "Retrieval-polished response generation for chatbot," *IEEE Access*, vol. 8, pp. 123 882–123 890, 2020.

[50] P. Gupta, J. P. Bigham, Y. Tsvetkov, and A. Pavel, "Controlling dialogue generation with semantic exemplars," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 3018–3029.

[51] Y. Su, Y. Wang, D. Cai, S. Baker, A. Korhonen, and N. Collier, "Prototype-to-style: Dialogue generation with style-aware editing on retrieval memory," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2152–2161, 2021.

[52] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3784–3803.

[53] P. Ren, Z. Chen, Z. Ren, E. Kanoulas, C. Monz, and M. De Rijke, "Conversations with search engines: Serp-based conversational response generation," *ACM Transactions on Information Systems (TOIS)*, vol. 39, no. 4, pp. 1–29, 2021.

[54] Y. Zhang, S. Sun, X. Gao, Y. Fang, C. Brockett, M. Galley, J. Gao, and B. Dolan, "Retgen: A joint framework for retrieval and grounded text generation modeling," 2022.