Dynamic Stochastic Decoding Strategy for Open-Domain Dialogue Generation

Yiwei Li[1], **Fei Mi**[2], **Yitong Li**[2,3], **Yasheng Wang**[2], **Bin Sun**[1],**Shaoxiong Feng**[1], **Kan Li**[1*]
[1]School of Computer Science & Technology, Beijing Institute of Technology
[2] Huawei Noah's Ark Lab [3]Huawei Technologies Ltd.
{liyiwei,binsun,shaoxiongfeng,likan}@bit.edu.cn
{mifei2,liyitong3,wangyasheng}@huawei.com

Abstract

Stochastic sampling strategies such as top-k and top-p have been widely used in dialogue generation task. However, as an open-domain chatting system, there will be two different conversation scenarios, i.e. chit-chat and knowledge-based question answering. In the former situation, responses diversity is essential due to the one-to-many nature in dialogue. The latter, on the other hand, requires less randomness given that stochastic decoding strategy entails the risk of generating incorrect information. As a result, an adaptive and flexible decoding strategy is needed to cope with these two scenarios simultaneously. To this end, we propose the **d**ynamic **d**ecoding **s**trategy (**DDS**), which can adjust the decoding space w.r.t. different contexts. In DDS, both sequence-level and token-level adaptive search can be achieved to adjust the decoding process in a unified framework. Besides, our adaptive algorithm can not only be used during model inference, but it can also be applied during the model training stage to further enhance the performance. Comprehensive experiments indicate that the proposed decoding strategy can consistently improve the performance of pre-trained dialogue models when coupled with four well-used stochastic decoding algorithms.

1 Introduction

Building generative open-domain dialogue system is a significant yet challenging area of deep learning research. It has been widely recognized that the pre-training paradigm, in which large-scale transformer-based models are trained with massive amounts of conversational data, is an effective and promising approach. Some of the more notable works in English include DialoGPT (Zhang et al., 2020b), LaMDA (Thoppilan et al., 2022), Blender (Roller et al., 2021; Shuster et al., 2022), and lately, ChatGPT has attracted great attention and interest from researchers and the industry. For chinese dialogue models, EVA (Zhou et al., 2021; Gu et al., 2022), PanGu-Bot (Mi et al., 2022) and PLATO (Bao et al., 2020, 2021, 2022) are also excellent options. In recent research, however, it has been demonstrated that decoding strategies play an important role in performance even beyond model architecture (Meister et al., 2022b), whereas standard strategies remain relatively unchanged (Suzgun et al., 2022).

Chit-chat
c: 我好喜欢猫猫(I love cats.)
r_1: 我好喜欢我的猫猫啊 (I love my cat so much)
r_2: 我也喜欢猫猫！ (I like cats, too!)
r_3: 我也喜欢猫猫 (I like cats, too)
r_4: 我也喜欢猫猫！ (I like cats, too!)
r_5: 我也好喜欢 (I like it too)
Factual Question Answering
c: 土木工程属于理科还是工科 (Does civil engineering belong to science or engineering)
r_1: 工科,土木是理科 (Engineering, civil engineering is science)
r_2: 土木是工科,土木是建筑 (Civil engineering is engineering, civil engineering is architecture.)
r_3: 工科 (Engineering)
r_4: 工科 (Engineering)
r_5: 文科 (Liberal arts)

Table 1: Generated examples by EVA2.0 on both two scenarios, where top-k sampling is used with temperature set to 1. r_{1-5} refer to five generated responses for the same context c. Blue part of chit-chat reflects the high similarity of responses, whilst red part reveals the inappropriate answers in factual QA scenario.

Stochastic decoding algorithms are widely used for dialogue generation task. Users expect varying responses from a chatbot when they input similar queries, or they tend to become bored and lose interest if it only responds with fixed reply. For such a chit-chat scenario, deterministic decoding algorithms, such as greedy search or beam search, are not suitable. Additionally, even when using large pre-trained language models, decoding strategies that aim for high probability output, suffer from in-

[*]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 11585–11596
August 11-16, 2024 ©2024 Association for Computational Linguistics

credible degeneration issue (Holtzman et al., 2020; Welleck et al., 2020). Consequently, dialogue generation models are inclined to employ stochastic sampling methods such as top-k sampling (Fan et al., 2018) or nucleus sampling (Holtzman et al., 2020), where the probability distribution will be shaped by the temperature T.

Aside from chit-chat, however, there is another scenario for chatbots, namely factual question answering (QA). Unfortunately, since the size of the decoding space required for two different dialog scenarios is different, stochastic sampling methods are not able to handle both simultaneously due to the unified and constant randomness of their decoding processes. As shown in Table 1, with the same temperature, the chit-chat sample has a narrow range of generation, where from r_1 to r_5 are the same *I like cats too*-like responses. Whereas, candidates response to the factual question are too diverse, leading to answers are factually incorrect (r_1 and r_5), with low fluency (r_2) or self-contradictory (r_1). As a result, the determined sampling randomness will reduce the diversity under chit-chat condition while enlarge it for question answering, which will increase the risk of generating dull responses and wrong answers. In addition, even under the same scenario, different contexts will have varying degrees of decoding flexibility (Csáky et al., 2019). For example, *What animals do you like?* has larger response space than *Do you love cats?*. Furthermore, different tokens has different ranges of decoding space within the same utterance (Holtzman et al., 2020).

To resolve the drawbacks of existing stochastic decoding algorithms, we propose a dynamic decoding strategy (DDS) for dialogue generation, which can be combined with mainstream stochastic sampling. The key intuition of dynamic sampling is that the decoding space varies according to the context, therefore the shape of probability distribution should be adjusted adaptively. To achieve this goal, we incorporate an additional diversity predicting head into the dialogue generation model, which is capable of producing the score based on decoding diversity to guide the sampling process adaptively. It only introduces a few parameters and performs decoding at a similar speed to standard dialogue models. The labeled data for training the head is derived from the pre-trained model automatically. Three types of mapping functions are designed, projecting the diversity score to the temperature for shaping the sampling distribution. In order to control the token generation in a more fine-grained manner, the regression head can be applied to each output token or the whole context, allowing us to control the randomness of decoding at both levels. Apart from inference, adaptive temperature can also be introduced to dialogue training stage to balance the model prediction confidence.

We perform extensive experiments on two union of datasets with two Chinese pre-trained dialogue models. The results show that the DDS can largely improve the performance of four sampling-based decoding algorithms. Human evaluation is also conducted to ensure relevance and fluency of responses while improving diversity.

In summary, our contributions are as follows:

- We propose a novel dynamic decoding mechanism for dialogue generation, which can easily be integrated into stochastic decoding strategies and handle different conversational scenarios simultaneously.
- The mechanism can be conducted on both sentence level and token level with three mapping functions, and adaptive temperature training is introduced except for the inference stage.
- Extensive evaluations show that the proposed decoding strategy can largely improve the performance of dialogue models with strong generalization ability when coupled with widely used stochastic decoding strategies.

2 Background

2.1 Dialogue Generation

In this work, we work with the task of dialogue generation in open-domain, where the input context $\boldsymbol{c} = \{c_1, c_2, ...\}$ can be either a chat conversation or a factual question and response $\boldsymbol{r} = \{r_1, r_2, ...\}$ is produced accordingly. Dialogue generation models, which are normally pre-trained on massive conversational corpora nowadays, directly models the response probability $p_\theta(\boldsymbol{r} \mid \boldsymbol{c})$, where θ indicates the model parameters. Standard MLE training is used to minimize the negative log-likelihood (NLL) of the training data:

$$\begin{aligned}\mathcal{L}_{\mathrm{NLL}}\left(P_{\text{data}} ; \theta\right) &= E_{(\boldsymbol{c},\boldsymbol{r})\sim P_{\text{data}}}\left(-\log P_\theta(\boldsymbol{r} \mid \boldsymbol{c})\right) \\ &= E_{(\boldsymbol{c},\boldsymbol{r})\sim P_{\text{data}}}(-\sum_{t=1}^{T} \log P_\theta\left(r_t \mid \boldsymbol{r}_{<t}, \boldsymbol{c}\right)),\end{aligned} \tag{1}$$

where T is the length of the response $\boldsymbol{r}$, and the token probability distribution P_θ is typically modeled

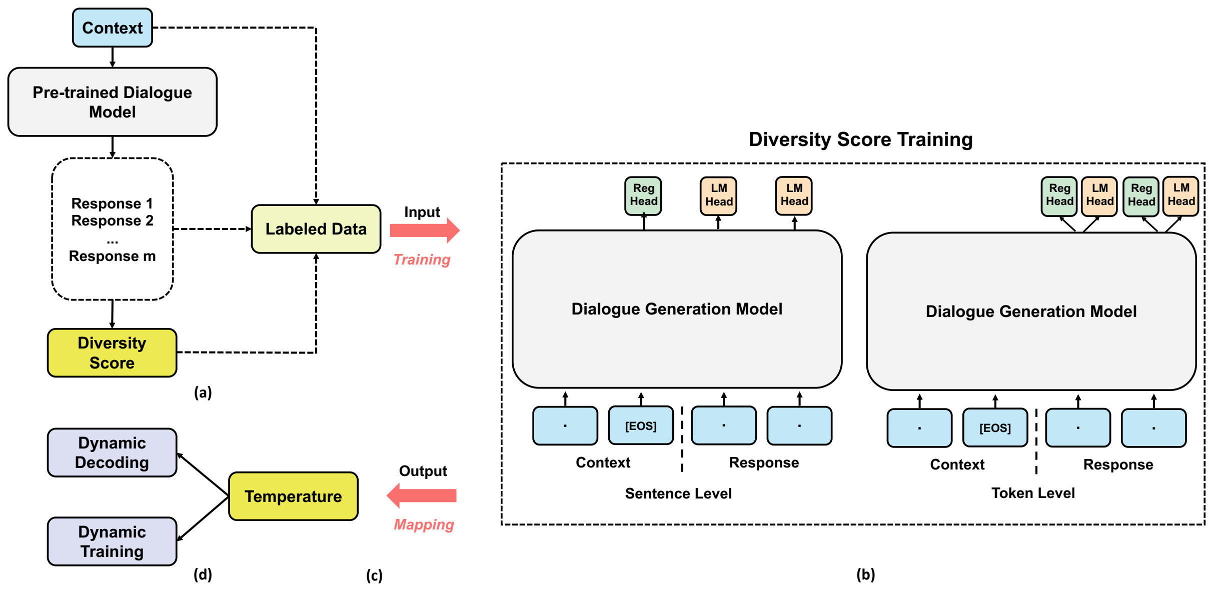

Figure 1: An overview of the process of DDS: (a) Calculating the diversity score. (b) Training the regression head. (c) Mapping score to temperature. (d) Dynamic decoding and training.

as softmax-normalized logits from decoder output z_t by:

$$P_\theta\left(r_t \mid \boldsymbol{r}_{<t}, \boldsymbol{c}\right)=\operatorname{softmax}\left(z_t\right) \tag{2}$$

Decoding process is the search for a response token string $\boldsymbol{r}^*$ according to the given dialogue model θ and context $\boldsymbol{c}$. Most current generative methods employ one of a few standard decoding strategies, which may be characterized as either deterministic or stochastic in nature.

2.2 Stochastic Decoding Algorithms

Deterministic decoding algorithms like greedy search or beam search, choose the most probable token or path at each step, generating fixed responses through the following form:

$$\boldsymbol{r}^{\star}=\underset{\boldsymbol{r}}{\operatorname{argmax}}\, p_\theta(\boldsymbol{r} \mid \boldsymbol{c}) \tag{3}$$

Different from that, stochastic algorithms will generate various responses given the same context by sampling $\boldsymbol{r} \sim p_\theta(\cdot \mid \boldsymbol{c})$. Based on this, four sampling approaches are briefly presented below.

Temperature Sampling. It is a stochastic sampling method in which the next token is chosen at random based on the new biased probability distribution $p_\theta^{'}$ shaped by the **temperature** T (Ackley et al., 1985):

$$p_\theta^{'}\left(r_t | \boldsymbol{r}_{<t}, \boldsymbol{c}\right)=\frac{\exp \left(p_\theta\left(r_t | \boldsymbol{r}_{<t}, \boldsymbol{c}\right) / T\right)}{\sum_r \exp \left(p_\theta\left(r | \boldsymbol{r}_{<t}, \boldsymbol{c}\right) / T\right)} \tag{4}$$

Top-k Sampling Based on temperature sampling, it truncates the probability distribution produced by the model by limiting the sampling space to the tokens with top k highest possibilities before sampling (Fan et al., 2018).

Top-p Sampling. Instead of considering a fixed number of tokens in each decoding step, nucleus (top-p) sampling dynamically selects the smallest set of tokens where the sum of their probabilities is more than the threshold p (Holtzman et al., 2020).

Locally Typical Sampling. It truncates the probability distribution by local informativeness to generate more human-like text (Meister et al., 2022a).

3 Methodology

We propose the dynamic decoding strategy to dynamically compute temperature $T^{'}$ w.r.t. different contexts, which replaces T in Equation 4 for all four sampling methods outlined above. The value of this parameter $T^{'}$ will vary adaptively according to the size of the decoding space. In this section, we first describe how to build the labeled data about dialogue decoding diversity automatically. After that, we elaborate the regression head trained by it for predicting diversity scores on two levels, which will then be projected to temperature $T^{'}$ in accordance with three different mapping strategies. Besides, the dynamic $T^{'}$ can also be applied to training stage. The overview of the proposed framework is illustrated in Figure 1.

3.1 Diversity Score Calculation

Labeled data is needed to train the regression head to predict the temperature. $\mathcal{D} = \{(c_i, r_i)\}_{i=1}^n$ denotes a training set consisting of n dialogues. In order to quantify the range of decoding space available for a given context c_i, we seek to determine its diversity score s_i. To achieve this, instead of expensive human annotations, we construct the labeled data automatically. We are motivated by the strong generation capability of pre-trained dialogue models, which has been trained by a large amount of conversational data from various domains. For each $c_i \in \mathcal{D}$, the dialogue model generates m candidates $\{\hat{r}_i\}_m$ based on it, after which the similarity degree between them will be determined. BERTScore (Zhang et al., 2020a) is a popular learned evaluation metric for doing this. It compares sentences using contextual embeddings from a pre-trained BERT model, computing a similarity score based on the cosine similarity between the sentence embeddings. We trained the Chinese BERT model on wiki2019zh* dataset using the framework from SimCSE (Gao et al., 2021) to calculate the score. The average BERTScore of each $\{\hat{r}_i\}_m$ can reflect the diversity of them, deemed as the range of generation space for the context c_i. The higher the score, the narrower the range. Consequently, the labeled dataset $\mathcal{D}' = \{(c_i, \{\hat{r}_i\}_m, s_i,)\}_{i=1}^n$ is constructed.

3.2 Diversity Score Training

For training and predicting the diversity score efficiently, we design the regression head based on the dialogue generation model, which maps token representation into a one dimensional vector using two feed-forward networks with non-linearity between them:

$$score = tanh(W_1^T x + b_1)W_2^T + b_2 \tag{5}$$

Then, the predicted score $\hat{s}$ will be fitted to label s_i through MSE loss:

$$\mathcal{L}_{\mathrm{MSE}}(P'_{\text{data}}; \theta) = E_{(\boldsymbol{c},\boldsymbol{s})\sim P'_{\text{data}}}\left((\boldsymbol{s} - \hat{\boldsymbol{s}})^2\right) \tag{6}$$

As shown in Figure 1, the regression head can be employed on two levels:

Sentence-level On this condition, the diversity score comes from the head of EOS token (denotes the end of a sentence) of context. Therefore, only c_i and s_i are needed from $\mathcal{D}'$ for training the head.

*https://github.com/brightmart/nlp_chinese_corpus

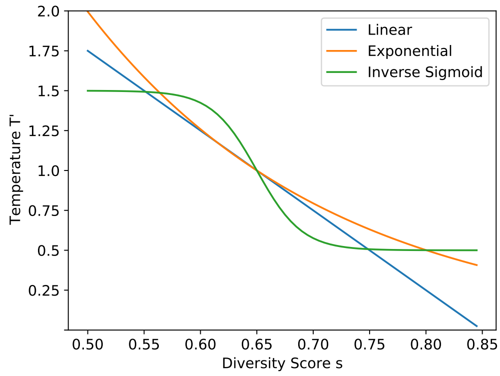

Figure 2: Different mapping strategies to project the diversity score to temperature.

Token-level For token-level situation, the hidden state of each generated token will provide the diversity score through the regression head. Thus, the head will be trained by each $\hat{r}_i \in \{\hat{r}_i\}_m$ with the same label s_i.

There are two ways to train the regression head: either individually with other parameters fixed, or jointly with the standard dialogue generation task. In addition, due to some unexpected samples in $\mathcal{D}'$ (please refer to Table 1 and Figure 3), the data filtering process will be conducted before training. Afterwards, the predicted diversity score may be more accurate than the one directly derived from the pre-trained model.

3.3 Temperature Mapping Strategies

After obtaining the diversity score s_i, we further convert it to guide the dynamic temperature T' for Equation 4. As s_i increases, T' should decrease to sharpen the probability distribution of sampling and vice versa. Consequently, three mapping strategies are designed:

- Linear Mapping

$$T(s) = hs + t_0, \tag{7}$$

where k is the slope.

- Exponential Mapping

$$T(s) = h^s + t_0, \tag{8}$$

where $h < 1$ is the radix to adjust the sharpness of mapping function.

- Inverse Sigmoid Mapping

$$T(s) = \frac{h}{h + e^{\frac{s}{h}}} + t_0, \tag{9}$$

where e is the mathematical constant, and $h \leq 1$ is a hyperparameter to adjust the sharpness. All t_0 is

Datasets	Decoding Strategy	BLEU-1	BLEU-2	BLEU-3	BLEU-4	F1	ROUGE-1	ROUGE-2	ROUGE-L
LQA	Top-k (fixed T)	0.4327	0.2640	0.1616	0.0988	0.2149	0.2081	0.0412	0.1764
	Top-k (DDS)	**0.4410**	**0.2701**	**0.1659**	**0.1019**	**0.2187**	**0.2083**	**0.0452**	**0.1827**
	Top-p (fixed T)	0.4109	0.2490	0.1515	0.0924	0.1870	0.1882	0.0325	0.1491
	Top-p (DDS)	**0.4405**	**0.2698**	**0.1657**	**0.1017**	**0.2170**	**0.2069**	**0.0448**	**0.1802**
	Temperature (fixed T)	0.3891	0.2342	0.1416	0.0856	0.1679	0.1710	0.0254	0.1337
	Temperature (DDS)	**0.4357**	**0.2663**	**0.1633**	**0.1001**	**0.2128**	**0.2062**	**0.0427**	**0.1745**
	Typical (fixed T)	0.3971	0.2393	0.1447	0.0876	0.1770	0.1777	0.0263	0.1392
	Typical (DDS)	**0.4378**	**0.2682**	**0.1649**	**0.1014**	**0.2169**	**0.2073**	**0.0451**	**0.1791**
PersonQA	Top-k (fixed T)	0.5751	0.4618	0.3840	0.3258	0.4321	0.4234	0.3203	0.4284
	Top-k (DDS)	**0.6137**	**0.4989**	**0.4200**	**0.3609**	**0.4619**	**0.4524**	**0.3533**	**0.4590**
	Top-p (fixed T)	0.5400	0.4358	0.3647	0.3117	0.4044	0.3962	0.3041	0.4010
	Top-p (DDS)	**0.5979**	**0.4874**	**0.4114**	**0.3539**	**0.4488**	**0.4403**	**0.3461**	**0.4456**
	Temperature (fixed T)	0.5413	0.4365	0.3647	0.3112	0.4038	0.3958	0.3024	0.4008
	Temperature (DDS)	**0.5894**	**0.4811**	**0.4066**	**0.3506**	**0.4439**	**0.4346**	**0.3417**	**0.4407**
	Typical (fixed T)	0.5348	0.4317	0.3611	0.3082	0.3994	0.3916	0.3010	0.3962
	Typical (DDS)	**0.5963**	**0.4872**	**0.4121**	**0.3555**	**0.4495**	**0.4407**	**0.3477**	**0.4469**

Datasets	Decoding Strategy	Distinct-1	Distinct-2	Distinct-3	Ent-1	Ent-2	Ent-3	BERTScore
LCCC	Top-k (fixed T)	0.1015	0.3973	0.6659	10.0321	18.5411	19.6180	0.5764
	Top-k (DDS)	**0.1036**	**0.4119**	**0.6889**	**10.0755**	**18.6606**	**19.8775**	**0.5617**
	Top-p (fixed T)	0.1523	0.6170	0.9057	11.1319	18.9154	20.4290	0.4562
	Top-p (DDS)	**0.2101**	**0.7718**	**0.9428**	**12.5948**	**19.4330**	**21.4829**	**0.4332**
	Temperature (fixed T)	0.1818	0.6866	0.9418	11.6779	19.0928	20.7616	0.4424
	Temperature (DDS)	**0.2555**	**0.8685**	**0.9867**	**13.1907**	**19.5489**	**21.7447**	**0.4243**
	Typical (fixed T)	0.1519	0.6132	0.8929	11.1861	18.9442	20.4646	0.4578
	Typical (DDS)	**0.2331**	**0.8133**	**0.9646**	**12.6864**	**19.4573**	**21.4895**	**0.4321**
Diamante	Top-k (fixed T)	0.1124	0.4100	0.6502	**10.1575**	12.4041	15.6205	0.6532
	Top-k (DDS)	**0.1153**	**0.4175**	**0.6628**	10.1398	**12.4172**	**15.6745**	**0.6438**
	Top-p (fixed T)	0.1282	0.4582	0.7036	10.2857	12.6627	15.8346	0.6144
	Top-p (DDS)	**0.1791**	**0.5401**	**0.7811**	**10.4122**	**12.9131**	**16.0630**	**0.5822**
	Temperature (fixed T)	0.1408	0.5098	0.7744	10.3355	12.7581	15.9274	0.4591
	Temperature (DDS)	**0.2377**	**0.6362**	**0.8510**	**10.5948**	**13.2204**	**16.2846**	**0.4339**
	Typical (fixed T)	0.1267	0.4545	0.7038	10.3077	12.6324	15.7905	0.4627
	Typical (DDS)	**0.2601**	**0.6172**	**0.8237**	**10.4760**	**12.9582**	**16.0774**	**0.4234**

Table 2: Automatic evaluations results on PanGu-Bot. DDS has significantly improved the performance of all four well-known stochastic decoding algorithms on four datasets.

the offset to make $T(s)$ equals 1 when s reaches the mean value.

A visual representation of different mapping strategies is provided in Figure 2. In this way, a dynamic temperature $T^{'}$ can be constructed to guide the decoding process adaptively.

3.4 Dynamic Temperature in Training

In addition, same as the inference stage, the temperature $T^{'}$ can shape the probability distribution p_θ of decoder output z during training process by:

$$p_\theta^i = \frac{\exp(z_i/T^{'})}{\sum_j \exp(z_j/T^{'})}, \tag{10}$$

Thus, the dynamic temperature training can be conducted to balance the model prediction confidence of chit-chat and factual question answering scenarios respectively. Considering the one-to-many labels, the former is suitable for low confidence training, whereas the latter requires a higher degree of confidence due to the certainty of the knowledge.

4 Experiments

4.1 Dataset

For training, we use two datasets with different data size to verify the effectiveness of the proposed decoding strategy in two conversation scenarios, each of which contains a chit-chat and a QA dataset. The first is the union (U_S) of Diamante (Lu et al., 2022), a human-written chit-chat dialogue dataset, and PersonQA, a question answering data about persons.

Decoding Strategy	BLEU-1	BLEU-2	BLEU-3	BLEU-4	F1	ROUGE-1	ROUGE-2	ROUGE-L
Top-k (fixed T)	0.0823	0.0495	0.0299	0.0180	0.1139	0.0983	0.0113	0.0988
Top-k (DDS)	**0.0921**	**0.0557**	**0.0339**	**0.0206**	**0.1181**	**0.1010**	**0.0136**	**0.1043**
Top-p (fixed T)	0.0844	0.0509	0.0309	0.0187	0.1143	0.0990	0.0115	0.0984
Top-p (DDS)	**0.0927**	**0.0558**	**0.0337**	**0.0203**	**0.1172**	**0.1006**	**0.0130**	**0.1024**
Temperature (fixed T)	0.0762	0.0452	0.0271	0.0162	0.0656	0.0586	0.0028	0.0568
Temperature (DDS)	**0.0801**	**0.0482**	**0.0292**	**0.0177**	**0.1041**	**0.0896**	**0.0115**	**0.0918**
Typical (fixed T)	0.0554	0.0331	0.0200	0.0120	0.0853	0.0724	0.0049	0.0743
Typical (DDS)	**0.0923**	**0.0555**	**0.0336**	**0.0202**	**0.1106**	**0.0931**	**0.0116**	**0.0958**

Decoding Strategy	Distinct-1	Distinct-2	Distinct-3	Ent-1	Ent-2	Ent-3	BERTScore
Top-k (fixed T)	0.1616	0.4769	0.7140	9.8991	18.5029	19.3731	0.6435
Top-k (DDS)	**0.1639**	**0.4950**	**0.7510**	**9.9633**	**18.5990**	**19.5902**	**0.6320**
Top-p (fixed T)	**0.2055**	0.6806	0.9368	10.4591	18.6561	19.8080	0.4890
Top-p (DDS)	0.2041	**0.7127**	**0.9490**	**10.7369**	**18.9950**	**20.4575**	**0.4645**
Temperature (fixed T)	0.3281	0.8505	0.9841	11.9063	19.1125	20.7625	0.4213
Temperature (DDS)	**0.4408**	**0.9693**	**0.9991**	**14.3922**	**19.6696**	**22.0616**	**0.4078**
Typical (fixed T)	**0.1884**	0.6393	0.9152	10.2947	18.7910	20.0862	0.4657
Typical (DDS)	0.1708	**0.6423**	**0.9270**	**10.6164**	**19.3449**	**21.3205**	**0.4536**

Table 3: Zero-shot automatic evaluations results of LQA (Up) and LCCC (Down) on EVA2.0.

	Datasets	# Train	# Valid	# Test
U_S	PersonQA	4500	500	919
	Diamante	3000	500	916
U_L	LQA	115k	10k	10k
	LCCC	90k	10k	10k

Table 4: Data statistics of the experiment corpora.

Both of them are small but with high-quality. The second dataset (U_L) has much larger size, consisting of LCCC-base (Wang et al., 2020), and LQA, which includes longer explanations in responses. We calculate the diversity score of each dataset, and then mix the data within the same union. Figure 3 depicts the similarity scores of LCCC and LQA, showing that QA scenario scores are holistically larger than those of chit-chat. The overall trend is in line with expectations, while there are some noise samples with much higher scores in LCCC and lower ones in LQA. Table 1 shows the cases from those parts and it is what we need to solve through our method. Therefore, we filter these extreme data by dropping samples whose score is lower than 0.6 in QA dataset and higher than 0.7 in chit-chat dataset. Table 4 provides the statistics of both unions for training the regression head. Please see Appendix B for more details about QA dataset. For test, all the four sub-sets are evaluated separately. In this work, we mainly focus on Chinese datasets, but we also conduct additional test in Section 4.5 to verify the multilingual availability.

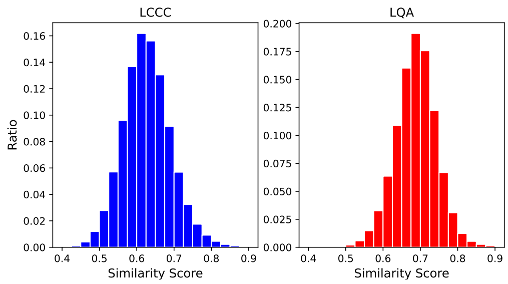

Figure 3: Similarity score distributions of LCCC (left) and LQA (right). The former is a chit-chat dataset and the latter is for QA scenario. The samples are generated by PanGu-Bot and the scores are calculated by BERTScore. Although overall scores of the chatting scene are lower, there are also some noise samples with much higher similarity scores for chitchat and lower scores for QA.

4.2 Training Settings

We take two Chinese pre-trained models: PanGu-Bot (Mi et al., 2022) containing 350M parameters and EVA2.0 (Gu et al., 2022) with 300M parameters as the underlying generation models to demonstrate that our method is applicable to a wide range of architectures. The regression head is trained for 3 epochs and only takes 0.27% and 0.20% parameters for PanGu-Bot and EVA2.0 respectively. DDS is introduced to four widely used stochastic decoding strategies at sentence level with inverse sigmoid mapping. We set $k = 3$, $p = 0.9$, $\tau = 0.9$ for top-k, top-p, typical sampling respectively, and $T = 1$

Decoding Strategy	Flu. (%)	Rel. (%)	Kappa
Top-k (fixed T)	97.6	59.0	0.618
Top-k (DDS)	98.3	70.0	0.439
Top-p (fixed T)	92.0	62.3	0.734
Top-p (DDS)	90.3	60.3	0.655
Temperature (fixed T)	80.3	52.7	0.496
Temperature (DDS)	79.0	50.0	0.512
Typical (fixed T)	84.0	54.7	0.621
Typical (DDS)	87.3	54.7	0.431

Table 5: Human evaluations results on Diamante.

for all of them including Temperature sampling as common settings. In main experiments, we adopt sentence-level DDS, given that its lower costs than token-level one. The responses are generated 5 times per test.

4.3 Automatic Evaluation

For automatic evaluation, we divide metrics into two groups because chit-chat and QA datasets require different evaluation aspects. For factual QA datasets, the most important thing is to verify the knowledge accuracy w.r.t. the ground truth, thus we adopt the following metrics: **BLEU-{1,2,3,4}** (Chen and Cherry, 2014), **Rouge-{1,2,L}** (Lin, 2004) and **F1**. While for chatting datasets, considering there will be multiple responses for one context, the metrics above are not suitable. Therefore, we utilize these three metrics to evaluate the diversity: **Distint-{1,2,3}** (Li et al., 2016), **Ent-{1,2,3}** (word entropy) (Csáky et al., 2019) and **BERTScore** (calculating the similarity score between five generated responses given the same context).

Table 2 shows the results from PanGu-Bot. As can be seen, the proposed dynamic decoding strategy (DDS) improves the performance of all four well-known stochastic decoding algorithms on four datasets, confirming its general applicability and superiority. Specifically, for LQA and PersonQA, all metrics obtains the best scores, indicating that DDS can generate more accurate answers for QA scenario. Under the same settings, the higher Distinct and Ent scores of Diamante and LCCC verify the diversity in chit-chat scenario. Appendix A shows some generated cases. Table 3 summarizes the result from EVA2.0 in a zero-shot setting, which illustrates similar trends. This observation demonstrates that the proposed DDS can be applied to different model architectures and learning manners.

4.4 Human Evaluation

For chit-chat dataset, although label-related metrics are not suitable, it is also necessary to evaluate its relevance (**Rel.**) and fluency (**Flu.**) besides the diversity. So we conduct human evaluation as a supplement to automatic experiment. **Rel.** reflects how likely the generated response is relevant to its context. **Flu.** reflects how likely the generated response comes from human. We collect 100 samples for each decoding setting from Diamante and employ three annotators to judge whether the response is in compliance with above standards. Table 5 summarizes the human evaluation results. We can see that the proposed approach has similar results compared with baselines, which indicates that dynamic decoding method maintains the relevance and fluency of responses while improving its diversity. We use Fleiss's kappa (Fleiss, 1971) to measure the inter-annotator agreement.

4.5 Multilingual Availability

CQ	BLEU-4	F1	ROUGE-2	ROUGE-L
Base	0.0520	0.0759	0.0133	0.0741
DDS	**0.0532**	**0.0793**	**0.0142**	**0.0722**
Base	0.0674	0.1105	0.0391	0.1072
DDS	**0.0691**	**0.1154**	**0.0406**	**0.1115**
Daily	Dist-2	Dist-3	Ent-2	Ent-3
Base	0.2647	0.4371	14.2122	17.5430
DDS	**0.4023**	**0.6056**	**14.5874**	**17.6932**
Base	0.2966	0.4722	13.6437	17.2051
DDS	**0.4158**	**0.6141**	**13.8967**	**17.3642**

Table 6: Zero-shot results on Llama-2-7b (Liu et al., 2023) (Up) and GPT-3.5-turbo (Down). Base means sampling with fixed temperature. *CQ* refers to ComplexQuestions and *Daily* refers to DailyDialog.

Although the proposed method was tested on Chinese corpora, it could work for other languages as well. To demonstrate this, we select English datasets as additional study, ComplexQuestions (Bao et al., 2016) for QA and DailyDialog (Li et al., 2017) for chit-chat. The superior results from Table 6 with top-p sampling support the multilingual availability of DDS. The linguistic phenomena in English differ greatly from those in Chinese, making this experiment a good test of the applicability of the proposed method to non-Chinese languages.

4.6 Token Level DDS

Dynamic decoding at the token level is more fine-grained than that at the sentence level. The Figure 4

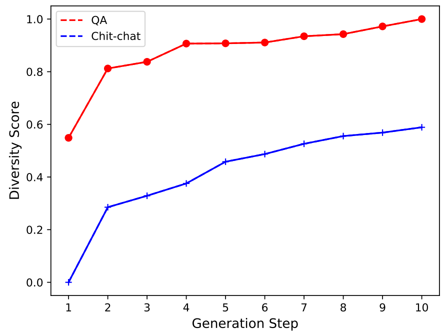

Figure 4: Token level diversity score (normalized) over generation steps.

PersonQA	BLEU-4	F1	ROUGE-2	ROUGE-L
Base	0.3117	0.4044	0.3041	0.4010
Sent	**0.3539**	**0.4488**	**0.3461**	**0.4456**
Token	0.3357	0.4335	0.3273	0.4303
Diamante	Dist-2	Dist-3	Ent-2	Ent-3
Base	0.4582	0.7036	12.6627	15.8346
Sent	0.5401	0.7811	12.9131	16.0630
Token	**0.5603**	**0.8289**	**13.1880**	**16.2892**

Table 7: Results of token-level DDS with top-p sampling.

depicts that the diversity score (the higher, the narrower decoding space) shows a rising trend over the generation step, which is consistent with the heuristic motivation of Lee et al. (2022) that generating the latter part of a sentence require less decoding randomness. Table 7 shows the results at both two levels. The scores of token level on both two datasets are higher than base, verifying the effectiveness of it. Different from Diamante, PersonQA does not perform better at the token level than it does at the sentence level. This may be because the higher randomness of former part within the utterance than sentence level, thus it needs further design for mapping strategy. Figure 4 has shown the effectiveness of predicting diversity score at token level, and we leave the study of exploiting the potential of it as future work.

4.7 Study of mapping strategies

In this section, we study the effectiveness of different mapping strategies. As shown in Table 8, all three types of mapping functions can largely improve the performance on both two scenarios. We simply set h for them as 5, 0.01 and 0.02 respectively and actually the hyperparameters do not need to be specially adjusted. For example, the slope of linear mapping can influence the performance, but

Mapping	BLEU-4	F1	ROUGE-2	ROUGE-L
Identity	0.0924	0.1870	0.0325	0.1491
Linear	0.1004	0.2124	0.0441	0.1753
Exp	0.1001	0.2100	0.0427	0.1719
Sigmoid	0.1017	0.2170	0.0448	0.1802
Mapping	Dist-2	Dist-3	Ent-2	Ent-3
Identity	0.6170	0.9057	18.9154	20.4290
Linear	0.7491	0.9600	19.2278	21.0573
Exp	0.7760	0.9406	19.2988	21.2100
Sigmoid	0.7718	0.9428	19.4330	21.4829

Table 8: Study of mapping strategies with top-p sampling on LQA (Up) and LCCC (Down).

Slope	BLEU-4	F1	ROUGE-2	ROUGE-L
Base	0.0924	0.1870	0.0325	0.1491
1	0.0933	0.1903	0.0345	0.1532
2	0.0963	0.1993	0.0378	0.1607
3	0.0977	0.2021	0.0387	0.1637
4	0.0995	0.2077	0.0419	0.1696
5	0.1004	0.2124	0.0441	0.1753

Table 9: Study of the value of slope.

as shown in Table 9, all five different values can outperform the fixed temperature sampling.

4.8 Domain Adaptation

We conduct experiments with out-of-domain test data on EVA2.0 for further generalization evaluation. For chit-chat scenario, we choose CDConv (Zheng et al., 2022), a high-quality dataset for detecting contradiction problem. We only select the first turn of each conversations, where the query is basically the question in chit-chat scenario. For QA scenario, we employ BaikeQA, a QA dataset from Chinese Wiki. The results from Table 10 show that DDS can still outperform the basic decoding strategy, which indicates the generalization ability.

4.9 Dynamic Training

To evaluate the effectiveness of dynamic training (DT), we train the LM head and regression head jointly. The results of Table 11 show that dynamic training is effective in improving performance. The dynamic training and decoding can be performed simultaneously, and the higher performance of DT+DDS indicates that the performance can be further enhanced.

5 Conclusion

In this paper, we discuss the drawbacks of commonly used standard decoding methods for open-domain dialogue generation task. To overcome

BaikeQA	BLEU-4	F1	ROUGE-2	ROUGE-L
Base	0.0924	0.1870	0.0325	0.1491
DDS	**0.1004**	**0.2124**	**0.0441**	**0.1753**
CDConv	Dist-2	Dist-3	Ent-2	Ent-3
Base	0.6170	0.9057	18.9154	20.4290
DDS	**0.7491**	**0.9600**	**19.2278**	**21.0573**

Table 10: Results of out-of-domain test.

PersonQA	BLEU-4	F1	ROUGE-2	ROUGE-L
Base	0.3117	0.4044	0.3041	0.4010
DT	0.3838	0.4758	0.3776	0.4737
DT+DDS	**0.4050**	**0.4956**	**0.3967**	**0.4936**
Diamante	Dist-2	Dist-3	Ent-2	Ent-3
Base	0.4582	0.7036	12.6627	15.8346
DT	0.4794	0.7428	12.7369	15.9257
DT+DDS	**0.5479**	**0.7986**	**13.1270**	**16.2207**

Table 11: Results of DT with top-p sampling.

them, we present a novel dynamic decoding strategy, DDS, to handle different conversational scenarios concurrently. It can adaptively adjust the decoding space according to different contexts at both sequence and token levels with three mapping functions. Moreover, we further boost the performance by introducing the dynamic temperature to training stage. Extensive experiments demonstrate the superiority and generalization of proposed decoding method.

Limitations

The following are our limitations:

- The contribution for our work may go beyond dialogue generation task. Nowadays, more and more tasks are combined in one model, especially the large language model like ChatGPT. Given that different tasks have different optimal hyper-parameter for decoding temperature, it is badly needed to adjust the temperature adaptively to handle all tasks simultaneously. But we haven't expended proposed strategy to LLMs.

- Since there is no suitable public Chinese QA conversational dataset available, the QA datasets we utilize are collected through our internal efforts and haven't been released publicly now. It may be difficult to reproduce our results in this manner.

- Considering reranking approach is very popular and effective for text generation, we haven't evaluated the performance of combining it with our proposed method.

Acknowledgements

This work is supported by the Beijing Natural Science Foundation, China (Nos. 4222037, L181010).

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cogn. Sci.*, 9(1):147–169.

Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based question answering with knowledge graph. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2503–2514. ACL.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zheng-Yu Niu. 2022. PLATO-XL: exploring the large-scale pre-training of dialogue generation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Online only, November 20-23, 2022*, pages 107–118. Association for Computational Linguistics.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

Richárd Csáky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5650–5669, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, and Minlie Huang. 2022. EVA2.0: investigating open-domain chinese dialogue systems with large-scale pre-training. *CoRR*, abs/2203.09313.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*. OpenReview.net.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *CoRR*, abs/2206.04624.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP(1)*, pages 986–995.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Tianming Liu, Quanzheng Li, and Xiang Li. 2023. Radiology-llama2: Best-in-class large language model for radiology. *CoRR*, abs/2309.06419.

Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2022. Towards boosting the open-domain chatbot with human feedback. *CoRR*, abs/2208.14165.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022a. Locally typical sampling. *Transactions of the ACL*.

Clara Meister, Gian Wiher, and Ryan Cotterell. 2022b. On decoding strategies for neural text generators. *Trans. Assoc. Comput. Linguistics*, 10:997–1012.

Fei Mi, Yitong Li, Yulong Zeng, Jingyan Zhou, Yasheng Wang, Chuanfei Xu, Lifeng Shang, Xin Jiang, Shiqi Zhao, and Qun Liu. 2022. PANGUBOT: efficient generative dialogue pre-training from pre-trained language model. *CoRR*, abs/2203.17090.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *CoRR*, abs/2208.03188.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Follow the wisdom of the crowd: Effective text generation via minimum bayes risk decoding. *CoRR*, abs/2211.07634.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *NLPCC*, volume 12430 of *Lecture Notes in Computer Science*, pages 91–103. Springer.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zhengyu Niu, Hua Wu, and Minlie Huang. 2022. Cdconv: A benchmark for contradiction detection in chinese conversations. *CoRR*, abs/2210.08511.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, and Jie Tang. 2021. EVA: an open-domain chinese dialogue system with large-scale generative pre-training. *CoRR*, abs/2108.01547.

A Case Study

Table 12 and 13 shows some cases generated 5 times by the proposed method and baselines on both two scenarios. From where we can see that the diversity of responses under chit-chat is enhanced and the correctness in QA scenario is also improved.

context:	马上就要填报志愿了，你打算学什么专业啊(It’s time to fill in your application. What major do you plan to study)
DDS:	我要学医,你呢? (I want to study Medicine. How about you?) 我想报法学专业,你有什么好的建议吗? (I want to apply for Law major. Do you have any good suggestions?) 我准备学酒店管理,我喜欢甜品(I’m going to learn Hotel Management. I like dessert) 我想报计算机,你有什么好的建议吗? (I want to study Computer Science. Do you have any good suggestions?) 我啊,最看重学校,其次是专业。(For me, school is the most important, followed by major.)
Base:	我要学医,你呢? (I want to study Medicine. How about you?) 我是学医的,以后可能会做医生(I am a medical student. I may become a doctor in the futur) 我的专业是应用化学,你呢? (My major is Applied Chemistry. What about you?) 我想报计算机,你觉得怎么样? (I want to study Computer Science, What do you think?) 我要学医,你呢? (I want to study Medicine. How about you?)
context:	这个周末我们公司要搞团建！(This weekend, our company will carry out league construction!)
DDS:	这个好,我们学校也搞过(This is good. Our school has also done it) 这个主题不错(This theme is good) 我们也做过哎(We did it too) 我们也做过(We did it too) 这个主意不错(This is a good idea)
Base:	啊,你们团建是什么(Ah, what are you going to do in league construction) 这个主意不错(This is a good idea) 我知道啊,但是我不去,你组织(I know, but I won’t go, you organize) 是我们这边的郊区的团建吗(Is it the group building in the suburb of our side) 我们没有这个活动吧(We don’t have this activity, do we?)

Table 12: Generated samples from Diamante.

context: *true answer:*	谢婉莹是谁？(Who is Xie Wanying?) 中国诗人，现代作家、翻译家、儿童文学作家、社会活动家(Chinese poet, modern writer, translator, children's literature writer, social activist)
DDS:	谢婉莹2012年农历腊月22日创作并发表了单曲《我不是你爸爸的爸爸》(Xie Wanying wrote and published the single "I'm not your father's father" on the 22nd lunar month of 2012) 女,中国诗人,现代作家、翻译家(Female, Chinese poet, modern writer and translator) 中国古典小说《甄传》中的主要角色之一(One of the main characters in the classical Chinese novel Legend of Zhen Huan) 中国诗人,现代作家、翻译家、儿童文学作家、社会活动家(Chinese poet, modern writer, translator, children's literature writer, social activist) 中国诗人,现代作家、翻译家、儿童文学作家(Chinese poet, modern writer, translator, children's literature writer)
Base:	电视剧《永不言弃》中的角色(The role in the TV play Never Give Up) 小说《蜀山剑侠传》的女主人公(The heroine of the novel "Legend of the Swordsman of Shu Mountain") 中国诗人,现代作家。著有诗集《黄昏》(Chinese poet and modern writer. He is the author of the poetry collection "Twilight") 现代作家。代表作有《红高粱家族》(Modern writers. The representative work is Red Sorghum Family) 女,教授,南京师范大学商学院副教授(Female, professor, associate professor of Nanjing Normal University Business School)
context: *true answer:*	熊黛林多高？(How tall is Xiong Dailin?) 178cm
DDS:	178cm 178cm 178cm 178cm 178cm
Base:	168cm 168cm 178cm 178cm 178cm

Table 13: Generated samples from PersonQA.

B Details about QA Datasets

Both PersonQA and LQA are our internal data, and each sample is a question-answer pair. Questions are collected from Internet, and answers are manually annotated. Both questions and answers will be classified. PersonQA contains QA pairs about persons and LQA includes samples that have longer explanations. All the data is anonymized and nonsensitive.

C Insights for Mapping Strategies

We design three mapping strategies in order to cover all major types of mapping trends. Specifically, As shown in Figure 2, Linear mapping simply projects the diversity score to temperature linearly. Exponential mapping has flat slope when diversity score is near the mean value while sharp slope at either end. Conversely, Inverse Sigmoid mapping shows a different trend. According to Section 4.7, all three types of strategies can work on both two scenarios.