EpLSA: Synergy of Expert-prefix Mixtures and Task-Oriented Latent Space Adaptation for Diverse Generative Reasoning

Fujun Zhang[1,2,3], **Xiangdong Su**[1,2,3](✉), **Jiang Li**[1,2,3], **Rong Yan**[1,2,3], **Guanglai Gao**[1,2,3]
[1]College of Computer Science, Inner Mongolia University, Hohhot, China
[2]National & Local Joint Engineering Research Center of
Intelligent Information Processing Technology for Mongolian
[3]Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
zfjimu@163.com, cssxd@imu.edu.cn, lijiangimu@gmail.com, csyanr@imu.edu.cn, csggl@imu.edu.cn

Abstract

Existing models for diverse generative reasoning still struggle to generate multiple unique and plausible results. Through an in-depth examination, we argue that it is critical to leverage a mixture of experts as prefixes to enhance the diversity of generated results and make task-oriented adaptation in the latent space of the generation models to improve the quality of the responses. At this point, we propose EpLSA, an innovative model based on the synergy of expert-prefix mixtures and task-oriented latent space adaptation for diverse generative reasoning. Specifically, we use expert-prefixes mixtures to encourage the model to create multiple responses with different semantics and design a loss function to address the problem that the semantics is interfered by the expert-prefixes. Meanwhile, we design a task-oriented adaptation block to make the pre-trained encoder within the generation model more effectively adapted to the pre-trained decoder in the latent space, thus further improving the quality of the generated text. Extensive experiments on three different types of generative reasoning tasks demonstrate that EpLSA outperforms existing baseline models in terms of both the quality and diversity of the generated outputs. Our code is publicly available at https://github.com/IMU-MachineLearningSXD/EpLSA.

Keywords: diverse generative reasoning, mixture of expert, task-oriented

1. Introduction

Diverse generative reasoning aims to generate multiple semantically distinct and reasonable outputs according to the same context, like abductive commonsense reasoning, where there are multiple possible intermediate hypotheses. Figure 1 shows an example of abductive commonsense reasoning. Given the cause "Mickey was bored." and the effect "Then, they played for the next hour.", there are multiple possible explanations for the intermediate hypothesis. As proved, generating multiple outputs with different semantics presents unique challenges in diverse generative reasoning. Therefore, this paper investigates diverse generative reasoning and expects to improve the quality and diversity of the generated text.

There are many models developed for diverse generative reasoning due to their importance in NLP applications. Among them, pre-trained language models have been successful in performing commonsense inference by implicitly learning relational patterns from large-scale corpora (Trinh and Le, 2018). ClarET (Zhou et al., 2022) proposes a general pre-trained framework for generative reasoning tasks. MoKGE (Yu et al., 2022) also uses a pre-trained Transformer as the backbone network to diversify the generative reasoning. COLD Decoding (Qin et al., 2022) proposes a decoding framework that unifies the constrained generation as the

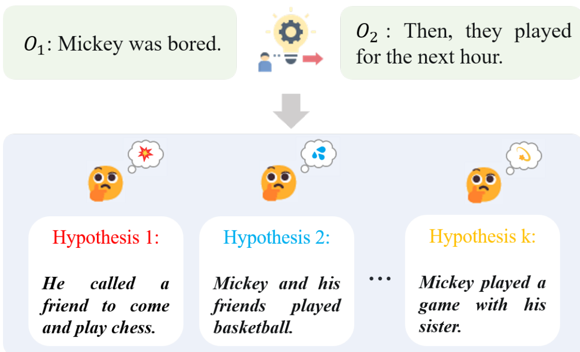

Figure 1: An example of abductive commonsense reasoning. It aims to generate explanatory hypotheses given two observations: O_1 as the cause and O_2 as the effect.

specific constraints through an energy function. To improve diversity, the latest methods for generating diverse sequences attempt to introduce uncertainty by incorporating random noise into latent variables (Gupta et al., 2018) or by using alternative search algorithms (Vijayakumar et al., 2018; Fan et al., 2018). The mixture of experts models has also recently started to be widely used for diversity generation (He et al., 2018; Cho et al., 2019; Shen et al., 2019; Yu et al., 2022).

After examining the examples from the models with the pre-trained models and the mixture of ex-

LREC-COLING 2024, pages 6216–6227
20-25 May, 2024.

pert noise prefixes, we find that the models focusing on promoting diversity lead to the result that the generated texts do not match the semantics of the source text, while the models aiming to improve the quality of generation tend to produce duplicate text when multiple samples are expected. There are two reasons for such problems. The first reason is that these models introduce noise prefixes that interfere with the semantics of the source text, thus leading to poor inference. The second reason is that the pre-trained models, such as BART (Lewis et al., 2020), are not trained for the specific tasks, which means that the latent representations of the BART encoder for different types of generative reasoning tasks do not match well with the BART decoder through limited fine-tuning alone. In different generative reasoning tasks, the source text and the target text will show different types of relationships, such as the causal relationship between the source text and the target text in the Abductive Commonsense Reasoning task, and the summarization relationship between the source text and the target text in the Story Ending Generation task. Therefore, to enhance the diverse generative reasoning, it is necessary to design a specified loss function to correct the semantic information after the introduction of the noisy prefix and make the latent representations of the encoder better adapted to the decoder according to task types.

Based on the above analysis, we propose EpLSA, a novel method for diverse generative reasoning tasks. In EpLSA, we employ a mixture of expert-prefixes (MoE) module to extend the semantics of the source text, in which each expert-prefix represents different semantic perspectives of the source text. We design a loss function to correct the semantic information after the introduction of the noisy prefix (&3.2). We introduce a task adapter in EpLSA to address the problem that the latent representation of the encoder of pre-trained models for different types of tasks can't be better adapted to the decoder (&3.3). We utilize a hard-EM algorithm to train EpLSA. Unlike other MoE models, we use the minimum loss to choose the optimal expert-prefix (&3.4). We conduct experiments on three different types of generative inference tasks and found that our model outperformed the strong baseline in terms of diversity and generation quality, demonstrating the effectiveness of our approach.

Our contributions can be summarized as follows:

- We propose a mixture of expert-prefixes in EpLSA to improve the diversity in generative reasoning tasks. Different from the previous works, we design a loss function to correct the semantic information after the introduction of the noisy expert-prefixes.

- We introduce a task-oriented adapter in EpLSA, which allows pre-trained models to be better adapted to different tasks through the adapter.

- The proposed EpLSA for diverse generative reasoning outperforms all baseline models in terms of both the diversity and quality of the generated outputs on three different types of generative reasoning tasks.

2. Related work

2.1. Diversity Text Generation

Multiple output generation had a wide range of applications in machine translation (Shen et al., 2019), question generation (Cho et al., 2019), dialogue systems (Dou et al., 2021), story generation (Yu et al., 2021), and paraphrase generation (Gupta et al., 2018).

To improve generation diversity, various methods were developed by exploring different perspectives. Some research focused on generating uncertainty by introducing random noise (Gupta et al., 2018) or changing latent variables (Lachaux et al., 2020), thereby increasing generation diversity. Shi et al. (2018) used inverse reinforcement learning methods for unconditional diversified text generation. The mixture of experts model was also used to enhance generation diversity. Cho et al. (2019) divided diversified generation into two stages: selection and generation. Intuitively, at the selection stage, each latent variable indicated which part of the source sequence was important. The generation, phase was biased towards their focus for generation. Shen et al. (2019) used a mixture of experts module to improve machine translation diversity, where each source input was assigned to a minimum-loss predictor. Wen et al. (2023) proposed an Equal-size Hard Expectation–Maximization algorithm to train a multi-decoder model for diverse dialogue generation. In addition, sampling-based decoding was one of the most effective solutions to improve generation diversity. Truncated sampling (Fan et al., 2018) limited the range of values that can be generated by the sampling process based on a predetermined cutoff value. Nucleus sampling (Holtzman et al., 2020) sampled the next token from a dynamic core unit that contains most of the probability mass, rather than decoding text by maximizing the likelihood. Some of these methods do not reason about the semantic information of the source text from different perspectives, and some do not correct the semantic information of the source text after introducing noise prefixes.

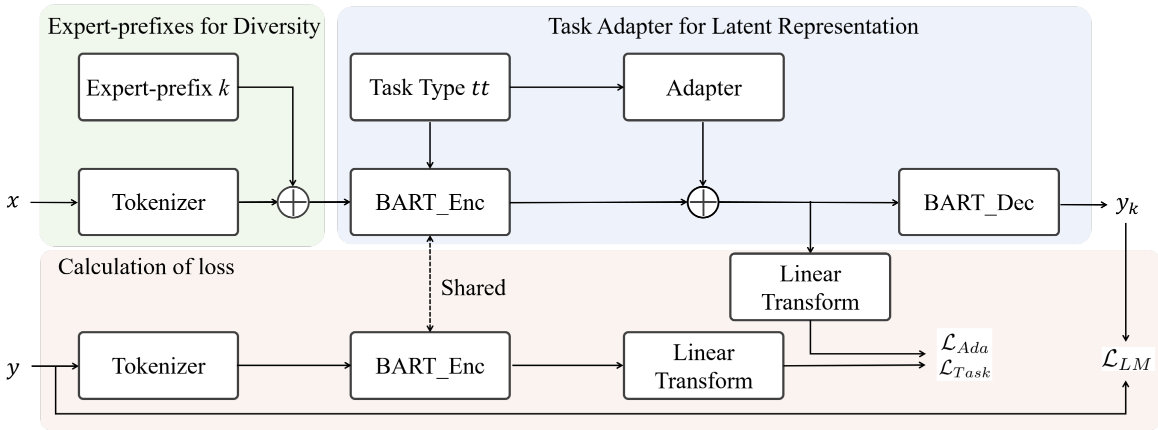

Figure 2: Overview of EpLSA. Expert-prefixes for diversity: we focus on k perspectives of the source text through k expert-prefixes (&3.2); Task adapter for latent representation: we make the BART encoder more effectively adapted to the BART decoder in the latent space (&3.3).

2.2. Generative Reasoning

In recent years, commonsense reasoning tasks have received widespread attention. GRF (Ji et al., 2020) proposed a method that utilizes the structure and semantic information of external knowledge bases by performing dynamic multi-hop reasoning on relationship paths. MoKGE (Yu et al., 2022) was the first work to boost diversity in natural language generation by diversifying knowledge reasoning on commonsense knowledge graphs. ClarET (Zhou et al., 2022) proposed a pre-training framework for event-centered reasoning by learning relevant perceptual contexts to event transformers from event-rich textual corpora. COLD Decoding (Qin et al., 2022) was an energy-based Constrained Text Generation with Langevin Dynamics. Arabshahi et al. (2021) also explored generative reasoning in commonsense scenarios, but the domain of their approach is limited. Chain-of-thought prompting techniques were used to conduct step-by-step reasoning by eliciting intermediate steps from large language models (Wei et al., 2022; Creswell et al., 2022). However, none of the existing methods took into account the adaptation of the encoder and decoder to different types of tasks.

3. Methodology

3.1. Overview of EpLSA

Generative reasoning tasks are characterized by having multiple possible reasoning results that correspond to the same given reasoning premise, which means one-to-many generation. Given a source input x, a set of outputs $Y = (y_1, y_2 ... y_k)$ is obtained. Our goal is to model a conditional distribution for the target outputs: $p(y|x)$ that assigns high values to $\{p(y_1|x), \cdots, p(y_k|x)\}$ for k mappings, i.e., $\{x \to y_1, \cdots, x \to y_k\}$.

To improve the diversity and quality of the generated text, this paper proposes a model with the synergy of expert-prefix mixtures and task-oriented latent space adaptation (EpLSA), as shown in Figure 2. The expert-prefix provides a semantics aspect prefix for diversity generation, and the task adapter improves the latent representation for better generation quality. We model expert-prefixes mixtures as a hard mixture of experts (hard-MoE) (Jacobs et al., 1991; Shen et al., 2019). We use the BART encoder and BART decoder for information encoding and decoding.

3.2. Expert-prefixes for Diversity

To perform diverse generations, we explore different semantic perspectives of the source text. Inspired by the mixture of experts (MoE) approach, we regard experts as prefixes, and reason from different semantic perspectives of the source text by mixing expert-prefixes (ep). As shown in Figure 2, we include k expert-prefixes with a length of l before each input text sequence, thus providing different inferred views of the source text.

Splicing noisy expert-prefixes with the source text sequence brings multiple inference perspectives while causing disturbance to the semantics of the source text. Therefore, we design a loss function to correct the semantic information of the source text. Specifically, we introduce the task type, denoted as tt. We randomly initialize the tt and let it be updated during training. The semantic information is corrected by learning the similarity between the task type tt and the semantic difference of (ep, x) and y. To facilitate the similarity calculation, we define a linear transformation function $LT(h)$, which converts the hidden state of the BART encoder to the sentence-level semantic information similar to

Bert's [cls] (Devlin et al., 2019).

$$\mathbf{LT}_t = Adapter(tt)\ , \tag{1}$$
$$\mathbf{LT}_x = LT(\text{BART\_Enc}(ep, x))\ , \tag{2}$$
$$\mathbf{LT}_y = LT(\text{BART\_Enc}(y))\ . \tag{3}$$

The training loss (here only for one expert-prefix) is

$$\mathcal{L}_{Task} = 1 - \cos(\mathbf{LT}_x + \mathbf{LT}_t, \mathbf{LT}_y), \tag{4}$$

where $cos(\cdot,\cdot)$ is the cosine similarity. By this method, we alleviate the problem that the semantics is interfered by the expert-prefixes for better diverse reasoning.

3.3. Task Adapter for Latent Representation

In diverse generative reasoning, the pre-trained models are not trained specifically for downstream tasks, so the latent representations of the BART encoder for different types of generative reasoning tasks do not match well with the BART decoder. To allow the model to better adapt to different generative reasoning tasks, we introduce a task adapter denoted as $Adapter$. The $Adapter$ includes only a position embedding layer, leading to a small increase in parameters compared to the original pre-trained model. We refer to the relationship between the source text and the target text as the task type tt, and use tt as the input of $Adapter$. After that, the hidden state of the BART encoder and the output of $Adapter$ is used as the input of the BART decoder.

$$\mathbf{LR} = \text{BART\_Enc(X)} + Adapter(tt), \tag{5}$$
$$\mathbf{output} = \text{BART\_Dec}(\mathbf{LR}). \tag{6}$$

To enhance the compatibility between the encoder-adapter combination and the decoder, we define the training loss for $Adapter$ (here only for one expert-prefix) :

$$\mathbf{LT}_A = LT(Adapter(tt)), \tag{7}$$
$$\mathcal{L}_{Ada} = max(0, \lambda + d(\mathbf{LT}_x + \mathbf{LT}_A, \mathbf{LT}_y)), \tag{8}$$

where $d(\cdot,\cdot)$ represents the distance in semantic space, which is the Euclidean distance used in this work. λ is a hyperparameter that is used to balance the difference between the source semantic representation and the target semantic representation. It is worth noting that our task adapter module is applicable to all models with encoder-decoder structures.

3.4. Overall objective

Taking the expert-prefixes ep, source text x and task type tt as model inputs and generating the output sequence y, we adopt the cross-entropy loss, which can be denoted as:

$$\mathcal{L}_{LM} = -\log p(y|ep, x, tt) = \sum_{t=1}^{|y|} \log p(y_t|ep, x, tt, y < t). \tag{9}$$

Algorithm 1 Training
(N: Dataset size, K: Number of expert-prefixes)

Input: $D = \left\{(ep_K, x^{(i)}, y^{(i)}, tt)\right\}_{i=1}^{N}$
Output: $\mathcal{L}$
1: **for** each $i \in [1, N]$ **do**
2: /*E-step select best expert-prefix.*/
3: **for** each $z \in [1, K]$ **do**
4: $\mathcal{L}_{LM}^{(i)z} = -\log p(y|ep, x, tt)$
5: **end for**
6: $z^{(best)i} = \arg\min \mathcal{L}_{LM}^{(i)z}$
7: /*M-step updates the parameters with gradients of the best expert-prefix from E-step.*/
8: $\mathcal{L} = \mathcal{L}_{LM}^{z^{(best)i}} + \alpha\mathcal{L}_{Task}^{z^{(best)i}} + \beta\mathcal{L}_{Ada}^{z^{(best)i}}$
9: $\theta = \theta - \nabla_\theta \mathcal{L}$
10: **end for**

The final loss we need to optimize is a linear combination

$$\mathcal{L} = \mathcal{L}_{LM} + \alpha\mathcal{L}_{Task} + \beta\mathcal{L}_{Ada}, \tag{10}$$

where α and β are hyperparameters set according to different tasks.

3.5. Model Training and Inference

Training Stage: Ideally, different expert-prefixes represent different reasoning perspectives, allowing for diverse reasoning. However, in the training phase, only one reasoning perspective should be dominant for a given input premise (Shen et al., 2019). Unlike other mixture of experts models that select a guidance expert, we select the optimal expert-prefix based on the $\mathcal{L}_{LM}$. Specifically, we employ a hard mixture model with hard-EM algorithm (Dempster et al., 1977; Shen et al., 2019) and select the best expert-prefix with the minimum $\mathcal{L}_{LM}$ as the prefix during the training process. The specific training process be expressed as:

E-STEP (line 3-6 in Alg. 1) we sample all the expert-prefixes and calculate their $\mathcal{L}_{LM}$ using the current parameters θ; We then select the best expert-prefix with the minimum $\mathcal{L}_{LM}$.

M-STEP (line 8-9 in Alg. 1) we only use the gradient of the expert-prefix selected by E-STEP to update the parameters.

Independently parameterizing each expert-prefix would lead to a dramatic increase in the number of parameters. Therefore, we follow the parameter sharing mode used by Cho et al. (2019); Shen et al. (2019); Yu et al. (2022). This only requires a negligible increase in parameters over the models that do not use MoE.

Inference Stage: To generate k different reasoning results on the test set, we follow the method proposed by (Shen et al., 2019). By enumerating all expert-prefixes, we decode each token through $\hat{y}_t = \arg\max p(y|\hat{y}_{1:t-1}, ep, x, tt)$, where we require each expert-prefix to represent different perspectives of the source text. The decoding process is efficient, easily parallelized, and can accommodate a variety of decoding strategies. We use Nucleus sampling at $p = 0.95$ (Holtzman et al., 2020).

4. Experiments

4.1. Tasks and Datasets

Abductive Commonsense Reasoning (αNLG) : It aims to generate explanatory hypotheses when two observations are given: O_1 is the cause and O_2 is the effect. We use the $\mathcal{ART}$ benchmark dataset (Bhagavatula et al., 2020), following the data split (Yu et al., 2022) with 50,481/1,779/3,560 in training/dev/test. Each example in the $\mathcal{ART}$ dataset has 1 to 5 references.

Explanation Generation (EG) : Its purpose is to provide explanations when counterfactual statements are given (Wang et al., 2019). We use the benchmark dataset ComVE from SemEval-2020 Task 4 (Li et al., 2020). We follow the data split Yu et al. (2022) with 10000/997/1000 in training/dev/test. All examples in the dataset have 3 references.

Story Ending Generation (SEG) : It is to generate a reasonable ending given a four-sentence story context. The stories come from ROCStories corpus (Mostafazadeh et al., 2016). We follow the data split (Guan et al., 2019) with 90000/4081/4081 in training/dev/test. All examples in the dataset only have 1 reference.

4.2. Baseline Methods

When we perform diversified reasoning, which means one-to-many text generation, we exclude baseline methods that can't produce multiple outputs mentioned in related work and only compare with methods that can generate diverse outputs, e.g., Ji et al. (2020); Zhou et al. (2022); Qin et al. (2022).

BART-base (Lewis et al., 2020) is a pre-trained language generation model based on the Transformer structure. We fine-tune the model on the abductive commonsense reasoning, explanation generation, and story ending generation tasks. Then using Nucleus sampling (Holtzman et al., 2020) also known as **Top-p sampling** and Truncated sampling (Fan et al., 2018) also known as **Top-k sampling** for sampling in the generation phase.

CVAE-SVG (Gupta et al., 2018) is a conditional VAE model that can produce multiple outputs based on an original sentence as input.

MoE-based method (Shen et al., 2019; Cho et al., 2019): Mixture models provide an alternative approach to generating diverse outputs by sampling different mixture components. We compare two mixture of experts (MoE) implementations by Shen et al. (2019) and Cho et al. (2019). We refer to them as MoE-Shen (Shen et al., 2019) and MoE-Cho (Cho et al., 2019).

MoKGE (Yu et al., 2022) is the first work to boost diversity in NLG by diversifying knowledge reasoning on commonsense knowledge graphs. MoKGE uses both embed and prefix to implement mixture of experts. We refer to them as MoKGE_embed and MoKGE_prompt. It is also the current SOTA for the abductive commonsense reasoning and explanation generation tasks.

4.3. Implementation Details

We initialize EpLSA and baseline model use BART-base (Lewis et al., 2020), which is one of the state-of-the-art pre-trained Transformer models for natural language generation (Gehrmann et al., 2021).

For model training, we use Adam (Kingma and Ba, 2015) with a batch size of 15, gradient accumulation steps of 4, the learning rate of 3e-5, learning rate warm-up over the first 10,000 steps, and linear decay of the learning rate. Our model is trained by one GTX 1080Ti GPU with 11GB memory, and implemented on PyTorch with the Huggingface's Transformer (Wolf et al., 2020).

4.4. Automatic Evaluation

We evaluate the performance of different generation models from two aspects: quality and diversity.

Quality metric(↑). We compare the highest accuracy between all generated sequences in the Top-K list with the target sequence to measure generation quality (Ott et al., 2018; Vijayakumar et al., 2018). Concretely, we generate K hypotheses $\left\{\widehat{Y}(1), \cdots \widehat{Y}(K)\right\}$ from each source X and keep the hypothesis $\widehat{Y}^{best}$ that achieves the best sentence level metric with the target Y. Then, we calculate a corpus-level metric with the greedily-selected hypotheses $\{Y^{(i)best}\}_{i=1}^{N}$ and references $\{Y^{(i)}\}_{i=1}^{N}$. We use BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) to evaluate the abductive commonsense reasoning and the explanation generation tasks and use BLEU-1 and BLEU-2 (Papineni et al., 2002) to evaluate the story ending generation task.

Models	αNLG				EG			
	SB-3(↓)	SB-4(↓)	B-4(↑)	R-L(↑)	SB-3(↓)	SB-4(↓)	B-4(↑)	R-L(↑)
CVAE-SVG (Gupta et al., 2018)	57.02	53.40	11.91	36.46	54.81	49.28	15.62	41.23
Top-k sample (Fan et al., 2018)	52.11	47.75	14.01	38.98	61.39	56.93	17.48	42.44
Top-p sample (Holtzman et al., 2020)	56.32	52.44	13.53	38.42	63.43	59.23	17.68	42.60
MoE_embed (Cho et al., 2019)	29.02	24.19	14.31	38.91	33.64	28.21	18.66	43.72
MoE_prompt (Shen et al., 2019)	28.05	23.18	14.26	38.78	33.42	28.40	18.91	43.71
MoKGE_embed (Yu et al., 2022)	29.17	24.04	13.74	38.06	35.36	29.71	19.13	43.70
MoKGE_prompt (Yu et al., 2022)	27.40	22.43	14.12	38.41	30.93	25.30	19.01	43.83
EpLSA (Ours)	**23.18**	**17.82**	**15.25**	**40.35**	**27.93**	**21.33**	**19.40**	**44.02**

Table 1: Diversity and quality evaluation on the test set of αNLG and EG. Each model is required to generate three outputs. We use the generation results from Yu et al. (2022). Metrics: SB-3/4: Self-BLEU-3/4 (↓), B-4: BLEU-4(↑), R-L: ROUGE-L (↑). (↑)/(↓) means the higher/lower score the better.

Models	SEG	
	SB-3/4 (↓)	B-1/2 (↑)
CVAE-SVG	61.78/58.18	30.65/14.60
Top-k sample	52.54/48.00	31.33/15.34
Top-p sample	51.83/47.32	31.26/15.37
MoE_embed	36.77/31.75	32.08/16.13
MoE_prompt	31.71/27.38	32.12/16.20
MoKGE_embed	26.02/20.65	29.90/13.76
MoKGE_prompt	30.00/25.66	30.08/13.85
EpLSA (Ours)	**24.84/19.55**	**32.18/16.26**

Table 2: Diversity and quality evaluation on the test of SEG. Each model is required to generate three outputs. The above results were obtained by using the open-source code of the paper. Metrics: SB-3/4: Self-BLEU-3/4 (↓), B-1: BLEU-1(↑), B-2:BLEU-2 (↑). (↑)/(↓) means the higher/lower score the better.

Models	αNLG	
	SB-3/4 (↓)	B-3/4 (↑)
ChatGPT	23.62/**17.01**	1.61/19.32
EpLSA (Ours)	**23.18**/17.82	**15.25/40.35**

Table 3: Diversity and quality evaluation compare with ChatGPT on the test of α **NLG**. Each model is required to generate three outputs.

Pairwise metric(↓). Referred as self- (Zhu et al., 2018) or pairwise- (Ott et al., 2018) metric, it measures the within-distribution similarity. This metric computes the average of sentence-level metrics between all pairwise combinations of hypotheses $\{Y^{(1)}, \cdots, Y^{(K)}\}$ generated from each source sequence x. In this paper, we use Self-BLEU-3 and Self-BLEU-4. A low pairwise metric indicates high diversity between generated hypotheses.

4.5. Experimental Results

To evaluate the effectiveness of EpLSA, we perform experiments on three different types of generative reasoning tasks. We present the results of the automatic assessment for abductive commonsense reasoning and explanation generation in Table 1 and the results of the story ending generation in Table 2.

MoKGE (Yu et al., 2022) is the current SOTA for diverse abductive commonsense reasoning and the diverse explanation generation tasks. On the abductive commonsense reasoning task, EpLSA achieves the best results in terms of both generative diversity and generative quality among all baseline methods. EpLSA can further boost diversity by about 4.22% and 4.61% on Self-BLEU-3 and Self-BLEU-4, compared with the MoKGE. Moreover, EpLSA also enhances the generated quality by approximately 1.13% and 1.94% on BLEU-4 and ROUGE-L, compared with the MoKGE. On the explanation generation task, EpLSA achieves competitive results in all baseline models. Compared with MoKGE, EpLSA achieves improvements of 3.00%, 3.97%, 0.39% and 0.19%, respectively. For story ending generation task, EpLSA also achieves the best results in all baseline models. Compared with MoKGE, EpLSA achieves improvements of 1.18%, 1.10%, 2.28% and 2.50%, respectively. The above results confirm that EpLSA improves the quality and diversity of generated text. This is because different expert-prefixes focus on different semantic perspectives of the source text and the introduction of the loss function corrects the semantic information after introducing expert-prefixes, increasing the diversity while ensuring the quality of the generation. The task adapter makes the pre-trained encoder within the generation model more effectively adapted to the pre-trained decoder in the latent space for better diverse reasoning.

In addition, we compared our method with ChatGPT (OpenAI, 2023). The experimental data are shown in the Table 3. Since the number of expert-prefixes is set to 3 in this paper, we also let Chat-

Models	Size	αNLG		EG		SEG	
		Diversity	Reasonability	Diversity	Reasonability	Diversity	Reasonability
Nucleus sampling	**139M**	3.01±0.60	3.20±0.62	2.83±0.44	3.64±0.51	2.98±0.30	3.56±0.42
MoE_embed	140M	3.71±0.21	3.51±0.43	3.46±0.41	3.82±0.42	3.62±0.43	4.10±0.22
MoE_prompt	140M	3.78±0.35	3.32±0.27	3.54±0.52	3.88±0.36	3.85±0.36	4.15±0.18
MoKGE_embed	145M	3.79±0.45	3.49±0.38	3.54±0.39	3.92±0.26	4.02±0.27	3.45±0.21
MoKGE_prompt	145M	3.93±0.26	3.25±0.36	3.88±0.27	3.91±0.29	3.88±0.36	3.51±0.30
EpLSA (Ours)	140M	**4.22**±0.30	**4.12**±0.28	**4.15**±0.35	**4.08**±0.33	**4.16**±0.36	**4.20**±0.26

Table 4: Human evaluation results on three datasets.

Models	αNLG				EG			
	SB-3(↓)	SB-4(↓)	B-4(↑)	R-L(↑)	SB-3(↓)	SB-4(↓)	B-4(↑)	R-L(↑)
EpLSA (Ours)	**23.18**	**17.82**	**15.25**	**40.35**	**27.93**	**21.33**	**19.40**	**44.02**
$w/o\ L_{Task}$	26.80	22.09	15.17	40.11	31.77	24.91	19.09	43.82
$w/o\ Adapter$	32.20	24.89	14.73	30.01	32.62	23.34	18.57	43.09
w/o MoE	44.35	38.91	14.61	39.66	57.91	52.51	18.44	43.37
BART-base (Top-p sampling)	56.32	52.44	13.53	38.42	63.43	59.23	17.68	42.60

Table 5: Ablation study of the proposed model. When not using MoE (line –w/o MoE), we set the beam as three to generate three outputs. Metrics: SB-3/4: Self-BLEU-3/4 (↓), B-4: BLEU-4(↑), R-L: ROUGE-L (↑). (↑)/(↓) means the higher/lower score the better.

GPT generate three different answers for a fair diversity comparison. As shown in the table, the performance of our approach is close to that of ChatGPT, which proves that our method is effective in diversity generation. We significantly outperform ChatGPT on the quality assessment metrics Bleu-4 and ROUGE-L. This is due to the fact that ChatGPT is not fine-tuned in our dataset and the answers are very different from those in the test set resulting in lower metrics. The manual review of the generated results reveals that the answers generated by ChatGPT are logical and easy to understand. Therefore, Bleu-4 and ROUGE-L do not reflect the quality of ChatGPT generation. Overall, our method achieves a similar performance to ChatGPT.

4.6. Human Evaluation

Automatic diversity evaluation does not reflect content-level diversity and contextual logical reasonability. Therefore, we conducted an extensive human evaluation to evaluate the quality and diversity of outputs produced by different models. We recruited 20 annotators and evaluated 100 sentences randomly selected from the test set of each pair of models. The diversity and reasonability scores are normalized to the range from 0 to 5 and the results are shown in Table 4. Higher scores represent better diversity and rationality.

4.7. Ablation Study

We conduct ablation studies to assess the effectiveness of the various model components, as summarized in Table 5. Our results demonstrate that each component plays a crucial role in achieving optimal performance. Particularly, removing the mixture of expert-prefixes module (w/o MoE) resulted in a significant decrease in the diversity reasoning ability of the model. It indicates that the mixture of expert-prefixes module is effective in performing high-quality inference from multiple perspectives. In addition, the removal of the loss function that corrects semantic information ($w/o\ L_{Task}$) and the task adapter module ($w/o\ Adapter$) both reduce the generation quality and diversity to some extent. This is because the loss function helps to correct semantic representation after introducing expert-prefix and task adapter to improve the latent representation for better diversity reasoning. There is also a significant improvement in generation quality and diversity when we do not use the MoE module compared to BART-base. This observation implies that the combination of task type and task adapter effectively adapts the pre-trained model to different tasks.

4.8. Impact of the Sampling Mothed

We carry out numerous experiments to examine the impact of different sampling methods on diverse text generation. We use Self-BLEU-3/4 (↓) to evaluate diversity and BLEU-4 (↑) to evaluate

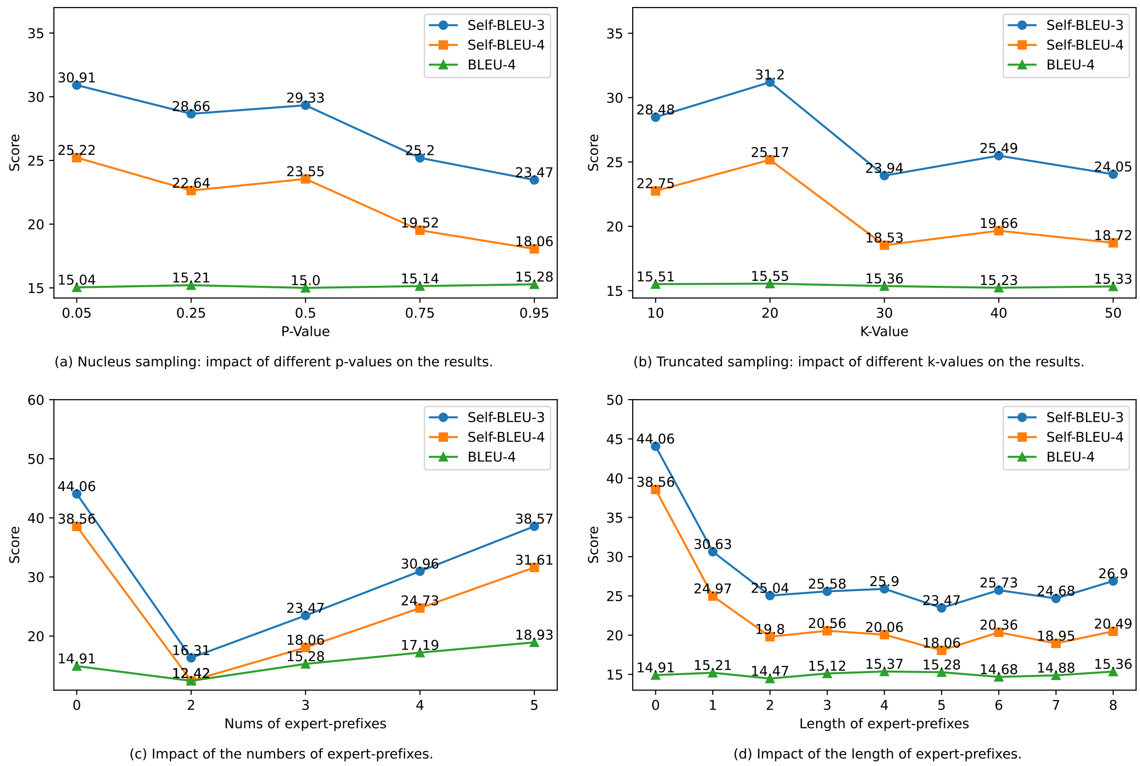

Figure 3: Effect of different parameters on the test set of αNLG.

generation quality. Specifically, we explore the effects of different parameters of Truncated sampling (Fan et al., 2018) and Nucleus sampling (Holtzman et al., 2020), as summarized in Figure 3. Overall, it illustrates that the different parameters have a small effect on the generation quality, with large fluctuations in the generation diversity. Notably, we obtain similar results with Truncated sampling at $k = 30$ or $k = 50$ and Nucleus sampling at $p = 0.95$. These observations suggest that the selection of an appropriate sampling method is also crucial for achieving diverse text generation.

4.9. Impact of the Numbers of Expert-prefixes

We conduct experiments to analyze the effect of the number of expert-prefixes on diversity and generation quality, as shown in Figure 3 (c). We found that the increase in the number of expert-prefixes negatively impacted the diversity metrics while improving the quality of the generation. Upon closer examination of the generated examples, it becomes evident that the more expert prefixes there are, the greater the diversity of the generated answers. The reason behind the decline in the diversity indicator can be attributed to the presence of common elements within the generated answers, such as shared names or locations. It is necessary to understand that the method employed to calculate the diversity metric, Self-BLEU, magnifies the impact of even subtle similarities. As a result, the metric responds significantly even if there are very few shared elements, resulting in a deteriorating indicator.

Since all baseline models generate three different answers, we also set the number of expert-prefixes to 3 in our experiments to maintain fairness. It is worth noting that when the number of mixtures of expert-prefixes is equal to 1, it is equivalent to directly performing the M-step of the hard-EM algorithm, making the mixture of experts system irrelevant at this point.

4.10. Impact of the Length of Expert-prefixes

These expert-prefixes represent different perspectives on the semantics of the source text that we should focus on. To ensure that the length of the expert-prefixes is appropriate to avoid a large interference with the semantics of the source text, we conduct extensive experiments, the results of which are shown in Figure 3 (d). Our results show that the length of the expert-prefixes can't highlight the semantic perspective that needs to be focused on if it is too short, while the length of the expert-prefixes can have an excessive impact on the semantics of the source text if it is too long. The optimal length of the expert-prefixes is different on different datasets, so we need to set it according to different task types. In this paper, we set the length of expert-prefixes to 5.

EpLSA consistently outperforms existing models when the length of expert-prefixes is greater than 1. It is worth noting that the optimal results are achieved when the length of the expert prefix is between 15% and 25% of the average length of the source text.

5. Conclusion

In this paper, we propose EpLSA, an innovative model based on the synergy of expert-prefix mixtures and task-oriented latent space adaptation for diverse generative reasoning. We introduce expert-prefix mixtures to encourage the model to create multiple responses with different semantics, where each expert-prefix focuses on a different perspective of the source text. Meanwhile, we define a loss function to correct semantic information after introducing expert-prefixes. In addition, the task adapter makes the pre-trained encoder within the generation model more effectively adapted to the pre-trained decoder in the latent space for better generation quality. Our experiments show that EpLSA outperforms existing baseline models in terms of diversity and generation quality across three different types of generative reasoning tasks.

6. Acknowledgements

This work was funded by National Natural Science Foundation of China (Grant No. 62366036), National Education Science Planning Project (Grant No. BIX230343), Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (Grant No. 2022YFHH0077), The Central Government Fund for Promoting Local Scientific and Technological Development (Grant No. 2022ZY0198), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (Grant No. NJYT24033), Inner Mongolia Autonomous Region Science and Technology Planning Project (Grant No. 2023YFSH0017), Joint Fund of Scientific Research for the Public Hospitals of Inner Mongolia Academy of Medical Sciences (Grant No.2023GLLH0035), Natural Science Foundation of Inner Mongolia (Grant No. 2021BS06004). The work by Fujun Zhang was supported by the Research and Innovation Projects for Graduate Students in Inner Mongolia Autonomous Region.

7. Bibliographical References

Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria, and Tom M. Mitchell. 2021. Conversational neuro-symbolic commonsense reasoning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 4902–4911. AAAI Press.

Chandra Bhagavatula and Ronan Le Bras and Chaitanya Malaviya and Keisuke Sakaguchi and Ari Holtzman and Hannah Rashkin and Doug Downey and Wen-tau Yih and Yejin Choi. 2020. *Abductive Commonsense Reasoning*. OpenReview.net.

Jaemin Cho, Min Joon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3119–3129. Association for Computational Linguistics.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *CoRR*, abs/2205.09712.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2021. Multitalk: A highly-branching dialog testbed for diverse conversations. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event,* *February 2-9, 2021*, pages 12760–12767. AAAI Press.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6473–6480. AAAI Press.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5149–5156. AAAI Press.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2018. Sequence to sequence mixture model for diverse machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 583–592. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 725–736. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. 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*.

Marie-Anne Lachaux, Armand Joulin, and Guillaume Lample. 2020. Target conditioning for one-to-many generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2853–2862. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Junyi Li and Bin Wang and Haiyan Ding. 2020. *Lijunyi at SemEval-2020 Task 4: An ALBERT Model Based Maximum Ensemble with Different Training Sizes and Depths for Commonsense Validation and Explanation*. International Committee for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

OpenAI. 2023. Chatgpt. Accessed: 2023-02-08.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3953–3962. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. COLD decoding: Energy-based constrained text generation with langevin dynamics. In *NeurIPS*.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International conference on machine learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5719–5728. PMLR.

Zhan Shi, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2018. Toward diverse text generation with inverse reinforcement learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4361–4367. ijcai.org.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7371–7379. AAAI Press.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Yuqiao Wen, Yongchang Hao, Yanshuai Cao, and Lili Mou. 2023. An equal-size hard EM algorithm for diverse dialogue generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao, and Meng Jiang. 2022. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1896–1906. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Tong Zhao, Zhichun Guo, and Meng Jiang. 2021. Sentence-permuted paragraph generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5051–5062. Association for Computational Linguistics.

Yucheng Zhou, Tao Shen, Xiubo Geng, Guodong Long, and Daxin Jiang. 2022. Claret: Pre-training a correlation-aware context-to-event transformer for event-centric generation and classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2559–2575. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development* *in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.