MVP: Multi-task Supervised Pre-training for Natural Language Generation

Tianyi Tang[1,4] **Junyi Li**[1,3] **Wayne Xin Zhao**[1,4 ✉] **Ji-Rong Wen**[1,2,4]
[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] School of Information, Renmin University of China
[3] DIRO, Université de Montréal
[4] Beijing Key Laboratory of Big Data Management and Analysis Methods
`steventianyitang@outlook.com` `lijunyi@ruc.edu.cn` `batmanfly@gmail.com`

Abstract

Pre-trained language models (PLMs) have achieved notable success in natural language generation (NLG) tasks. Up to now, most of the PLMs are pre-trained in an unsupervised manner using large-scale general corpus. In the meanwhile, an increasing number of models pre-trained with less labeled data showcase superior performance compared to unsupervised models. Motivated by the success of supervised pre-training, we propose **M**ulti-task super**V**ised **P**re-training (**MVP**) for natural language generation. For pre-training the text generation model MVP, we collect a labeled pre-training corpus from 45 datasets over seven generation tasks. For each task, we further pre-train specific soft prompts to stimulate the model capacity in performing a specific task. Extensive experiments have demonstrated the effectiveness of our supervised pre-training in a number of NLG tasks, and our general methods achieve state-of-the-art performance on 12 of 17 datasets.

1 Introduction

Natural language generation (NLG, also known as text generation) is a crucial capacity for language intelligence, which aims to generate texts that are credible and readable to humans [41]. Since the emergence of the *pre-training—fine-tuning* paradigm, pre-trained language models (PLMs) have dominated the mainstream approaches for NLG, and extensive evidence shows that PLMs can produce highly fluent texts by a large model pre-trained on massive text corpus.

Till now, the majority of PLMs are pre-trained in an unsupervised (self-supervised) manner, based on large-scale general corpus. The basic idea is to leverage intrinsic data correlations as supervision signals of pre-training objectives. For example, T5 utilizes the C4 corpus of approximately 750GB as the pre-training corpus and employs a denoising objective that enforces the model to recover corrupted text spans successively [114]. Pre-trained on unlabeled text data, models can capture certain types of semantic knowledge (*e.g.,* knowledge facts) and generalize to new tasks to some extent [59]. However, unsupervised pre-training may incorporate irrelevant or noisy information that affects the performance of downstream tasks [40]. Moreover, unsupervised pre-training causes models to acquire knowledge at a slower rate as model size increases [159].

In the meanwhile, more and more large-scale labeled datasets have become accessible [29, 85]. There is growing evidence that pre-training with labeled data can further improve the performance of PLMs, both in the field of computer vision [54, 36] and natural language processing [97, 85]. These promising developments motivate us to consider supervised pre-training of language generation models with labeled data. The advantages of supervised pre-training for natural language generation

✉ Corresponding author

Preprint. Under review.

are at least twofold. First, it enables the explicit learning of task-specific characteristics or semantics, which is typically infeasible by learning from the general text-to-text relationship as in unsupervised pre-training. Second, supervised pre-training can alleviate the discrepancy between unsupervised pre-training and supervised fine-tuning [85]. It has been demonstrated that supervised pre-trained models can achieve competitive or superior performance compared to their unsupervised counterparts [27, 94], even with significantly less labeled data.

Inspired by the recent success of supervised pre-training, we propose **M**ulti-task super**V**ised **P**re-training (**MVP**) for natural language generation by leveraging a variety of supervised text generation datasets. Specially, we collect large-scale labeled pre-training corpus, consisting of 32 million examples from 45 datasets over seven generation tasks. Since recent research shows that an extensive scale of multi-task pre-training [3, 6] is the key to generalizing to new tasks for large PLMs, we combine these labeled datasets for multi-task pre-training.

To develop the text generation model, we adopt a Transformer-based [139] sequence-to-sequence model as the pre-training backbone. However, different tasks may "neutralize" the ability learned through other tasks [53]. To mitigate this potential issue, we learn task-specific soft prompts based on the MVP model. Following the structure of Prefix-tuning [79], our prompts are inserted in a layer-wise way. Task-specific pre-training enables prompts to "store" specialized knowledge in the corresponding task and stimulate the MVP model's capacity in performing such a task.

In this paper, we mainly investigate the following research questions:

- *How does supervised pre-training perform for NLG tasks?* Our supervised pre-trained MVP can effectively learn task-specific knowledge during pre-training, compared to unsupervised pre-trained BART. In full tuning experiments, the proposed MVP model with task-specific prompts achieves state-of-the-art performance on 12 out of 17 datasets. In parameter-efficient tuning experiments, only with tuned prompts, our frozen MVP model is superior to the frozen BART, which further verifies the importance of supervised pre-training.
- *Can supervised pre-trained models generalize to unseen tasks?* To examine the generalizability of our model, we conduct experiments on unseen language generation and understanding tasks. The experimental results demonstrate that our supervised MVP model has a strong generalization ability for unseen tasks. In the meantime, integrating MVP with existing task-specific methods yields superior performance compared to BART-based counterparts, indicating that MVP can also be utilized to enhance existing methods (*e.g.,* parameter initialization).

For reproducing and reusing our models, we have released our models (*e.g.,* MVP, task-specific prompts and multi-task variants), intermediate results (*e.g.,* the generated texts), and codes for pre-training and fine-tuning at the link `https://github.com/RUCAIBox/MVP`.

2 Related Work

Pre-trained Models. Pre-trained models have achieved exceptional success in a wide range of tasks, and the majority of them are pre-trained in an unsupervised manner [112, 31, 74, 114]. For example, with large-scale unsupervised plain texts as pre-training corpus, GPT series [112, 113, 14] employ language modeling as the pre-training task, *i.e.,* predicting the next token conditioned on previous tokens; BART [74] learns to recover the original text from corrupted text which has been altered by arbitrary noise transformations. GPT-3 and BART utilize 570GB and 160GB of plain text as pre-training corpus, respectively. In the meanwhile, the computer vision community benefits a lot from the labeled dataset ImageNet [29]. Influential models, such as ResNet [54], EfficientNet [136], and ViT [36], leverage ImageNet for pre-training. Inspired by the success of leveraging labeled data for pre-training, machine translation researchers explore supervised pre-training [97, 85, 155, 108]. [85] attempt to pre-train a translation model mRASP with parallel data in multiple languages. Despite having much fewer pre-trained data, mRASP still achieves better performance than translation models pre-trained in an unsupervised manner [27, 94]. In this paper, we propose to pre-train a universal NLG model with a large-scale collection of labeled datasets (23GB).

Multi-task Learning. Our supervised pre-training process can also be viewed as a formulation of multi-task learning (MTL), a method that combines multiple tasks into a single training process [26, 150]. A model trained with MTL can benefit from helpful knowledge of relevant tasks, resulting in

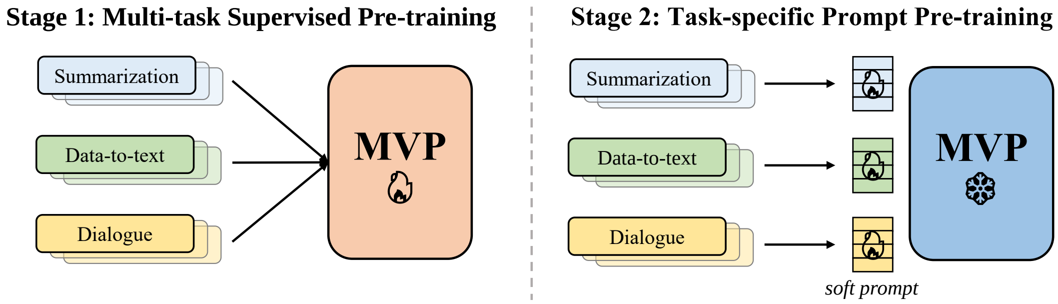

Figure 1: The overview of the pre-training process of our MVP model and task-specific prompts. In the first stage, we utilize labeled datasets from seven tasks to jointly pre-train the model. In the second stage, we freeze the MVP and pre-train specific prompts for each task using intra-task datasets.

improved performance [98, 132]. Recently, MT-DNN [92] and Muppet [3] collect tens of datasets in the multi-task procedure and achieve better performance in downstream tasks. The *pre-finetuning* proposed in Muppet [3] shares a similar idea as our multi-task supervised pre-training. [6] further combine the denoising pre-training task of T5 and multi-task learning to pre-train a new model ExT5. MTL has also contributed to sub-fields of text generation, such as open-ended dialogue system [160, 9], task-oriented dialogue system [131], text style transfer [16], and question answering [63]. At the same time, researchers explore the transferability of models trained on multi-task datasets [99]. FLAN [145], T0 [125], and ZeroPrompt [153] investigate the zero-shot generalization abilities of large PLMs trained on numerous datasets with well-designed prompts. [157] develop a benchmark CrossFit to study the few-shot learning ability of models.

Prompt Learning. Prompt learning is a thriving method in the field of natural language processing. Prompt learning converts fine-tuning text into a format similar to pre-training to leverage implicit pre-training knowledge and alleviate the discrepancy between pre-training and fine-tuning [89]. GPT-2 [113] and T5 [114] add human-written task prompts (instructions) to the input text. For instance, T5 pre-pends "*Summarize:*" to the input document for summarization tasks. GPT-3 [14] further combines several demonstrations to input to learn task patterns, which is called *in-context learning*. Some researchers also design elaborate prompts or demonstrations for each task and dataset and investigate their effectiveness and robustness [145, 125, 153, 99]. Even so, whether models can truly understand the semantic meanings of prompts looks worthy of further investigation [144]. To overcome the constraints of manually constructed prompts, researchers develop continuous (soft) prompts that can be optimized in the continuous space [73, 90, 111, 137]. Prefix-tuning [79] and P-tuning v2 [91] increase the number of parameters in prompts and employ prompting in each Transformer layer. Considering the random initialization of soft prompts, [50] propose PPT to pre-train continuous prompts using unlabeled data. SPoT [141], PTG [77], and UnifiedSKG [152] learn the prompts on related tasks and transfer them to new tasks.

3 The MVP Model

This section introduces our **MVP** model: a **M**ulti-task super**V**ised **P**re-trained model for natural language generation. We first collect 45 labeled datasets from diverse NLG tasks as our pre-training corpus and unify input and output in a text-to-text format. Then, we pre-train our MVP model using the pre-training corpus, *i.e.,* a mixture of labeled data from various tasks. We further learn the task-specific prompts to stimulate the MVP model in performing a certain task.

3.1 Data Collection

The natural language generation (NLG) task aims to generate a sequence of tokens $Y = (y_1, y_2, \dots, y_n)$ conditioned on input data $\mathcal{X}$ (*e.g.,* one or more pieces of text and structured data) [78]. Typically, NLG tasks are categorized according to the data format of $\mathcal{X}$ and Y. For example, text summarization condenses a long document into a brief text containing essential information; data-to-text generation produces descriptive text about structured input; and dialogue system creates pertinent responses given multiple dialog turns.

In this paper, we collect 45 labeled datasets from 7 representative NLG tasks [2], including data-to-text generation, open-ended dialogue system, question answering, question generation, task-oriented dialogue system, text summarization, and story generation. These datasets come from various domains and are of different sizes. Some datasets are elaborately hand-crafted and thus relatively small in size, while others are created for large-scale weak supervision. Despite originating from various tasks, these diverse labeled datasets contain rich task-specific supervision signals for establishing global sequence-to-sequence mapping relations. The detailed descriptions and statistics of these tasks and datasets for pre-training can be found in Table 6 in Appendix B.1.

To adapt these datasets for multi-task pre-training, we transform all tasks into a unified text-to-text format, *i.e.,* converting different input data $\mathcal{X}$ into text format. For instance, we linearize structured data (*e.g.,* knowledge graph or table) by concatenating triples or key-value pairs using the special token "`[SEP]`" for data-to-text generation, and we utilize the special token "`[X_SEP]`" to separate answer and paragraph for question generation. The transformed input format of each task can be found in Appendix E. To enrich our datasets, we reverse the input and output of dual tasks for obtaining new datasets (*e.g.,* story generation and summarization, question generation, and question answering). We also eliminate pre-training examples overlapping with evaluation data to avoid data leakage (more details in Appendix B.2). Finally, we obtain a 25GB supervised pre-training corpus containing 32M examples (*i.e.,* pairs of $\langle \mathcal{X}, Y \rangle$).

In addition, we do not include the datasets of paraphrase generation, text style transfer, and natural language understanding (NLU) during the pre-training phase. We leave them to evaluate the generalization ability of our methods. We reserve some common datasets (*e.g.,* CNN/DailyMail [127] and XSum [104]) for downstream fine-tuning and do not use them during pre-training. The details of the datasets used for fine-tuning evaluation are provided in Table 7 in Appendix B.1.

3.2 Model Architecture

We then pre-train our MVP model in a text-to-text format based on the supervised pre-training corpus. Our model is built upon a Transformer encoder-decoder architecture [139]. In the first stage, our model is trained using a mixture of NLG datasets with human-written instructions [3]. For example, we use "*Summarize:*" as the prefix instruction for summarization tasks. This process is similar to instruction tuning in FLAN [145]. The difference is that we only keep one instruction for each task. As shown by [125], a single instruction can typically lead to positive performance, whereas adding multiple instructions does not always improve performance and requires considerable human effort. The detailed instructions for each task can be found in Appendix E.

However, it is not easy to conduct effective multi-task training since these tasks may compete with one another and thus "blur out" the features extracted by individual tasks [53]. To address this issue, in the second stage we freeze our model and train a set of task-specific soft prompts (*i.e.,* continuous vectors) using a mixture of corresponding intra-task datasets (*i.e.,* datasets under the same task [4]). These soft prompts, which are not shared between tasks, encode the task-specific semantic knowledge to alleviate the blurring-out problem.

Specifically, we employ the standard Transformer encoder-decoder [139] as our backbone, a universal architecture suited for both NLU and NLG tasks. Compared to decoder-only architectures such as GPT-3 [14] and prefix LMs such as UniLM [35], the encoder-decoder architecture is more effective for text generation tasks [114]. As for task-specific prompts, we insert continuous vectors at each Transformer layer, following Prefix-tuning [79]. Compared to prompt tuning [73], which only adds trainable embeddings to the input word embeddings, layer-wise prompting is more effective and stable [91], especially for NLG tasks [79].

[2]We do not consider incorporating machine translation tasks while focusing on English-only tasks in this work.

[3]To avoid ambiguity with continuous prompts, we designate human-written prompts as *instructions*.

[4]For instance, we train summarization-specific prompts using summarization datasets (*e.g.,* Newsroom [48], WikiHow [66], and MSNews [87]).

3.3 Training Details

Our MVP backbone conforms to the model size of $\text{BART}_{\text{LARGE}}$ [74], *i.e.,* a Transformer with 12 layers in both the encoder and decoder. The hidden size is $1,024$, and the inner hidden size of the feed-forward network is $4,096$. We employ the same byte-level byte-pair-encoding tokenizer as BART, and the vocabulary size is $50,267$. The whole backbone consists of approximately 406M parameters and is initialized by $\text{BART}_{\text{LARGE}}$. We pre-train the model with a batch size of $8,192$. We leverage our collected 45 datasets for multi-task learning using a temperature-scaled mixing strategy [114] with mixing rate $T = 2$ to mitigate the disparity between tasks and datasets.

The task-specific prompts follow the schema as Prefix-tuning [79], which prepends trainable continuous vectors to the keys and values of the multi-head attention module at each layer. The prompt length is set to 100, and we utilize the MLP reparameterization function with a medium hidden size of 800 to improve the training robustness and performance [79]. Hence, every task prompts have approximately 62M parameters. Then, we freeze the backbone and train seven groups of task-specific prompts, each of which corresponds to a different task. The batch size is set to $8,192$, and we leverage the mixing strategy with rate $T = 2$.

In the two stages, the maximum length of the input and output sequences is both set to 1024 for supporting examples to contain more tokens. We optimize the model with a constant learning rate of 3×10^{-5} using standard sequence-to-sequence cross-entropy loss. We apply AdamW [95] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1 \times 10^{-6}$ to improve training stability [93]. The weight decay is 0.1. For the test, we select the checkpoint with the highest validation performance. All the experiments utilize 32 NVIDIA Tesla V100 32GB GPU cards (4 nodes) on Ubuntu 18.04.5 LTS. We implement the code using the PLM library `Hugging Face` [149] and text generation toolkit `TextBox` [76].

In summary, we pre-train a 406M backbone and seven groups of 62M task-specific prompts. For downstream tasks, we can either directly utilize the MVP backbone (406M) or integrate the backbone with one group of task-specific prompts (468M).

4 Experiment Results

In this section, we mainly investigate the first research question we proposed: *How does supervised pre-training perform for NLG tasks?* Specifically, we apply our MVP model to new datasets from pre-trained generation tasks under *full tuning* and *parameter-efficient tuning* settings.

Under the full tuning setting, we fine-tune the entire model (including the backbone MVP and prompts), while for parameter-efficient tuning, we only fine-tune prompts but freeze the parameter weights of MVP. We apply the AdamW optimizer with default hyper-parameters and batch size of 192 in both settings. The learning rate is set to 3×10^{-5}. We optimize the model by the seq2seq loss with label smoothing [135] factor 0.1. We utilize the checkpoint with the best validation performance for test set inference. During inference, we set the beam size to 5 and the no repetitive ngram size to 3. For evaluation, we leverage the automatic generation metrics BLEU [109], ROUGE [84], and METEOR [7] to measure the quality of the generated text and employ Distinct [75] to evaluate its diversity. Details regarding fine-tuning and evaluation can be found in Appendix C.

In all of our experiments, we report the mean and standard deviation of our test set result over three random seeds $2,020$, $2,021$, and $2,022$. We also reproduce the results of baselines (*e.g.,* BART) to compare them with our models under the same configuration.

4.1 Full Tuning Performance

For full tuning, we select a number of widely-used datasets for each seen task. In this setting, we consider three competitive baselines (including MVP) and three model variants with different prompts. Table 1 shows the performance of one dataset for each task. Tables 8 and 9 in Appendix D list the results of other datasets and the GEM benchmark [43].

For the baselines, we compare three backbones based on different pre-training strategies:

- **$\text{BART}_{\text{LARGE}}$ [74]**: BART is a widely-used PLM for natural language generation. We use it to initialize the MVP model during pre-training.

Table 1: The main results on seven seen tasks under full tuning settings. The best and second-best results among all the methods are marked in **bold** and underlined, respectively. QG, QA, and TODS are short for question generation, question answering, and task-oriented dialogue system, respectively. B, R, D, and ME denote BLEU, ROUGE, Distinct, and METEOR. "–" means the SOTA paper does not compute the corresponding result. These setups and abbreviations are the same below. [a] [120] [b] [62] [c] [8] [d] [87] [e] [151] [f] [51] [g] [21] [h] [55]

	Summarization (CNN/DM)			**Data-to-text (WebNLG)**			**QG (SQuAD)**			**QA (CoQA)**	
	R-1	R-2	R-L	B-4	ME	R-L	B-4	ME	R-L	F1	EM
SOTA	**47.16**[a]	**22.55**	**43.87**	66.14[b]	47.25	76.10	**25.97**[c]	**27.33**	**53.43**	73.0[d]	84.5[e]
BART	$\underline{44.47}_{0.10}$	$21.50_{0.14}$	$\underline{41.35}_{0.08}$	$67.33_{0.06}$	$47.78_{0.07}$	$\underline{76.83}_{0.04}$	$25.08_{0.13}$	$26.73_{0.18}$	$52.55_{0.07}$	$74.00_{0.17}$	$84.07_{0.21}$
Single	$44.35_{0.16}$	$21.51_{0.11}$	$41.21_{0.19}$	$67.40_{0.24}$	$47.80_{0.03}$	$76.72_{0.20}$	$25.77_{0.15}$	$27.14_{0.09}$	$53.02_{0.11}$	$74.90_{0.20}$	$84.57_{0.06}$
MVP	$44.45_{0.05}$	$21.44_{0.12}$	$41.34_{0.08}$	$67.32_{0.10}$	$\mathbf{47.94}_{0.13}$	$76.70_{0.26}$	$\underline{25.91}_{0.07}$	$\underline{27.22}_{0.10}$	$\underline{53.08}_{0.16}$	$\underline{75.50}_{0.20}$	$\underline{85.07}_{0.21}$
MVP+R	$44.46_{0.07}$	$\underline{21.59}_{0.11}$	$41.31_{0.06}$	$\underline{67.41}_{0.12}$	$\underline{47.92}_{0.03}$	$76.82_{0.14}$	$25.31_{0.13}$	$26.69_{0.04}$	$52.68_{0.03}$	$75.17_{0.21}$	$84.83_{0.12}$
MVP+S	$44.30_{0.07}$	$21.45_{0.03}$	$41.15_{0.08}$	$\mathbf{67.63}_{0.13}$	$47.90_{0.05}$	$\mathbf{76.87}_{0.16}$	$25.62_{0.02}$	$26.98_{0.10}$	$52.98_{0.09}$	$\mathbf{75.70}_{0.00}$	$\mathbf{85.40}_{0.10}$
MVP+M	$44.33_{0.03}$	$21.42_{0.07}$	$41.19_{0.03}$	$67.15_{0.14}$	$47.72_{0.21}$	$76.55_{0.04}$	$25.29_{0.10}$	$26.69_{0.12}$	$52.81_{0.12}$	$75.00_{0.10}$	$84.73_{0.06}$
	Story Generation (ROCStories)				**Open-ended Dialogue (PersonaChat)**				**TODS (MultiWOZ)**		
	B-1	B-2	D-1	D-4	B-1	B-2	D-1	D-2	B-4	Success	Inform
SOTA	33.4[f]	15.4	–	69.3	48.2[g]	39.9	1.5	9.4	**20.50**[h]	**85.30**	**94.40**
BART	$\underline{33.79}_{0.13}$	$\underline{15.78}_{0.21}$	$3.43_{0.17}$	$78.76_{2.15}$	$\mathbf{49.58}_{1.12}$	$39.24_{0.90}$	$1.44_{0.09}$	$8.89_{0.57}$	$20.17_{0.63}$	$75.40_{1.22}$	$84.40_{1.15}$
Single	$33.42_{0.22}$	$15.47_{0.17}$	$\mathbf{3.61}_{0.04}$	$\mathbf{81.06}_{0.63}$	$\underline{49.57}_{0.03}$	$\underline{40.32}_{0.11}$	$1.31_{0.07}$	$7.90_{0.67}$	$\underline{20.44}_{0.63}$	$74.13_{1.33}$	$82.23_{0.92}$
MVP	$\mathbf{33.96}_{0.08}$	$\mathbf{15.96}_{0.05}$	$3.17_{0.15}$	$76.11_{1.38}$	$49.56_{0.44}$	$\mathbf{40.41}_{0.10}$	$1.55_{0.07}$	$10.20_{0.46}$	$20.34_{0.37}$	$75.47_{0.40}$	$84.07_{0.15}$
MVP+R	$33.19_{0.20}$	$15.47_{0.08}$	$3.14_{0.12}$	$75.14_{0.96}$	$47.68_{0.16}$	$39.90_{0.10}$	$\underline{1.68}_{0.13}$	$\underline{10.80}_{1.04}$	$19.23_{0.15}$	$75.00_{0.70}$	$83.40_{0.87}$
MVP+S	$33.53_{0.03}$	$15.48_{0.10}$	$\underline{3.56}_{0.10}$	$\underline{80.29}_{0.60}$	$47.70_{0.34}$	$39.91_{0.21}$	$\mathbf{1.77}_{0.04}$	$\mathbf{11.72}_{0.23}$	$19.79_{0.47}$	$\underline{77.23}_{0.21}$	$\underline{85.13}_{1.06}$
MVP+M	$33.42_{0.02}$	$15.52_{0.06}$	$3.13_{0.36}$	$76.10_{3.56}$	$47.94_{0.07}$	$40.00_{0.07}$	$1.50_{0.17}$	$9.44_{1.54}$	$19.54_{0.32}$	$75.00_{2.36}$	$83.23_{2.01}$

- **Single-task pre-training (Single)**: We individually train a single model for each task using intra-task datasets following the same pre-training settings in multi-task training. For instance, we pre-train a summarization model using summarization datasets (*e.g.,* Newsroom, WikiHow, and MSNews). Therefore, we have seven single-task pre-trained models in total.
- **Multi-task pre-training (MVP)**: This is our MVP model, which is trained on a mixture of labeled datasets from seven tasks.

For the prompt-based variants, we integrate our MVP model with three different prompts:

- **Randomly initialized prompts (MVP+R)**: The prompts are randomly initialized without pre-training.
- **Single-task pre-trained prompts (MVP+S)**: This is the primary method in our paper, as introduced in Section 3.2. We pre-train specific prompts for each task.
- **Multi-Task pre-trained prompts (MVP+M)**: We only pre-train one group of prompts, using the same mixed datasets as the backbone pre-training.

Besides these baselines and variants, we further collect the state-of-the-art (SOTA) results from their original papers for comparison. From the results in Table 1, we can see that:

First, supervised pre-training (*i.e.,* Single and MVP) achieves better performance than unsupervised pre-trained model BART. For example, in question answering, Single and MVP achieve $+1.17\%$ and $+1.70\%$ gain on F1 score compared with BART. This observation demonstrates the effectiveness of our supervised pre-training. With labeled datasets, supervised pre-training enables the model acquire more task-specific information during pre-training, thus leading to improved results on downstream tasks. Regarding single-task (Single) and multi-task pre-training (MVP), our MVP model outperforms single-task counterparts on 15 of 22 metrics (combing all the metrics for different tasks). This result indicates that the proposed multi-task learning approach can enhance single-task performance by learning transferable semantic information across tasks.

Second, task-specific prompt learning is effective for some tasks, such as data-to-text generation and question answering. MVP with single-task prompt pre-training (MVP+S) consistently outperforms the other two methods. Compared with MVP+M, MVP+S can alleviate the "blurring-out" issue of multi-task learning, *i.e.,* different tasks may "neutralize" the ability learned by others [53]. Pre-trained on intra-task datasets, task-specific prompts can acquire specialized knowledge of each task and stimulate the capacity of the MVP model in performing a certain task. On the other hand, we find that

Table 2: The main results on seven seen tasks under parameter-efficient settings. We also include the results of BART and MVP under full tuning (denoted as FT) settings for comparison.

	Summarization (CNN/DM)			**Data-to-text (WebNLG)**			**QG (SQuAD)**			**QA (CoQA)**	
	R-1	R-2	R-L	B-4	ME	R-L	B-4	ME	R-L	F1	EM
FT BART	$44.47_{0.10}$	$21.50_{0.14}$	$41.35_{0.08}$	$67.33_{0.06}$	$47.78_{0.07}$	$76.83_{0.04}$	$25.08_{0.13}$	$26.73_{0.18}$	$52.55_{0.07}$	$74.00_{0.17}$	$84.07_{0.21}$
FT MVP	$44.45_{0.05}$	$21.44_{0.12}$	$41.34_{0.08}$	$67.32_{0.10}$	$47.94_{0.13}$	$76.70_{0.26}$	$25.91_{0.07}$	$27.22_{0.10}$	$53.08_{0.16}$	$75.50_{0.20}$	$85.07_{0.21}$
BART+R	$42.47_{0.16}$	$19.79_{0.15}$	$39.10_{0.17}$	$65.15_{0.66}$	$46.74_{0.18}$	$76.06_{0.34}$	$24.12_{0.16}$	$25.95_{0.11}$	$51.84_{0.26}$	$70.80_{1.31}$	$81.10_{1.21}$
MVP+R	$42.88_{0.10}$	$\underline{20.25}_{0.04}$	$\underline{39.66}_{0.06}$	$66.02_{0.58}$	$47.10_{0.13}$	$75.75_{0.39}$	$25.07_{0.04}$	$26.49_{0.13}$	$52.63_{0.06}$	$73.97_{0.64}$	$83.93_{0.49}$
MVP+S	$\underline{42.89}_{0.11}$	$20.21_{0.08}$	$39.58_{0.11}$	$\underline{66.68}_{0.08}$	$\mathbf{47.45}_{0.04}$	$\mathbf{76.23}_{0.12}$	$\mathbf{25.24}_{0.04}$	$\mathbf{26.63}_{0.05}$	$\underline{52.68}_{0.05}$	$\mathbf{75.03}_{0.06}$	$\mathbf{84.63}_{0.15}$
MVP+M	$\mathbf{43.03}_{0.06}$	$\mathbf{20.34}_{0.05}$	$\mathbf{39.74}_{0.07}$	$\mathbf{66.70}_{0.35}$	$\underline{47.41}_{0.23}$	$\underline{76.17}_{0.28}$	$\underline{25.21}_{0.03}$	$\underline{26.50}_{0.02}$	$\mathbf{52.68}_{0.03}$	$\underline{74.43}_{0.06}$	$\underline{84.00}_{0.10}$

	Story Generation (ROCStories)				**Open-ended Dialogue (PersonaChat)**				**TODS (MultiWOZ)**		
	B-1	B-2	D-1	D-4	B-1	B-2	D-1	D-2	B-4	Success	Inform
FT BART	$33.79_{0.13}$	$15.78_{0.21}$	$3.43_{0.17}$	$78.76_{2.15}$	$49.58_{1.12}$	$39.24_{0.90}$	$1.44_{0.09}$	$8.89_{0.57}$	$20.17_{0.63}$	$75.40_{1.22}$	$84.40_{1.15}$
FT MVP	$33.96_{0.08}$	$15.96_{0.05}$	$3.17_{0.15}$	$76.11_{1.38}$	$49.56_{0.44}$	$40.41_{0.10}$	$1.55_{0.07}$	$10.20_{0.46}$	$20.34_{0.37}$	$75.47_{0.40}$	$84.07_{0.15}$
BART+R	$32.47_{0.31}$	$14.96_{0.22}$	$2.85_{0.00}$	$69.19_{0.30}$	$44.88_{1.21}$	$36.97_{2.27}$	$1.30_{0.05}$	$6.43_{0.39}$	$17.63_{0.28}$	$65.00_{2.52}$	$72.33_{2.73}$
MVP+R	$32.61_{0.30}$	$15.08_{0.20}$	$\underline{2.96}_{0.06}$	$\mathbf{70.70}_{0.36}$	$\mathbf{47.24}_{0.47}$	$\underline{37.62}_{1.39}$	$1.31_{0.04}$	$6.86_{0.15}$	$\underline{18.86}_{0.50}$	$66.93_{3.18}$	$74.60_{4.23}$
MVP+S	$\mathbf{32.87}_{0.06}$	$\underline{15.08}_{0.11}$	$2.85_{0.11}$	$\underline{69.91}_{1.06}$	$46.90_{0.23}$	$\mathbf{39.42}_{0.19}$	$\underline{1.33}_{0.02}$	$\underline{6.94}_{0.36}$	$18.80_{0.54}$	$\underline{67.83}_{3.65}$	$\underline{74.20}_{3.70}$
MVP+M	$\underline{32.69}_{0.07}$	$\mathbf{15.25}_{0.04}$	$\mathbf{2.97}_{0.04}$	$69.81_{0.66}$	$\underline{46.95}_{0.32}$	$37.41_{1.92}$	$\mathbf{1.34}_{0.05}$	$\mathbf{7.00}_{0.25}$	$\mathbf{19.22}_{0.25}$	$\mathbf{69.63}_{2.21}$	$\mathbf{76.23}_{3.02}$

equipping our MVP model with prompts decreases its performance on certain tasks, such as question generation, from 25.91% to 25.29% on BLEU-4. We speculate that this is due to the different degrees of convergence between the pre-trained backbone and prompts, *i.e.,* the prompts have been trained to perform specific tasks, while the backbone has been trained to be more applicable to different tasks.

Finally, our MVP models (MVP and MVP+R/S/M) produce comparable or better performance compared with current SOTA approaches on data-to-text generation, question answering, story generation, and open-ended dialogue tasks, which shows a strong text generation capability. As for the remaining three tasks, the SOTA models incorporate specific techniques tailored to tasks (*e.g.,* re-ranking framework [120], self-training [8], and various task-specific objectives [55]), which yield better performance than our models. In contrast, the results of our models are still encouraging, considering we adopt only a simple architecture and a unified objective.

4.2 Parameter-Efficient Tuning Performance

In the lightweight fine-tuning setting, we only tune the prompts while freezing the backbone MVP model. We compare our model to three baselines:

- **Prefix-tuning** [79]: Prefix-tuning is a popular prompt-based lightweight tuning method for text generation. We employ $\text{BART}_{\text{LARGE}}$ as its backbone, denoted as **BART+R**.
- **Only tuning randomly initialized prompts (MVP+R)**: This baseline only tunes the randomly initialized prompts of MVP+R (Section 4.1), and it shares a similar idea with Prefix-tuning.
- **Only tuning single-task pre-trained prompts (MVP+S)**: This baseline only tunes the single-task pre-trained prompts of MVP+S (Section 4.1). These pre-trained prompts contain task-specific knowledge and can serve as a better initialization than the random ones.
- **Only tuning multi-task pre-trained prompts (MVP+M)**: This baseline tunes the multi-task pre-trained prompts of MVP+M (Section 4.1). Such an idea has been used in SPoT [141].

From the experimental results in Table 2, we can see that:

First, the good performance of thr MVP model in lightweight settings further demonstrates the effectiveness of supervised pre-training. By comparing two randomly initialized prompting methods (BART+R and MVP+R), we can see that MVP+R achieves superior performance to BART+R due to its multi-task supervised backbone. Furthermore, when initialized with pre-trained prompts, MVP+S and MVP+M achieve improved results over MVP+R, which is consistent with the findings of SPoT [141]. When compared with MVP+M, MVP+S performs marginally better, indicating that task-specific prompts are useful to improve the model in specific generation tasks.

Surprisingly, our lightweight MVP+S can even outperform fully tuned BART on question generation and question answering tasks, showcasing the effectiveness of the proposed supervised pre-training approach. Another note is that lightweight prompting methods [73, 141] that work on NLU tasks

Table 3: The main results of unseen NLG tasks, including text style transfer and paraphrase generation. Accuracy is calculated by a pre-trained TextCNN to evaluate the style strength, and HM denotes the harmonic mean of BLEU-4 and style accuracy [70]. [a] [133] [b] [19] [c] [70]

	Paraphrase Generation (Quora)				
	B-4	R-1	R-2	R-L	ME
SOTA	47.3[a]	73.3	54.1	75.1	49.7
BART + AESOP	$48.35_{0.70}$	$74.16_{0.47}$	$55.25_{0.74}$	$75.84_{0.42}$	$50.60_{0.49}$
MVP + AESOP	$\mathbf{49.86}_{0.21}$	$\mathbf{74.93}_{0.05}$	$\mathbf{56.55}_{0.13}$	$\mathbf{76.56}_{0.09}$	$\mathbf{52.27}_{0.21}$

	Style Transfer (GYAFC E&M)			Style Transfer (GYAFC F&R)		
	B-4	Accuracy	HM	B-4	Accuracy	HM
SOTA	76.52[b]	93.7[c]	83.9[c]	**80.29**[b]	92.0[c]	85.2[c]
BART + SC & BLEU	$76.93_{0.55}$	$94.37_{0.87}$	$84.74_{0.05}$	$80.11_{0.29}$	$92.29_{0.37}$	$85.77_{0.10}$
MVP + SC & BLEU	$\mathbf{77.01}_{0.20}$	$\mathbf{94.66}_{0.36}$	$\mathbf{84.92}_{0.04}$	$79.70_{0.25}$	$\mathbf{93.07}_{0.28}$	$\mathbf{85.87}_{0.27}$

Table 4: The main results of NLU tasks on the GLUE benchmark. We evaluate the results on the official website `https://gluebenchmark.com/`.

	CoLA Matt.	SST-2 Accuracy	MRPC F1/Accuracy	STS-B P/S Corr.	QQP F1/Accuracy	MNLI m./mm.	QNLI Accuracy	RTE Accuracy	Average
BART	$\mathbf{60.30}_{3.20}$	$96.30_{0.10}$	$90.47_{1.46}$ / $86.70_{2.10}$	$90.97_{0.15}$ / $90.30_{0.10}$	$73.03_{0.23}$ / $89.87_{0.12}$	$\mathbf{90.03}_{0.21}$ / $\mathbf{89.27}_{0.15}$	$94.60_{0.10}$	$79.83_{2.76}$	$85.17_{1.05}$
MVP	$59.87_{1.04}$	$\mathbf{96.43}_{0.32}$	$\mathbf{92.07}_{0.25}$ / $\mathbf{89.43}_{0.29}$	$\mathbf{91.37}_{0.15}$ / $\mathbf{90.90}_{0.35}$	$\mathbf{73.20}_{0.10}$ / $\mathbf{90.13}_{0.12}$	$89.70_{0.00}$ / $88.73_{0.25}$	$\mathbf{95.10}_{0.26}$	$\mathbf{82.87}_{0.58}$	$\mathbf{85.88}_{0.37}$

cannot achieve competitive performances when compared to full tuning methods on NLG tasks. Therefore, it is necessary to design lightweight tuning models specially for generation tasks.

5 Generalization Ability

In this section, we concentrate on the second question: *Can supervised pre-trained models generalize to unseen tasks?* We employ our MVP model to unseen NLG and NLU tasks to verify the generalizability of our model.

Generalization to Unseen NLG Tasks. According to [30], every NLG task can be considered as compression (*e.g.,* summarization), transduction (*e.g.,* translation), or creation (*e.g.,* story generation) tasks. Since we do not include any transduction tasks during pre-training, we evaluate our MVP model using two unseen transduction NLG tasks: paraphrase generation and text style transfer.

We follow the methods which can achieve SOTA performance for these two tasks, *i.e.,* AESOP [133] for paraphrase generation and SC & BLEU [70] for text style transfer. We replace their backbone BART with our MVP model for comparison. The experimental setup remains the same as described in Section 4, and details are listed in Appendix C. From the results in Table 3, we can see that our model outperforms BART on 10 out of 11 metrics and achieves a new SOTA performance, which verifies the strong generalizability of our model. In addition, combining our MVP model with existing SOTA methods yields better performance than BART-based counterparts. This observation suggests that our MVP model can adapt to existing methods effectively by providing superior parameter initialization.

Generalization to Unseen NLU Tasks. Although MVP is designed specially for NLG tasks, we also evaluate its performance on unseen NLU tasks using the widely-used GLUE benchmark [142]. We compare our model to $\text{BART}_{\text{LARGE}}$ using its original sequence classification method [74]. The detailed settings can be found in Appendix C. According to the results presented in Table 4, our MVP model outperforms BART on 9 of 12 metrics and has superior overall performance. This result indicates the strong generalization ability of our MVP model on unseen NLU tasks. It further demonstrates that the model learned through supervised pre-training is not limited to generation tasks and can instead improve the overall semantic representations.

6 Discussion

In this section, we discuss this work and compare it to existing methods.

Supervised Pre-training for NLG Models. Unsupervised pre-training has been extensively investigated for natural language understanding [31, 93] and generation [74, 14], aiming to learn universal language representations that can adapt to a variety of tasks. In spite of its effectiveness, increasing evidence indicates that a general solution cannot always generate the best task-specific representations in comparison to a supervised approach [85, 6]. In the meantime, the growing availability of labeled data makes it feasible to conduct large-scale supervised pre-training. Inspired by the success of

Table 5: Comparison of our paper with existing works. MTL model denotes whether the work utilizes multi-task learning to train a backbone model, similar to MTL prompts. Usage mode is the primary way for applying a model to downstream tasks. Open-source refers to whether the work has released models to the public. #NLU and #NLG are the numbers of NLU and NLG tasks for evaluation.

Methods	MTL model	MTL prompts	Usage mode	Open source	#NLU	#NLG
FLAN [145]	✓	✗	zero-shot	✗	9	2
T0 [125]	✓	✗	zero-shot	✓	4	0
Muppet [3]	✓	✗	fine-tune	✓	3	1
ExT5 [6]	✓	✗	fine-tune	✗	8	6
SPoT [141]	✗	✓	fine-tune	✗	6	0
MVP (ours)	✓	✓	fine-tune	✓	3	11

supervised pre-training [36, 85], we present an important attempt to pre-train a more capable PLM for NLG tasks using labeled data. Our experiments have shown that such an approach works not only for seen tasks but also for unseen tasks (including NLU tasks), indicating that supervised pre-training is a promising, general, and effective method for various tasks.

Differences with Existing Methods. To the best of our knowledge, existing supervised pre-training works mainly focus on NLU tasks [3, 6] or some specific NLG tasks [85, 108]. Given the superior performance yielded by supervised pre-training approaches, it is important to explore supervised pre-training for deriving both *effective* and *general* NLG models. Our work makes a significant contribution in this direction, achieving SOTA performance with a single model on 12 of 17 datasets. Compared with its strong counterpart ExT5 [6], our MVP model outperforms it in 26 out of 27 metrics (detailed in Appendix D.2). In order to better understand the difference between our paper with previous supervised (multi-task) pre-training works, we present a detailed comparison in Table 5. As we can see, our work conducts the evaluation with the greatest number of NLG tasks, adopts multi-tasking to learn model and task-specific prompts, and makes abundant resources available.

Applicability. To facilitate the use of our MVP model, we have released both the code and the pre-trained models. To apply our model, we consider two different settings for seen and unseen tasks. For seen tasks, one can utilize the MVP model or integrate it with task prompts. For unseen tasks, besides the above methods, one can also pre-train prompts using task-specific labeled data. Also, our model can serve as an effective parameter initialization for adapting to existing methods and diverse tasks, as described in Section 5. Furthermore, we release all the intermediate results (*e.g.,* the generated texts) of our model on evaluation tasks. These results provide valuable data resources for understanding and analyzing the task capacity of PLMs. In addition, the pre-trained task-specific prompts can be used to study the task similarity and their effect on the multi-task pre-training.

Limitations. Despite our efforts to collect as many generation tasks and datasets as possible, we only evaluate the generation quality and generalization ability of our models on a small number of tasks and datasets. The interpretability and robustness of our models require further analysis. Besides, there exists subjectivity when collecting intra-task datasets, albeit our attempts to employ widely-recognized categorizations from the literature. Due to limitation of computing power, we do not study the performance of our method at different model scales. The effectiveness of multi-task pre-training from scratch, similar to ExT5 [6], also merits an in-depth study. Regarding evaluation methods, we only consider basic automatic metrics such as BLEU [109] and ROUGE [84]. However, there is still a certain gap between these metrics and human judgments [124].

7 Conclusion

In this paper, we propose a multi-task supervised pre-trained model MVP with task-specific prompts for NLG tasks. Extensive experiments have demonstrated that: 1) supervised pre-training is beneficial for NLG tasks. Our MVP model outperforms the unsupervised pre-trained model BART on examined tasks and even achieves SOTA performance on 12 out of 17 datasets; 2) supervised pre-trained models have strong generalization ability on unseen generation and understanding tasks. We hope that the open-sourced MVP models will facilitate future work on supervised pre-training and contribute to the advancement of NLG research.

References

[1] URL https://github.com/markriedl/WikiPlots.

[2] Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.278. URL https://aclanthology.org/2021.naacl-main.278.

[3] Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.468. URL https://aclanthology.org/2021.emnlp-main.468.

[4] Huda Alamri, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Jue Wang, Irfan Essa, Dhruv Batra, Devi Parikh, Anoop Cherian, Tim K Marks, et al. Audio visual scene-aware dialog (avsd) challenge at dstc7. *arXiv preprint arXiv:1806.00525*, 2018. URL http://arxiv.org/abs/1806.00525.

[5] Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.424. URL https://aclanthology.org/2020.acl-main.424.

[6] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. Ext5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Vzh1BFUCiIX.

[7] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

[8] Hangbo Bao, Li Dong, Wenhui Wang, Nan Yang, and Furu Wei. s2s-ft: Fine-tuning pretrained transformer encoders for sequence-to-sequence learning. *arXiv preprint arXiv:2110.13640*, 2021. URL https://arxiv.org/abs/2110.13640.

[9] Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.9. URL https://aclanthology.org/2020.acl-main.9.

[10] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

[11] Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. The fifth pascal recognizing textual entailment challenge. In *In Proc Text Analysis Conference (TAC'09*, 2009.

[12] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://aclanthology.org/D13-1160`.

[13] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. URL `http://arxiv.org/abs/2108.07258`.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

[15] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL `https://aclanthology.org/D18-1547`.

[16] Pawel Bujnowski, Kseniia Ryzhova, Hyungtak Choi, Katarzyna Witkowska, Jaroslaw Piersa, Tymoteusz Krumholc, and Katarzyna Beksa. An empirical study on multi-task learning for text style transfer and paraphrase generation. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 50–63, Online, December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-industry.6. URL `https://aclanthology.org/2020.coling-industry.6`.

[17] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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)*, pages 4516–4525, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1459. URL `https://aclanthology.org/D19-1459`.

[18] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL `https://aclanthology.org/S17-2001`.

[19] Kunal Chawla and Diyi Yang. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2340–2354, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.212. URL `https://aclanthology.org/2020.findings-emnlp.212`.

[20] Mingda Chen, Sam Wiseman, and Kevin Gimpel. WikiTableT: A large-scale data-to-text dataset for generating Wikipedia article sections. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 193–209, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.17. URL `https://aclanthology.org/2021.findings-acl.17`.

[21] Wei Chen, Yeyun Gong, Song Wang, Bolun Yao, Weizhen Qi, Zhongyu Wei, Xiaowu Hu, Bartuer Zhou, Yi Mao, Weizhu Chen, Biao Cheng, and Nan Duan. DialogVED: A pre-trained latent variable encoder-decoder model for dialog response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4852–4864, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.acl-long.333`.

[22] Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.708. URL `https://aclanthology.org/2020.acl-main.708`.

[23] Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.697. URL `https://aclanthology.org/2020.emnlp-main.697`.

[24] Liying Cheng, Dekun Wu, Lidong Bing, Yan Zhang, Zhanming Jie, Wei Lu, and Luo Si. ENT-DESC: Entity description generation by exploring knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1187–1197, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.90. URL `https://aclanthology.org/2020.emnlp-main.90`.

[25] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL `https://aclanthology.org/D18-1241`.

[26] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM, 2008. doi: 10.1145/1390156.1390177. URL `https://doi.org/10.1145/1390156.1390177`.

[27] Alexis CONNEAU and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf`.

[28] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 248–255, Los Alamitos, CA, USA, jun 2009. IEEE Computer Society. doi: 10.1109/CVPR.2009.5206848. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR.2009.5206848`.

[30] Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.599. URL `https://aclanthology.org/2021.emnlp-main.599`.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

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

[33] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. Evaluating prerequisite qualities for learning end-to-end dialog systems. In *4th International Conference on Learning Representations, ICLR 2016*, 2016. URL http://arxiv.org/abs/1511.06931.

[34] William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

[35] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c20bb2d9a50d5ac1f713f8b34d9aac5a-Paper.pdf.

[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[37] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5526. URL https://aclanthology.org/W17-5526.

[38] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5506. URL https://aclanthology.org/W17-5506.

[39] Angela Fan, Mike Lewis, and Yann Dauphin. 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, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

[40] Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, and Yue Gao. Rethinking supervised pre-training for better downstream transferring. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Jjcv9MTqhcq.

[41] Cristina Garbacea and Qiaozhu Mei. Neural language generation: Formulation, methods, and evaluation. *arXiv preprint arXiv:2007.15780*, 2020. URL http://arxiv.org/abs/2007.15780.

[42] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1017. URL https://aclanthology.org/P17-1017.

[43] Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.10. URL `https://aclanthology.org/2021.gem-1.10`.

[44] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, June 2007. Association for Computational Linguistics. URL `https://aclanthology.org/W07-1401`.

[45] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL `https://aclanthology.org/D19-5409`.

[46] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, pages 1891–1895. ISCA, 2019. doi: 10.21437/Interspeech.2019-3079. URL `https://doi.org/10.21437/Interspeech.2019-3079`.

[47] David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34, 2003.

[48] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1065. URL `https://aclanthology.org/N18-1065`.

[49] Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. ChainCQG: Flow-aware conversational question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2061–2070, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.177. URL `https://aclanthology.org/2021.eacl-main.177`.

[50] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.acl-long.576`.

[51] Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. Long text generation by modeling sentence-level and discourse-level coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6379–6393, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.499. URL `https://aclanthology.org/2021.acl-long.499`.

[52] R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7, 2006.

[53] Han He and Jinho D. Choi. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.451. URL https://aclanthology.org/2021.emnlp-main.451.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.90. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90.

[55] Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. GALAXY: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *arXiv preprint arXiv:2111.14592*, 2021. URL https://arxiv.org/abs/2111.14592.

[56] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf.

[57] Xinyu Hua and Lu Wang. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.57. URL https://aclanthology.org/2020.emnlp-main.57.

[58] Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.709. URL https://aclanthology.org/2020.acl-main.709.

[59] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57.

[60] Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. GenWiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2398–2409, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.217. URL https://aclanthology.org/2020.coling-main.217.

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

[62] Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. JointGT: Graph-text joint representation learning for text generation from knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*

2021, pages 2526–2538, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.223. URL https://aclanthology.org/2021.findings-acl.223.

[63] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.171. URL https://aclanthology.org/2020.findings-emnlp.171.

[64] Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023.

[65] Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. Text Generation from Knowledge Graphs with Graph Transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1238. URL https://aclanthology.org/N19-1238.

[66] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018. URL http://arxiv.org/abs/1810.09305.

[67] Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:329–345, 2020. doi: 10.1162/tacl_a_00318. URL https://aclanthology.org/2020.tacl-1.22.

[68] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

[69] Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.360. URL https://aclanthology.org/2020.findings-emnlp.360.

[70] Huiyuan Lai, Antonio Toral, and Malvina Nissim. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.62. URL https://aclanthology.org/2021.acl-short.62.

[71] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1128. URL https://aclanthology.org/D16-1128.

[72] Sungjin Lee, Hannes Schulz, Adam Atkinson, Jianfeng Gao, Kaheer Suleman, Layla El Asri, Mahmoud Adada, Minlie Huang, Shikhar Sharma, Wendy Tay, and Xiujun Li. Multi-domain task-completion dialog challenge. In *Dialog System Technology Challenges*, volume 8, 2019. URL http://workshop.colips.org/dstc7/dstc8/DTSC8_multidomain_task_proposal.pdf.

[73] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243.

[74] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

[75] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014.

[76] Junyi Li, Tianyi Tang, Gaole He, Jinhao Jiang, Xiaoxuan Hu, Puzhao Xie, Zhipeng Chen, Zhuohao Yu, Wayne Xin Zhao, and Ji-Rong Wen. TextBox: A unified, modularized, and extensible framework for text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 30–39, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.4. URL https://aclanthology.org/2021.acl-demo.4.

[77] Junyi Li, Tianyi Tang, Jian-Yun Nie, Ji-Rong Wen, and Wayne Xin Zhao. Learning to transfer prompts for text generation. *arXiv preprint arXiv:2205.01543*, 2022. URL http://arxiv.org/abs/2205.01543.

[78] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. A survey of pretrained language models based text generation. *arXiv preprint arXiv:2201.05273*, 2022. URL https://arxiv.org/abs/2201.05273.

[79] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

[80] Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*, 2018. URL http://arxiv.org/abs/1807.11125.

[81] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1099.

[82] Percy Liang, Michael Jordan, and Dan Klein. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1011.

[83] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November 2020. Association for Computational

Linguistics. doi: 10.18653/v1/2020.findings-emnlp.165. URL `https://aclanthology.org/2020.findings-emnlp.165`.

[84] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

[85] Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.210. URL `https://aclanthology.org/2020.emnlp-main.210`.

[86] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1275`.

[87] Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, Pengcheng Wang, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, Ruofei Zhang, Winnie Wu, Ming Zhou, and Nan Duan. GLGE: A new general language generation evaluation benchmark. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 408–420, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.36. URL `https://aclanthology.org/2021.findings-acl.36`.

[88] Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.106. URL `https://aclanthology.org/2021.findings-emnlp.106`.

[89] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. URL `http://arxiv.org/abs/2107.13586`.

[90] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021. URL `http://arxiv.org/abs/2103.10385`.

[91] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.acl-short.8`.

[92] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1441. URL `https://aclanthology.org/P19-1441`.

[93] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL `http://arxiv.org/abs/1910.10683`.

[94] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

[95] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[96] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. *Gender Bias in Neural Natural Language Processing*, pages 189–202. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-62077-6_14. URL https://doi.org/10.1007/978-3-030-62077-6_14.

[97] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/20c86a628232a67e7bd46f76fba7ce12-Paper.pdf.

[98] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018. URL http://arxiv.org/abs/1806.08730.

[99] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.244.

[100] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1081. URL https://aclanthology.org/P19-1081.

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

[102] Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1163. URL https://aclanthology.org/P17-1163.

[103] Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.37. URL https://aclanthology.org/2021.naacl-main.37.

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

[105] Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. Enriching and controlling global semantics for text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9443–9456, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.744. URL https://aclanthology.org/2021.emnlp-main.744.

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

[107] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5525. URL https://aclanthology.org/W17-5525.

[108] Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.21. URL https://aclanthology.org/2021.acl-long.21.

[109] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[110] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/5c04925674920eb58467fb52ce4ef728-Paper.pdf.

[111] Guanghui Qin and Jason Eisner. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.410. URL https://aclanthology.org/2021.naacl-main.410.

[112] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[113] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[114] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[115] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

[116] Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1012. URL https://aclanthology.org/N18-1012.

[117] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534.

[118] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. volume 34, pages 8689–8696, Apr. 2020. doi: 10.1609/aaai.v34i05.6394. URL https://ojs.aaai.org/index.php/AAAI/article/view/6394.

[119] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. volume 34, pages 8689–8696, Apr. 2020. doi: 10.1609/aaai.v34i05.6394. URL https://ojs.aaai.org/index.php/AAAI/article/view/6394.

[120] Mathieu Ravaut, Shafiq Joty, and Nancy Chen. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.309.

[121] Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019. doi: 10.1162/tacl_a_00266. URL https://aclanthology.org/Q19-1016.

[122] Pedro Rodriguez, Paul Crook, Seungwhan Moon, and Zhiguang Wang. Information seeking in the spirit of learning: A dataset for conversational curiosity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8153–8172, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.655. URL https://aclanthology.org/2020.emnlp-main.655.

[123] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1044. URL https://aclanthology.org/D15-1044.

[124] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2), jan 2022. ISSN 0360-0300. doi: 10.1145/3485766. URL https://doi.org/10.1145/3485766.

[125] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

[126] Maarten Sap, Eric Horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker. Recollection versus imagination: Exploring human memory and cognition via neural language models. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1970–1978, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.178. URL https://aclanthology.org/2020.acl-main.178.

[127] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://aclanthology.org/P17-1099.

[128] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

[129] Karl Stratos. Mutual information maximization for simple and accurate part-of-speech induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1095–1104, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1113. URL https://aclanthology.org/N19-1113.

[130] Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.76. URL https://aclanthology.org/2021.findings-emnlp.76.

[131] Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.319.

[132] Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B18WgG-CZ.

[133] Jiao Sun, Xuezhe Ma, and Nanyun Peng. AESOP: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.420. URL https://aclanthology.org/2021.emnlp-main.420.

[134] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019. doi: 10.1162/tacl_a_00264. URL https://aclanthology.org/Q19-1014.

[135] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.308. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.308.

[136] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

[137] Tianyi Tang, Junyi Li, and Wayne Xin Zhao. Context-tuning: Learning contextualized prompts for natural language generation. *arXiv preprint arXiv:2201.08670*, 2022. URL http://arxiv.org/abs/2103.10385.

[138] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL https://aclanthology.org/W17-2623.

[139] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[140] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, Los Alamitos, CA, USA, jun 2015. IEEE Computer Society. doi: 10.1109/CVPR.2015.7299087. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7299087.

[141] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.346.

[142] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

[143] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290. URL https://aclanthology.org/Q19-1040.

[144] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*, 2021. URL http://arxiv.org/abs/2109.01247.

[145] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

[146] Anuradha Welivita, Yubo Xie, and Pearl Pu. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.96. URL https://aclanthology.org/2021.emnlp-main.96.

[147] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1042.

[148] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

[149] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

[150] Joseph Worsham and Jugal Kalita. Multi-task learning for natural language processing in the 2020s: Where are we going? *Pattern Recognition Letters*, 136:120–126, 2020. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2020.05.031. URL https://www.sciencedirect.com/science/article/pii/S0167865520302087.

[151] Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3997–4003. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/553. URL https://doi.org/10.24963/ijcai.2020/553. Main track.

[152] Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*, 2022. URL http://arxiv.org/abs/2201.05966.

[153] Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. Zeroprompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. *arXiv preprint arXiv:2201.06910*, 2022. URL http://arxiv.org/abs/2201.06910.

[154] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016. doi: 10.1162/tacl_a_00107. URL https://aclanthology.org/Q16-1029.

[155] Jian Yang, Shuming Ma, Dongdong Zhang, ShuangZhi Wu, Zhoujun Li, and Ming Zhou. Alternating language modeling for cross-lingual pre-training. volume 34, pages 9386–9393, Apr. 2020. doi: 10.1609/aaai.v34i05.6480. URL https://ojs.aaai.org/index.php/AAAI/article/view/6480.

[156] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

[157] Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.572. URL https://aclanthology.org/2021.emnlp-main.572.

[158] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205.

[159] Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.90. URL `https://aclanthology.org/2021.acl-long.90`.

[160] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.30. URL `https://aclanthology.org/2020.acl-demos.30`.

[161] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1076. URL `https://aclanthology.org/D18-1076`.

[162] Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.474. URL `https://aclanthology.org/2021.naacl-main.474`.

A Broader Impacts

In this paper, we pre-trained a language model MVP using labeled NLG datasets. According to the research [10, 13], PLMs tend to "remember" what they have "seen" in pre-training corpus. This could result in the reproduction of undesirable biases from pre-training data on downstream tasks. Training data intervention could be a solution to alleviate this issue [96]. It is also interesting to investigate whether supervised pre-training produces fewer biases than unsupervised pre-training in the future.

Environmental impact is another factor we should consider. We have attempted a more efficient pre-training strategy and released our PLM for future work. In contrast to large PLMs with tens of billions of parameters, such as T5 [114] and GPT-3 [14], we pre-train only a small model with hundreds of millions of parameters. In addition, we utilize supervised pre-training data and initialize our model with pre-trained BART, both of which improve the convergence of our model. Ultimately, our model is pre-trained for about $20,000$ steps, whereas BART of the same size is pre-trained for $500,000$ steps. We also released our pre-trained model and task-specific prompts for reproducing our results and future work at the link `https://github.com/RUCAIBox/MVP`.

B Tasks and Datasets

B.1 Description of Tasks and Datasets

We provide the details of the tasks and datasets used in our paper for pre-training and fine-tuning in Tables 6 and 7. If the dataset for pre-training does not have a valid set, we divide 10% of the training set for validation.

We list the licenses for all datasets if them have. All datasets are publicly available. The majority of them can be directly downloaded from GitHub or Google Drive. ROCStories [101] and CommonGen [83] can be obtained after filling out a form. GYAFC [116] is accessible after requesting Yahoo and the authors of the dataset.

The tasks and datasets we use in this paper are as follows:

- **Data-to-text generation** aims to generate descriptive text about structured data, such as the knowledge graph and the table. We use the following datasets for pre-training:
  1. AGENDA [65];
  2. ENT-DESC [24];
  3. GenWiki [60];
  4. LogicNLG [22];
  5. TEKGEN [2];
  6. WEATHERGOV [82];
  7. WikiTableT [20].

  We utilize the following datasets for fine-tuning evaluation:
  1. WebNLG [42], we utilize the version 2.1;
  2. WikiBio [71].
- **Open-ended dialogue system**, also known as chatbots, is focused on daily communication. We use the following datasets for pre-training:
  1. Cleaned OpenSubtitles Dialogs (Cleaned OS Dialogs) [146], which is a cleaned variant of OpenSubtitles Dialogs [86];
  2. CMU Document Grounded Conversations (CMUDog) [161];
  3. Curiosity [122];
  4. DREAM [134];
  5. Empathetic Dialogues [117];
  6. Movie Dialog [33];
  7. MuTual [129];
  8. OpenDialKG [100];
  9. Topical-Chat [46];

10. Wizard of Wikipedia [32].

We utilize the following datasets for fine-tuning evaluation:

1. DailyDialog [81];
2. DSTC7-AVSD [4];
3. PersonaChat [158].

- **Paraphrase generation** involves rewriting a sentence with the same semantic meaning but a different syntactic or lexical form. We utilize the following datasets for fine-tuning evaluation:

  1. Quora (also known as QQP-Pos) [67], which is a subset of Quora Question Pairs [5].

- **Question answering** requires the model to answer a question based on optional background information. Note that we conduct this task in a generative way in our paper. We use the following datasets for pre-training:

  1. HotpotQA [156];
  2. MS MARCO [106];
  3. MSQG [87], since it is designed for QG, we reverse the question and answer to enrich QA examples;
  4. NarrativeQA [64];
  5. Natural Questions [68];
  6. NewsQA [138];
  7. QuAC [25];
  8. TriviaQA [61];
  9. WebQuestions [12].

  We utilize the following datasets for fine-tuning evaluation:

  1. CoQA [121];
  2. SQuAD [115], we utilize the version 1.1.

- **Question generation** generates a coherent question given a passage and its corresponding answer. We use the following datasets for pre-training:

  1. HotpotQA [156];
  2. MS MARCO [106];
  3. MSQG [87];
  4. NarrativeQA [64];
  5. NewsQA [138];
  6. QuAC [25];

  Most of them are QA tasks, and we invert the question and answer to enrich QG examples.

  We utilize the following datasets for fine-tuning evaluation:

  1. CoQA [121];
  2. SQuAD [115], we utilize the version 1.1.

- **Story generation** creates a long and informative text with a short title. We use the following datasets for pre-training:

  1. ChangeMyView [57];
  2. English Gigaword [123];
  3. Hippocorpus [126];
  4. WikiPlots [1];
  5. WritingPrompts [39], we split the original training set for pre-training and corresponding validation.

  Considering English Gigaword is a large summarization dataset, we use the summary as the title to generate the passage in turn to enrich the examples of story generation.

  We utilize the following datasets for fine-tuning evaluation:

  1. ROCStories [101];

[5] `https://www.kaggle.com/c/quora-question-pairs`

2. WritingPrompts [39], we use the sets created by [51] (who split the original valid and test sets for training, validation, and testing) to fine-tune our model for a fair comparison.

- **Task-oriented dialogue system** meets real-life needs of users, such as restaurant reservations and airplane bookings. We use the datasets for pre-training, following [131]:
  1. CamRest676 [147];
  2. Frames [37];
  3. KVRET [38];
  4. MetaLWOZ [72];
  5. MSR-E2E [80];
  6. MultiWOZ [15];
  7. Schema-Guided [118];
  8. TaskMaster [17];
  9. WOZ [102].

  We utilize the following datasets for fine-tuning evaluation:
  1. MultiWOZ [15], we utilize the version 2.0;
- **Text style transfer** modifies the style (*e.g.,* sentiment and formality) of given texts while retaining their style-independent content. We utilize the following datasets for fine-tuning evaluation:
  1. GYAFC [116], which has two sub-domains "Entertainment and Music" (E&M) and "Family and Relationships" (F&R).
- **Text summarization** condenses a long document into a brief text while retaining the essential details. We use the following datasets for pre-training:
  1. English Gigaword [47], we use the variant provided by [123];
  2. MediaSum [162];
  3. MSNews [87];
  4. Newsroom [48];
  5. WikiHow [66].

  We utilize the following datasets for fine-tuning evaluation:
  1. CNN/DailyMail [56], we use the variant provided by [127];
  2. SAMSum [45];
  3. XSum [104].

To better compare with ExT5 [6], we utilize the language generation benchmark GEM [43] for fine-tuning evaluation. GEM includes five tasks:

- **Commonsense generation**:
  1. CommonGen (CG) [83].
- **Data-to-text generation**:
  1. DART [103];
  2. E2E NLG cleaned [107];
  3. ToTTo [130];
  4. WebNLG [42].
- **Dialogue system**:
  1. Schema-Guided Dialog (SGD) [119].
- **Text simplification**:
  1. WikiAuto + Turk/ASSET (WiA-T/A) [58, 154, 5].
- **Text summarization**:
  1. Wiki-Lingua (WLE) [69].

To test the generalization ability of our model, we also utilize the natural language standing benchmark GLUE [142], which is composed of three tasks:

- **Natural language inference**:
  1. MNLI [148];
  2. QNLI [115, 142];
  3. RTE [28, 52, 44, 11].
- **Paraphrase detection**:
  1. MRPC [34];
  2. QQP [5];
  3. STS-B [18].
- **Text classification**:
  1. CoLA [143];
  2. SST-2 [128].

B.2 Data Leakage

Since our model is pre-trained on a large number of labeled datasets, it may have "seen" examples from fine-tuning test sets during pre-training, which leads to an unfair comparison with other methods. Hence, we eliminate the pre-training examples that share n-gram overlap with either of the test datasets. Following [14], n is the 5^{th} percentile example length in words, and the maximum value of n is set to 13. Finally, we have removed $17,848$ examples from the pre-training datasets. The number of "cleaned" examples for each dataset can be found in Table 6.

Table 6: The statistics and licenses of datasets for pre-training our MVP model. The #Train, #Valid, and #Test denote the number of examples in the train, valid, and test sets, respectively. Cleaned #Train represents the number of training examples after filtering. Input and Output are the average number of words (split by space) in the input and output sequences, respectively. These setups and abbreviations are the same below.

Dataset	#Train	Cleaned #Train	#Valid	#Test	Input	Output	License
AGENDA	38,720	38,720	1,000	1,000	52.1	141.2	N/A
ENT-DESC	88,652	88,652	11,081	11,081	279.9	31.0	N/A
GenWiki	681,436	681,436	75,716	1,000	21.4	29.5	MIT
LogicNLG	28,450	28,450	4,260	4,305	178.4	14.2	MIT
TEKGEN	6,310,061	6,307,995	788,746	796,982	17.0	21.2	CC BY-SA 2.0
WEATHERGOV	25,000	25,000	1,000	3,528	148.7	30.6	N/A
WikiTableT	1,453,794	1,452,778	4,533	4,351	81.0	99.7	MIT
Cleaned OS Dialogs	13,355,487	13,355,368	1,483,944	-	75.5	16.7	N/A
CMUDoG	82,818	82,818	5,555	14,510	433.0	12.2	N/A
Curiosity	64,930	64,551	8,539	8,495	144.4	20.2	CC BY-NC 4.0
DREAM	14,264	14,242	4,709	4,766	75.6	13.6	N/A
Empathetic Dialogues	64,636	64,636	9,308	8,426	52.7	12.9	CC BY-NC 4.0
Movie Dialog	762,751	762,711	8,216	8,066	126.9	44.0	N/A
MuTual	33,691	33,691	4,090	3,248	53.6	14.5	N/A
OpenDialKG	69,680	69,680	7,743	-	54.2	12.4	CC BY-NC 4.0
Topical-Chat	179,750	179,750	22,295	22,452	223.3	20.0	CDLA-Sharing-1.0
Wizard of Wikipedia	148,357	147,702	15,767	15,564	297.0	16.7	MIT
HotpotQA	90,447	87,815	7,405	-	187.9	2.2	CC BY-SA 4.0
MS MARCO	681,445	681,226	77,580	-	68.7	13.3	N/A
MSQG	198,058	198,029	11,008	-	48.1	3.7	CC BY-SA 4.0
NarrativeQA	65,494	65,494	6,922	21,114	584.1	4.2	Apache 2.0
Natural Questions	96,676	96,676	10,693	6,490	9.0	2.1	CC BY-SA 3.0
NewsQA	97,850	97,700	5,486	5,396	726.8	5.0	MIT
QuAC	83,568	83,485	31,906	-	487.9	12.5	CC BY-SA 4.0
TriviaQA	78,785	78,785	8,837	11,313	14.0	2.0	Apache 2.0
WebQuestions	8,933	8,933	4,863	4,863	6.7	2.4	CC BY 4.0
HotpotQA	90,440	87,808	6,972	-	79.6	19.8	CC BY-SA 4.0
MS MARCO	681,445	681,226	77,580	-	75.9	6.0	N/A
MSQG	198,058	198,029	11,008	11,022	45.9	6.0	CC BY-SA 4.0
NarrativeQA	65,494	65,494	6,922	21,114	579.7	8.6	Apache 2.0
NewsQA	97,850	97,700	5,486	5,396	724.2	7.6	MIT
QuAC	69,109	69,026	26,301	-	496.7	6.5	CC BY-SA 4.0
ChangeMyView	42,462	42,459	6,480	7,562	17.9	104.1	MIT
English Gigaword	3,803,957	3,802,620	189,651	1,951	8.8	33.3	MIT
Hippocorpus	6,168	6,168	686	-	34.1	262.6	CDLA-Permissive 2.0
WikiPlots	101,642	101,641	11,294	-	3.4	338.5	N/A
WritingPrompts	272,600	272,518	15,620	15,138	28.4	630.8	MIT
CamRest676	4,872	4,872	616	-	55.3	9.4	N/A
Frames	26,631	26,631	2,106	-	116.1	13.0	MIT
KVRET	14,136	14,136	1,616	-	30.5	9.3	N/A
MetaLWOZ	176,073	176,073	17,912	-	45.6	8.0	N/A
MSR-E2E	103,362	103,362	5,235	-	51.3	12.8	Microsoft
Schema-Guided	494,946	494,933	73,089	-	120.8	12.5	CC BY-SA 4.0
TaskMaster	249,664	249,662	20,680	-	95.6	12.0	CC BY 4.0
WOZ	6,364	6,359	1,260	-	47.0	10.6	N/A
English Gigaword	3,803,957	3,802,620	189,651	1,951	33.3	8.8	MIT
MediaSum	443,596	442,021	10,000	10,000	1641.0	14.4	N/A
MSNews	136,082	135,937	7,496	7,562	309.9	9.8	CC BY-SA 4.0
Newsroom	995,041	989,351	108,837	108,862	642.4	26.7	N/A
WikiHow	157,252	157,247	5,599	5,577	502.6	45.6	CC BY-NC-SA

Table 7: The statistics and licenses of datasets for fine-tuning our MVP model. The license of the MNLI dataset is composed of OANC, CC BY-SA 3.0, and CC BY 3.0. The license of the CoQA dataset is composed of CC BY-SA 4.0, MSR-LA, and Apache 2.0. The license of the WiA-A/T datasets is composed of CC BY-NC 3.0, CC BY-NC 4.0, and GNU General Public License v3.0.

Task	Dataset	#Train	#Valid	#Test	Input	Output	License
Commonsen generation	CommonGen	67,389	993	–	5.5	11.6	MIT
Data-to-text generation	DART	62,659	2,768	–	27.5	21.5	MIT
	E2E	33,525	4,299	–	9.5	20.6	CC BY-SA 4.0
	ToTTo	120,761	7,700	–	37.8	18.0	CC BY-SA 3.0
	WebNLG	34,338	4,313	4,222	18.0	19.9	CC BY-NA-SA 4.0
	WebNLG (GEM)	35,426	1,667	–	17.7	22.7	CC BY-NA-SA 4.0
	WikiBio	582,659	72,831	72,831	81.6	26.1	CC BY-SA 3.0
Open-ended dialogue	DailyDialog	76,052	7,069	6,740	72.5	13.9	CC BY-NC-SA 4.0
	DSTC7-AVSD	76,590	17,870	1,710	148.2	11.5	MIT
	PersonaChat	122,499	14,602	14,056	132.1	11.9	MIT
	SGD	164,982	10,000	–	134.7	11.3	CC BY-SA 4.0
Natural language inference	MNLI-m	392,702	9,815	9,796	29.8	–	Mixed
	MNLI-mm		9,832	9,847			
	QNLI	104,743	5,463	5,463	36.6	–	CC BY-SA 4.0
	RTE	2,490	277	3,000	51.0	–	N/A
Paraphrase generation	Quora	137,185	3,000	3,000	10.9	10.8	N/A
Paraphrase detection	MRPC	3,668	408	1,725	43.8	–	N/A
	QQP	363,846	40,430	390,965	22.3	–	N/A
	STS-B	5,749	1,500	1,379	20.3	–	N/A
Question answering	CoQA	107,286	31,621	–	349.4	2.6	Mixed
	SQuAD	75,722	10,570	11,877	156.2	3.6	CC BY-SA 4.0
Question generation	CoQA	107,286	31,621	–	346.6	5.5	Mixed
	SQuAD	75,722	10,570	11,877	148.3	11.6	CC BY-SA 4.0
Story generation	ROCStories	176,688	9,816	4,909	9.0	40.7	N/A
	WritingPrompts	53,516	4,000	2,000	25.5	150.4	MIT
Task-oriented dialogue	MultiWOZ	170,220	22,074	22,116	128.3	11.3	MIT
Text classification	CoLA	8,551	1,043	1,063	7.7	–	N/A
	SST-2	67,349	872	1,821	9.8	–	N/A
Text simplification	WiA-A	483,801	20,000	359	26.2	21.5	Mixed
	WiA-T			359			
Text style transfer	GYAFC-E&M	52,595	11,508	1,416	9.9	10.6	N/A
	GYAFC-F&R	51,967	11,152	1,332	10.7	11.3	
Text summarization	CNN/DailyMail	287,227	13,368	11,490	679.8	48.3	MIT
	SAMSum	14,732	818	819	103.4	20.3	CC BY-NC-ND 4.0
	WLE	99,020	28,614	–	367.6	33.4	CC0 1.0
	XSum	204,045	11,332	11,334	373.7	21.1	MIT

C Fine-tuning and Evaluation Details

In this section, we introduce the details for fine-tuning and evaluating each downstream task.

For the experiments in Section 4 (Tables 1 and 2), and Appendix D.1 (Table 8), the fine-tuning details are introduced in Section 4, and the evaluation details are presented as follows:

- For data-to-text generation tasks, we use BLEU(-4), ROUGE-L, and METEOR for evaluation. We use the script provided by [23] [6];
- For open-ended dialogue system tasks, we use BLEU-1, BLEU-2, Distinct-1, and Distinct-2 for evaluation. For DSTC7-AVSD we also utilize CIDEr [140]. We employ NLTK 3.5 with smoothing function 7 to compute BLEU for PersonaChat and DailyDialog, and utilize the script [7] to evaluate DSTC7-AVSD;
- For question answering tasks, we use Exact Match (EM) and Macro-averaged F1 score (F1) for evaluation. We use the provided script for CoQA [8] and SQuAD [9].
- For question generation tasks, we use BLEU-4, ROUGE-L, and METEOR for evaluation. We use the script provided by [35] [10];
- For story generation, we employ nucleus sampling with $p = 0.9$ and temperature of 0.7 following [51]. We use corpus BLEU-1, BLEU-2, Distinct-1, and Distinct-4 for evaluation. We use NLTK 3.5 to calculate corpus BLEU following [51];
- For task-oriented dialogue system tasks, we use BLEU(-4), inform (rate), success (rate), and combined score for evaluation. Inform and success are two specially designed accuracy metrics for task-oriented dialogue system, and the combined score is defined as $(\text{Inform} + \text{Success}) \times 0.5 +$ BLEU [15]. We use the script provided by [131] [11];
- For text summarization tasks, we use ROUGE-1, ROUGE-2, and ROUGE-L for evaluation. We use the toolkit `files2rouge` [12].

For the experiments in Section 5 (Tables 3 and 4), the fine-tuning and evaluation details are as follows:

- For paraphrase generation tasks, we employ the fine-tuning and evaluation scripts provided by AESOP [133] [13]. The evaluation metrics are BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR.
- For text style transfer tasks, we employ the fine-tuning and evaluation scripts provided by SC & BLEU [70] [14]. We conduct the informal-to-formal transfer and train the model on the data from both the E&M and F&R domains following [70]. The evaluation metrics are BLEU-4, accuracy, and HM. Accuracy is calculated by a pre-trained TextCNN to evaluate the style strength, and HM denotes the harmonic mean of BLEU-4 and style accuracy [70].
- For GLUE tasks, we utilize the fine-tuning code provided by Hugging Face [15]. The hyper-parameters are consistent with original BART [74] [16]. The evaluation is computed by the official website [17].

[6] `https://github.com/wenhuchen/Data-to-text-Evaluation-Metric`

[7] `https://github.com/lemuria-wchen/DialogVED/blob/main/src/utils/evaluate.py`

[8] `https://github.com/PaddlePaddle/ERNIE/blob/repro/ernie-gen/eval/tasks/coqa/eval.py`

[9] `https://github.com/allenai/bi-att-flow/blob/master/squad/evaluate-v1.1.py`

[10] `https://github.com/microsoft/unilm/blob/master/unilm-v1/src/qg/eval.py`

[11] `https://github.com/awslabs/pptod/blob/main/E2E_TOD/eval.py`

[12] `https://github.com/pltrdy/files2rouge`

[13] `https://github.com/PlusLabNLP/AESOP`

[14] `https://github.com/laihuiyuan/pre-trained-formality-transfer`

[15] `https://github.com/huggingface/transformers/tree/main/examples/pytorch/text-classification`

[16] `https://github.com/facebookresearch/fairseq/blob/main/examples/bart/README.glue.md`

[17] `https://gluebenchmark.com/`

Table 8: The results on six seen tasks under full tuning settings. WikiBio is the dataset of the data-to-text generation task. [a] [105] [b] [88] [c] [49] [d] [51] [e] [21] [f] [23] [g] [114]

	Summarization (XSum)			Summarization (SAMSum)			QG (CoQA)		
	R-1	R-2	R-L	R-1	R-2	R-L	B-4	ME	R-L
SOTA	**49.57**[a]	**25.08**	**41.81**	**54.3**[b]	$\underline{29.3}$	45.2	15.78[c]	40.15	50.98
BART	$\underline{45.80}_{0.10}$	$\underline{22.56}_{0.02}$	$\underline{37.50}_{0.03}$	$53.64_{0.17}$	$29.18_{0.31}$	$\underline{49.58}_{0.36}$	$22.24_{0.14}$	$24.00_{0.08}$	$54.10_{0.18}$
MVP	$45.67_{0.04}$	$22.53_{0.06}$	$37.41_{0.03}$	$\underline{53.82}_{0.19}$	$29.03_{0.43}$	$49.32_{0.16}$	$\mathbf{22.52}_{0.21}$	$\underline{24.20}_{0.16}$	$\underline{54.50}_{0.22}$
MVP+S	$45.59_{0.06}$	$22.54_{0.04}$	$37.39_{0.03}$	$53.78_{0.06}$	$\mathbf{29.42}_{0.03}$	$\mathbf{49.60}_{0.09}$	$\underline{22.45}_{0.15}$	$24.07_{0.13}$	$\mathbf{54.63}_{0.22}$

	Story Generation (WritingPrompts)				Open-ended Dialogue (DailyDialog)				WikiBio
	B-1	B-2	D-1	D-4	B-1	B-2	D-1	D-2	B-4
SOTA	22.4[d]	8.4	–	31.3	46.1[e]	40.7	4.1	22.2	45.1[f]
BART	$\mathbf{31.85}_{0.47}$	$\underline{12.53}_{0.22}$	$1.99_{0.17}$	$61.90_{2.52}$	$\underline{51.85}_{0.14}$	$\underline{43.59}_{0.11}$	$\underline{6.47}_{0.06}$	$\underline{35.69}_{0.42}$	$\underline{48.37}_{0.19}$
MVP	$\underline{31.81}_{0.53}$	$\mathbf{12.80}_{0.05}$	$\underline{2.58}_{0.14}$	$\underline{69.45}_{1.24}$	$\mathbf{52.34}_{0.35}$	$\mathbf{43.93}_{0.30}$	$6.39_{0.06}$	$35.65_{0.33}$	$\mathbf{48.42}_{0.23}$
MVP+S	$29.18_{0.32}$	$11.11_{0.16}$	$\mathbf{3.71}_{0.32}$	$\mathbf{78.02}_{3.36}$	$51.04_{0.57}$	$42.87_{0.48}$	$\mathbf{6.70}_{0.14}$	$\mathbf{36.84}_{0.49}$	$48.19_{0.09}$

	Open-ended Dialogue (DSTC7-AVSD)							QA (SQuAD)	
	B-1	B-2	B-3	B-4	ME	R-L	CIDEr	F1	EM
SOTA	83.2[e]	$\underline{70.5}$	$\underline{59.8}$	$\underline{50.6}$	$\underline{31.4}$	63.8	1.391	**91.26**[g]	**96.22**
BART	$82.48_{0.52}$	$69.40_{0.40}$	$58.57_{0.40}$	$49.33_{0.40}$	$31.39_{0.39}$	$64.07_{0.25}$	$1.401_{0.01}$	$84.95_{0.34}$	$91.98_{0.11}$
MVP	$\underline{83.37}_{0.64}$	$\underline{70.50}_{0.47}$	$59.73_{0.47}$	$50.42_{0.32}$	$\mathbf{31.59}_{0.33}$	$\mathbf{64.54}_{0.47}$	$\mathbf{1.429}_{0.01}$	$86.44_{0.21}$	$93.04_{0.08}$
MVP+S	$\mathbf{83.76}_{0.07}$	$\mathbf{70.80}_{0.22}$	$\mathbf{60.03}_{0.26}$	$\mathbf{50.65}_{0.32}$	$31.07_{0.28}$	$\underline{64.10}_{0.19}$	$\underline{1.403}_{0.01}$	$\underline{86.78}_{0.16}$	$\underline{93.21}_{0.05}$

For the experiments of the GEM benchmark in Appendix D.2 (Table 9), the fine-tuning settings are the same as those described in Section 4. We use BLEU-4, ROUGE-2, and METEOR for evaluation. We use the GEM evaluation scripts [18].

D Additional Results

In this section, we provide additional results of our MVP model and other baselines.

D.1 Results of Common Datasets

We also conduct experiments on various common datasets under full tuning settings. Due to space limits in Section 4, these results are shown in Table 8. We can see that these results share a similar trend to those in Section 4, and we achieve SOTA performances in 20 of 27 metrics.

D.2 Results on the GEM Benchmark

To better compare with ExT5 [6], we conduct experiments on the GEM benchmark [43]. For "unseen" commonsense generation and text simplification tasks, we utilize prompts of data-to-text generation and summarization, respectively. The results are presented in Table 9. Note that, because the fine-tuning and dataset hyper-parameters of ExT5 and GEM are unavailable, the results of some datasets we reproduced differ from the original papers [6, 43]. Regardless, our MVP models outperform ExT5 in 26 out of 27 metrics.

D.3 Results without Fine-tuning

Considering our MVP model has already been pre-trained on several tasks, we conduct experiments on these "seen" tasks without fine-tuning our model. To some degree, this setting can be viewed as zero-shot learning. Nonetheless, it does not conform to the definition of *true zero-shot* settings [110]. To avoid controversy, we refer to this as *without fine-tuning*.

We include T0-3B [125] as our baseline. The results are listed in Table 10. All tasks demonstrate that methods without fine-tuning perform significantly worse than those with full tuning settings. This

[18] `https://github.com/GEM-benchmark/GEM-metrics`

Table 9: The results on the GEM benchmark under full tuning settings. We utilize the large version of T5.1.1 and ExT5, and all the results of them are from [6].

	Data-to-text (DART)			Data-to-text (E2E)			Data-to-text (ToTTo)		
	B-4	R-2	ME	B-4	R-2	ME	B-4	R-2	ME
T5.1.1	34.31	45.22	36.3	**42.57**	46.60	38.2	39.79	49.90	36.8
ExT5	36.62	48.14	37.6	42.25	46.70	38.1	40.14	50.33	36.9
BART	38.89$_{0.35}$	48.76$_{0.07}$	38.31$_{0.28}$	37.24$_{0.43}$	47.76$_{0.18}$	39.24$_{0.21}$	50.39$_{0.04}$	55.14$_{0.12}$	41.11$_{0.05}$
MVP	**39.13**$_{0.06}$	**48.92**$_{0.11}$	**38.53**$_{0.13}$	37.38$_{0.17}$	**47.96**$_{0.09}$	**39.39**$_{0.21}$	50.58$_{0.12}$	55.24$_{0.12}$	41.27$_{0.04}$
MVP+S	38.83$_{0.17}$	48.49$_{0.18}$	38.41$_{0.05}$	37.32$_{0.35}$	47.40$_{0.25}$	38.90$_{0.11}$	**50.69**$_{0.13}$	**55.52**$_{0.05}$	**41.29**$_{0.07}$
	Data-to-text (WebNLG)			**CommonGen (CG)**			**Dialogue (SGD)**		
	B-4	R-2	ME	B-4	R-2	ME	B-4	R-2	ME
T5.1.1	31.67	43.31	34.4	8.38	17.01	20.2	33.15	36.17	32.4
ExT5	35.03	48.17	36.5	9.68	19.04	21.4	34.74	37.77	33.0
BART	46.67$_{0.30}$	58.63$_{0.44}$	42.19$_{0.07}$	32.68$_{0.85}$	37.16$_{0.28}$	32.81$_{0.24}$	45.14$_{0.11}$	47.93$_{0.05}$	38.19$_{0.11}$
MVP	**47.03**$_{0.06}$	59.00$_{0.19}$	**42.34**$_{0.08}$	32.59$_{0.92}$	37.71$_{0.59}$	33.00$_{0.07}$	**45.63**$_{0.10}$	**48.29**$_{0.14}$	**38.48**$_{0.08}$
MVP+S	**47.03**$_{0.24}$	**59.03**$_{0.20}$	42.28$_{0.09}$	**34.10**$_{0.35}$	**37.87**$_{0.58}$	**33.11**$_{0.11}$	45.24$_{0.11}$	48.25$_{0.20}$	38.47$_{0.40}$
	Simplification (WiA-A)			**Simplification (WiA-T)**			**Summarization (WLE)**		
	B-4	R-2	ME	B-4	R-2	ME	B-4	R-2	ME
T5.1.1	29.30	38.37	30.1	42.12	50.52	36.2	15.55	20.47	19.6
ExT5	29.23	37.98	30.0	41.39	50.38	35.8	16.64	21.16	20.4
BART	71.07$_{0.21}$	**71.09**$_{0.06}$	47.46$_{1.69}$	90.81$_{0.24}$	83.36$_{0.07}$	**57.58**$_{0.19}$	18.51$_{0.01}$	22.57$_{0.05}$	21.78$_{0.09}$
MVP	**71.55**$_{0.18}$	70.88$_{0.17}$	**48.19**$_{0.13}$	**91.73**$_{0.20}$	83.46$_{0.08}$	57.34$_{0.06}$	**18.80**$_{0.10}$	**22.84**$_{0.08}$	21.95$_{0.06}$
MVP+S	70.37$_{0.23}$	70.65$_{0.07}$	47.70$_{0.22}$	91.12$_{0.30}$	**83.59**$_{0.23}$	56.95$_{0.25}$	18.52$_{0.12}$	22.57$_{0.06}$	**22.02**$_{0.04}$

Table 10: The results on seven seen tasks without fine-tuning. We also include the results of BART and MVP under full tuning (denoted as FT) settings for comparison. Given that T0 has been pre-trained on the CNN/DailyMail dataset, we exclude their results to provide a fair comparison (denoted as "–").

	Summarization (CNN/DM)			Data-to-text (WebNLG)			QG (SQuAD)			QA (CoQA)	
	R-1	R-2	R-L	B-4	ME	R-L	B-4	ME	R-L	F1	EM
FT BART	44.47$_{0.10}$	21.50$_{0.14}$	41.35$_{0.08}$	67.33$_{0.06}$	47.78$_{0.07}$	76.83$_{0.04}$	25.08$_{0.13}$	26.73$_{0.18}$	52.55$_{0.07}$	74.00$_{0.17}$	84.07$_{0.21}$
FT MVP	44.45$_{0.05}$	21.44$_{0.12}$	41.34$_{0.08}$	67.32$_{0.10}$	47.94$_{0.13}$	76.70$_{0.26}$	25.91$_{0.07}$	27.22$_{0.10}$	53.08$_{0.16}$	75.50$_{0.20}$	85.07$_{0.21}$
T0	–	–	–	1.40	10.20	18.43	3.06	12.43	14.91	6.60	13.30
MVP	**29.50**	**11.29**	**25.92**	34.42	31.33	52.33	2.90	13.94	15.48	18.20	29.40
MVP+S	25.60	9.51	22.67	**39.43**	**34.32**	**55.34**	**2.96**	**15.23**	**18.23**	**37.30**	**52.40**

	Story Generation (ROCStories)				Open-ended Dialogue (PersonaChat)				TODS (MultiWOZ)		
	B-1	B-2	D-1	D-4	B-1	B-2	D-1	D-2	B-4	Success	Inform
FT BART	33.79$_{0.13}$	15.78$_{0.21}$	3.43$_{0.17}$	78.76$_{2.15}$	49.58$_{1.12}$	39.24$_{0.90}$	1.44$_{0.09}$	8.89$_{0.57}$	20.17$_{0.63}$	75.40$_{1.22}$	84.40$_{1.15}$
FT MVP	33.96$_{0.08}$	15.96$_{0.05}$	3.17$_{0.15}$	76.11$_{1.38}$	49.56$_{0.44}$	40.41$_{0.10}$	1.55$_{0.07}$	10.20$_{0.46}$	20.34$_{0.37}$	75.47$_{0.40}$	84.07$_{0.15}$
T0	8.69	3.02	4.37	35.49	23.20	23.57	2.56	12.06	0.02	**2.50**	22.10
MVP	1.01	0.31	**7.18**	**86.26**	35.54	32.71	**2.87**	**16.38**	**3.08**	**2.50**	**22.20**
MVP+S	**10.52**	**3.54**	2.13	69.55	**37.04**	**33.38**	2.66	14.84	0.38	**2.50**	22.10

suggests that zero-shot strategies that are effective for NLU tasks may not produce satisfactory results for NLG tasks. Even though our model has acquired task knowledge, it struggles to perform well in a new domain without being fine-tuned. Thus, we focus mainly on full tuning settings in this paper.

E Qualitative Examples

In this section, we showcase the linearized inputs, task instructions, and corresponding outputs of a single dataset for tasks in Section 4. We provide the results of BART, MVP, and MVP+S under full tuning settings. To minimize human intervention, we select the first and second instances of the test set with the random seed $2,020$.

Table 11: The first instance from the CNN/Daily Mail dataset. Task instructions are labeled in *italic*. The setting is the same below.

Input

Summarize: Marseille, France (CNN)The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that "so far no videos were used in the crash investigation." He added, "A person who has such a video needs to immediately give it to the investigators." Robin's comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps. All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. The two publications described the supposed video, but did not post it on their websites. The publications said that they watched the video, which was found by a source close to the investigation. "One can hear cries of 'My God' in several languages," Paris Match reported. "Metallic banging can also be heard more than three times, perhaps of the pilot trying to open the cockpit door with a heavy object. Towards the end, after a heavy shake, stronger than the others, the screaming intensifies. Then nothing." "It is a very disturbing scene," said Julian Reichelt, editor-in-chief of Bild online. An official with France's accident investigation agency, the BEA, said the agency is not aware of any such video. Lt. Col. Jean-Marc Menichini, a French Gendarmerie spokesman in charge of communications on rescue efforts around the Germanwings crash site, told CNN that the reports were "completely wrong" and "unwarranted." Cell phones have been collected at the site, he said, but that they "hadn't been exploited yet." Menichini said he believed the cell phones would need to be sent to the Criminal Research Institute in Rosny sous-Bois, near Paris, in order to be analyzed by specialized technicians working hand-in-hand with investigators. But none of the cell phones found so far have been sent to the institute, Menichini said. Asked whether staff involved in the search could have leaked a memory card to the media, Menichini answered with a categorical "no." Reichelt told "Erin Burnett: Outfront" that he had watched the video and stood by the report, saying Bild and Paris Match are "very confident" that the clip is real. He noted that investigators only revealed they'd recovered cell phones from the crash site after Bild and Paris Match published their reports. "That is something we did not know before. ... Overall we can say many things of the investigation weren't revealed by the investigation at the beginning," he said. What was mental state of Germanwings co-pilot? German airline Lufthansa confirmed Tuesday that co-pilot Andreas Lubitz had battled depression years before he took the controls of Germanwings Flight 9525, which he's accused of deliberately crashing last week in the French Alps. Lubitz told his Lufthansa flight training school in 2009 that he had a "previous episode of severe depression," the airline said Tuesday. Email correspondence between Lubitz and the school discovered in an internal investigation, Lufthansa said, included medical documents he submitted in connection with resuming his flight training. The announcement indicates that Lufthansa, the parent company of Germanwings, knew of Lubitz's battle with depression, allowed him to continue training and ultimately put him in the cockpit. Lufthansa, whose CEO Carsten Spohr previously said Lubitz was 100% fit to fly, described its statement Tuesday as a "swift and seamless clarification" and said it was sharing the information and documents – including training and medical records – with public prosecutors. Spohr traveled to the crash site Wednesday, where recovery teams have been working for the past week to recover human remains and plane debris scattered across a steep mountainside. He saw the crisis center set up in Seyne-les-Alpes, laid a wreath in the village of Le Vernet, closer to the crash site, where grieving families have left flowers at a simple stone memorial. Menichini told CNN late Tuesday that no visible human remains were left at the site but recovery teams would keep searching. French President Francois Hollande, speaking Tuesday, said that it should be possible to identify all the victims using DNA analysis by the end of the week, sooner than authorities had previously suggested. In the meantime, the recovery of the victims' personal belongings will start Wednesday, Menichini said. Among those personal belongings could be more cell phones belonging to the 144 passengers and six crew on board. Check out the latest from our correspondents. The details about Lubitz's correspondence with the flight school during his training were among several developments as investigators continued to delve into what caused the crash and Lubitz's possible motive for downing the jet. A Lufthansa spokesperson told CNN on Tuesday that Lubitz had a valid medical certificate, had passed all his examinations and "held all the licenses required." Earlier, a spokesman for the prosecutor's office in Dusseldorf, Christoph Kumpa, said medical records reveal Lubitz suffered from suicidal tendencies at some point before his aviation career and underwent psychotherapy before he got his pilot's license. Kumpa emphasized there's no evidence suggesting Lubitz was suicidal or acting aggressively before the crash. Investigators are looking into whether Lubitz feared his medical condition would cause him to lose his pilot's license, a European government official briefed on the investigation told CNN on Tuesday. While flying was "a big part of his life," the source said, it's only one theory being considered. Another source, a law enforcement official briefed on the investigation, also told CNN that authorities believe the primary motive for Lubitz to bring down the plane was that he feared he would not be allowed to fly because of his medical problems. Lubitz's girlfriend told investigators he had seen an eye doctor and a neuropsychologist, both of whom deemed him unfit to work recently and concluded he had psychological issues, the European government official said. But no matter what details emerge about his previous mental health struggles, there's more to the story, said Brian Russell, a forensic psychologist. "Psychology can explain why somebody would turn rage inward on themselves about the fact that maybe they weren't going to keep doing their job and they're upset about that and so they're suicidal," he said. "But there is no mental illness that explains why somebody then feels entitled to also take that rage and turn it outward on 149 other people who had nothing to do with the person's problems." Germanwings crash compensation: What we know. Who was the captain of Germanwings Flight 9525? CNN's Margot Haddad reported from Marseille and Pamela Brown from Dusseldorf, while Laura Smith-Spark wrote from London. CNN's Frederik Pleitgen, Pamela Boykoff, Antonia Mortensen, Sandrine Amiel and Anna-Maja Rappard contributed to this report.

Gold

Match are "very confident" the video clip is real, an editor says. Andreas Lubitz had informed his Lufthansa training school of an episode of severe depression, airline says.

BART

Prosecutor: "A person who has such a video needs to immediately give it to the investigators" French Gendarmerie spokesman: Cell phones have been collected at the site, but haven't been exploited.

MVP

Prosecutor: "So far no videos were used in the crash investigation" An official with France's accident investigation agency says the agency is not aware of any video.

MVP+S

Prosecutor: "So far no videos were used in the crash investigation" Two publications claim to have found a cell phone video of the crash. "It is a very disturbing scene," says editor-in-chief of Bild online.

Table 12: The second instance from the CNN/Daily Mail dataset.

Input	
Summarize: The Palestinian Authority officially became the 123rd member of the International Criminal Court on Wednesday, a step that gives the court jurisdiction over alleged crimes in Palestinian territories. The formal accession was marked with a ceremony at The Hague, in the Netherlands, where the court is based. The Palestinians signed the ICC's founding Rome Statute in January, when they also accepted its jurisdiction over alleged crimes committed "in the occupied Palestinian territory, including East Jerusalem, since June 13, 2014." Later that month, the ICC opened a preliminary examination into the situation in Palestinian territories, paving the way for possible war crimes investigations against Israelis. As members of the court, Palestinians may be subject to counter-charges as well. Israel and the United States, neither of which is an ICC member, opposed the Palestinians' efforts to join the body. But Palestinian Foreign Minister Riad al-Malki, speaking at Wednesday's ceremony, said it was a move toward greater justice. "As Palestine formally becomes a State Party to the Rome Statute today, the world is also a step closer to ending a long era of impunity and injustice," he said, according to an ICC news release. "Indeed, today brings us closer to our shared goals of justice and peace." Judge Kuniko Ozaki, a vice president of the ICC, said acceding to the treaty was just the first step for the Palestinians. "As the Rome Statute today enters into force for the State of Palestine, Palestine acquires all the rights as well as responsibilities that come with being a State Party to the Statute. These are substantive commitments, which cannot be taken lightly," she said. Rights group Human Rights Watch welcomed the development. "Governments seeking to penalize Palestine for joining the ICC should immediately end their pressure, and countries that support universal acceptance of the court's treaty should speak out to welcome its membership," said Balkees Jarrah, international justice counsel for the group. "What's objectionable is the attempts to undermine international justice, not Palestine's decision to join a treaty to which over 100 countries around the world are members." In January, when the preliminary ICC examination was opened, Israeli Prime Minister Benjamin Netanyahu described it as an outrage, saying the court was overstepping its boundaries. The United States also said it "strongly" disagreed with the court's decision. "As we have said repeatedly, we do not believe that Palestine is a state and therefore we do not believe that it is eligible to join the ICC," the State Department said in a statement. It urged the warring sides to resolve their differences through direct negotiations. "We will continue to oppose actions against Israel at the ICC as counterproductive to the cause of peace," it said. But the ICC begs to differ with the definition of a state for its purposes and refers to the territories as "Palestine." While a preliminary examination is not a formal investigation, it allows the court to review evidence and determine whether to investigate suspects on both sides. Prosecutor Fatou Bensouda said her office would "conduct its analysis in full independence and impartiality." The war between Israel and Hamas militants in Gaza last summer left more than 2,000 people dead. The inquiry will include alleged war crimes committed since June. The International Criminal Court was set up in 2002 to prosecute genocide, crimes against humanity and war crimes. CNN's Vasco Cotovio, Kareem Khadder and Faith Karimi contributed to this report.	
Gold	
Membership gives the ICC jurisdiction over alleged crimes committed in Palestinian territories since last June. Israel and the United States opposed the move, which could open the door to war crimes investigations against Israelis.	
BART	
Palestinian Authority becomes 123rd member of the International Criminal Court. The move gives the court jurisdiction over alleged crimes in Palestinian territories. Israel and the United States opposed the Palestinians' efforts to join the body.	
MVP	
"Today brings us closer to our shared goals of justice and peace," foreign minister says. The Palestinians signed the ICC's founding Rome Statute in January. The move gives the court jurisdiction over alleged crimes in Palestinian territories.	
MVP+S	
"Today brings us closer to our shared goals of justice and peace," foreign minister says. The United States says it "strongly" disagrees with the decision. The Palestinian Authority is the 123rd member of the International Criminal Court.	

Table 13: The first instance from the WebNLG dataset, which has two gold target sentences.

<table>
<tr><td>Input
Describe the following data: Abilene,_Texas | cityServed | Abilene_Regional_Airport</td></tr>
<tr><td>Gold
Abilene, Texas is served by the Abilene regional airport.
Abilene Regional Airport serves the city of Abilene in Texas.</td></tr>
<tr><td>BART
Abilene Regional Airport serves the city of Abilene in Texas.</td></tr>
<tr><td>MVP
Abilene Regional Airport serves the city of Abilene, Texas.</td></tr>
<tr><td>MVP+S
Abilene Regional Airport serves the city of Abilene, Texas.</td></tr>
</table>

Table 14: The second instance from the WebNLG dataset, which has three gold target sentences.

<table>
<tr><td>Input
Describe the following data: "Madrid, Paracuellos de Jarama, San Sebastián de los Reyes and Alcobendas" | location | Adolfo_Suárez_Madrid–Barajas_Airport</td></tr>
<tr><td>Gold
Adolfo Suárez Madrid–Barajas Airport can be found in Madrid, Paracuellos de Jarama, San Sebastián de los Reyes and Alcobendas.
Adolfo Suarez Madrid-Barajas airport is located at Madrid, Paracuellos de Jarama, San Sebastián de los Reyes and Alcobendas.
Adolfo Suarez Madrid-Barajas Airport is located in Madrid, Paracuellos de Jarama, San Sebastian de los Reyes and Alcobendas.</td></tr>
<tr><td>BART
Adolfo Suárez Madrid–Barajas Airport can be found in Madrid, Paracuellos de Jarama, San Sebastián de los Reyes and Alcobendas.</td></tr>
<tr><td>MVP
Adolfo Suárez Madrid–Barajas Airport can be found in Madrid, Paracuellos de Jarama, San Sebastián de los Reyes and Alcobendas.</td></tr>
<tr><td>MVP+S
Adolfo Suárez Madrid–Barajas Airport is located in Madrid, Paracuellos de Jarama, San Sebastián de los Reyes and Alcobendas.</td></tr>
</table>

Table 15: The first instance from the SQuAD dataset.

Input *Generate the question based on the answer:* Saint Bernadette Soubirous [SEP] Architecturally , the school has a Catholic character . Atop the Main Building ' s gold dome is a golden statue of the Virgin Mary . Immediately in front of the Main Building and facing it , is a copper statue of Christ with arms upraised with the legend " Venite Ad Me Omnes " . Next to the Main Building is the Basilica of the Sacred Heart . Immediately behind the basilica is the Grotto , a Marian place of prayer and reflection . It is a replica of the grotto at Lourdes , France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858 . At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome) , is a simple , modern stone statue of Mary .
Gold To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France ?
BART Who is believed to have appeared to the Virgin Mary at Lourdes ?
MVP Who did the Virgin Mary appear to in Lourdes ?
MVP+S The Grotto is a replica of the grotto at Lourdes , France where the Virgin Mary appeared to whom ?

Table 16: The second instance from the SQuAD dataset.

Input *Generate the question based on the answer:* a copper statue of Christ [SEP] Architecturally , the school has a Catholic character . Atop the Main Building ' s gold dome is a golden statue of the Virgin Mary . Immediately in front of the Main Building and facing it , is a copper statue of Christ with arms upraised with the legend " Venite Ad Me Omnes " . Next to the Main Building is the Basilica of the Sacred Heart . Immediately behind the basilica is the Grotto , a Marian place of prayer and reflection . It is a replica of the grotto at Lourdes , France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858 . At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome) , is a simple , modern stone statue of Mary .
Gold What is in front of the Notre Dame Main Building ?
BART What is in front of the Main Building and facing it ?
MVP What is immediately in front of the Main Building ?
MVP+S What is immediately in front of the Main Building ?

Table 17: The first instance from the CoQA dataset.

Input *Answer the following question:* what color was cotton ? [X_SEP] once upon a time , in a barn near a farm house , there lived a little white kitten named cotton . cotton lived high up in a nice warm place above the barn where all of the farmer ’ s horses slept . but cotton wasn ’ t alone in her little home above the barn , oh no . she shared her hay bed with her mommy and 5 other sisters . all of her sisters were cute and fluffy , like cotton . but she was the only white one in the bunch . the rest of her sisters were all orange with beautiful white tiger stripes like cotton ’ s mommy . being different made cotton quite sad . she often wished she looked like the rest of her family . so one day , when cotton found a can of the old farmer ’ s orange paint , she used it to paint herself like them . when her mommy and sisters found her they started laughing . " what are you doing , cotton ? ! " " i only wanted to be more like you " . cotton ’ s mommy rubbed her face on cotton ’ s and said " oh cotton , but your fur is so pretty and special , like you . we would never want you to be any other way " . and with that , cotton ’ s mommy picked her up and dropped her into a big bucket of water . when cotton came out she was herself again . her sisters licked her face until cotton ’ s fur was all all dry . " don ’ t ever do that again , cotton ! " they all cried . " next time you might mess up that pretty white fur of yours and we wouldn ’ t want that ! " then cotton thought , " i change my mind . i like being special " .
Gold white
BART white
MVP white
MVP+S white

Table 18: The second instance from the CoQA dataset.

Input *Answer the following question:* what color was cotton ? [SEP] white [X_SEP] where did she live ? [X_SEP] once upon a time , in a barn near a farm house , there lived a little white kitten named cotton . cotton lived high up in a nice warm place above the barn where all of the farmer ’ s horses slept . but cotton wasn ’ t alone in her little home above the barn , oh no . she shared her hay bed with her mommy and 5 other sisters . all of her sisters were cute and fluffy , like cotton . but she was the only white one in the bunch . the rest of her sisters were all orange with beautiful white tiger stripes like cotton ’ s mommy . being different made cotton quite sad . she often wished she looked like the rest of her family . so one day , when cotton found a can of the old farmer ’ s orange paint , she used it to paint herself like them . when her mommy and sisters found her they started laughing . " what are you doing , cotton ? ! " " i only wanted to be more like you " . cotton ’ s mommy rubbed her face on cotton ’ s and said " oh cotton , but your fur is so pretty and special , like you . we would never want you to be any other way " . and with that , cotton ’ s mommy picked her up and dropped her into a big bucket of water . when cotton came out she was herself again . her sisters licked her face until cotton ’ s fur was all all dry . " don ’ t ever do that again , cotton ! " they all cried . " next time you might mess up that pretty white fur of yours and we wouldn ’ t want that ! " then cotton thought , " i change my mind . i like being special " .
Gold in a barn
BART in a barn
MVP in a barn
MVP+S in a barn

Table 19: The first instance from the ROCStories dataset.

Input *Given the story title:* male was out jogging one morning .
Gold the weather was crisp and cool . male felt good and energetic . he decided to keep jogging longer than normal . male went several more miles out of his way .
BART he saw a man struggling to walk across a field . male stopped to help him . he saw a guy struggling to walking across a fields . the man thanked male profusely !
MVP he ran into a man who was holding a knife . male was scared , so he ran away . the man threatened male . male ran back into his house .
MVP+S he noticed a sound coming from a tree . male went to investigate . he found a small baby bird stuck in the trunk . male took the bird to the vet and they released it .

Table 20: The second instance from the ROCStories dataset.

Input *Given the story title:* ken was driving around in the snow .
Gold he needed to get home from work . he was driving slowly to avoid accidents . unfortunately the roads were too slick and ken lost control . his tires lost traction and he hit a tree .
BART he hit a patch of black ice . ken slid and hurt his knee . he went to the doctor for it . krampus was sent home with a concussion .
MVP he was in his car and it was very cold . ken started to back up . the car started to make a loud noise . kena had to back out and call for help .
MVP+S he was having trouble starting his car . he had to push it a couple of blocks . his car would n't start . ken had to call a tow truck for help .

Table 21: The first instance from the PersonaChat dataset.

Input *Given the dialog:* i love to meet new people . [SEP] i have a turtle named timothy . [SEP] my favorite sport is ultimate frisbee . [SEP] my parents are living in bora bora . [SEP] autumn is my favorite season . [X_SEP] hello , how are you doing tonight ?
Gold i am well an loving this interaction how are you ?
BART i am doing well , how are you ?
MVP i am doing well . how are you ?
MVP+S i am doing well . how are you ?

Table 22: The second instance from the PersonaChat dataset.

Input *Given the dialog:* i just bought a brand new house . [SEP] i like to dance at the club . [SEP] i run a dog obedience school . [SEP] i have a big sweet tooth . [SEP] i like taking and posting selkies . [X_SEP] hello , how are you doing tonight ? [SEP] i am well an loving this interaction how are you ?
Gold i am great . i just got back from the club .
BART i am doing well , thanks for asking . what do you do for fun ?
MVP i am doing well . i just got home from dancing at the club .
MVP+S i am doing well . i just bought a new house .

Table 23: The first instance from the MultiWOZ dataset. The task-oriented dialogue is composed of dialogue state tracking, dialogue action learning and system response generation.

Input *Given the task dialog:* Belief state [X_SEP] i would like a taxi from saint john 's college to pizza hut fen ditton . *Given the task dialog:* Dialogue action [X_SEP] [db_nores] [X_SEP] i would like a taxi from saint john 's college to pizza hut fen ditton . *Given the task dialog:* System response [X_SEP] [db_nores] [X_SEP] i would like a taxi from saint john 's college to pizza hut fen ditton .
Gold [taxi] destination pizza hut fen ditton departure saint john 's college [taxi] [request] leave arrive what time do you want to leave and what time do you want to arrive by ?
BART [taxi] destination pizza express fen ditton departure saint johns college [taxi] [request] leave what time would you like to leave or arrive ?
MVP [taxi] destination pizza hut fen ditton departure saint john 's college [taxi] [request] leave sure ! when would you like to arrive ?
MVP+S [taxi] destination pizza hut fen ditton departure saint john 's college [taxi] [request] leave arrive what time would you like to leave ?

Table 24: The second instance from the MultiWOZ dataset.

Input *Given the task dialog:* Belief state [X_SEP] i would like a taxi from saint john 's college to pizza hut fen ditton . [SEP] what time do you want to leave and what time do you want to arrive by ? [SEP] i want to leave after 17:15 . *Given the task dialog:* Dialogue action [X_SEP] [db_nores] [X_SEP] i would like a taxi from saint john 's college to pizza hut fen ditton . [SEP] what time do you want to leave and what time do you want to arrive by ? [SEP] i want to leave after 17:15 . *Given the task dialog:* System response [X_SEP] [db_nores] [X_SEP] i would like a taxi from saint john 's college to pizza hut fen ditton . [SEP] what time do you want to leave and what time do you want to arrive by ? [SEP] i want to leave after 17:15 .
Gold [taxi] destination pizza hut fen ditton departure saint john 's college leave 17:15 [taxi] [inform] car phone booking completed ! your taxi will be [value_car] contact number is [value_phone]
BART [taxi] destination pizza hut fen ditton departure saint john 's college leave 17:15 [taxi] [inform] car phone booking completed ! booked car type : [value_car] contact number : [value_phone]
MVP [taxi] destination pizza hut fen ditton departure saint john 's college leave 17:15 [taxi] [inform] car phone booking completed ! booked car type : [value_car] contact number : [value_phone]
MVP+S [taxi] destination pizza hut fen ditton departure saint john 's college leave 17:15 [taxi] [inform] car phone booking completed ! booked car type : [value_car] contact number : [value_phone]