Task Compass: Scaling Multi-task Pre-training with Task Prefix

Zhuosheng Zhang[1*], **Shuohang Wang**[2], **Yichong Xu**[2], **Yuwei Fang**[2], **Wenhao Yu**[3*], **Yang Liu**[2], **Hai Zhao**[1], **Chenguang Zhu**[2] **and Michael Zeng**[2]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Microsoft Cognitive Services Research, Redmond, WA, USA
[3]University of Notre Dame, Notre Dame, IN, USA
[1]zhangzs@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn;
[2]{shuowa, yicxu, yuwfan, yaliu10, chezhu, nzeng}@microsoft.com; [3]wyu1@nd.edu

Abstract

Leveraging task-aware annotated data as supervised signals to assist with self-supervised learning on large-scale unlabeled data has become a new trend in pre-training language models. Existing studies show that multi-task learning with large-scale supervised tasks suffers from negative effects across tasks. To tackle the challenge, we propose a task prefix guided multi-task pre-training framework to explore the relationships among tasks. We conduct extensive experiments on 40 datasets, which show that our model can not only serve as the strong foundation backbone for a wide range of tasks but also be feasible as a probing tool for analyzing task relationships. The task relationships reflected by the prefixes align transfer learning performance between tasks. They also suggest directions for data augmentation with complementary tasks, which help our model achieve human-parity results on commonsense reasoning leaderboards. Code is available at https://github.com/cooelf/CompassMTL

1 Introduction

Recent years have witnessed a growing interest in leveraging a unified pre-trained language model (PrLM) to solve a wide range of natural language processing tasks (Tay et al., 2022; Chowdhery et al., 2022; Xie et al., 2022; Zhang et al., 2022). The pre-training recipe of a PrLM is driving from self-supervised learning (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019; Lan et al., 2020; Clark et al., 2020) to multi-task learning (MTL) with a mixture of standard self-supervised tasks and various supervised tasks,

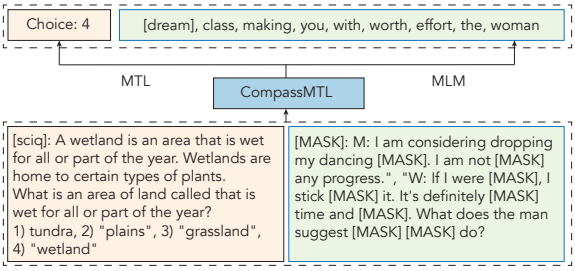

Figure 1: Input-output view. We append a task prefix for each data sequence to capture common patterns from the dataset and require the model to predict some randomly masked prefixes to capture task differences.

which takes advantage of learning from both large-scale unlabeled corpus and high-quality human-labeled datasets (Raffel et al., 2019; Aribandi et al., 2021).[1] Benefitting from supervision from related tasks, MTL approaches reduce the cost of curating deep learning models for an individual task and provide a shared representation that is generally applicable for a range of tasks (Wu et al., 2020b).

In the research line of multi-task learning for PrLMs, a typical solution is to cast all tasks into a text-to-text format and utilize an encoder-decoder PrLM such as T5 to predict the target sequences (Raffel et al., 2019; Aribandi et al., 2021). Despite the extensive efforts on leveraging supervised tasks in strengthening PrLMs, the latest trend is extreme scaling of task numbers, with little attention paid to the relationships between tasks (Sanh et al., 2021; Wei et al., 2021). Aribandi et al. (2021) investigated co-training transfer effects amongst task-families and empirically found that tasks in different families may have side effects between each other, e.g., summarization tasks generally seem to hurt performance on other task families such as dialogue system (Mehri et al., 2020),

* Work done when Zhuosheng Zhang and Wenhao Yu interned at Microsoft Cognitive Services Research group. This work was partially supported by Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

[1]Since multi-task pre-training is often implemented as an additional large-scale learning stage between language model pre-training and fine-tuning, it is also known as multi-task pre-fine-tuning in literature (Aghajanyan et al., 2021).

natural language inference (Bowman et al., 2015), and commonsense reasoning (Lourie et al., 2021).

When the task number scales up, the training of PrLMs would be more vulnerable to negative transfer due to the severe inconsistency of domain and data distribution between tasks (Wu et al., 2020b; Padmakumar et al., 2022). As one of the key concepts underlying MTL, task relationships potentially provide an effective basis for employing PrLMs in a more effective and interpretable way.

To handle the issue of negative transfer during multi-task learning, early studies have taken task relationships into account by employing a dual-process model architecture that is composed of a shared encoder and task-specific layers. The two parts are supposed to integrate the common features of all the learning tasks and explore the task relationship in a predefined manner (Zheng et al., 2019; Liu et al., 2019a; Bai et al., 2020; Ma et al., 2021), respectively. However, these methods require additional modifications to model architecture and increase the model complexity and computation cost. Therefore, they are suboptimal for applying to PrLMs in terms of generality and computational bottlenecks.

All the considerations above lay down our goal to investigate simple yet effective ways to measure the task relationship without additional cost and keep the generality of PrLMs. In this work, we propose a prefix-guided multi-task learning framework (CompassMTL) to explore the mutual effects between tasks (Figure 1) and improve model performance with complementary tasks. Targeting natural language understanding (NLU) tasks, we employ a discriminative PrLM[2] as the backbone model and train the model on 40 tasks. Experimental results show that our model achieves human-parity performance on commonsense reasoning tasks. We further probe into the task relationship entailed in the tasks prefix representations, finding that the measured relationship highly correlates with task-to-task transfer performance, and it is also of referenced value for optimizing the PrLM on a target task with its complementary tasks during MTL, i.e., fewer tasks with better performance.

In summary, our contributions are three folds:

1) A unified discriminative multi-task PrLM for NLU tasks will be released as a strong counterpart for the dominant T5-based encoder-decoder PrLMs trained with MTL.

2) A probing tool of using task prefix to explore the task relationships in large-scale MTL. We observe that the task relationships reflected by the prefixes manifest a correlation with transfer learning performance, and they help our model achieve better results with complementary tasks.

3) State-of-the-art results on a variety of NLU tasks, especially human-parity benchmark performance on commonsense reasoning leaderboards, i.e., HellaSwag and αNLI.

2 Background and Related Work

2.1 Self-supervised Pre-training

PrLMs are commonly pre-trained on large-scale corpora and then used for fine-tuning individual tasks. One of the most widely-used pre-training tasks is masked language modeling (MLM), which first masks out some tokens from the input sentences and then trains the model to predict them by the rest tokens. There are derivatives of MLM including permuted language modeling in XLNet (Yang et al., 2019) and sequence-to-sequence MLM in MASS (Song et al., 2019) and T5 (Raffel et al., 2019). Beyond the general-purpose pre-training, domain-adaptive pre-training and task-adaptive pre-training have attracted attention in recent studies.

1) Domain-adaptive Pre-training. To incorporate specific in-domain knowledge, domain-aware pre-training is designed, which directly post-trains the original PrLMs using the domain-specific corpus. Popular models have been proposed in the dialogue domain (Whang et al., 2020; Wu et al., 2020a), as well as in the medical and science domains (Lee et al., 2020; Beltagy et al., 2019; Huang et al., 2019a; Yu et al., 2022).

2) Task-adaptive Pre-training. The goal of task-adaptive pre-training is to capture task-specific skills by devising the pre-training tasks. The popular application scenarios include logical reasoning and dialogue-related tasks Kumar et al. (2020); Gu et al. (2020); Zhang and Zhao (2021); Li et al. (2021). For example, Whang et al. (2021) proposed various utterance manipulation strategies, including utterance insertion, deletion, and retrieval, to maintain dialog coherence.

[2]Also known as encoder-only PrLMs. As this work focuses on NLU tasks, we find that encoder-only PrLMs are competitive based on our empirical studies though they may lose generalizability on natural language generation tasks.

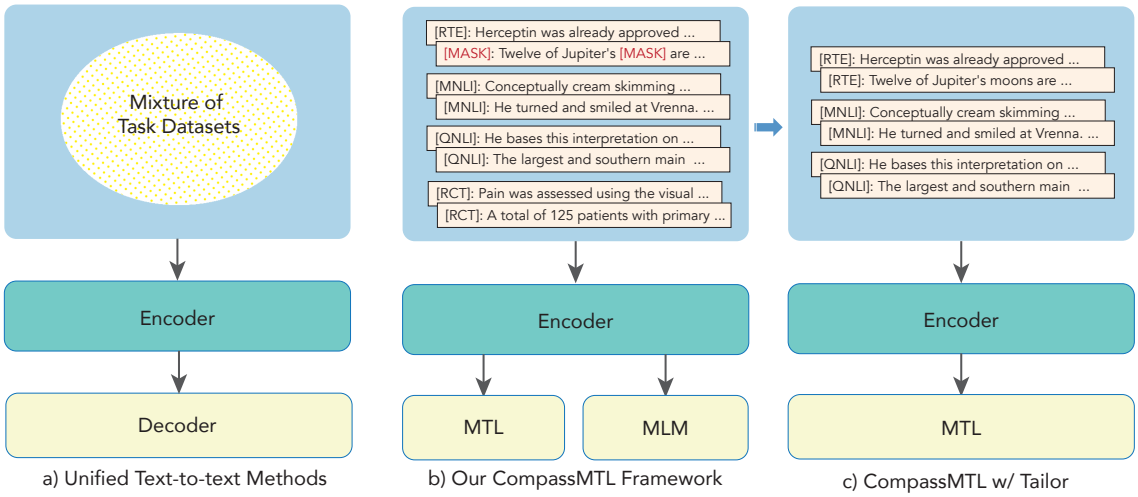

Figure 2: Comparison with existing paradigms of multi-task learning. Typical unified text-to-text methods include T5 (Raffel et al., 2019), ExT5 (Aribandi et al., 2021), FLAN (Wei et al., 2021), and T0 (Sanh et al., 2021).

2.2 Multi-task Learning for PrLMs

Our concerned MTL in the field of PrLMs is partially related to the studies of task-adaptive pre-training discussed above. The major difference is that the PrLMs in MTL are fed with human-annotated datasets instead of those automatically constructed ones for self-supervised tasks. Figure 2 overviews the paradigms of MTL PrLMs. Existing methods in this research line mostly vary in model architectures and training stages. For example, MT-DNN (Liu et al., 2019a) applied multi-task learning to train a shared model on all the target datasets in the fine-tuning stage, and there are several task-aware output modules to adapt the shared representations to each task. Recent studies, such as ExT5 (Aribandi et al., 2021), T0 (Sanh et al., 2021), and FLAN (Wei et al., 2021), commonly applied an Encoder-Decoder architecture and convert a variety of tasks into the same text-to-text format and train those tasks jointly (Figure 2-a). We argue that they are not the optimal solution considering the model complexity and the gap between original and transformed task formats, especially for natural language understanding tasks that are in a discriminative manner, e.g., classification, multiple-choice, etc. Actually, there are studies (McCann et al., 2018; Keskar et al., 2019; Li et al., 2020; Khashabi et al., 2020) that transform traditional tasks into other formats like reading comprehension or question answering and achieve better results than prior methods. These studies motivate us to explore superior model backbones and data formats, especially for the application in NLU tasks.

2.3 Modeling Task Relationships in MTL

Modeling task relationships is a classic topic in deep learning studies. Bingel and Søgaard (2017) studied the research question about what task relations make gains in traditional natural language processing tasks and investigated when and why MTL works in sequence labeling tasks such as chunking, sentence compression, POS tagging, keyphrase detection, etc. Wu et al. (2020b) found that task data alignment can significantly affect the performance of MTL and proposed architecture with a shared module for all tasks and a separate output module for each task.

Since these methods require additional modifications of model architecture, they are suboptimal for employment in PrLMs, considering computational bottlenecks and generality when task scaling. In the era of pre-trained models, Geva et al. (2021) analyzed the behavior transfer in PrLMs between related jointly-trained tasks such as QA and summarization and thus provided evidence for the extrapolation of skills as a consequence of multi-task training. ExT5 (Aribandi et al., 2021) evaluated the transfer performance among task families in a multi-task co-training setup and observed that negative transfer is common, especially when training across task families. Although there are recent studies that insert prompts to describe the task requirements in the data sequences (Liu et al., 2021; Su et al., 2022; Qin et al., 2021; Vu et al., 2022), it is still not clear whether the prompts help negative transfer or whether the prompts necessarily capture task relationships. In this work, we find that using task

prefixes along with the MLM for prefix prediction effectively indicates task relationships and helps MTL with fewer datasets but better performance.

3 Methodology

3.1 Task Format

According to prior studies (McCann et al., 2018; Keskar et al., 2019; Khashabi et al., 2020), the benchmark results on a task can be affected dramatically by training a model on different formats of the same dataset. In contrast to converting all tasks in a text-to-text manner, we choose to model our tasks in a multiple-choice-like format to minimize the format transformation for NLU tasks. Our transformation aims to ensure that each example in a task has a specific number of k candidate options during the multi-task training stage. The original pair-wise input texts are regarded as context and question in the view of the multiple-choice problem. If there is only one text given, then the question will be kept empty. For the outliers, the data will be processed as follows (Examples are provided in Appendix A.1).

1) If the number of candidate options $> k$, the redundant options will be randomly discarded;

2) If the number of candidate options $< k$, add "N/A" placeholder options.

3) If the ground truth is a list, randomly select a correct option from the gold list and randomly sample $k-1$ negative options from the held-out set[3] except the left items in the gold list.

4) If the ground truth is a list and there is an empty choice, construct the truth option manually. For example, "there is no violation"; the negative examples are constructed as the same as 3).

As a result, each training example will be formed as a sequence like {[Prefix]: context, question, option}, where [Prefix] indicates the task name in natural language such as [hellawag] prepended to each data example.

3.2 CompassMTL

Our model is encoder-only, which is based on the DeBERTa architecture (He et al., 2021). The model is trained by using both the supervised task objective and the standard self-supervised denoising objective as described below.

Suppose that we have a dataset $\mathcal{D} = \{(y_i, c_i, q_i, r)\}_{i=1}^{N}$, where c_i represents the context, q_i represents the question, r denotes a set of answer options $r = \{r_1, \ldots, r_k\}$, and y_i is the label. N is the number of training data. Each data example is formed as x = [CLS] [Prefix] c_i [SEP] q_i r_j [SEP],[4] $r_j \in r$. The goal is to learn a discriminator $g(\cdot, \cdot)$ from $\mathcal{D}$. For the supervised task, the loss function is: $\mathcal{L}_{mtl} = -\sum_{i=1}^{N}\sum_{j=1}^{k}\log(g(c_i, q_i \circ r_j))$.

At the inference phase, given any new context c_i, question q_i and options r, we use the discriminator to calculate $g(c_i, q_i \circ r_j)$ as their matching score where $\circ$ denotes concatenation. The option with the highest score is chosen as the answer for the i-th example.

Let $\hat{x}_i$ denote the masked sequence where a certain proportion of tokens in x_i are randomly replaced with a special [MASK] symbol. Using $\hat{x}_i$ as the input fed to the model in parallel with x, the self-supervised denoising objective is computed in the way of MLM: $\mathcal{L}_{mlm} = -\sum_{i=1}^{N}\sum_{j\in\mathcal{M}}\log p_\theta(t_{i,j} \mid \hat{x}_i)$, where $t_{i,j}$ is the j-th token in x_i and $\mathcal{M}$ denotes the index set of masked tokens for which the loss will be computed. To encourage the model to learn from both supervised and self-supervised signals, we combine $\mathcal{L}_{mtl}$ and $\mathcal{L}_{mlm}$ during training: $\mathcal{L} = \mathcal{L}_{mtl} + \lambda\mathcal{L}_{mlm}$ where λ is a hyper-parameter to balance the weight of the training objectives.

Compared with traditional MTL methods, CompassMTL is data-centric, without any modification of model architecture (Figure 2-b). It can be regarded as an efficient implementation of the traditional MTL method composed of a shared representation module and multiple task-aware modules. Since the data from the same datasets share the same task prefix, the prefix is supposed to reflect the common patterns from the dataset, which works in a similar operational principle to the shared representation module. During the training with our self-supervised objective, task prefixes will be randomly masked in a specific probability.[5] The model is required to distinguish the task prefixes and predict the right prefix according to the input data. Therefore, the task differences will also be necessarily captured.

3.3 Task Relationship Exploration

Regarding the task prefixes as the compass to navigate the task relationships, it is possible to use our framework to analyze the relevance of

[3] The held-out set is composed of all the candidate items in each gold list.

[4] The task prefixes are added to the model vocabulary as additional tokens to avoid tokenization.

[5] Each token in the input sequence will be masked in the same probability, including the task prefix and the rest tokens.

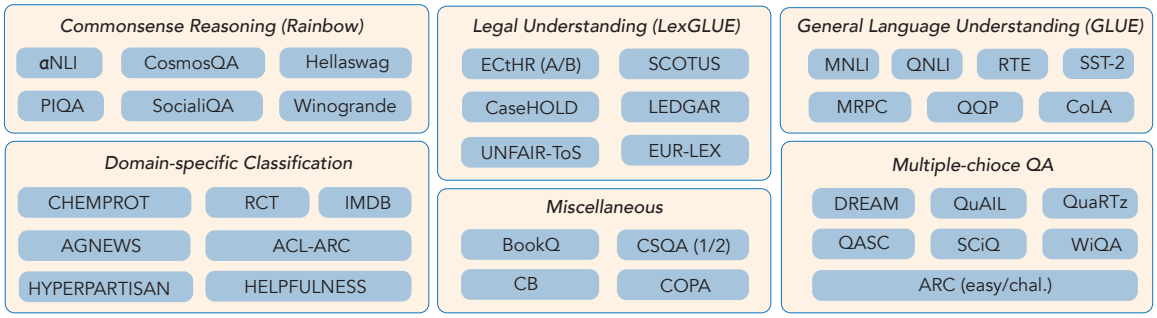

Figure 3: Task taxonomy used in this work.

tasks (Section 5.2). Our model for prefix probing experiments is slightly revised from CompassMTL, which only uses the MLM objective and is fed by the data without options to alleviate possible shortcuts in options. After the model is pre-trained with MTL, we fetch the prefix embeddings from the model embedding layer and calculate the Pearson correlation between each task pair with min-max normalization. Assuming that we have n tasks, the process will result in $n \times n$ correlation scores to indicate the task relationships.

For a target task, we can directly rank the top-related tasks according to the correclation scores and use those complementary tasks for MTL before fine-tuning a target task (Figure 2-c).

4 Experiments

4.1 Datasets

There are 40 datasets used for training our multi-task model, some of which are collected from GLUE (Wang et al., 2019b), SuperGLUE (Wang et al., 2019a), Rainbow (Lourie et al., 2021), and LexGLUE (Chalkidis et al., 2021). Figure 3 illustrates the composition of our task families.

GLUE GLUE (The General Language Understanding Evaluation benchmark) (Wang et al., 2019b) is a collection of 9 various tasks for sentence-level classification. We only use 8 of them: CoLA(Warstadt et al., 2019), SST-2 (Socher et al., 2013), MRPC (Dolan and Brockett, 2005), STS-B (Cer et al., 2017), QQP (Chen et al., 2018), QNLI (Rajpurkar et al., 2016), MNLI (Nangia et al., 2017) and RTE (Bentivogli et al., 2009).

Rainbow Rainbow (Lourie et al., 2021) is a suite of commonsense question answering tasks including αNLI (Bhagavatula et al., 2020), CosmosQA (Huang et al., 2019b), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), SocialIQA (Sap et al., 2019), Winogrande (Sakaguchi et al., 2020).

LexGLUE LexGLUE (Legal General Language Understanding Evaluation) (Chalkidis et al., 2021) is a collection of datasets for evaluating model performance across a diverse set of legal NLU tasks, which contain 7 subtasks, namely ECtHR (Task A), ECtHR (Task B), SCOTUS, EUR-LEX, LEDGAR, UNFAIR-ToS, and CaseHOLD.

Domain-specific Classification We use seven datasets that cover specific domains (biomedical and computer science publications, news, and reviews) following Gururangan et al. (2020). The datasets are CHEMPROT (Kringelum et al., 2016), RCT (Dernoncourt and Lee, 2017), ACL-ARC (Jurgens et al., 2018), HYPERPARTISAN (Kiesel et al., 2019), AGNEWS (Zhang et al., 2015), HELPFULNESS (McAuley et al., 2015), and IMDB (Maas et al., 2011).

Multiple-choice QA The datasets include DREAM (Sun et al., 2019), QuAIL (Rogers et al., 2020), QuaRTz (Tafjord et al., 2019), WiQA (Tandon et al., 2019), QASC (Khot et al., 2020), SCiQ (Welbl et al., 2017), ARC (Clark et al., 2018). We follow Sanh et al. (2021) to organize this task family.

Miscellaneous The other datasets are BookQ (Clark et al., 2019), CB (De Marneffe et al., 2019), CommonsenseQA v1/v2 (Talmor et al., 2019, 2021), and COPA (Roemmele et al., 2011). BoolQ, CB, and COPA are also collected in SuperGLUE (Wang et al., 2019a). We select those tasks as they can be easily transformed into our unified format.

4.2 Implementations

Our model is implemented using Pytorch and based on the Transformers Library (Wolf et al., 2019). To save computation, we initialize our model with the released checkpoints of DeBERTa-V3-Large, and the hyper-parameter setting generally follows DeBERTa (He et al., 2021). Our experiments

Model	Arch.	Tasks	Params.	αNLI	CosmosQA	HellaSwag	PIQA	SocialIQA	Winogrande	Average
UNICORN	Enc-Dec	6	770M	79.5	83.2	83.0	82.2	75.5	78.7	80.4
ExT5	Enc-Dec	107	770M	82.3	85.9	89.0	85.0	79.7	82.5	84.1
ExDeBERTa	Enc only	40	567M	87.9	85.3	83.6	85.5	79.6	87.0	84.8
CompassMTL	Enc only	40	567M	91.7	87.8	95.6	87.3	81.7	89.6	89.0
w/ Tailor	Enc only	14	567M	**92.5**	**88.8**	**96.1**	**88.3**	**82.2**	**90.5**	**89.7**

Table 1: Results on the Rainbow commonsense reasoning validation sets. The baseline models are $\text{UNICORN}_{\texttt{large}}$ (Lourie et al., 2021) and $\text{ExT5}_{\texttt{large}}$ (Aribandi et al., 2021). ExDeBERTa is our imitation of ExT5-style (Aribandi et al., 2021) MTL training by using DeBERTa backbone trained on 40 datasets with a multi-task objective of self-supervised denoising and supervised task objective, after which is transferred to each individual task. "w/ Tailor" denotes multi-task training with related datasets (14-subset) according to our discovery in Section 5.3.

Method	ECtHR (A)		ECtHR (B)		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ/m-F_1
BERT	71.2	63.6	79.7	73.4	68.3	58.3	71.4	57.2	87.6	81.8	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	71.6	62.0	71.9	**57.9**	87.9	82.3	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	**72.1**	57.4	88.2	83.1	95.5	80.3	72.6
Longformer	69.9	**64.7**	79.4	71.7	72.9	64.0	71.6	57.7	88.2	83.0	95.5	80.9	71.9
BigBird	70.0	62.9	78.8	70.9	72.8	62.0	71.5	56.8	87.8	82.6	95.7	81.3	70.8
Legal-BERT	70.0	64.0	80.4	**74.7**	76.4	66.5	**72.1**	57.4	88.2	83.0	96.0	83.0	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	76.6	65.9	70.7	56.6	**88.3**	83.0	96.0	82.3	75.4
ExDeBERTa	-	-	-	-	-	-	-	-	-	-	-	-	74.8
CompassMTL	71.7	60.7	80.6	73.2	**77.7**	**68.9**	67.2	42.1	88.1	82.3	**96.3**	**84.3**	76.1
w/ Tailor	**73.0**	**64.7**	**80.7**	72.3	76.3	68.6	66.9	44.9	**88.3**	**83.2**	96.2	83.2	**78.1**

Table 2: Results on LexGLUE test sets. The baseline results except ours in the last column are from Chalkidis et al. (2021). Since the LexGlue tasks except CaseHold are multi-label classification problems, the ExDeBERTa model is not directly applicable for those tasks without extra task-specific fine-tuning; thus, the results are not reported. "w/ Tailor" denotes multi-task training with the seven datasets in the same LexGLUE family.

are run on 8x32GB Tesla A100 GPUs. The maximum input sequence length is 512. Similar to Lourie et al. (2021), the implementation of CompassMTL includes two procedures. We first conduct multi-task pre-training on all the datasets and then continue to train on each target dataset alone to verify the performance. For multi-task pre-training, we use a peak learning rate of 6e-6 with a warm-up rate of 0.1. We run up to 6 epochs using a batch size of 128. The masking ratio of MLM is 0.25, and λ is set to 0.1. To avoid large-scale datasets dominating the pre-training, the training data is randomly sampled by a limit of $10k$ on the maximum dataset size according to Raffel et al. (2019). For fine-tuning experiments, the initial learning rate is selected in {3e-6, 6e-6, 8e-5} with a warm-up rate of 0.1. The batch size is selected in {16, 32}. The maximum number of epochs is chosen from {6,10}. More fine-tuning details are available in Appendix A.2.

4.3 Main Results

Our main results are reported on the Rainbow and LexGLUE benchmark datasets for comparisons with public methods. As the statistics shown in Tables 1-2, we see that CompassMTL models outperform the related public models in general. Specifically, it is observed that our encoder-only models yield better performance than the T5-based encoder-decoder models under similar model sizes. Further, the comparison in the second column discloses the potential to achieve comparable or better performance by multi-task learning with related tasks (w/ Tailor). How to find the related tasks and use them to enhance model performance will be discussed in the following section.

5 Analysis

5.1 Ablation Study

Table 3 presents our ablation study to dive into the effectiveness of different training objectives and the influence of task prefixes in our method. For the training objectives, MTL and MLM denote the training objectives of $\mathcal{L}_{mtl}$ and $\mathcal{L}_{mlm}$, respectively. The results suggest that both supervised and self-supervised tasks contribute to the overall model performance, and the supervised task is more beneficial than the self-supervised task in our study. Further, to inspect the role of the task prefixes, we

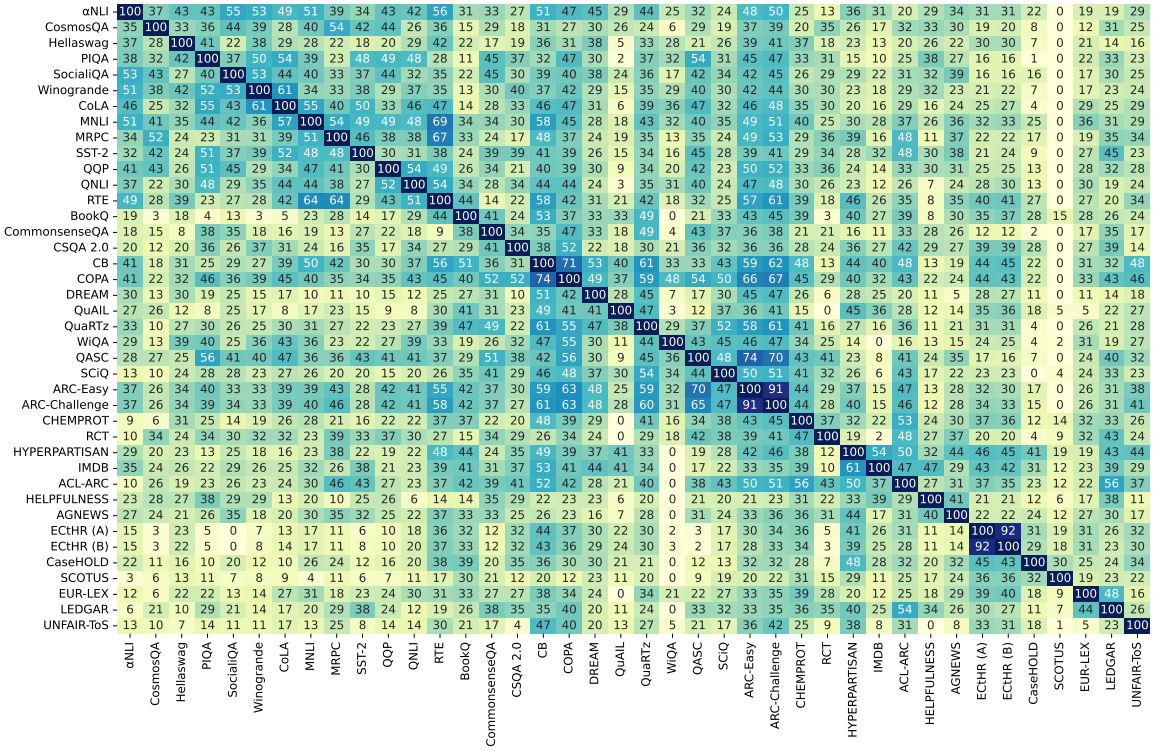

Figure 4: Heatmap of task relationships probed by prefix embeddings.

ablate the model with three conditions: 1) *must*: the prefixes are masked with the probability of 1.0; 2) *no*: the prefixes are masked with the probability of 0.0; 3) *only*: only prefixes will be masked, i.e., the prefix of each example will be masked, while the other tokens are left as original.[6] The results in Table 3 show that using prefixes ($\text{Prefix}_{\texttt{must}}$ and $\text{Prefix}_{\texttt{only}}$) indeed boosts the model performance generally.

Model	Accuracy
Single	84.6
CompassMTL	89.4
- MTL	85.0
- MLM	88.8
$\text{Prefix}_{\texttt{must}}$	89.3
$\text{Prefix}_{\texttt{no}}$	88.9
$\text{Prefix}_{\texttt{only}}$	89.1

Table 3: Ablation Study of the training objectives and task prefixes. We calculate the average accuracy scores on the development sets of all the 40 datasets.

5.2 Relationship Probing

Figure 4 illustrates the heatmap of task relationships probed by prefix embeddings. We see that the datasets inside the same task family (e.g., GLUE and Rainbow) correlate highly with each other. The LexGLUE tasks are less related to other tasks because the texts are mainly legal descriptions. In addition, the correlation scores also accord with the common practice of data augmentation. For example, the NLI datasets (MNLI, QNLI, RTE) share close relevance, and it is helpful to initialize parameters from an MNLI model to fine-tune RTE (Liu et al., 2019b; Qu et al., 2020).

We are interested in whether the probed relationship scores coordinate with the model performance transferred between tasks. We first obtain transfer accuracy between tasks in a dual-task training setup (Aribandi et al., 2021). Assume that we have 13 source tasks from GLUE and Rainbow tasks and 5 target tasks (αNLI, HellaSwag, MRPC, PIQA, QNLI, and RTE). We first train individual models using the mixture of training sets from each pair of source and target tasks, and then evaluate the model on the validation set of the target dataset. As a result, we have 5×13 transfer results. For each

[6]Note that if we ideally mask all the prefixes, the prefix tokens will not appear in the input sequence; thus, the prefix embeddings will not be updated. To avoid this issue, we follow the standard practice of training BERT-like models, where the masked tokens will experience extra processes: 1) 80% of the time, we replace masked input tokens with mask symbols; 2) 10% of the time, we replace masked input tokens with a random word; 3) The rest of the time (10% of the time) we keep the masked input tokens unchanged.

Dataset	RTE	MRPC	QNLI	HellaSwag	αNLI	Avg.
Probing	0.19	0.22	0.38	0.12	0.51	0.28
Length	-0.12	0.43	-0.17	0.04	-0.07	0.02
Vocab	0.37	-0.27	-0.001	0.09	0.31	0.10

Table 4: Pearson correlation between each relationship measure and the transfer accuracy.

Model	Tasks	RTE	MRPC	QNLI	HellaSwag	αNLI
Single	1	61.4	89.2	95.0	95.1	91.3
40-fullset	40	**92.8**	90.4	95.5	95.6	91.7
Top 5	5	92.4	**91.9**	95.3	95.6	91.6
Family	6/7	91.4	90.2	95.0	95.7	91.9
14-subset	14	91.8	90.3	**95.6**	**96.1**	**92.5**

Table 5: Complementary transfer results using different mixtures of datasets for MTL. The last three rows represent the mixture in different granularity inspired by our relationship probing.

target dataset, we calculate Pearson correlation between relationship scores and transfer accuracy among the source datasets. In Table 4, we find that the relationship scores are positively bound up with the transfer performance. The results indicate the potential to find related tasks by the relationship scores. In other words, the relationship scores essentially reflect task relationships.

Task relationships may also be reflected by shallow token distributions, such as vocabulary overlap or sentence length. To investigate if our relationship probing can be replaced by comparing the token distributions, we further analyze the correlation between the similarity of token distributions and dual-task transfer accuracy. For sentence length, we first calculate the absolute values of the average length difference between source and target datasets and then convert them to negative values (intuitively less difference in length, more close the relationship). The vocab overlap of the source and target datasets is also computed for comparison. The similarity between datasets reflects weak correlations with the transfer accuracy (2/5 and 3/5 datasets, respectively in Table 4). These results are less consistent than our probing method, which indicates that our method mines more complex patterns toward task relationships.

5.3 Complementary Transfer

To inspect whether using more datasets always leads to better performance and whether using the most related datasets can lead to competitive

Model	SQuADv1.1 EM	SQuADv1.1 F1	SQuADv2.0 EM	SQuADv2.0 F1	NER F1
Baseline	88.8	94.8	87.1	90.5	96.5
CompassMTL	89.7	95.1	88.5	91.3	96.9

Table 6: Results on the SQuAD v1.1/V2.0 and CoNLL2003 (NER) development sets. The evaluation metrics are Exact-Match (EM) and F1 scores.

Model	HellaSwag	αNLI
Human Performance	95.60	92.90
Previous SOTA	94.87	92.20
Our Results	95.94	92.80

Table 7: Leaderboard tests of HellaSwag and αNLI.

results. In this part, we conduct a complementary transfer analysis by selecting a group of datasets to train an MTL model and fine-tuning the model on target datasets. Four choices of dataset mixture are compared: 1) 40-fullset: the same as our basic setting of CompassMTL in this work; 2) Top-5 ranked dataset according to based on our probed relationship scores; 3) Family: the datasets belonged to the same family with the target dataset, i.e., 6 datasets for Rainbow tasks and 7 datasets for GLUE tasks; 4) 14-subset: the mixture of Rainbow and GLUE datasets.

Table 5 presents the comparison results. We observe that the top-5 ranked variant yields comparable, even better results than the others, which indicates that models trained with more datasets may not always bring benefits. The results also indicate that small-scale datasets (e.g., MRPC and RTE), which have relatively high average correlation scores with the other datasets, are more likely to benefit from the complementary transfer. With the tasks scaling up, the performance (family $\rightarrow$ 14-subset) may improve as more related tasks are involved in training.

5.4 Human-parity on Commonsense Reasoning Leaderboards

Table 7 presents our test evaluation on the official leaderboards of HellaSwag[7] and αNLI[8]. The submissions are based on the ensemble of three models selected according to Section 5.3. Compared with public methods that use much larger PrLMs, model ensemble, and knowledge

[7] https://leaderboard.allenai.org/hellaswag/submissions/public

[8] https://leaderboard.allenai.org/anli/submissions/public

Model	αNLI	CosmosQA	HellaSwag	PIQA	SocialIQA	Winogrande	Average
T5	68.5	69.6	56.6	67.7	65.1	62.4	65.0
UNICORN	65.3	72.8	56.2	73.3	66.1	61.8	65.9
CompassMTL	**69.1**	**72.6**	**57.7**	**73.6**	**66.6**	**64.9**	**67.4**

Table 8: Results on the Rainbow validation sets by using T5-base as the backbone model.

graphs, our models establish new state-of-the-art results and reach human-parity performance.

5.5 Beyond The Unified Format

To verify whether our model can be employed for tasks that are unavailable to be transformed into our unified format, we evaluate the effectiveness of CompassMTL by using the typical reading comprehension datasets SQuAD v1.1/2.0 (Rajpurkar et al., 2016, 2018) and named entity recognition (NER) dataset CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003), which represent extractive question answering and sequence labeling task formats, respectively. We first replicate the baselines for fine-tuning QA and NER tasks using the Transformers toolkit.[9] For comparison, we initialize the baseline parameters with our model weights to see if CompassMTL is better than the baselines. Results in Table 6 show that our model is generally effective across formats. The results also indicate that CompassMTL can serve as a strong off-the-shelf representation encoder that is applicable for new tasks without needing to be pre-trained again.

5.6 Implementation Using The T5 Backbone

Although our method is implemented by the encoder-only backbone to compete in NLU tasks, it is supposed to be generally applicable to other kinds of PrLMs, such as encoder-decoder T5. To verify the effectiveness, we employ the pre-trained T5-base model (Raffel et al., 2019) as the backbone. We use the Rainbow datasets for MTL and convert the data into text-to-text format following the standard processing for T5 training, with task prefixes inserted before each data sequence. The baselines are the single-task T5 trained on each individual task and UNICORN (Lourie et al., 2021) trained on the Rainbow datasets. Results in Table 8 verify that our method is generally effective.

[9] https://github.com/huggingface/transformers.

6 Conclusions

This work presents a task prefix guided multi-task method by making use of task prefix to explore the mutual effects between tasks and improve model performance with complementary tasks. Our released model can not only serve as the strong foundation backbone for a wide range of NLU tasks but also be used as a probing tool for analyzing task relationships. Our model shows generalizable advances over tasks in diverse formats and establishes human-parity results on commonsense reasoning tasks. Based on our pre-trained model, we find that the prefixes necessarily reflect task relationships, which correlate with transfer learning performance between tasks and suggest directions for data augmentation of complementary tasks. In summary, our work has the following prospects for future studies:

1) Collaborative multi-task learning of PrLMs. The recipe of using task prefixes in conjunction with prefix prediction in MLM training has shown effective for large-scale MTL pre-training.

2) Suggestive choice for data augmentation. The task relationships probed by the prefix embeddings have shown informative in finding the complementary tasks. Using complementary tasks helps obtain better performance for a target task, especially for small-scale task datasets.

3) Guidance for skill-aware model evaluation. The discovery of task relationships may help determine redundant datasets that assess similar patterns of models. Recently, there has been a trend to evaluate the comprehensive skills of deep learning models by using a large number of datasets (Srivastava et al., 2022), the selection of distinctive datasets can be guided by our relationship discovery criteria to avoid evaluation redundancy and save computation.

Limitations. We acknowledge the major limitation of this work is that our model may not readily apply to new tasks. It is based on the common assumption of MTL that the set of tasks is known at training time. Adaptation to new tasks could be future work.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*.

Lu Bai, Yew-Soon Ong, Tiantian He, and Abhishek Gupta. 2020. Multi-task gradient descent for multi-task learning. *Memetic Computing*, 12(4):355–369.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *ACL-PASCAL*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. 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. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs.

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

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Mor Geva, Uri Katz, Aviv Ben-Arie, and Jonathan Berant. 2021. What's in your head? emergent behaviour in multi-task transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8201–8215.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2020. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. *arXiv:2012.01775*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Kexin Huang, Jaan Altosaar, and R. Ranganath. 2019a. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019b. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Unifying question answering, text classification, and regression via span extraction. *arXiv preprint arXiv:1904.09286*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. 2016. Chemprot-3.0: a global chemical biology diseases mapping. *Database*, 2016.

Pawan Kumar, Dhanajit Brahma, Harish Karnick, and Piyush Rai. 2020. Deep attentive ranking networks for learning to order sentences. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8115–8122. AAAI Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, D. Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Lu Li, Chenliang Li, and Donghong Ji. 2021. Deep context modeling for multi-turn response selection in dialogue systems. *Information Processing & Management*, 58(1):102415.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified

MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. 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. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13480–13488.

Yanbao Ma, Hao Xu, Junzhou He, Kun Qian, and Tiebing Li. 2021. Adaptive transfer learning via fine-grained multi-task pre-training. In *2021 4th International Conference on Algorithms, Computing and Artificial Intelligence*, pages 1–5.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 43–52. ACM.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The RepEval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Vishakh Padmakumar, Leonard Lausen, Miguel Ballesteros, Sheng Zha, He He, and George Karypis. 2022. Exploring the role of task transferability in large-scale multi-task learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2542–2550, Seattle, United States. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Yujia Qin, Xiaozhi Wang, Yusheng Su, Yankai Lin, Ning Ding, Zhiyuan Liu, Juanzi Li, Lei Hou, Peng Li, Maosong Sun, et al. 2021. Exploring low-dimensional intrinsic task subspace via prompt tuning. *arXiv preprint arXiv:2110.07867*.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. 2020. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. In *International Conference on Learning Representations*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv: 1910.10683*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. 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. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8722–8731.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. 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. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, et al. 2022. On transferability of prompt tuning for natural language processing. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. QuaRTz: An open-domain dataset of qualitative relationship questions. 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 5941–5946, Hong Kong, China. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of ai through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. 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 6076–6085, Hong Kong, China. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2022. 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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

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

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. *INTERSPEECH*.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020a. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Sen Wu, Hongyang R. Zhang, and Christopher Ré. 2020b. Understanding and improving information transfer in multi-task learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Wenhao Yu, Chenguang Zhu, Yuwei Fang, Donghan Yu, Shuohang Wang, Yichong Xu, Michael Zeng, and Meng Jiang. 2022. Dict-bert: Enhancing language model pre-training with dictionary. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1907–1918.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2022. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. *arXiv preprint arXiv:2204.03508*.

Zhuosheng Zhang and Hai Zhao. 2021. Structural pre-training for dialogue comprehension. 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 5134–5145, Online. Association for Computational Linguistics.

Zimu Zheng, Yuqi Wang, Quanyu Dai, Huadi Zheng, and Dan Wang. 2019. Metadata-driven task relation discovery for multi-task learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4426–4432. ijcai.org.

Context	Question	Option(s)
[sciq] A wetland is an area that is wet for all or part of the year. Wetlands are home to certain types of plants.	What is an area of land called that is wet for all or part of the year?	["tundra", "plains", "grassland", "wetland"]
[commonsense_qa] revolving door	A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?	["bank", "library", "department store", "mall", ~~"new york"~~]
[dream] M: I am considering dropping my dancing class. I am not making any progress.", "W: If I were you, I stick with it. It's definitely worth time and effort.	What does the man suggest the woman do?	["Consult her dancing teacher.", "Take a more interesting class.", "Continue her dancing class.", "N/A"]
[scotus] The Interstate Commerce Commission, acting under § 19a of the Interstate Commerce Act, ordered the appellant to furnish certain inventories, schedules, maps and charts of its pipe line property ...	-	["Unions", "Economic Activity", "Judicial Power", "Federalism"]
[unfair_tos] you must provide accurate and complete data during the registration and update your registration data if it changes .	-	["there is no unfair contractual term", "Limitation of liability", "Unilateral termination", "Arbitration"]

Table 9: Examples of transformed datasets.

A Appendix

A.1 Examples of transformed datasets

Table 9 shows examples of transformed datasets. The first column presents the standard multiple-choice dataset, followed by four types of outlier datasets (Section 3.1) that are transformed into our unified format.

A.2 Fine-tuning Details

According to Section 3.1, our training datasets are converted into a multiple-choice-like format for multi-task pre-training. During fine-tuning, because our evaluated GLUE and Rainbow tasks for public comparisons are either single-label classification or multiple-choice tasks, the conversion would not affect the performance according to our preliminary experiments as the predictions can be easily mapped to the original formats by choosing the best-ranked options. For the other tasks, such as the multi-label classification tasks in LexGLUE, where the conversion will result in the clip of ground-true labels, we use the original datasets for fine-tuning and initialize the corresponding baseline models with our pre-trained weights after MTL. The criteria for choosing the baseline models for different types of tasks basically follows the standard practice in literature (He et al., 2021; Chalkidis et al., 2021).