Exploring and Predicting Transferability across NLP Tasks

Tu Vu[1]★ **Tong Wang**[2] **Tsendsuren Munkhdalai**[2] **Alessandro Sordoni**[2]
Adam Trischler[2] **Andrew Mattarella-Micke**[3] **Subhransu Maji**[1] **Mohit Iyyer**[1]

University of Massachusetts Amherst[1] Microsoft Research Montreal[2] Intuit AI[3]

{tuvu,smaji,miyyer}@cs.umass.edu
{tong.wang,tsendsuren.munkhdalai}@microsoft.com
{alsordo,adam.trischler}@microsoft.com
andrew_mattarella-micke@intuit.com

Abstract

Recent advances in NLP demonstrate the effectiveness of training large-scale language models and transferring them to downstream tasks. *Can fine-tuning these models on tasks other than language modeling further improve performance?* In this paper, we conduct an extensive study of the transferability between 33 NLP tasks across three broad classes of problems (text classification, question answering, and sequence labeling). Our results show that transfer learning is more beneficial than previously thought, especially when target task data is scarce, and can improve performance even with low-data source tasks that differ substantially from the target task (e.g., part-of-speech tagging transfers well to the DROP QA dataset). We also develop *task embeddings* that can be used to predict the most transferable source tasks for a given target task, and we validate their effectiveness in experiments controlled for source and target data size. Overall, our experiments reveal that factors such as data size, task and domain similarity, and task complexity all play a role in determining transferability.

1 Introduction

With the advent of methods such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), the dominant paradigm for developing NLP models has shifted to transfer learning: first, pretrain a large language model, and then fine-tune it on the target dataset. Prior work has explored whether fine-tuning on intermediate source tasks before the target task can further improve this pipeline (Phang et al., 2018), but the conditions for successful transfer remain opaque, and choosing arbitrary source tasks can even adversely impact downstream performance (Wang et al., 2019b). Our work has two main contributions: (1) we perform a large-scale empirical study across 33 different datasets to shed light on the transferability between NLP tasks, and (2) we develop *task embeddings* to predict which source tasks to use for a given target task.

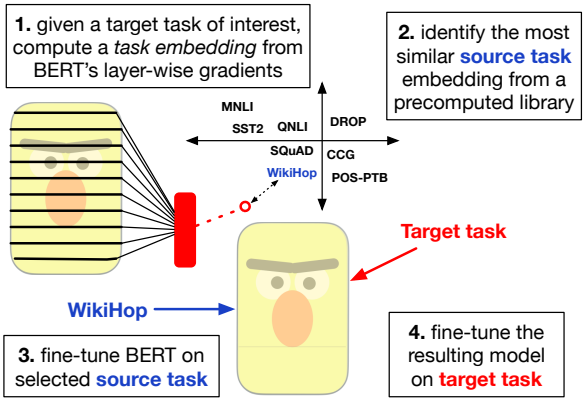

Figure 1: A demonstration of our task embedding pipeline. Given a target task, we first compute its task embedding and then identify the most similar source task embedding (in this example, WikiHop) from a precomputed library via cosine similarity. Finally, we perform intermediate fine-tuning of BERT on the selected source task before fine-tuning on the target task.[1]

Our study includes over 3,000 combinations of tasks and data regimes within and across three broad classes of problems (text classification, question answering, and sequence labeling), which is considerably more comprehensive than prior work (Wang et al., 2019a; Talmor and Berant, 2019a; Liu et al., 2019a). Our results show that transfer learning is more beneficial than previously thought (Wang et al., 2019b), especially for low-data target tasks, and even low-data source tasks that are on the surface very different than the target task can result in transfer gains. While previous work has recommended using the amount of labeled data as a criterion to select source

★ Part of this work was done during an internship at Microsoft Research.

[1]Credit to Jay Alammar for creating the BERT image.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7882–7926,
November 16–20, 2020.

tasks (Phang et al., 2018), our analysis suggests that the similarity between the source and target tasks and domains are crucial for successful transfer, particularly in data-constrained regimes.

Motivated by these results, we move on to a more practical research question: *given a particular target task, can we predict which source tasks (out of some predefined set) will yield the largest transfer learning improvement, especially in low-data settings?* We address this challenge by learning embeddings of tasks that encode their individual characteristics (Figure 1). More specifically, we process all examples from a dataset through BERT and compute a task embedding based on the model's gradients with respect to the task-specific loss, following recent meta-learning work in computer vision (Achille et al., 2019). We empirically demonstrate the practical value of these task embeddings for selecting source tasks (via simple cosine similarity) that effectively transfer to a given target task. To the best of our knowledge, this is the first work that builds explicit representations of NLP tasks to investigate transferability.

We publicly release our task library, which consists of pretrained models and task embeddings for the 33 NLP tasks we study, along with a codebase that computes task embeddings for new tasks and identifies source tasks that will likely yield positive transferability.[2]

2 Exploring task transferability

To shed light on the transferability between different NLP tasks,[3] we perform an empirical study with 33 tasks across three broad classes of problems: text classification/regression (CR), question answering (QA), and sequence labeling (SL).[4] In each experiment, we follow the STILTs pipeline of Phang et al. (2018) by taking a pretrained BERT model,[5] fine-tuning it on an intermediate *source* task, and then fine-tuning the resulting model on a *target* task. We explore in-class and out-of-class transfer in both data-rich and data-constrained regimes and demonstrate that positive transfer can occur in a more diverse array of settings than previously thought (Wang et al., 2019b).

[2]Library and code available at `http://github.com/tuvuumass/task-transferability`.

[3]We define a *task* as a (dataset, objective function) pair.

[4]We divide tasks into classes based on how they are modeled; there is considerable in-class linguistic diversity.

[5]We use BERT-Base Uncased, which has 12 layers, 768-d hidden size, 12 heads, and 110M total parameters.

Task	**|Train|**
text classification/regression (CR)	
SNLI (Bowman et al., 2015)	570K
MNLI (Williams et al., 2018)	393K
QQP (Iyer et al., 2017)	364K
QNLI (Wang et al., 2019b)	105K
SST-2 (Socher et al., 2013)	67K
SciTail (Khot et al., 2018)	27K
CoLA (Warstadt et al., 2019)	8.5K
STS-B (Cer et al., 2017)	7K
MRPC (Dolan and Brockett, 2005)	3.7K
RTE (Dagan et al., 2005, et seq.)	2.5K
WNLI (Levesque, 2011)	634
question answering (QA)	
SQuAD-2 (Rajpurkar et al., 2018)	162K
NewsQA (Trischler et al., 2017)	120K
HotpotQA (Yang et al., 2018)	113K
SQuAD-1 (Rajpurkar et al., 2016)	108K
DuoRC-p (Saha et al., 2018)	100K
DuoRC-s (Saha et al., 2018)	86K
DROP (Dua et al., 2019)	77K
WikiHop (Welbl et al., 2018)	51K
BoolQ (Clark et al., 2019)	16K
ComQA (Abujabal et al., 2019)	11K
CQ (Bao et al., 2016)	2K
sequence labeling (SL)	
ST (Bjerva et al., 2016)	43K
CCG (Hockenmaier and Steedman, 2007)	40K
Parent (Liu et al., 2019a)	40K
GParent (Liu et al., 2019a)	40K
GGParent (Liu et al., 2019a)	40K
POS-PTB (Marcus et al., 1993)	38K
GED (Yannakoudakis et al., 2011)	29K
NER (Tjong Kim Sang and De Meulder, 2003)	14K
POS-EWT (Silveira et al., 2014)	13K
Conj (Ficler and Goldberg, 2016)	13K
Chunk (Tjong Kim Sang and Buchholz, 2000)	9K

Table 1: Datasets used in our experiments, grouped by task class and sorted by training dataset size.

2.1 Experimental setup

We denote a dataset $D = \{(x^i, y^i)\}_{i=1}^n$, with n total examples of inputs x and associated outputs y. Each input x, which can be either a single text or a concatenation of multiple text segments (e.g., a question-passage pair), is represented as:

$$[\text{CLS}]\ w_1^1\ w_2^1\ \ldots w_{L_1}^1\ [\text{SEP}]\ w_1^2\ w_2^2\ \ldots w_{L_2}^2,$$

where w_j^i is token i of the j^{th} segment, [CLS] is a special symbol for classification output, and [SEP] is a special symbol to separate any text segments if they exist. Finally, each task is solved by applying a classification layer over either the final [CLS] token representation (for CR) or the entire sequence

of final layer token representations (for QA or SL). For both stages of fine-tuning, we follow Devlin et al. (2019) by backpropagating into all model parameters for a fixed number of epochs.[6] While individual task performance can likely be further improved with more involved hyperparameter tuning for each experimental setting, we standardize hyperparameters across each of the three classes to cut down on computational expense, following prior work (Phang et al., 2018; Wang et al., 2019b).

2.1.1 Datasets & data regimes

Table 1 lists the 33 datasets in our study.[7] We select these datasets by mostly following prior work: nine of the eleven CR tasks come from the GLUE benchmark (Wang et al., 2019b); all eleven QA tasks are from the MultiQA repository (Talmor and Berant, 2019b); and all eleven SL tasks were used by Liu et al. (2019a). We consider all possible pairs of source and target datasets;[8] while some training datasets contain overlapping examples (e.g., SQuAD-1 and 2), we evaluate our models on target development sets, which do not contain overlap.

For each (source, target) dataset pair, we perform transfer experiments in three data regimes to examine the impact of data size on SOURCE → TARGET transfer: FULL → FULL , FULL → LIMITED , and LIMITED → LIMITED. In the FULL training regime, all training data for the associated task is used for fine-tuning. In the LIMITED setting, we artificially limit the amount of training data by randomly selecting 1K training examples without replacement, following Phang et al. (2018); since fine-tuning BERT can be unstable on small datasets (Devlin et al., 2019), we perform 20 random restarts for each experiment and report the mean.[9]

We measure the impact of transfer learning by computing the *relative transfer gain* given a source task s and target task t. More concretely, if a baseline model that is directly fine-tuned on the target dataset (without any intermediate fine-tuning) achieves a performance of p_t, while a transferred model achieves a performance of $p_{s\to t}$, the relative transfer gain is defined as: $g_{s\to t} = \frac{p_{s\to t} - p_t}{p_t}$.

FULL → FULL			
↓src,tgt→	CR	QA	SL
CR	6.3 (11)	3.4 (10)	0.3 (10)
QA	3.2 (10)	9.5 (11)	0.3 (9)
SL	5.3 (8)	2.5 (10)	0.5 (11)
FULL → LIMITED			
	CR	QA	SL
CR	56.9 (11)	36.8 (10)	2.0 (10)
QA	44.3 (11)	63.3 (11)	5.3 (11)
SL	45.6 (11)	39.2 (6)	20.9 (11)
LIMITED → LIMITED			
	CR	QA	SL
CR	23.7 (11)	7.3 (11)	1.1 (11)
QA	37.3 (11)	49.3 (11)	4.2 (11)
SL	29.3 (10)	30.0 (8)	10.2 (11)

Table 2: A summary of our transfer results for each combination of the three task classes in the three data regimes. Each cell represents the relative gain of the *best* source task in the source class (row) for a given target task, averaged across all of target tasks in the target class (column). In parentheses, we additionally report the number of target tasks (out of 11) for which at least one source task results in a positive transfer gain. The diagonal cells indicate in-class transfer.

2.2 Analyzing the transfer results

Table 2 contains the results of our transfer experiments across each combination of classes and data regimes.[10] In each cell, we first compute the transfer gain of the *best* source task for each target task in a particular class, and then average across all target tasks in the same class. We summarize our findings as follows:

- Contrary to prior belief, transfer gains are possible even when the source dataset is small.
- Out-of-class transfer succeeds in many cases, some of which are unintuitive.
- Factors other than source dataset size, such as the similarity between source and target tasks, matter more in low-data regimes.

In the rest of this section, we analyze each of these three findings in more detail.

[6] We fine-tune all CR and QA tasks for three epochs, and SL tasks for six epochs, using the Transformers library (Wolf et al., 2019) and its recommended hyperparameters.

[7] Appendix A.1 contains more details about dataset characteristics and their associated evaluation metrics.

[8] All experiments conducted on a GPU cluster operating on renewable energy.

[9] See Appendix B for variance statistics. We resample 1K examples for each restart; for tasks with fewer than 1K training examples, we use the full training dataset.

[10] See Appendix B for tables for each individual task.

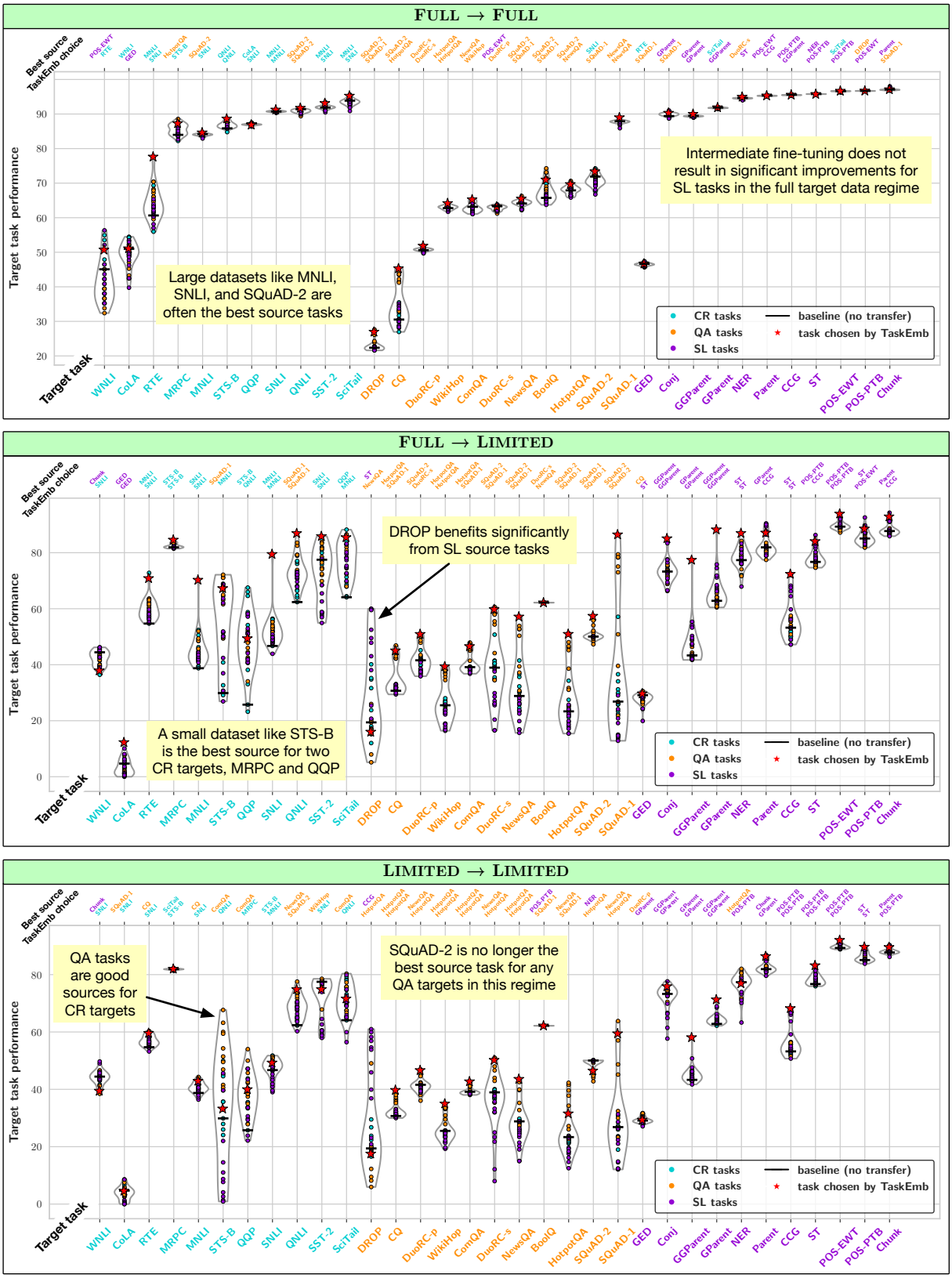

Figure 2: In these plots (best viewed in zoom with color), each violin corresponds to a target task in the specified data regime. Each point inside a violin represents an individual source task; its color denotes task class, and its y-coordinate denotes target task performance after transfer. Above each violin, we provide the *best* source task (highest point within the violin) and TASKEMB's top-ranked source task (the red star). The horizontal black line in each violin represents the baseline target task performance of BERT without intermediate fine-tuning. **TASKEMB generally selects source tasks that yield positive transfer, and often selects the best source task.**

In-class transfer: The diagonal of each block of Table 2 shows the results for in-class transfer, in which source tasks are from the same class as the target task. Across all three data regimes, most target tasks benefit from in-class transfer, and the average transfer gain is larger for CR and QA tasks than for SL tasks. Changing the data regimes significantly impacts the average transfer gain, which is lowest in the FULL → FULL regime (+5.4% average relative gain across all tasks) and highest in the FULL → LIMITED regime (+47.0%). In general, tasks with fewer training examples benefit the most from transfer, such as RTE (+17.0 accuracy points) and CQ (+14.9 F1), and the best source tasks in the FULL → FULL regime tend to be data-rich tasks such as MNLI, SNLI, and SQuAD-2 (Figure 2).[11]

Out-of-class transfer: We switch gears now to out-of-class transfer, in which the source task comes from a different class than the target task. The off-diagonal entries of each block of Table 2 summarize our results. In general, we observe that most tasks benefit from out-of-class transfer, although the magnitude of the transfer gains is lower than for in-class transfer, and that CR and QA tasks benefit more than SL tasks (similar to our in-class transfer results). While some of the results are intuitive (e.g., SQuAD is a good source task for QNLI, which is an entailment task built from QA pairs), others are more difficult to explain (using part-of-speech tagging as a source task for DROP results in huge transfer gains in limited target regimes).

Large source datasets are not always best for data-constrained target tasks: Phang et al. (2018) observe that source data size is a good heuristic to obtain positive transfer gain. In the FULL → LIMITED regime, we find to the contrary that the largest source datasets do not always result in the largest transfer gains. For CR tasks, MNLI/SNLI are the best sources for only four targets (three of which are entailment tasks), compared to seven in FULL → FULL . STS-B, which is much smaller than MNLI and SNLI, is the best source for MRPC and QQP, while MRPC, an even smaller dataset, is the best source for STS-B. As STS-B, QQP, and MRPC are all sentence similarity and paraphrase tasks, this result suggests that the similarity between the source and target tasks matters more for data-constrained targets. We observe similar task similarity patterns for QA (the best source for WikiHop is the other multi-hop QA task, HotpotQA) and SL (POS-PTB is the best source for POS-EWT, the only other POS tagging task). However, the large SQuAD-2 dataset is almost always the best source within QA. Another important factor especially apparent in our QA tasks is domain similarity (e.g., SQuAD and several other datasets were all built from Wikipedia).

When does transfer work with data-constrained sources? We now turn to the LIMITED → LIMITED regime, which eliminates the source data size confound. For CR, STS-B is the best source for six targets out of 11, including four entailment tasks (MNLI, QNLI, SNLI, SciTail), whereas MNLI/SNLI are the best sources for only two tasks (RTE, WNLI). This result suggests that source/target task similarity, which we found to be a factor for the FULL → LIMITED , is not the only important factor for effective transfer in data-constrained scenarios. We hypothesize that the complexity of the source task can also play a role: perhaps regression objectives (as used in STS-B) are more useful for transfer learning than classification objectives (MNLI/SNLI). Unknown factors may also play a role: in QA, SQuAD-2 is no longer the best source for any targets, while NewsQA is the best source for five tasks.

3 Predicting task transferability

The above analysis suggests that no single factor (e.g., data size, task and domain similarity, task complexity) is predictive of transfer gain across all of our settings. *Given a novel target task, how can we identify the single source task that maximizes transfer gain?* One straightforward but extremely expensive approach is to enumerate every possible (source, target) task combination. Work on multi-task learning within NLP offers a more practical alternative by developing feature-based models to identify task and dataset characteristics that are predictive of task synergies (Bingel and Søgaard, 2017). Here, we take a different approach, inspired by recent computer vision methods (Achille et al., 2019), by computing *task embeddings* from layer-wise gradients of BERT. Our approach generally outperforms baseline methods that use the data size heuristic (Phang et al., 2018) and the gradients of the learning curve (Bingel and Søgaard, 2017) in terms of selecting the most transferable source tasks across settings.

[11] As in Phang et al. (2018), we find that intermediate fine-tuning reduces variance across random restarts (Appendix B).

3.1 Task embedding methods

We develop two methods for computing task embeddings from BERT. The first, TEXTEMB, is computed by pooling BERT's representations across an entire dataset, and as such captures properties of the text and domain. The second, TASKEMB, relies on the correlation between the fine-tuning loss function and the parameters of BERT, and encodes more information about the type of knowledge and reasoning required to solve the task.

TEXTEMB: As our analysis indicates that domain similarity is a relevant factor for transfer, we first explore a simple method based on averaging BERT token-level representations of the inputs. Given a dataset D, we process each input sample x^i through the pretrained BERT model without any finetuning and compute $\boldsymbol{h}_x$, the average of final layer token-level representations. The final task embedding is the average of these pooled vectors over the entire dataset: $\sum_{x \in D} \frac{\boldsymbol{h}_x}{|D|}$. This method captures linguistic properties of the input text x and does not depend on the training labels y.

TASKEMB: Ideally, we want a way of capturing task similarity beyond just input properties represented by TEXTEMB. Following the methodology of TASK2VEC (Achille et al., 2019), which develops task embeddings for meta-learning over vision tasks, we create representations of tasks derived from the Fisher information matrix (or simply *Fisher*). The Fisher captures the curvature of the loss surface (the sensitivity of the loss to small perturbations of model parameters), which intuitively tells us which of the model parameters are most useful for the task and thus provides a rich source of knowledge about the task itself.

To begin, we fine-tune BERT on the training dataset of a given task; the model without the final task-specific layer forms our *feature extractor*. Next, we feed the entire training dataset into the model and compute the task embedding based on the Fisher of the feature extractor's parameters (weights) θ, i.e., the expected covariance of the gradients of the log-likelihood with respect to θ:

$$F_\theta = \mathop{\mathbb{E}}_{x,y \sim P_\theta(x,y)} \nabla_\theta \log P_\theta(y|x) \nabla_\theta \log P_\theta(y|x)^T.$$

In our experiments, we compute the *empirical Fisher*, which uses the training labels instead of sampling from $P_\theta(x, y)$:

$$F_\theta = \frac{1}{n} \sum_{i=1}^{n} \left[\nabla_\theta \log P_\theta(y^i|x^i) \nabla_\theta \log P_\theta(y^i|x^i)^T \right],$$

and only consider the diagonal entries to reduce computational complexity. Additionally, we consider the Fisher F_ϕ with respect to the feature extractor's outputs (activations) ϕ, which encodes useful features about the inputs to solve the task. The diagonal F_ϕ is averaged over the input tokens and over the entire dataset.[12]

We explore task embeddings derived from the diagonal Fisher of different components of BERT, including the token embeddings, multi-head attention, feed-forward network, and the layer output, performing layer-wise averaging. Since our base model is BERT, this method may result in high-dimensional task embeddings (from 768-d to millions of dimensions). While one can optionally perform dimensionality reduction (e.g., through PCA), all of our experiments are conducted directly on the original task embeddings.

3.2 Task embedding evaluation

We investigate whether a high similarity between two different task embeddings correlates with a high degree of transferability between those two tasks. Our evaluation centers around the meta-task of selecting the best source task for a given target task. Specifically, given a target task, we rank all the other source tasks in our library in descending order by the cosine similarity[13] between their task embeddings and the target task's embedding. This ranking is evaluated using two metrics: (1) the average rank ρ of the source task with the highest absolute transfer gain from Section 2's experiments, and (2) the Normalized Discounted Cumulative Gain (NDCG; Järvelin and Kekäläinen, 2002), a common information retrieval measure that evaluates the quality of the entire ranking, not just the rank of the best source task.[14] The NDCG

[12] While Fisher matrices are theoretically more comparable when the feature extractor is fixed during fine-tuning, as done in TASK2VEC, we find empirically that TASKEMB computed from a fine-tuned task-specific BERT result in better correlations to task transferability in data-constrained scenarios. We leave further exploration of this phenomenon to future work.

[13] We leave the exploration of asymmetric similarity metrics to future work.

[14] We use NDCG instead of Spearman correlation, as the latter penalizes top-ranked and bottom-ranked mismatches with the same weight.

at position p is defined as: $\text{NDCG}_p = \frac{\text{DCG}_p(R_{pred})}{\text{DCG}_p(R_{true})}$, where R_{pred}, R_{true} are the predicted and gold rankings of the source tasks, respectively; and $\text{DCG}_p(R) = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$, where rel_i is the relevance (target performance) of the source task with rank i in the evaluated ranking R.[15] An NDCG of 100% indicates a perfect ranking.

Aggregating similarity signals from embedding spaces: For our TASKEMB approach, we aggregate rankings from all of the different components of BERT rather than evaluate each component-specific ranking separately.[16] We expect that task embeddings derived from different components might contain complementary information about the task, which motivates this decision. Concretely, given a target task t, assume that $r_{1:c}$ are the rank scores assigned to a source task s by c different components of BERT. Then, the aggregated score is computed according to the reciprocal rank fusion algorithm (Cormack et al., 2009): $\text{RRF}(s) = \sum_{i=1}^{c} \frac{1}{60 + r_i}$. We also use this approach to aggregate rankings from TEXTEMB and TASKEMB, which results in TEXT + TASK.

3.3 Baseline methods

DATASIZE: To measure the effect of data size, we compare rankings derived from TEXTEMB and TASKEMB to DATASIZE, a heuristic baseline that ranks all source tasks by the number of training examples.

CURVEGRAD: We also consider CURVEGRAD, a baseline that uses the gradients of the loss curve of BERT for each task. Bingel and Søgaard (2017) find such learning curve features to be good predictors of gains from multi-task learning. They suggest that multi-task learning is more likely to work when the main tasks quickly plateau (small negative gradients) while the auxiliary tasks continue to improve (large negative gradients). Following the setup in Bingel and Søgaard (2017), we fine-tune BERT on each source task for a fixed number of steps (i.e., 10,000) and compute the gradients of the loss curve at 10, 20, 30, 50 and 70 percent of the fine-tuning process. Given a target task, we rank all the source tasks in descending order by the gradients and aggregate the rankings using the reciprocal rank fusion algorithm.

3.4 Source task selection experiments

The average performance of selecting the best source task across target tasks using different methods is shown in Table 3.[17] Here, we provide an overview and analysis of these results.

Baselines: DATASIZE is a good heuristic when the full source training data is available, but it struggles in all out-of-class transfer scenarios as well as on SL tasks, for which most datasets contain roughly the same number of examples (Table 1).[18] CURVEGRAD lags far behind DATASIZE in most cases, though its performance is better on SL tasks in the FULL → FULL regime. This indicates that CURVEGRAD cannot reliably predict the most transferable source tasks in our transfer scenarios.

TEXTEMB and TASKEMB improve transferability prediction: Table 3 shows that TEXTEMB performs better than DATASIZE on average, especially within the limited data regimes. Interestingly, TEXTEMB underperforms significantly on CR tasks compared to QA and SL. We theorize that this effect is partly due to the relative homogeneity of the QA and SL datasets (i.e., many QA datasets use Wikipedia while many SL tasks are extracted from the Penn Treebank) compared to the more diverse CR datasets. If TEXTEMB captures mainly domain similarity, then it may struggle when that is not a relevant transfer factor.

TASKEMB can substantially boost the quality of the rankings, frequently outperforming the other methods across different classes of problems, data regimes, and transfer scenarios. These results demonstrate that the task similarity between the computed embeddings is a robust predictor of effective transfer. The ensemble of TEXT + TASK

[15] In our experiments, we set p to the number of source tasks in each setting.

[16] We observe that rankings derived from certain components are more useful than others (e.g., token embeddings are crucial for classification), but aggregating across all components generally outperforms individual ones.

[17] In the LIMITED settings, we report the mean results across random restarts.

[18] All methods obtain a higher NDCG score on SL tasks in the FULL → FULL regime because there is little difference in target task performance between source tasks here (see Figure 2), and thus the rankings are not penalized heavily.

	FULL → FULL				FULL → LIMITED				LIMITED → LIMITED			
	in-class (10)		*all-class (32)*		*in-class (10)*		*all-class (32)*		*in-class (10)*		*all-class (32)*	
Method	ρ	NDCG	ρ	NDCG	ρ	NDCG	ρ	NDCG	ρ	NDCG	ρ	NDCG
classification / regression												
DATASIZE	3.6	80.4	8.5	74.7	3.8	62.9	9.8	54.6	-	-	-	-
CURVEGRAD	5.5	68.6	17.8	64.9	6.4	45.2	18.8	35.0	5.9	50.8	13.3	42.4
TEXTEMB	5.2	76.4	13.1	71.3	3.5	60.3	8.6	52.4	4.8	61.4	13.2	43.9
TASKEMB	2.8	82.3	6.2	76.7	3.4	68.2	**8.2**	60.9	**4.2**	62.6	11.6	**44.8**
TEXT+TASK	**2.6**	**83.3**	**5.6**	**78.0**	**3.3**	**69.5**	**8.2**	**62.0**	**4.2**	**62.7**	**11.4**	**44.8**
question answering												
DATASIZE	**3.2**	84.4	13.8	63.5	2.3	77.0	13.6	40.2	-	-	-	-
CURVEGRAD	8.3	64.8	15.7	55.0	8.2	49.1	16.7	32.8	6.8	53.4	15.3	40.1
TEXTEMB	4.1	81.1	6.8	79.7	2.7	77.6	4.1	77.0	4.1	65.6	7.6	66.5
TASKEMB	**3.2**	84.5	6.5	81.6	2.5	78.0	4.0	79.0	**3.6**	**67.1**	7.5	68.5
TEXT+TASK	**3.2**	**85.9**	**5.4**	**82.5**	**2.2**	**81.2**	**3.6**	**82.0**	**3.6**	66.5	**7.0**	**69.6**
sequence labeling												
DATASIZE	7.9	90.5	19.2	91.6	4.3	63.2	20.3	34.0	-	-	-	-
CURVEGRAD	5.6	92.6	14.6	92.8	8.0	40.7	17.9	30.8	7.0	53.2	18.6	40.8
TEXTEMB	3.7	95.0	10.4	**95.3**	3.9	65.1	8.5	61.1	5.0	67.2	10.1	63.8
TASKEMB	3.4	95.7	**9.6**	95.2	**2.7**	**80.5**	4.4	76.3	**2.5**	82.1	5.5	**76.9**
TEXT+TASK	**3.3**	**96.0**	**9.6**	95.2	**2.7**	80.3	**4.2**	**78.4**	**2.5**	**82.5**	**5.3**	**76.9**

Table 3: To evaluate our embedding methods, we measure the average rank (ρ) that they assign to the best source task (i.e., the one that results in the largest transfer gain) across target tasks, as well as the average NDCG measure of the overall ranking's quality. In parentheses, we show the number of source tasks in each setting. Combining the complementary signals in TASKEMB and TEXTEMB consistently decreases ρ (lower is better) and increases NDCG across all settings, and both methods in isolation generally perform better than the baseline methods.

results in further slight improvements, but the small magnitude of these gains suggests that TASKEMB partially encodes domain similarity. For LIMITED → LIMITED, where the DATASIZE heuristic does not apply, TASKEMB still performs strongly, although not as well as in the full source data regimes. Figure 2 shows that TASKEMB usually selects the best or near the best available source task for a given target task across data regimes.

Understanding the task embedding spaces: What kind of information is encoded by TASKEMB and TEXTEMB? Figure 3 visualizes the different task spaces in the FULL → FULL regime using the Fruchterman-Reingold force-directed placement algorithm (Fruchterman and Reingold, 1991).[19]

The task space of TEXTEMB (Figure 3, top) shows that datasets with similar sources are near one another: in QA, tasks built from web snippets are closely linked (CQ and ComQA), while in SL, tasks extracted from Penn Treebank are clustered together (CCG, POS-PTB, Parent, GParent, GGParent, Chunk, and Conj). Additionally, the SQuAD datasets are strongly linked to QNLI, which was created by converting SQuAD questions. TASKEMB captures domain information to some extent (Figure 3, bottom), but it also encodes task similarity: for example, POS-PTB is closer to POS-EWT, another part-of-speech tagging task that uses a different data source. Neither method captures some unintuitive cases in low-data regimes, such as STS-B's high transferability to CR target tasks, or that DROP benefits most from SL tasks in low-data regimes (see Tables 9, 10, 27, and 28 in Appendix B). Our methods clearly do not capture all of the factors that influence task transferability, which motivates the future development of more sophisticated task embedding methods.

4 Related Work

We build on existing work in exploring and predicting transferability across tasks.

Transferability between NLP tasks: Sharing knowledge across different tasks, as in multi-

[19] An alternative to dimensionality reduction algorithms for better preservation of the data's topology; see Appendix A.2.

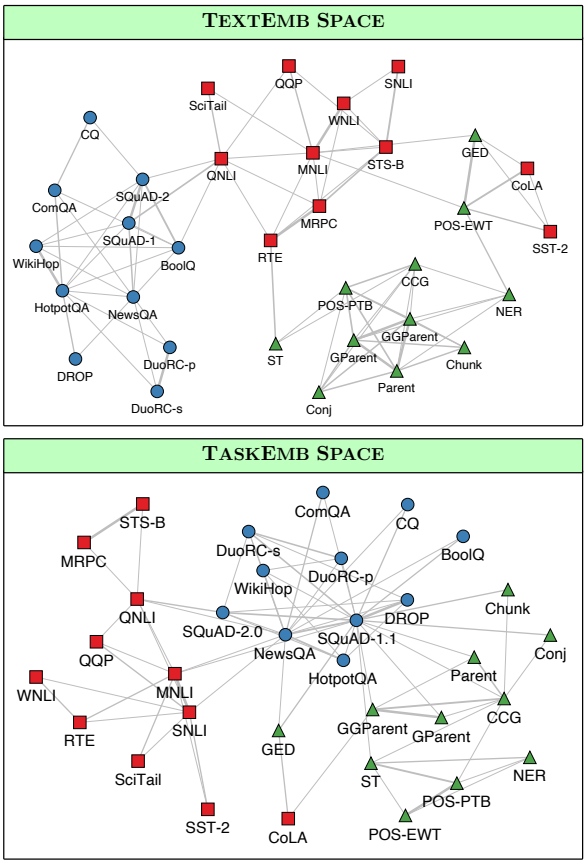

Figure 3: A 2D visualization of the task spaces of TEXTEMB and TASKEMB. TEXTEMB captures domain similarity (e.g., the Penn Treebank SL tasks are highly interconnected), while TASKEMB focuses more on task similarity (the two part-of-speech tagging tasks are interconnected despite their domain dissimilarity).

task/transfer learning, often improves over standard single-task learning (Ruder, 2017). Within multi-task learning, several works (e.g., Luong et al., 2016; Liu et al., 2019b; Raffel et al., 2020) combine multiple tasks for better regularization and transfer. More related to our work, Phang et al. (2018) explore intermediate fine-tuning and find that transferring from data-rich source tasks boosts target task performance for text classification, while Liu et al. (2019a) observe transfer gains between related sequence labeling tasks. Expanding from single to multi-source transfer, Talmor and Berant (2019a) show that pretraining on multiple datasets improves generalization on QA tasks. Nevertheless, exploiting synergies between tasks remains difficult, with many combinations of tasks negatively impacting downstream performance (Bingel and Søgaard, 2017; McCann et al., 2018; Wang et al., 2019a), and the factors that determine successful transfer still remain murky. Concurrent work indicates that intermediate tasks that require high-level inference and reasoning abilities tend to work best (Pruksachatkun et al., 2020).

Identifying beneficial task relationships: To predict transferable tasks, some methods (Martínez Alonso and Plank, 2017; Bingel and Søgaard, 2017) rely on features derived from dataset characteristics and learning curves. However, manually designing such features is time-consuming and may not generalize well across classes of problems (Kerinec et al., 2018). Recent work on *task embeddings* in computer vision offers a more principled way to encode tasks for meta-learning (Zamir et al., 2018; Achille et al., 2019; Yan et al., 2020). Taskonomy (Zamir et al., 2018) models the underlying structure among tasks to reduce the need for supervision, while Task2Vec (Achille et al., 2019) uses a frozen feature extractor pretrained on ImageNet to represent tasks in a topological space (analogous to our approach's reliance on BERT). Finally, recent work in NLP augments a generative model with an embedding space for modeling latent skills (Cao and Yogatama, 2020).

5 Conclusion

We conduct a large-scale empirical study of the transferability between 33 NLP tasks across three broad classes of problems. We show that the benefits of transfer learning are more pronounced than previously thought, especially when target training data is limited, and we develop methods that learn vector representations of tasks that can be used to reason about the relationships between them. These *task embeddings* allow us to predict source tasks that will likely improve target task performance. Our analysis suggests that data size, the similarity between the source and target tasks and domains, and task complexity are crucial for effective transfer, particularly in data-constrained regimes.

Acknowledgments

We thank Yoshua Bengio and researchers at Microsoft Research Montreal for valuable feedback on this project. We also thank the anonymous reviewers, Kalpesh Krishna, Nader Akoury, Shiv Shankar, and the rest of the UMass NLP group for their helpful comments. We are grateful to Alon Talmor and Nelson Liu for sharing the QA and SL datasets. Finally, we thank Peter Potash for additional experimentation efforts. Vu and Iyyer were supported by an Intuit AI Award for this project.

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2019. ComQA: A community-sourced dataset for complex factoid question answering with paraphrase clusters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 307–317.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 242–247.

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2019)*, pages 6430–6439.

Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based question answering with knowledge graph. In *Proceedings of the International Conference on Computational Linguistics (COLING 2016)*, pages 2503–2514.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 164–169.

Johannes Bjerva, Barbara Plank, and Johan Bos. 2016. Semantic tagging with deep residual networks. In *Proceedings of the International Conference on Computational Linguistics (COLING 2016)*, pages 3531–3541.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 632–642.

Kris Cao and Dani Yogatama. 2020. Modelling latent skills for multitask language generation. *arXiv preprint arXiv:2002.09543*.

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.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 2924–2936.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, page 758–759.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment (MLCW 2005)*, page 177–190.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 4171–4186.

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

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 2368–2378.

Jessica Ficler and Yoav Goldberg. 2016. Coordination annotation extension in the Penn tree bank. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 834–842.

Thomas M. J. Fruchterman and Edward M. Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and Experience (SPE 1991)*, 21(11):1129–1164.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn treebank. *Computational Linguistics (CL 2007)*, 33(3):355–396.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First Quora Dataset Release: Question pairs.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS 2002)*, 20(4):422–446.

Emma Kerinec, Chloé Braud, and Anders Søgaard. 2018. When does deep multi-task learning work for loosely related document classification tasks? In *Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (EMNLP Workshop BlackboxNLP 2018)*, pages 1–8.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the Conference on Artificial Intelligence (AAAI 2018)*.

Hector Levesque. 2011. The winograd schema challenge. In *Proceedings of the AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning (AAAI Spring Symposium 2011)*, volume 46, page 47.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 1073–1094.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Multi-task deep neural networks for natural language understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 4487–4496.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *Proceedings of the International Conference on Learning Representations (ICLR 2016)*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics (CL 1993)*, 19(2):313–330.

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 44–53.

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*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*, pages 2227–2237.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 5231–5247.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR 2020)*, 21(140):1–67.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 2383–2392.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1181–1191.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 1683–1693.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2897–2904.

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 (EMNLP 2013)*, pages 1631–1642.

Alon Talmor and Jonathan Berant. 2019a. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 4911–4921.

Alon Talmor and Jonathan Berant. 2019b. MultiQA repository.

Alon Talmor, Mor Geva, and Jonathan Berant. 2017. Evaluating semantic parsing against a simple web-based question answering model. In *Proceedings of the Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 161–167.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Proceedings of the Conference on Computational Natural Language Learning and the Learning Language in Logic Workshop (CoNLL-LLL 2000)*, pages 127–132.

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 Conference on Natural Language Learning (CoNLL 2003)*, pages 142–147.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the Workshop on Representation Learning for NLP (RepL4NLP 2017)*, pages 191–200.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019a. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 4465–4476.

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. *Proceedings of the International Conference on Learning Representations (ICLR 2019)*.

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

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics (TACL 2018)*, 6:287–302.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*, pages 1112–1122.

Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xi Yan, David Acuna, and Sanja Fidler. 2020. Neural data server: A large-scale search engine for transfer learning data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 2369–2380.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 180–189.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 3712–3722.

Appendices

A Additional details for experimental setup

A.1 Tasks & datasets

In this work, we experiment with 33 datasets across three broad classes of problems (text classification/regression, question answering, and sequence labeling). Below, we briefly describe the datasets, and summarize their characteristics in Table 4.

Text classification/regression (eleven tasks): We use the nine GLUE datasets (Wang et al., 2019b), including grammatical acceptability judgments (**CoLA**; Warstadt et al., 2019); sentiment analysis (**SST-2**; Socher et al., 2013); paraphrase identification (**MRPC**; Dolan and Brockett, 2005); semantic similarity with STS-Benchmark (**STS-B**; Cer et al., 2017) and Quora Question Pairs[20] (**QQP**); natural language inference (NLI) with Multi-Genre NLI (**MNLI**; Williams et al., 2018), SQuAD (Rajpurkar et al., 2016) converted into Question-answering NLI (**QNLI**; Wang et al., 2019b), Recognizing Textual Entailment 1,2,3,5 (**RTE**; Dagan et al., 2005, et seq.), and the Winograd Schema Challenge (Levesque, 2011) recast as Winograd NLI (**WNLI**). Additionally, we include the Stanford NLI dataset (**SNLI**; Bowman et al., 2015) and the science QA dataset (Khot et al., 2018) converted into NLI (**SciTail**). We report F1 scores for QQP and MRPC, Spearman correlations for STS-B, and accuracy scores for the other tasks. For MNLI, we report the average score on the "matched" and "mismatched" development sets.

Question answering (eleven tasks): We use eleven QA datasets from the MultiQA (Talmor and Berant, 2019a) repository[21], including the Stanford Question Answering datasets **SQuAD-1** and **SQuAD-2** (Rajpurkar et al., 2016, 2018); **NewsQA** (Trischler et al., 2017); **HotpotQA** (Yang et al., 2018) – the version where the context includes 10 paragraphs retrieved by an information retrieval system; Natural Yes/No Questions dataset (**BoolQ**; Clark et al., 2019); Discrete Reasoning Over Paragraphs dataset (**DROP**; Dua et al., 2019) – we only use the extractive examples in the original dataset but evaluate on the entire development set, following Talmor and Berant (2019a); **WikiHop** (Welbl et al., 2018); DuoRC Self (**DuoRC-s**) and DuoRC Paraphrase (**DuoRC-p**) datasets (Saha et al., 2018) where the questions are taken from either the same version or a different version of the document from which the questions were asked, respectively; ComplexQuestions (**CQ**; Bao et al., 2016; Talmor et al., 2017); and **ComQA** (Abujabal et al., 2019) – contexts are not provided but the questions are augmented with web snippets retrieved from Google search engine (Talmor and Berant, 2019a). We report F1 scores for all QA tasks.

Sequence labeling (eleven tasks): We experiment with eleven sequence labeling tasks used by Liu et al. (2019a), including CCG supertagging with CCGbank (**CCG**; Hockenmaier and Steedman, 2007); part-of-speech tagging with the Penn Treebank (**POS-PTB**; Marcus et al., 1993) and the Universal Dependencies English Web Treebank (**POS-EWT**; Silveira et al., 2014); syntactic constituency ancestor tagging, i.e., predicting the constituent label of the parent (**Parent**), grandparent (**GParent**), and great-grandparent (**GGParent**) of each word in the PTB phrase-structure tree; semantic tagging task (**ST**; Bjerva et al., 2016; Abzianidze et al., 2017); syntactic chunking with the CoNLL 2000 shared task dataset (**Chunk**; Tjong Kim Sang and Buchholz, 2000); named entity recognition with the CoNLL 2003 shared task dataset (**NER**; Tjong Kim Sang and De Meulder, 2003); grammatical error detection with the First Certificate in English dataset (**GED**; Yannakoudakis et al., 2011; Rei and Yannakoudakis, 2016); and conjunct identification, i.e., identifying the tokens that comprise the conjuncts in a coordination construction, with the coordination annotated PTB dataset (**Conj**; Ficler and Goldberg, 2016). We report F1 scores for all SL tasks.

[20] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

[21] https://github.com/alontalmor/MultiQA

Task	| Train |	Task type	Domain
text classification/regression (CR)			
SNLI (Bowman et al., 2015)	570K	NLI	misc.
MNLI (Williams et al., 2018)	393K	NLI	misc.
QQP (Iyer et al., 2017)	364K	paraphrase identification	social QA
QNLI (Wang et al., 2019b)	105K	QA-NLI	Wikipedia
SST-2 (Socher et al., 2013)	67K	sentiment analysis	movie reviews
SciTail (Khot et al., 2018)	27K	NLI	science QA
CoLA (Warstadt et al., 2019)	8.5K	grammatical acceptability	misc.
STS-B (Cer et al., 2017)	7K	semantic similarity	misc.
MRPC (Dolan and Brockett, 2005)	3.7K	paraphrase identification	news
RTE (Dagan et al., 2005, et seq.)	2.5K	NLI	news, Wikipedia
WNLI (Levesque, 2011)	634	coreference NLI	fiction books
question answering (QA)			
SQuAD-2 (Rajpurkar et al., 2018)	162K	QA	Wikipedia, crowd
NewsQA (Trischler et al., 2017)	120K	QA	news, crowd
HotpotQA (Yang et al., 2018)	113K	multi-hop QA	Wikipedia, crowd
SQuAD-1 (Rajpurkar et al., 2016)	108K	QA	Wikipedia, crowd
DuoRC-p (Saha et al., 2018)	100K	paraphrased QA	Wikipedia/IMDB, crowd
DuoRC-s (Saha et al., 2018)	86K	paraphrased QA	Wikipedia/IMDB, crowd
DROP (Dua et al., 2019)	77K	multi-hop quantitative reasoning	Wikipedia, crowd
WikiHop (Welbl et al., 2018)	51K	multi-hop QA	Wikipedia, KB
BoolQ (Clark et al., 2019)	16K	natural yes/no QA	Wikipedia, web queries
ComQA (Abujabal et al., 2019)	11K	factoid QA w/ paraphrases	snippets, WikiAnswers
CQ (Bao et al., 2016)	2K	knowledge-based QA	snippets, web queries/KB
sequence labeling (SL)			
ST (Bjerva et al., 2016)	43K	semantic tagging	Groningen Meaning Bank
CCG (Hockenmaier and Steedman, 2007)	40K	CCG supertagging	Penn Treebank
Parent (Liu et al., 2019a)	40K	syntactic tagging	Penn Treebank
GParent (Liu et al., 2019a)	40K	syntactic tagging	Penn Treebank
GGParent (Liu et al., 2019a)	40K	syntactic tagging	Penn Treebank
POS-PTB (Marcus et al., 1993)	38K	part-of-speech tagging	Penn Treebank
GED (Yannakoudakis et al., 2011)	29K	grammatical error detection	misc.
NER (Tjong Kim Sang and De Meulder, 2003)	14K	named entity recognition	news
POS-EWT (Silveira et al., 2014)	13K	part-of-speech tagging	Web Treebank
Conj (Ficler and Goldberg, 2016)	13K	conjunct identification	Penn Treebank
Chunk (Tjong Kim Sang and Buchholz, 2000)	9K	syntactic chunking	Penn Treebank

Table 4: Datasets used in our experiments and their characteristics, grouped by task class and sorted by training dataset size.

A.2 Fruchterman-Reingold force-directed placement algorithm

The Fruchterman-Reingold force-directed placement algorithm (Fruchterman and Reingold, 1991) simulates a space of nodes (in our setup, tasks) as a system of atomic particles/celestial bodies, exerting attractive forces on one another. In our setup, the algorithm resembles molecular/planetary simulations: the transferability between tasks specify the forces that are used to place the tasks towards each other in order to minimize the energy of the system. The force between a pair of tasks (t_1, t_2) is defined as: $f(t_1, t_2) = \frac{1}{r_{\to t_2}(t_1)} + \frac{1}{r_{\to t_1}(t_2)}$, where $r_{\to t}(s)$ is the rank of the source task s in the list of source tasks to transfer to the target task t.

B Full results for fine-tuning and transfer learning across tasks

For both fine-tuning and transfer learning, we use the same architecture across tasks, apart from the task-specific output layer. The feature extractor, i.e., BERT, is pretrained while the task-specific output layer is randomly initialized for each task. All the parameters are fine-tuned end-to-end. An alternative approach is to keep the feature extractor frozen during fine-tuning. We find that fine-tuning the whole model for a given task leads to better performance in most cases, except for WNLI and DROP, possibly because of their adversarial nature (see Tables 5, 6, and 7). In our experiments, we follow the fine-tuning recipe of (Devlin et al., 2019), i.e., only fine-tuning for a fixed number of t epochs for each class of problems. We develop our infrastructure using the HuggingFace's Transformers (Wolf et al., 2019) and its recommended hyperparameters for each class.

We show the full results for fine-tuning and transfer learning across tasks from Table 5 to Table 34. Below, we describe the setting for these tables in more detail:

In Tables 5, 6, and 7, we report the results of fine-tuning BERT (without any intermediate fine-tuning) on the 33 NLP tasks studied in this work. We perform experiments in two data regimes: FULL and LIMITED. In the FULL regime, all training data for the associated task is used while in the LIMITED setting, we artificially limit the amount of training data by randomly selecting 1K training examples without replacement, following Phang et al. (2018). For each experiment in the LIMITED regime, we perform 20 random restarts (1K examples are resampled for each restart) and report the mean and standard deviation. We show the results after each training epoch t.

For our transfer experiments, we consider every possible pair of (source, target) tasks within and across classes of problems in the three data regimes described in 2.1.1, which results in 3267 combinations of tasks and data regimes. We follow the transfer recipe of Phang et al. (2018) by first fine-tuning BERT on the source task (intermediate fine-tuning) before fine-tuning on the target task. For both stages, we only perform training for a fixed number t of epochs, following previous work (Devlin et al., 2019; Phang et al., 2018). For each task, we use the same value of t as in our fine-tuning experiments.

From Table 8 to Table 16, we show our in-class transfer results for each combination of (source, target) tasks, in which source tasks come from the same class as the target task. In each table, rows denote source tasks while columns denote target tasks. Each cell represents the target task performance of the transferred model from the associated source task to the associated target task. The orange-colored cells along the diagonal indicate the results of fine-tuning BERT on target tasks without any intermediate fine-tuning. Positive transfers are shown in blue and the best results are highlighted in bold (blue). For transfer results in the LIMITED setting, we report the mean and standard deviation across 20 random restarts.

Finally, from Table 17 to Table 34, we present our out-of-class transfer results, in which source tasks come from a different class than the target task. In each table, results are shown in a similar way as above, except that the orange-colored row Baseline shows the results of fine-tuning BERT on target tasks without any intermediate fine-tuning.

Task	FULL						LIMITED		
	frozen BERT			*unfrozen BERT*			*unfrozen BERT*		
	t = 1	t = 2	t = 3	t = 1	t = 2	t = 3	t = 1	t = 2	t = 3
CoLA	0.0	0.0	0.0	48.1	51.3	51.0	1.0 ± 2.3	4.0 ± 7.4	4.7 ± 8.2
SST-2	51.0	51.5	51.9	91.4	92.1	91.9	61.5 ± 7.9	74.3 ± 8.2	77.5 ± 6.3
MRPC	81.2	81.2	81.2	81.2	82.4	84.0	70.4 ± 26.2	81.8 ± 0.6	81.9 ± 0.7
STS-B	68.0	68.3	68.4	76.7	85.4	85.9	3.6 ± 9.5	22.8 ± 10.5	29.9 ± 10.5
QQP	0.2	13.9	16.9	86.0	87.0	87.3	9.5 ± 15.5	12.1 ± 15.9	25.7 ± 25.1
MNLI	40.9	40.2	40.8	83.1	84.3	84.2	33.7 ± 3.1	37.5 ± 3.4	38.7 ± 3.2
QNLI	65.9	66.0	66.0	90.3	91.3	91.4	58.0 ± 9.4	61.0 ± 9.9	62.4 ± 9.5
RTE	53.8	53.1	51.3	56.0	58.1	60.6	50.7 ± 3.8	54.6 ± 3.4	54.7 ± 3.2
WNLI	56.3	56.3	56.3	52.1	46.5	45.1	47.9 ± 5.6	45.6 ± 6.0	44.4 ± 6.3
SNLI	42.2	43.4	44.9	90.3	90.8	90.7	40.2 ± 4.5	45.1 ± 4.9	46.7 ± 4.5
SciTail	49.6	49.6	49.6	92.3	93.7	93.9	52.5 ± 6.3	60.1 ± 12.5	64.1 ± 13.6

Table 5: Fine-tuning results for classification/regression tasks.

Task	FULL						LIMITED		
	frozen BERT			*unfrozen BERT*			*unfrozen BERT*		
	t = 1	t = 2	t = 3	t = 1	t = 2	t = 3	t = 1	t = 2	t = 3
SQuAD-1	10.6	12.1	13.0	86.8	87.7	87.9	12.5 ± 1.0	20.8 ± 4.6	26.8 ± 6.0
SQuAD-2	49.8	49.8	49.8	68.4	70.4	71.9	50.0 ± 0.1	50.0 ± 0.1	50.1 ± 0.1
NewsQA	9.4	10.4	10.6	64.7	64.9	64.1	15.6 ± 3.4	26.5 ± 4.7	28.8 ± 4.9
HotpotQA	5.9	6.8	7.0	66.1	68.2	67.9	12.8 ± 2.4	21.6 ± 3.9	23.3 ± 4.0
BoolQ	62.1	62.2	62.2	62.2	66.4	65.7	62.2 ± 0.0	62.2 ± 0.1	62.2 ± 0.0
DROP	42.9	51.7	54.1	22.4	21.5	22.4	6.8 ± 4.4	13.5 ± 10.0	19.4 ± 11.8
WikiHop	10.1	11.4	11.6	60.0	62.3	62.8	18.3 ± 4.0	24.8 ± 4.9	25.5 ± 4.7
DuoRC-p	42.1	42.1	42.1	50.3	50.3	50.6	42.1 ± 0.0	42.2 ± 0.2	41.6 ± 1.1
DuoRC-s	4.6	5.6	5.8	66.2	64.4	63.3	22.2 ± 11.0	37.5 ± 3.5	38.9 ± 3.3
CQ	15.4	15.4	15.9	26.3	25.0	30.5	28.0 ± 3.3	29.6 ± 2.1	30.7 ± 2.5
ComQA	20.5	20.5	20.5	53.3	61.6	63.2	33.0 ± 2.4	36.0 ± 1.8	39.1 ± 1.2

Table 6: Fine-tuning results for question answering tasks.

Task	FULL						LIMITED		
	frozen BERT			*unfrozen BERT*			*unfrozen BERT*		
	t = 2	t = 4	t = 6	t = 2	t = 4	t = 6	t = 2	t = 4	t = 6
CCG	39.7	44.9	48.1	95.2	95.5	95.6	11.1 ± 6.1	45.2 ± 3.9	53.2 ± 1.6
POS-PTB	61.7	74.0	76.4	96.6	96.6	96.7	46.5 ± 2.8	80.5 ± 1.1	85.1 ± 0.9
POS-EWT	33.5	46.0	49.1	96.2	96.5	96.6	65.4 ± 3.0	86.8 ± 0.6	89.3 ± 0.4
Parent	37.9	58.1	61.5	95.1	95.3	95.4	61.1 ± 4.0	77.0 ± 1.0	81.9 ± 0.9
GParent	35.0	41.9	43.4	91.1	91.7	91.9	41.1 ± 1.4	58.0 ± 1.7	62.8 ± 1.3
GGParent	25.9	30.9	31.7	88.3	89.3	89.5	25.6 ± 3.1	37.9 ± 1.7	43.3 ± 1.7
ST	51.2	66.1	69.2	95.5	95.7	95.8	38.6 ± 1.1	71.3 ± 1.6	76.7 ± 0.9
Chunk	11.9	16.6	18.4	96.4	96.8	97.1	68.1 ± 2.4	85.0 ± 0.7	87.7 ± 0.5
NER	4.7	7.7	9.2	93.8	94.3	94.7	58.4 ± 7.3	73.5 ± 1.6	77.4 ± 1.5
GED	16.8	18.4	18.8	44.2	46.9	46.6	17.3 ± 1.2	27.4 ± 1.4	29.1 ± 1.3
Conj	14.7	19.8	21.1	88.6	89.9	89.4	40.6 ± 6.0	69.2 ± 2.4	73.3 ± 1.6

Table 7: Fine-tuning results for sequence labeling tasks.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	SNLI	SciTail
CoLA	51.0	92.2	86.6	86.4	**87.5**	84.2	91.4	60.3	**54.9**	90.5	93.8
SST-2	54.2	91.9	84.2	86.9	87.0	84.1	91.3	56.0	53.5	90.9	93.5
MRPC	51.0	92.3	84.0	87.1	87.1	84.4	91.3	61.7	47.9	90.9	93.5
STS-B	48.8	91.9	87.3	85.9	86.4	84.0	90.4	65.0	35.2	90.9	92.1
QQP	49.4	92.0	**87.7**	88.5	87.3	84.2	90.7	61.7	36.6	90.9	92.9
MNLI	50.0	**93.5**	87.6	87.0	87.1	84.2	**91.5**	**77.6**	40.8	**91.2**	**95.6**
QNLI	49.9	92.5	86.6	**88.6**	86.6	84.4	91.4	70.4	38.0	91.1	94.5
RTE	52.1	92.1	83.9	87.0	86.8	84.4	91.3	60.6	50.7	91.0	93.5
WNLI	**54.5**	91.7	84.2	84.8	87.0	84.2	91.4	60.6	45.1	90.9	93.6
SNLI	54.2	93.1	86.8	87.5	86.9	**84.6**	90.4	**77.6**	39.4	90.7	95.2
SciTail	50.8	91.9	82.2	88.1	86.6	84.3	91.0	69.3	46.5	91.0	93.9

Table 8: In-class transfer results for classification/regression tasks in the FULL → FULL regime.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	SNLI	SciTail
CoLA	4.7 ± 8.2	74.4 ± 5.9	82.0 ± 0.5	32.7 ± 10.6	38.2 ± 28.8	39.3 ± 2.6	66.7 ± 6.1	56.4 ± 2.7	40.1 ± 8.3	47.4 ± 2.8	68.6 ± 15.7
SST-2	1.3 ± 2.8	77.5 ± 6.3	81.9 ± 0.7	29.1 ± 12.7	33.1 ± 23.2	43.6 ± 3.4	66.4 ± 7.0	55.0 ± 2.8	39.7 ± 5.6	49.3 ± 2.8	64.5 ± 14.9
MRPC	1.2 ± 4.3	68.4 ± 11.3	81.9 ± 0.7	**71.2 ± 6.7**	54.2 ± 22.0	46.3 ± 2.0	73.5 ± 1.6	59.2 ± 1.7	38.7 ± 6.4	51.9 ± 2.5	84.7 ± 1.0
STS-B	2.3 ± 5.2	75.8 ± 7.4	**84.6 ± 0.5**	29.9 ± 10.5	**67.5 ± 1.4**	49.2 ± 1.2	76.7 ± 0.5	62.2 ± 1.9	**44.6 ± 8.5**	55.4 ± 1.7	86.4 ± 1.1
QQP	**7.7 ± 9.0**	82.0 ± 2.3	83.5 ± 1.2	67.4 ± 8.3	25.7 ± 25.1	52.4 ± 3.0	**77.1 ± 1.3**	62.8 ± 2.2	36.4 ± 6.5	56.4 ± 2.8	**88.2 ± 1.4**
MNLI	1.0 ± 2.2	85.0 ± 0.8	84.0 ± 1.0	67.3 ± 6.3	66.0 ± 3.6	38.7 ± 3.2	76.0 ± 1.6	**72.8 ± 2.0**	39.4 ± 5.6	**79.5 ± 3.5**	85.5 ± 2.2
QNLI	1.2 ± 2.7	80.0 ± 8.9	83.8 ± 1.8	68.3 ± 10.3	49.4 ± 26.3	48.5 ± 3.3	62.4 ± 9.5	60.3 ± 2.7	39.2 ± 7.4	56.3 ± 3.2	84.0 ± 3.9
RTE	5.2 ± 7.6	77.1 ± 8.0	82.4 ± 1.0	40.8 ± 14.0	40.6 ± 30.4	41.4 ± 5.3	64.8 ± 9.5	54.7 ± 3.2	43.6 ± 7.8	50.5 ± 2.7	71.3 ± 16.7
WNLI	4.2 ± 7.8	74.2 ± 10.1	81.9 ± 0.6	30.7 ± 13.7	23.2 ± 24.6	39.5 ± 2.6	64.0 ± 8.3	56.6 ± 2.2	44.4 ± 6.3	48.3 ± 4.2	67.9 ± 13.6
SNLI	1.5 ± 3.4	**85.9 ± 1.3**	82.1 ± 0.9	68.9 ± 2.2	64.6 ± 4.1	**70.3 ± 4.9**	72.7 ± 3.8	70.8 ± 4.9	37.9 ± 4.5	46.7 ± 4.5	82.9 ± 2.7
SciTail	6.5 ± 9.5	81.0 ± 5.8	83.0 ± 1.1	67.7 ± 8.2	58.8 ± 22.0	50.6 ± 4.3	70.7 ± 5.9	63.3 ± 3.8	42.3 ± 6.0	56.1 ± 3.6	64.1 ± 13.6

Table 9: In-class transfer results for classification/regression tasks in the FULL → LIMITED regime.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	SNLI	SciTail
CoLA	4.7 ± 8.2	74.8 ± 6.1	81.9 ± 0.7	24.1 ± 10.3	28.0 ± 27.3	38.4 ± 3.2	62.3 ± 9.5	54.8 ± 3.0	43.7 ± 6.4	47.1 ± 3.9	65.2 ± 13.6
SST-2	4.2 ± 7.3	77.5 ± 6.3	81.9 ± 0.7	27.9 ± 10.8	33.4 ± 26.5	39.1 ± 3.4	63.8 ± 8.9	55.9 ± 3.5	43.9 ± 6.4	47.8 ± 3.6	65.3 ± 13.9
MRPC	2.5 ± 5.2	75.2 ± 8.1	81.9 ± 0.7	45.2 ± 11.8	40.0 ± 28.3	41.2 ± 3.8	68.8 ± 5.8	57.2 ± 3.9	41.7 ± 7.9	51.3 ± 2.7	73.1 ± 14.8
STS-B	**6.7 ± 8.1**	76.7 ± 6.8	82.0 ± 0.7	29.9 ± 10.5	**43.8 ± 23.2**	**43.9 ± 2.2**	**73.2 ± 1.1**	58.6 ± 2.6	39.2 ± 6.1	**51.8 ± 2.7**	**79.3 ± 6.6**
QQP	3.2 ± 5.4	76.6 ± 8.3	82.1 ± 0.8	35.7 ± 12.1	25.7 ± 25.1	40.4 ± 4.1	65.5 ± 8.1	55.5 ± 3.9	39.7 ± 8.6	49.8 ± 2.7	69.3 ± 16.2
MNLI	3.7 ± 5.5	75.3 ± 9.6	82.1 ± 0.7	35.7 ± 12.6	33.6 ± 30.0	38.7 ± 3.2	64.9 ± 9.9	55.5 ± 3.5	**46.3 ± 8.1**	49.3 ± 2.9	69.8 ± 14.8
QNLI	4.9 ± 8.7	**78.3 ± 6.9**	81.8 ± 0.8	33.2 ± 14.8	35.4 ± 28.0	40.4 ± 4.2	62.4 ± 9.5	55.7 ± 4.2	43.1 ± 6.4	48.3 ± 3.6	71.6 ± 14.3
RTE	5.0 ± 8.2	77.4 ± 6.1	82.1 ± 0.8	32.9 ± 14.1	35.5 ± 28.8	40.4 ± 4.3	65.1 ± 8.3	54.7 ± 3.2	43.0 ± 7.4	48.2 ± 3.0	67.6 ± 14.8
WNLI	3.8 ± 5.8	74.9 ± 8.5	81.9 ± 0.6	**49.9 ± 11.7**	40.2 ± 24.2	42.6 ± 2.3	70.2 ± 2.6	57.9 ± 1.5	44.4 ± 6.3	51.6 ± 3.0	78.5 ± 9.1
SNLI	4.6 ± 7.8	74.9 ± 9.5	81.8 ± 0.5	44.6 ± 18.2	39.7 ± 25.7	42.9 ± 2.8	68.6 ± 3.3	**59.6 ± 2.8**	39.4 ± 7.1	46.7 ± 4.5	77.9 ± 9.2
SciTail	5.8 ± 9.8	77.5 ± 5.3	**82.2 ± 0.9**	26.0 ± 11.8	33.8 ± 32.4	40.2 ± 4.9	64.8 ± 8.6	54.5 ± 2.8	44.9 ± 6.9	47.2 ± 2.6	64.1 ± 13.6

Table 10: In-class transfer results for classification/regression tasks in the LIMITED → LIMITED regime.

Task	SQuAD-1	SQuAD-2	NewsQA	HotpotQA	BoolQ	DROP	WikiHop	DuoRC-p	DuoRC-s	CQ	ComQA
SQuAD-1	87.9	73.4	65.5	70.1	71.0	26.9	63.7	51.1	62.9	45.2	64.8
SQuAD-2	87.8	71.9	**66.3**	**70.6**	**74.3**	**27.7**	63.6	51.2	62.9	**45.4**	64.4
NewsQA	**89.0**	**73.8**	64.1	69.7	73.0	27.4	63.6	50.7	61.8	41.2	**65.3**
HotpotQA	88.6	72.8	64.8	67.9	73.1	26.1	**64.2**	50.2	62.0	45.3	63.3
BoolQ	87.8	70.3	64.5	68.0	65.7	22.2	63.0	50.8	62.1	33.0	63.6
DROP	88.1	71.8	65.6	69.6	69.0	22.4	63.7	50.8	63.0	41.5	65.2
WikiHop	87.4	69.2	63.7	68.4	68.3	21.8	62.8	50.1	61.2	43.5	**65.3**
DuoRC-p	88.1	71.7	64.6	68.4	71.5	23.9	63.3	50.6	63.1	44.1	65.1
DuoRC-s	88.5	72.6	64.5	69.0	71.1	24.3	63.9	**51.8**	63.3	43.6	62.1
CQ	87.6	69.8	64.8	67.9	68.3	22.1	63.1	50.8	63.3	30.5	64.6
ComQA	86.7	69.7	63.9	66.4	67.5	21.6	62.4	50.4	63.2	42.2	63.2

Table 11: In-class transfer results for question answering tasks in the FULL → FULL regime.

Task	SQuAD-1	SQuAD-2	NewsQA	HotpotQA	BoolQ	DROP	WikiHop	DuoRC-p	DuoRC-s	CQ	ComQA
SQuAD-1	26.8 ± 6.0	**57.4 ± 1.1**	57.1 ± 0.4	50.9 ± 0.5	62.2 ± 0.0	16.8 ± 0.8	38.1 ± 1.1	50.2 ± 0.9	59.8 ± 1.0	45.0 ± 2.1	46.6 ± 1.0
SQuAD-2	**86.5 ± 0.3**	50.1 ± 0.1	**57.2 ± 0.4**	**51.2 ± 0.5**	62.2 ± 0.0	**26.0 ± 4.0**	37.7 ± 1.0	**51.0 ± 1.1**	**60.6 ± 0.8**	44.5 ± 1.7	46.3 ± 0.7
NewsQA	79.4 ± 0.7	55.8 ± 0.9	28.8 ± 4.9	48.0 ± 0.5	62.2 ± 0.0	16.0 ± 1.8	38.1 ± 0.6	49.9 ± 0.6	57.9 ± 0.7	43.0 ± 2.3	47.4 ± 1.1
HotpotQA	78.4 ± 0.4	54.1 ± 0.9	52.8 ± 0.4	23.3 ± 4.0	62.2 ± 0.1	20.0 ± 2.8	**39.4 ± 0.8**	48.7 ± 0.8	55.5 ± 1.2	**46.9 ± 2.0**	**47.9 ± 1.0**
BoolQ	26.6 ± 6.8	50.1 ± 0.0	26.3 ± 3.9	31.0 ± 4.1	62.2 ± 0.0	15.3 ± 12.2	18.9 ± 3.3	41.2 ± 1.2	34.5 ± 3.3	31.9 ± 2.0	38.9 ± 1.4
DROP	73.0 ± 0.4	48.6 ± 1.7	50.3 ± 0.4	46.1 ± 0.4	62.2 ± 0.0	19.4 ± 11.8	35.7 ± 0.9	47.8 ± 0.9	54.4 ± 1.0	42.5 ± 2.0	45.1 ± 1.4
WikiHop	50.9 ± 2.3	49.4 ± 0.7	39.4 ± 0.9	38.6 ± 0.7	62.2 ± 0.1	15.4 ± 6.3	25.5 ± 4.7	43.5 ± 0.7	44.2 ± 0.9	42.4 ± 1.8	45.8 ± 1.2
DuoRC-p	75.1 ± 0.4	51.4 ± 0.8	52.7 ± 0.4	45.2 ± 0.7	62.2 ± 0.0	16.2 ± 1.8	37.0 ± 0.7	41.6 ± 1.1	58.2 ± 0.9	42.2 ± 1.9	45.0 ± 0.9
DuoRC-s	78.3 ± 0.4	52.1 ± 1.0	53.9 ± 0.4	46.6 ± 0.5	62.2 ± 0.1	17.1 ± 1.3	36.7 ± 0.7	50.9 ± 0.5	38.9 ± 3.3	43.8 ± 2.2	45.6 ± 1.3
CQ	21.8 ± 2.4	49.3 ± 0.6	30.8 ± 0.8	25.3 ± 1.0	62.2 ± 0.0	5.2 ± 0.5	24.1 ± 2.8	37.2 ± 1.1	34.6 ± 1.8	30.7 ± 2.5	41.4 ± 0.8
ComQA	39.6 ± 3.8	47.3 ± 2.0	37.2 ± 0.7	31.7 ± 0.7	62.2 ± 0.0	8.0 ± 6.8	34.5 ± 0.8	38.4 ± 1.1	38.0 ± 1.0	42.3 ± 1.6	39.1 ± 1.2

Table 12: In-class transfer results for question answering tasks in the FULL → LIMITED regime.

Task	SQuAD-1	SQuAD-2	NewsQA	HotpotQA	BoolQ	DROP	WikiHop	DuoRC-p	DuoRC-s	CQ	ComQA
SQuAD-1	26.8 ± 6.0	42.8 ± 2.9	35.3 ± 2.5	31.5 ± 2.1	62.2 ± 0.0	9.5 ± 0.5	27.9 ± 3.2	44.6 ± 0.7	42.9 ± 1.8	33.2 ± 1.8	39.7 ± 1.0
SQuAD-2	48.7 ± 1.6	50.1 ± 0.1	39.9 ± 1.0	34.3 ± 3.2	62.2 ± 0.0	17.8 ± 5.6	29.5 ± 2.2	45.0 ± 0.7	46.6 ± 1.7	32.2 ± 2.4	39.5 ± 1.0
NewsQA	**63.8 ± 1.1**	45.8 ± 1.7	28.8 ± 4.9	**42.3 ± 0.5**	62.2 ± 0.0	17.2 ± 3.8	33.5 ± 0.8	**47.0 ± 0.7**	**51.0 ± 1.0**	38.0 ± 2.3	**42.7 ± 1.1**
HotpotQA	59.4 ± 1.0	46.5 ± 1.4	**43.6 ± 0.8**	23.3 ± 4.0	62.2 ± 0.0	17.5 ± 7.8	**35.0 ± 0.8**	46.6 ± 0.6	50.2 ± 0.9	**39.7 ± 1.7**	**42.7 ± 1.4**
BoolQ	32.4 ± 7.8	50.0 ± 0.1	25.3 ± 2.8	26.0 ± 4.3	62.2 ± 0.0	**49.1 ± 14.4**	23.1 ± 4.4	42.1 ± 0.7	35.4 ± 5.6	31.4 ± 2.5	38.7 ± 1.1
DROP	28.5 ± 5.1	50.1 ± 0.0	27.6 ± 3.1	22.7 ± 1.9	62.2 ± 0.0	19.4 ± 11.8	23.6 ± 4.1	40.9 ± 0.9	38.2 ± 2.8	32.3 ± 2.0	38.6 ± 1.4
WikiHop	45.1 ± 2.5	46.2 ± 1.8	39.7 ± 1.0	37.8 ± 1.0	62.2 ± 0.0	12.3 ± 5.3	25.5 ± 4.7	42.4 ± 0.9	44.0 ± 1.6	37.3 ± 1.8	42.4 ± 1.1
DuoRC-p	57.1 ± 1.1	44.1 ± 1.9	42.5 ± 0.7	39.6 ± 0.8	62.0 ± 0.6	20.2 ± 4.9	33.1 ± 0.9	41.6 ± 1.1	48.0 ± 1.0	36.4 ± 2.8	42.3 ± 1.4
DuoRC-s	59.5 ± 1.6	44.7 ± 1.7	43.5 ± 0.5	41.6 ± 0.7	62.2 ± 0.0	19.9 ± 4.3	33.2 ± 1.3	46.7 ± 0.9	38.9 ± 3.3	35.2 ± 2.5	41.1 ± 0.9
CQ	23.3 ± 2.8	49.2 ± 1.9	27.5 ± 1.8	22.6 ± 1.1	62.2 ± 0.0	8.2 ± 5.1	21.7 ± 2.9	36.1 ± 1.7	32.1 ± 3.6	30.7 ± 2.5	40.8 ± 1.4
ComQA	30.0 ± 2.6	46.5 ± 3.5	32.8 ± 0.9	27.2 ± 1.2	62.2 ± 0.0	6.0 ± 0.4	31.0 ± 1.3	38.0 ± 1.2	35.5 ± 2.3	35.7 ± 1.5	39.1 ± 1.2

Table 13: In-class transfer results for question answering tasks in the LIMITED → LIMITED regime.

Task	CCG	POS-PTB	POS-EWT	Parent	GParent	GGParent	ST	Chunk	NER	GED	Conj
CCG	95.6	96.7	96.4	95.3	91.8	89.6	95.8	97.7	94.0	45.8	90.3
POS-PTB	**95.7**	96.7	**96.7**	95.3	91.7	89.1	95.7	97.0	94.6	46.5	90.2
POS-EWT	95.6	96.7	96.6	**95.5**	91.9	89.3	95.8	97.0	94.6	46.1	89.9
Parent	95.6	96.7	96.6	95.4	91.9	89.8	95.8	**98.0**	94.5	46.6	90.3
GParent	95.6	96.7	96.6	95.1	91.9	**90.0**	95.8	97.6	94.6	46.5	**91.0**
GGParent	95.5	96.6	96.5	95.4	91.9	89.5	95.8	97.5	94.5	46.5	90.8
ST	95.5	96.6	96.5	95.1	91.6	89.3	95.8	96.9	**94.9**	46.2	88.7
Chunk	95.6	96.7	96.5	95.2	91.8	89.5	95.7	97.1	94.6	46.4	89.7
NER	95.4	96.7	96.6	95.2	91.7	89.1	95.8	97.0	94.7	**47.3**	90.3
GED	95.5	96.7	96.6	95.2	91.7	89.3	95.8	97.0	94.7	46.6	90.2
Conj	95.4	96.7	96.6	95.4	91.9	89.7	95.8	97.0	94.5	46.2	89.4

Table 14: In-class transfer results for sequence labeling tasks in the FULL → FULL regime.

Task	CCG	POS-PTB	POS-EWT	Parent	GParent	GGParent	ST	Chunk	NER	GED	Conj
CCG	53.2 ± 1.6	89.8 ± 0.8	91.9 ± 0.2	87.1 ± 1.3	74.5 ± 0.5	54.0 ± 1.3	84.0 ± 0.8	92.9 ± 0.1	67.9 ± 3.3	24.3 ± 1.4	73.2 ± 1.3
POS-PTB	72.0 ± 0.5	85.1 ± 0.9	**93.9 ± 0.2**	87.7 ± 0.7	68.4 ± 1.0	49.5 ± 1.1	**86.3 ± 0.2**	91.2 ± 0.5	83.3 ± 1.2	28.8 ± 0.9	69.5 ± 2.0
POS-EWT	68.2 ± 0.7	88.5 ± 0.6	89.3 ± 0.4	86.4 ± 1.0	66.2 ± 1.0	47.5 ± 1.2	83.4 ± 0.8	91.8 ± 0.3	81.3 ± 1.3	29.1 ± 0.9	70.9 ± 2.7
Parent	66.2 ± 1.0	88.5 ± 0.9	92.6 ± 0.3	81.9 ± 0.9	75.8 ± 0.7	55.4 ± 1.7	82.4 ± 0.7	**94.3 ± 0.3**	78.3 ± 4.0	28.7 ± 0.9	76.5 ± 3.8
GParent	64.5 ± 3.0	87.2 ± 1.0	90.8 ± 0.3	**90.5 ± 0.2**	62.8 ± 1.3	**77.4 ± 0.6**	81.6 ± 0.6	92.0 ± 0.3	76.4 ± 1.9	24.2 ± 1.7	83.4 ± 0.6
GGParent	59.7 ± 3.0	82.8 ± 1.7	89.8 ± 0.3	89.4 ± 0.4	**88.2 ± 0.3**	43.3 ± 1.7	78.7 ± 1.1	91.0 ± 0.4	76.8 ± 2.0	19.9 ± 1.0	**85.1 ± 0.6**
ST	**72.4 ± 0.7**	**92.6 ± 0.4**	93.2 ± 0.2	87.4 ± 0.3	71.2 ± 0.7	50.3 ± 1.2	76.7 ± 0.9	91.1 ± 0.3	**87.0 ± 0.6**	**29.7 ± 0.6**	66.6 ± 2.7
Chunk	67.5 ± 1.0	88.9 ± 0.6	92.0 ± 0.3	90.0 ± 0.2	71.9 ± 0.8	53.3 ± 1.2	83.3 ± 0.8	87.7 ± 0.5	76.1 ± 2.2	28.7 ± 1.9	77.5 ± 0.9
NER	47.2 ± 3.4	83.1 ± 1.3	90.1 ± 0.6	79.0 ± 1.6	62.7 ± 1.7	42.0 ± 3.4	78.0 ± 1.5	85.9 ± 1.0	77.4 ± 1.5	29.4 ± 0.8	72.6 ± 1.9
GED	56.1 ± 1.6	87.0 ± 0.7	89.9 ± 0.4	82.3 ± 0.8	66.7 ± 1.1	47.1 ± 1.6	80.2 ± 0.7	88.2 ± 0.4	79.9 ± 1.5	29.1 ± 1.3	70.6 ± 2.4
Conj	48.9 ± 4.0	84.3 ± 1.0	89.1 ± 0.6	79.5 ± 0.8	67.9 ± 1.2	49.1 ± 1.9	76.6 ± 1.4	87.4 ± 0.7	77.4 ± 3.6	28.1 ± 1.4	73.3 ± 1.6

Table 15: In-class transfer results for sequence labeling tasks in the FULL → LIMITED regime.

Task	CCG	POS-PTB	POS-EWT	Parent	GParent	GGParent	ST	Chunk	NER	GED	Conj
CCG	53.2 ± 1.6	88.6 ± 0.4	90.8 ± 0.4	85.7 ± 0.3	64.9 ± 1.4	44.2 ± 3.8	80.0 ± 1.4	89.3 ± 0.2	63.3 ± 4.7	27.8 ± 2.1	66.3 ± 4.5
POS-PTB	**68.2 ± 0.8**	85.1 ± 0.9	**92.0 ± 0.1**	85.7 ± 0.5	65.3 ± 1.7	43.6 ± 2.4	**83.2 ± 0.6**	89.5 ± 0.3	76.9 ± 2.2	28.5 ± 1.8	61.5 ± 9.9
POS-EWT	66.7 ± 0.7	88.7 ± 1.3	89.3 ± 0.4	86.0 ± 0.9	65.0 ± 1.4	43.0 ± 3.3	82.6 ± 0.9	90.2 ± 0.5	77.7 ± 2.5	27.1 ± 0.8	57.7 ± 7.6
Parent	66.0 ± 2.0	88.5 ± 0.9	91.5 ± 0.2	81.9 ± 0.9	68.5 ± 0.8	47.6 ± 2.5	80.5 ± 1.2	**90.3 ± 0.3**	74.0 ± 2.6	29.1 ± 2.3	66.7 ± 4.0
GParent	63.7 ± 1.4	87.9 ± 0.6	90.7 ± 0.4	86.4 ± 0.5	62.8 ± 1.3	**58.1 ± 1.5**	80.0 ± 0.9	89.9 ± 0.3	70.1 ± 3.6	29.3 ± 1.6	75.9 ± 2.7
GGParent	59.2 ± 3.1	87.1 ± 1.4	90.2 ± 0.3	84.6 ± 0.5	**71.3 ± 0.5**	43.3 ± 1.7	78.4 ± 1.1	89.2 ± 0.3	73.3 ± 2.6	**29.9 ± 1.4**	**77.6 ± 1.4**
ST	67.5 ± 1.0	**89.6 ± 0.9**	91.7 ± 0.2	86.1 ± 0.5	66.2 ± 2.0	46.2 ± 1.9	76.7 ± 0.9	90.0 ± 0.4	77.5 ± 1.5	28.5 ± 1.5	64.9 ± 5.7
Chunk	66.7 ± 1.2	88.7 ± 0.9	91.5 ± 0.2	**86.9 ± 0.4**	69.0 ± 1.0	50.8 ± 1.2	81.4 ± 0.5	87.7 ± 0.5	71.1 ± 2.9	28.6 ± 1.7	72.6 ± 3.6
NER	50.7 ± 2.9	83.8 ± 1.5	89.7 ± 0.5	79.6 ± 1.9	63.1 ± 1.7	41.7 ± 2.4	79.0 ± 2.0	86.2 ± 1.3	77.4 ± 1.5	29.6 ± 2.0	69.9 ± 3.5
GED	54.3 ± 3.1	85.4 ± 1.0	89.5 ± 0.5	81.6 ± 1.2	64.5 ± 1.8	45.2 ± 2.2	78.0 ± 1.0	87.9 ± 0.4	**78.7 ± 2.4**	29.1 ± 1.3	75.2 ± 1.6
Conj	55.0 ± 1.8	85.2 ± 1.1	89.3 ± 0.3	81.0 ± 1.7	65.6 ± 2.1	44.7 ± 2.1	77.2 ± 1.9	87.3 ± 0.7	77.3 ± 3.4	29.5 ± 1.4	73.3 ± 1.6

Table 16: In-class transfer results for sequence labeling tasks in the LIMITED → LIMITED regime.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	SNLI	SciTail
Baseline	51.0	91.9	84.0	85.9	87.3	84.2	91.4	60.6	45.1	90.7	93.9
SQuAD-1	52.4	92.1	87.0	**88.5**	87.0	83.8	91.3	64.6	39.4	90.7	94.4
SQuAD-2	47.1	91.9	87.4	87.2	87.1	**84.6**	**91.7**	67.9	45.1	**90.9**	94.7
NewsQA	45.2	91.4	86.9	87.6	86.9	84.0	91.3	63.2	36.6	90.4	93.9
HotpotQA	43.3	92.1	**88.6**	86.9	86.8	83.8	91.1	66.1	39.4	90.8	94.2
BoolQ	51.0	92.1	86.3	85.8	**87.4**	83.9	90.5	59.6	32.4	90.7	93.7
DROP	**53.4**	92.3	87.0	87.9	87.1	84.3	91.1	**70.4**	42.3	90.7	**94.9**
WikiHop	49.2	91.9	84.6	86.8	86.8	83.7	90.7	66.1	38.0	90.7	93.5
DuoRC-p	42.4	92.2	86.3	87.3	86.7	83.4	90.9	62.8	36.6	90.5	92.5
DuoRC-s	48.8	91.5	86.4	87.9	87.1	83.6	90.8	67.1	42.3	90.6	93.9
CQ	52.1	91.9	85.4	86.9	86.9	84.0	90.6	68.2	45.1	90.8	93.6
ComQA	49.5	**92.4**	83.9	86.4	86.9	83.5	89.4	63.5	33.8	90.6	92.6

Table 17: Out-of-class transfer results from question answering tasks to classification/regression tasks in the FULL → FULL regime.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	SNLI	SciTail
Baseline	4.7 ± 8.2	77.5 ± 6.3	81.9 ± 0.7	29.9 ± 10.5	25.7 ± 25.1	38.7 ± 3.2	62.4 ± 9.5	54.7 ± 3.2	44.4 ± 6.3	46.7 ± 4.5	64.1 ± 13.6
SQuAD-1	3.3 ± 5.5	83.3 ± 1.4	83.4 ± 1.1	**72.1 ± 7.0**	47.8 ± 25.8	49.4 ± 2.3	**86.9 ± 0.6**	63.1 ± 1.7	41.7 ± 7.4	53.8 ± 2.1	**86.6 ± 1.2**
SQuAD-2	2.9 ± 6.5	**83.8 ± 1.6**	82.1 ± 0.8	65.0 ± 12.1	42.1 ± 32.1	**52.0 ± 6.4**	83.6 ± 2.2	61.6 ± 2.2	44.4 ± 7.0	**55.3 ± 4.1**	67.9 ± 16.9
NewsQA	1.8 ± 3.8	81.4 ± 3.1	83.6 ± 1.3	67.1 ± 6.4	52.2 ± 25.1	48.5 ± 3.2	79.2 ± 6.0	63.5 ± 3.0	43.5 ± 6.0	54.3 ± 2.5	83.0 ± 8.0
HotpotQA	1.9 ± 4.3	72.4 ± 8.3	**83.8 ± 1.4**	49.7 ± 10.9	34.1 ± 32.7	41.9 ± 2.8	73.8 ± 11.4	58.7 ± 3.0	45.3 ± 5.7	51.6 ± 4.8	74.4 ± 13.2
BoolQ	**7.7 ± 9.3**	76.5 ± 4.6	81.7 ± 0.6	49.4 ± 18.0	46.0 ± 25.3	42.0 ± 2.1	72.3 ± 2.4	57.6 ± 2.5	39.9 ± 6.7	47.8 ± 4.3	72.4 ± 11.7
DROP	6.0 ± 8.8	81.8 ± 1.9	82.4 ± 0.7	64.5 ± 10.4	49.7 ± 26.2	45.6 ± 1.8	78.9 ± 1.2	**63.6 ± 1.9**	43.5 ± 7.8	52.7 ± 2.5	82.0 ± 8.1
WikiHop	0.3 ± 2.3	69.9 ± 9.1	82.3 ± 0.7	63.1 ± 5.7	57.5 ± 20.4	44.5 ± 1.5	71.9 ± 1.8	62.1 ± 2.2	41.5 ± 6.3	53.2 ± 1.8	83.0 ± 1.4
DuoRC-p	0.9 ± 3.0	74.1 ± 5.2	83.2 ± 1.3	71.0 ± 6.5	41.3 ± 30.6	44.1 ± 2.3	79.3 ± 4.4	60.3 ± 3.3	**45.4 ± 6.0**	52.0 ± 2.6	69.7 ± 14.8
DuoRC-s	3.2 ± 5.6	78.5 ± 4.6	83.5 ± 1.5	66.7 ± 5.8	44.5 ± 29.6	45.7 ± 2.4	82.5 ± 1.4	61.1 ± 2.2	42.9 ± 6.6	52.9 ± 2.7	72.6 ± 14.3
CQ	5.6 ± 7.3	74.7 ± 7.6	81.8 ± 0.7	61.6 ± 9.3	42.6 ± 30.6	44.8 ± 2.1	71.8 ± 1.6	61.3 ± 2.7	39.6 ± 5.5	53.5 ± 2.8	78.2 ± 11.1
ComQA	1.2 ± 3.0	72.1 ± 6.8	81.7 ± 0.4	51.5 ± 19.1	**58.3 ± 13.9**	41.4 ± 2.3	68.1 ± 2.2	59.0 ± 1.9	39.6 ± 8.1	51.9 ± 1.7	80.9 ± 8.5

Table 18: Out-of-class transfer results from question answering tasks to classification/regression tasks in the FULL → LIMITED regime.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	SNLI	SciTail
Baseline	4.7 ± 8.2	77.5 ± 6.3	81.9 ± 0.7	29.9 ± 10.5	25.7 ± 25.1	38.7 ± 3.2	62.4 ± 9.5	54.7 ± 3.2	44.4 ± 6.3	46.7 ± 4.5	64.1 ± 13.6
SQuAD-1	**8.6** ± **10.0**	74.7 ± 7.2	81.8 ± 0.7	54.3 ± 8.9	38.5 ± 28.6	39.9 ± 3.2	73.9 ± 2.1	56.2 ± 3.5	45.8 ± 7.6	48.4 ± 3.6	70.4 ± 13.7
SQuAD-2	7.5 ± 10.4	77.3 ± 5.8	81.8 ± 0.8	51.4 ± 9.3	38.4 ± 29.0	41.8 ± 2.5	74.8 ± 1.8	56.9 ± 3.2	45.0 ± 5.7	49.3 ± 3.7	71.4 ± 15.3
NewsQA	3.3 ± 5.6	76.6 ± 6.3	82.0 ± 0.7	59.4 ± 8.2	45.7 ± 24.6	42.3 ± 2.6	**77.6** ± **1.1**	59.0 ± 2.7	43.7 ± 7.5	49.9 ± 2.6	77.3 ± 11.2
HotpotQA	5.8 ± 8.4	77.8 ± 3.6	81.8 ± 0.5	63.2 ± 8.5	42.6 ± 28.6	42.2 ± 2.5	72.5 ± 5.6	59.4 ± 1.3	44.3 ± 7.7	50.9 ± 3.5	75.8 ± 12.5
BoolQ	5.3 ± 7.3	77.2 ± 6.1	81.8 ± 0.7	40.6 ± 18.9	42.2 ± 28.2	39.9 ± 3.2	68.7 ± 5.0	56.5 ± 2.5	44.7 ± 7.9	47.3 ± 3.8	69.1 ± 13.3
DROP	4.5 ± 7.2	78.1 ± 5.7	**82.1** ± **0.9**	41.5 ± 11.7	35.0 ± 27.2	39.4 ± 3.1	67.4 ± 5.8	55.3 ± 2.9	41.3 ± 5.9	48.0 ± 4.0	67.5 ± 15.0
WikiHop	4.4 ± 7.5	**78.5** ± **3.5**	81.9 ± 0.7	46.9 ± 13.6	37.5 ± 30.2	40.9 ± 2.6	70.2 ± 1.7	58.0 ± 2.6	43.5 ± 8.4	50.0 ± 3.3	75.5 ± 13.5
DuoRC-p	3.6 ± 6.8	77.0 ± 4.0	81.7 ± 0.6	39.6 ± 11.1	27.8 ± 25.4	39.6 ± 3.0	69.2 ± 7.2	56.5 ± 3.2	46.4 ± 7.3	47.9 ± 3.7	66.7 ± 11.9
DuoRC-s	3.0 ± 5.2	75.8 ± 5.7	81.9 ± 0.7	49.5 ± 12.7	29.8 ± 27.6	40.1 ± 3.3	68.8 ± 9.8	55.9 ± 2.5	**46.6** ± **8.0**	48.0 ± 3.5	64.6 ± 13.8
CQ	2.7 ± 5.9	67.8 ± 5.5	82.0 ± 0.7	60.0 ± 16.0	50.0 ± 20.8	**44.3** ± **2.3**	71.7 ± 1.4	**60.3** ± **2.2**	41.2 ± 7.0	51.0 ± 2.7	75.8 ± 12.3
ComQA	3.1 ± 6.5	76.2 ± 5.3	81.8 ± 0.7	**67.7** ± **7.2**	**54.0** ± **15.3**	43.1 ± 2.2	74.0 ± 1.6	60.2 ± 1.9	38.5 ± 8.0	**51.4** ± **2.5**	**80.3** ± **8.4**

Table 19: Out-of-class transfer results from question answering tasks to classification/regression tasks in the LIMITED → LIMITED regime.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	SNLI	SciTail
Baseline	51.0	91.9	84.0	85.9	87.3	84.2	91.4	60.6	45.1	90.7	93.9
CCG	46.2	90.5	83.7	86.3	86.4	83.4	90.2	61.7	35.2	90.6	93.3
POS-PTB	39.7	91.2	85.7	86.2	86.9	82.9	90.3	61.7	42.3	**90.8**	91.9
POS-EWT	49.4	92.0	84.6	**86.9**	87.2	84.1	90.9	63.2	**56.3**	90.6	92.9
Parent	47.7	91.9	84.7	86.1	87.0	84.0	90.4	**65.3**	35.2	**90.8**	92.8
GParent	49.9	91.7	83.5	85.9	86.9	84.0	89.9	60.3	52.1	90.6	92.9
GGParent	49.2	91.4	84.3	86.2	86.9	83.3	90.9	57.0	43.7	90.3	90.9
ST	42.5	91.7	84.3	85.8	86.9	83.8	90.0	62.8	35.2	90.7	93.3
Chunk	48.6	90.9	85.1	86.1	86.9	84.0	91.0	62.1	46.5	**90.8**	93.6
NER	52.9	91.1	85.5	86.4	87.0	84.1	90.9	61.4	38.0	**90.8**	93.3
GED	51.1	91.5	82.7	86.2	87.2	84.1	90.7	58.1	40.8	90.6	92.8
Conj	**53.4**	**92.2**	**86.5**	86.5	87.2	83.9	90.4	63.9	38.0	90.7	**94.0**

Table 20: Out-of-class transfer results from sequence labeling tasks to classification/regression tasks in the FULL → FULL regime.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	SNLI	SciTail
Baseline	4.7 ± 8.2	77.5 ± 6.3	81.9 ± 0.7	29.9 ± 10.5	25.7 ± 25.1	38.7 ± 3.2	62.4 ± 9.5	54.7 ± 3.2	44.4 ± 6.3	46.7 ± 4.5	64.1 ± 13.6
CCG	0.0 ± 0.0	57.6 ± 3.2	81.5 ± 0.6	63.6 ± 10.4	57.2 ± 14.9	42.8 ± 2.0	**73.2 ± 1.1**	56.9 ± 2.6	44.7 ± 7.7	48.8 ± 3.7	78.3 ± 10.3
POS-PTB	0.2 ± 1.7	56.9 ± 4.0	**82.3 ± 0.6**	**68.7 ± 12.0**	52.9 ± 22.8	42.9 ± 1.0	73.0 ± 0.7	58.1 ± 1.7	42.7 ± 7.3	49.9 ± 2.3	80.1 ± 8.2
POS-EWT	0.9 ± 1.9	62.6 ± 4.2	81.9 ± 0.4	50.1 ± 12.3	44.1 ± 27.1	42.0 ± 2.5	72.1 ± 1.0	56.3 ± 3.1	**46.2 ± 7.1**	48.9 ± 3.7	78.0 ± 11.1
Parent	10.1 ± 5.6	58.7 ± 2.9	82.0 ± 0.7	51.9 ± 16.4	51.3 ± 13.7	43.5 ± 2.8	72.2 ± 1.1	**59.7 ± 2.4**	42.5 ± 6.6	49.7 ± 2.9	79.5 ± 9.0
GParent	8.1 ± 7.0	58.3 ± 3.2	81.7 ± 0.5	42.4 ± 20.2	51.5 ± 19.4	42.0 ± 1.8	70.9 ± 1.7	58.0 ± 3.2	44.5 ± 6.7	48.0 ± 3.2	77.3 ± 10.2
GGParent	6.3 ± 5.7	54.9 ± 2.3	**82.3 ± 0.9**	30.7 ± 16.9	41.9 ± 24.7	40.8 ± 1.9	68.1 ± 3.2	57.3 ± 3.2	42.6 ± 8.0	43.9 ± 3.8	74.8 ± 9.7
ST	1.3 ± 2.5	58.6 ± 3.0	**82.3 ± 0.7**	62.1 ± 20.5	**58.2 ± 15.5**	**44.3 ± 1.5**	71.3 ± 1.0	57.4 ± 2.0	45.1 ± 5.9	**50.8 ± 1.5**	**83.2 ± 1.7**
Chunk	0.5 ± 1.6	58.8 ± 5.3	81.8 ± 0.6	37.0 ± 27.4	51.0 ± 22.6	43.5 ± 2.2	72.1 ± 1.6	55.1 ± 3.5	**46.2 ± 7.8**	49.8 ± 3.7	75.3 ± 13.1
NER	3.6 ± 5.6	**77.8 ± 5.8**	81.7 ± 0.5	26.9 ± 18.8	50.9 ± 21.4	42.6 ± 2.9	67.8 ± 6.8	55.9 ± 2.1	45.8 ± 7.1	48.4 ± 3.0	72.7 ± 14.7
GED	**12.3 ± 11.8**	65.5 ± 8.0	81.5 ± 0.4	50.4 ± 11.1	40.7 ± 24.7	41.1 ± 2.0	69.6 ± 1.7	56.9 ± 2.4	39.2 ± 6.9	49.0 ± 3.6	69.9 ± 14.9
Conj	5.6 ± 8.3	68.5 ± 4.5	82.1 ± 0.8	51.6 ± 15.0	40.8 ± 30.1	42.3 ± 3.0	72.4 ± 2.1	58.2 ± 1.8	42.7 ± 5.7	48.9 ± 2.9	74.8 ± 14.6

Table 21: Out-of-class transfer results from sequence labeling tasks to classification/regression tasks in the FULL → LIMITED regime.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	SNLI	SciTail
Baseline	4.7 ± 8.2	77.5 ± 6.3	81.9 ± 0.7	29.9 ± 10.5	25.7 ± 25.1	38.7 ± 3.2	62.4 ± 9.5	54.7 ± 3.2	44.4 ± 6.3	46.7 ± 4.5	64.1 ± 13.6
CCG	0.2 ± 1.0	57.9 ± 4.5	81.4 ± 0.4	1.4 ± 16.1	30.6 ± 28.1	36.4 ± 3.2	65.0 ± 8.4	53.2 ± 4.0	49.0 ± 6.1	39.0 ± 4.6	56.4 ± 9.6
POS-PTB	0.3 ± 1.4	58.8 ± 3.7	**82.1 ± 0.6**	0.9 ± 12.1	29.3 ± 27.5	37.9 ± 3.6	63.6 ± 8.1	54.6 ± 4.0	45.7 ± 7.6	41.9 ± 5.0	65.2 ± 13.8
POS-EWT	0.3 ± 0.7	60.0 ± 6.5	81.6 ± 0.5	4.1 ± 7.3	23.9 ± 25.3	38.1 ± 3.2	68.0 ± 4.0	56.6 ± 2.4	46.1 ± 7.2	41.0 ± 4.3	65.8 ± 14.1
Parent	0.0 ± 0.9	62.6 ± 5.8	81.5 ± 0.6	7.7 ± 20.9	39.4 ± 29.1	40.8 ± 3.2	67.8 ± 5.9	58.8 ± 3.1	44.5 ± 7.7	45.2 ± 4.4	**79.9 ± 7.7**
GParent	1.3 ± 3.9	60.6 ± 4.7	81.7 ± 0.6	22.0 ± 23.3	47.1 ± 27.3	40.1 ± 2.8	69.9 ± 4.6	56.6 ± 2.0	45.4 ± 7.0	44.1 ± 4.3	72.5 ± 14.3
GGParent	1.0 ± 2.9	64.5 ± 5.2	81.4 ± 0.4	11.0 ± 16.6	40.3 ± 30.0	40.3 ± 3.2	66.3 ± 8.0	56.9 ± 3.1	44.4 ± 6.7	46.4 ± 5.1	71.0 ± 15.0
ST	0.7 ± 3.0	59.2 ± 4.7	81.6 ± 0.4	2.6 ± 11.6	22.1 ± 24.1	37.2 ± 4.0	60.3 ± 8.3	54.5 ± 3.7	45.3 ± 6.2	39.6 ± 4.3	59.9 ± 11.7
Chunk	0.0 ± 0.0	60.5 ± 3.8	81.6 ± 0.5	8.7 ± 24.8	**47.2 ± 24.7**	40.6 ± 4.2	68.6 ± 5.3	**59.3 ± 2.6**	**49.7 ± 8.2**	43.3 ± 6.4	74.8 ± 12.0
NER	4.9 ± 6.1	76.8 ± 2.7	81.7 ± 0.6	14.5 ± 23.9	43.5 ± 26.5	41.8 ± 2.8	70.5 ± 3.7	57.2 ± 3.2	43.7 ± 6.8	46.8 ± 5.1	70.6 ± 14.1
GED	**8.5 ± 10.2**	74.5 ± 8.7	81.9 ± 0.6	39.7 ± 14.5	33.6 ± 28.3	39.4 ± 2.8	64.2 ± 6.6	56.0 ± 2.6	43.6 ± 6.2	47.8 ± 4.1	69.0 ± 14.4
Conj	4.6 ± 6.5	73.9 ± 6.0	82.0 ± 0.6	**45.6 ± 14.0**	**47.2 ± 27.4**	**43.0 ± 2.7**	**70.7 ± 4.1**	58.2 ± 2.1	43.2 ± 5.8	**49.2 ± 4.0**	74.6 ± 16.0

Table 22: Out-of-class transfer results from sequence labeling tasks to classification/regression tasks in the LIMITED → LIMITED regime.

Task	SQuAD-1	SQuAD-2	NewsQA	HotpotQA	BoolQ	DROP	WikiHop	DuoRC-p	DuoRC-s	CQ	ComQA
Baseline	87.9	71.9	64.1	67.9	65.7	22.4	62.8	50.6	63.3	30.5	63.2
CoLA	87.8	70.1	64.6	68.2	64.9	22.3	62.9	51.0	**63.8**	30.0	62.7
SST-2	87.7	71.3	64.9	68.3	68.0	22.2	63.1	**51.1**	63.2	28.1	62.2
MRPC	87.8	67.7	63.8	66.4	66.4	22.4	62.5	51.0	63.1	26.9	62.5
STS-B	87.9	70.1	64.0	66.2	64.9	22.1	**63.4**	51.0	62.4	29.7	62.9
QQP	87.9	71.5	64.0	68.8	64.9	22.1	63.2	50.5	62.0	33.2	61.4
MNLI	87.4	72.8	64.9	68.7	**69.8**	22.7	63.3	50.7	62.6	**35.5**	61.6
QNLI	**88.2**	73.4	64.7	69.0	66.9	22.5	63.3	50.5	62.8	33.6	62.0
RTE	87.9	71.4	64.0	68.1	64.2	**22.8**	63.1	50.8	63.7	31.7	62.6
WNLI	87.9	70.3	64.3	67.9	65.3	22.3	62.3	50.7	63.7	32.5	61.9
SNLI	88.0	**74.3**	**65.1**	68.7	68.2	22.4	62.8	50.9	62.9	28.8	62.1
SciTail	87.9	71.3	64.5	**69.4**	68.3	22.7	63.3	51.0	63.0	33.0	61.6

Table 23: Out-of-class transfer results from classification/regression tasks to question answering tasks in the FULL → FULL regime.

Task	SQuAD-1	SQuAD-2	NewsQA	HotpotQA	BoolQ	DROP	WikiHop	DuoRC-p	DuoRC-s	CQ	ComQA
Baseline	26.8 ± 6.0	50.1 ± 0.1	28.8 ± 4.9	23.3 ± 4.0	62.2 ± 0.0	19.4 ± 11.8	25.5 ± 4.7	41.6 ± 1.1	38.9 ± 3.3	30.7 ± 2.5	39.1 ± 1.2
CoLA	26.2 ± 6.3	50.0 ± 0.1	30.4 ± 5.0	24.2 ± 3.4	62.2 ± 0.0	20.5 ± 13.8	27.3 ± 3.7	42.2 ± 1.2	41.5 ± 3.3	31.2 ± 1.5	39.2 ± 1.2
SST-2	22.4 ± 6.1	50.1 ± 0.0	30.4 ± 3.9	25.5 ± 4.6	62.2 ± 0.0	34.7 ± 14.3	26.7 ± 3.5	41.8 ± 1.2	39.8 ± 2.7	30.9 ± 2.6	38.7 ± 1.5
MRPC	23.4 ± 4.8	50.1 ± 0.0	25.7 ± 3.9	21.2 ± 2.1	62.2 ± 0.0	12.0 ± 11.1	23.9 ± 3.3	39.7 ± 1.7	35.0 ± 5.3	31.8 ± 2.3	38.6 ± 1.1
STS-B	34.1 ± 4.2	50.0 ± 0.0	24.6 ± 2.1	23.3 ± 3.0	62.2 ± 0.0	**40.1 ± 20.6**	23.7 ± 4.5	40.0 ± 1.8	35.5 ± 2.3	30.8 ± 2.3	38.2 ± 1.0
QQP	29.8 ± 6.6	50.0 ± 0.1	32.3 ± 3.6	31.3 ± 4.9	62.2 ± 0.0	17.3 ± 11.2	23.0 ± 4.5	42.0 ± 1.4	40.4 ± 3.2	**33.1 ± 2.0**	38.6 ± 1.5
MNLI	36.6 ± 2.6	50.1 ± 0.0	35.6 ± 2.8	27.5 ± 3.0	62.2 ± 0.0	15.2 ± 9.2	26.8 ± 2.9	42.7 ± 1.6	40.5 ± 3.1	32.7 ± 2.3	39.0 ± 1.5
QNLI	**57.1 ± 3.3**	**50.4 ± 0.5**	**41.5 ± 5.8**	**34.3 ± 7.2**	62.2 ± 0.0	31.6 ± 13.0	**28.0 ± 3.8**	**45.2 ± 1.7**	**50.7 ± 1.8**	32.9 ± 2.1	**39.4 ± 1.9**
RTE	25.9 ± 5.7	50.0 ± 0.1	28.7 ± 4.9	21.8 ± 4.4	62.2 ± 0.0	18.0 ± 11.6	24.8 ± 4.8	41.5 ± 1.2	39.1 ± 3.8	30.6 ± 1.8	38.9 ± 1.1
WNLI	25.8 ± 6.1	50.0 ± 0.1	30.1 ± 4.2	23.7 ± 3.7	62.2 ± 0.0	16.0 ± 10.0	26.2 ± 4.5	41.9 ± 0.8	39.2 ± 3.4	31.2 ± 2.2	38.8 ± 1.5
SNLI	31.2 ± 4.5	50.0 ± 0.1	36.7 ± 1.6	24.9 ± 3.0	62.2 ± 0.0	23.8 ± 12.6	26.0 ± 2.6	43.2 ± 1.4	41.3 ± 3.0	32.0 ± 2.0	39.3 ± 1.4
SciTail	29.9 ± 5.7	50.1 ± 0.0	28.7 ± 3.8	22.4 ± 3.6	62.2 ± 0.0	35.2 ± 16.0	23.1 ± 4.5	41.1 ± 2.1	40.4 ± 3.7	31.7 ± 1.9	38.7 ± 1.3

Table 24: Out-of-class transfer results from classification/regression tasks to question answering tasks in the FULL → LIMITED regime.

Task	SQuAD-1	SQuAD-2	NewsQA	HotpotQA	BoolQ	DROP	WikiHop	DuoRC-p	DuoRC-s	CQ	ComQA
Baseline	26.8 ± 6.0	50.1 ± 0.1	28.8 ± 4.9	23.3 ± 4.0	62.2 ± 0.0	19.4 ± 11.8	25.5 ± 4.7	41.6 ± 1.1	38.9 ± 3.3	30.7 ± 2.5	39.1 ± 1.2
CoLA	27.3 ± 6.0	50.0 ± 0.2	29.3 ± 4.6	**23.8 ± 4.2**	62.2 ± 0.0	17.0 ± 10.4	25.0 ± 4.5	**41.8 ± 1.1**	**40.1 ± 3.1**	31.2 ± 1.5	39.3 ± 1.1
SST-2	26.5 ± 6.0	50.1 ± 0.1	29.0 ± 4.6	23.4 ± 3.9	62.2 ± 0.0	22.4 ± 12.7	25.5 ± 4.5	41.5 ± 1.0	39.3 ± 3.2	30.9 ± 1.9	**39.4 ± 1.1**
MRPC	23.4 ± 4.5	50.0 ± 0.1	25.9 ± 3.6	21.2 ± 2.1	62.2 ± 0.0	18.8 ± 12.3	25.3 ± 4.5	41.2 ± 1.0	36.7 ± 3.9	31.4 ± 2.3	38.7 ± 1.7
STS-B	26.3 ± 4.5	50.1 ± 0.0	24.6 ± 2.5	21.5 ± 1.5	62.2 ± 0.0	26.8 ± 15.4	24.0 ± 4.5	41.2 ± 1.2	36.7 ± 3.7	31.5 ± 2.1	38.5 ± 1.3
QQP	19.0 ± 4.0	50.0 ± 0.1	26.9 ± 2.8	22.4 ± 2.5	62.2 ± 0.0	19.2 ± 11.3	24.4 ± 4.8	41.4 ± 1.1	37.3 ± 2.9	31.5 ± 2.1	38.7 ± 1.0
MNLI	26.7 ± 6.3	50.0 ± 0.1	28.2 ± 5.0	22.4 ± 3.8	62.2 ± 0.0	16.9 ± 11.0	25.0 ± 4.9	41.6 ± 1.4	39.4 ± 3.6	30.7 ± 1.7	38.7 ± 1.3
QNLI	**30.8 ± 4.9**	50.1 ± 0.0	28.4 ± 5.2	22.0 ± 4.1	62.2 ± 0.0	**29.5 ± 16.3**	24.9 ± 4.7	41.5 ± 1.2	37.7 ± 5.1	30.3 ± 2.9	38.7 ± 1.2
RTE	26.4 ± 5.6	50.0 ± 0.1	28.4 ± 4.5	22.7 ± 3.9	62.2 ± 0.0	18.4 ± 11.0	25.1 ± 5.2	41.5 ± 1.3	39.4 ± 3.0	30.8 ± 2.1	38.7 ± 1.2
WNLI	23.6 ± 4.6	50.1 ± 0.0	26.0 ± 3.8	21.6 ± 2.2	62.2 ± 0.0	23.8 ± 13.5	25.4 ± 4.1	41.2 ± 1.1	37.1 ± 2.9	**31.9 ± 2.1**	38.9 ± 1.2
SNLI	23.6 ± 5.7	50.0 ± 0.1	29.2 ± 4.5	23.4 ± 2.9	62.2 ± 0.0	16.6 ± 10.1	**25.6 ± 3.9**	**41.8 ± 1.0**	38.7 ± 3.6	30.7 ± 2.1	39.0 ± 1.3
SciTail	26.0 ± 6.1	50.0 ± 0.1	**29.8 ± 4.3**	22.8 ± 4.0	62.2 ± 0.0	19.6 ± 11.3	24.6 ± 4.9	**41.8 ± 1.1**	39.8 ± 3.2	31.2 ± 2.3	38.8 ± 1.3

Table 25: Out-of-class transfer results from classification/regression tasks to question answering tasks in the LIMITED → LIMITED regime.

Task	SQuAD-1	SQuAD-2	NewsQA	HotpotQA	BoolQ	DROP	WikiHop	DuoRC-p	DuoRC-s	CQ	ComQA
Baseline	87.9	71.9	64.1	67.9	65.7	22.4	62.8	50.6	63.3	30.5	63.2
CCG	87.0	68.1	63.8	66.3	65.5	22.0	62.2	49.7	62.1	30.5	61.1
POS-PTB	87.4	70.2	62.2	65.8	64.7	21.6	62.2	49.7	63.5	28.4	62.8
POS-EWT	85.9	66.7	62.6	66.2	65.4	22.0	62.6	50.2	**63.8**	33.7	61.5
Parent	87.4	69.5	64.4	67.9	66.4	21.9	63.1	**51.3**	63.3	34.3	62.3
GParent	87.6	70.2	64.1	67.9	65.8	**22.7**	61.9	50.5	62.8	**35.1**	62.2
GGParent	87.7	71.0	64.8	67.1	**67.0**	21.8	62.1	50.6	61.8	28.8	63.1
ST	87.6	70.7	62.6	68.0	63.7	21.9	61.7	50.3	63.2	30.2	61.6
Chunk	87.8	69.1	62.3	66.4	65.6	22.5	62.6	51.2	62.9	30.0	61.1
NER	**88.1**	70.0	63.7	67.0	66.6	22.5	62.6	51.1	63.6	34.6	62.6
GED	87.5	69.7	**65.0**	67.8	65.2	22.3	63.0	50.7	62.4	30.5	62.3
Conj	87.8	70.6	64.7	**68.3**	66.3	21.8	**63.2**	50.6	61.8	30.7	**64.3**

Table 26: Out-of-class transfer results from sequence labeling tasks to question answering tasks in the FULL → FULL regime.

Task	SQuAD-1	SQuAD-2	NewsQA	HotpotQA	BoolQ	DROP	WikiHop	DuoRC-p	DuoRC-s	CQ	ComQA
Baseline	26.8 ± 6.0	50.1 ± 0.1	28.8 ± 4.9	23.3 ± 4.0	62.2 ± 0.0	19.4 ± 11.8	25.5 ± 4.7	41.6 ± 1.1	38.9 ± 3.3	30.7 ± 2.5	39.1 ± 1.2
CCG	15.0 ± 0.9	50.1 ± 0.0	19.9 ± 2.5	19.9 ± 1.8	62.2 ± 0.0	59.7 ± 5.4	16.4 ± 1.7	36.0 ± 1.5	26.1 ± 7.7	31.1 ± 2.9	37.8 ± 1.7
POS-PTB	14.7 ± 0.8	50.1 ± 0.0	15.7 ± 2.5	15.4 ± 2.2	62.2 ± 0.0	59.5 ± 3.4	17.8 ± 1.7	35.9 ± 2.0	16.6 ± 7.6	29.9 ± 2.0	37.2 ± 1.2
POS-EWT	14.2 ± 1.1	50.1 ± 0.0	16.9 ± 2.4	17.2 ± 2.5	62.2 ± 0.0	**60.0 ± 4.0**	22.9 ± 2.0	36.7 ± 2.5	20.4 ± 9.7	32.2 ± 2.1	38.2 ± 1.4
Parent	19.1 ± 3.8	50.1 ± 0.1	23.8 ± 1.7	22.5 ± 2.0	62.2 ± 0.0	47.8 ± 14.9	22.2 ± 2.5	37.7 ± 1.7	29.8 ± 3.8	32.0 ± 2.4	38.7 ± 1.4
GParent	14.3 ± 0.7	50.1 ± 0.0	23.4 ± 2.2	19.6 ± 2.0	62.2 ± 0.0	49.5 ± 19.0	19.0 ± 1.7	38.0 ± 2.0	26.1 ± 6.0	31.6 ± 2.1	38.0 ± 1.1
GGParent	13.7 ± 0.4	50.1 ± 0.0	23.5 ± 2.7	17.9 ± 2.5	62.2 ± 0.0	38.1 ± 17.1	17.1 ± 1.4	37.9 ± 2.2	27.0 ± 6.1	32.0 ± 2.2	37.8 ± 1.7
ST	12.8 ± 0.6	50.1 ± 0.0	16.8 ± 3.3	15.5 ± 2.6	62.2 ± 0.0	**60.0 ± 3.9**	16.7 ± 1.3	36.8 ± 1.6	16.6 ± 7.4	29.4 ± 2.2	36.8 ± 1.4
Chunk	20.8 ± 4.7	50.1 ± 0.0	22.5 ± 3.2	21.7 ± 4.1	62.2 ± 0.0	52.5 ± 12.9	18.7 ± 3.5	37.5 ± 1.7	28.6 ± 6.6	31.7 ± 2.3	38.7 ± 1.4
NER	14.8 ± 1.2	50.1 ± 0.0	26.0 ± 4.0	18.7 ± 2.5	62.2 ± 0.0	25.4 ± 20.2	22.5 ± 4.9	38.5 ± 1.8	25.5 ± 9.8	31.0 ± 2.5	37.8 ± 1.8
GED	24.1 ± 4.7	50.1 ± 0.0	27.3 ± 3.4	24.0 ± 4.1	62.2 ± 0.0	43.1 ± 14.6	23.8 ± 4.8	41.2 ± 1.7	37.6 ± 2.9	31.8 ± 1.9	38.7 ± 1.7
Conj	**29.0 ± 8.9**	50.0 ± 0.2	28.3 ± 3.8	**25.7 ± 5.3**	62.2 ± 0.0	20.7 ± 15.5	**25.7 ± 3.4**	40.9 ± 2.0	38.8 ± 3.0	**32.8 ± 2.2**	38.9 ± 1.5

Table 27: Out-of-class transfer results from sequence labeling tasks to question answering tasks in the FULL → LIMITED regime.

Task	SQuAD-1	SQuAD-2	NewsQA	HotpotQA	BoolQ	DROP	WikiHop	DuoRC-p	DuoRC-s	CQ	ComQA
Baseline	26.8 ± 6.0	50.1 ± 0.1	28.8 ± 4.9	23.3 ± 4.0	62.2 ± 0.0	19.4 ± 11.8	25.5 ± 4.7	41.6 ± 1.1	38.9 ± 3.3	30.7 ± 2.5	39.1 ± 1.2
CCG	12.3 ± 1.0	50.1 ± 0.0	15.0 ± 2.8	12.5 ± 2.6	62.2 ± 0.0	**61.0 ± 2.0**	24.7 ± 1.4	40.5 ± 1.1	8.0 ± 7.1	30.7 ± 1.8	38.4 ± 0.9
POS-PTB	12.4 ± 1.1	50.1 ± 0.0	19.0 ± 2.6	16.2 ± 2.6	62.2 ± 0.0	59.9 ± 5.9	22.1 ± 2.2	39.0 ± 1.3	23.4 ± 9.6	30.0 ± 2.1	38.3 ± 1.4
POS-EWT	12.1 ± 0.7	50.1 ± 0.0	20.3 ± 4.4	18.2 ± 2.9	62.2 ± 0.0	53.8 ± 8.6	23.4 ± 2.2	39.7 ± 1.2	21.7 ± 10.7	31.4 ± 2.0	38.2 ± 1.4
Parent	14.5 ± 2.2	50.1 ± 0.0	21.4 ± 2.7	17.6 ± 2.2	62.2 ± 0.0	56.9 ± 10.7	21.3 ± 2.5	38.8 ± 0.9	25.0 ± 9.4	32.2 ± 2.2	38.3 ± 1.5
GParent	21.1 ± 5.2	50.1 ± 0.0	23.4 ± 1.9	19.6 ± 2.1	62.2 ± 0.0	54.6 ± 8.9	21.6 ± 2.6	38.9 ± 1.6	32.0 ± 3.2	**32.9 ± 1.6**	38.9 ± 1.4
GGParent	**31.3 ± 8.0**	49.6 ± 0.9	25.5 ± 3.4	23.2 ± 4.5	62.2 ± 0.0	36.9 ± 19.9	25.3 ± 2.3	40.2 ± 1.3	35.7 ± 2.2	31.8 ± 2.6	39.0 ± 1.7
ST	12.1 ± 0.5	50.1 ± 0.0	15.0 ± 3.5	14.7 ± 3.1	62.2 ± 0.0	58.2 ± 3.8	19.3 ± 2.0	39.8 ± 1.7	12.1 ± 8.9	30.1 ± 2.3	38.0 ± 1.3
Chunk	14.8 ± 3.6	50.1 ± 0.1	24.2 ± 0.9	18.4 ± 1.7	62.2 ± 0.0	45.9 ± 15.6	19.7 ± 2.7	39.4 ± 1.1	33.3 ± 2.8	30.5 ± 2.4	38.5 ± 1.5
NER	26.4 ± 7.2	50.1 ± 0.0	24.8 ± 2.7	19.2 ± 2.7	62.2 ± 0.0	39.9 ± 17.5	23.6 ± 4.3	39.9 ± 1.1	32.1 ± 4.3	31.2 ± 2.4	38.4 ± 1.4
GED	22.3 ± 5.9	50.1 ± 0.0	**28.9 ± 4.0**	23.1 ± 4.3	62.2 ± 0.0	23.1 ± 12.2	23.6 ± 5.0	41.1 ± 1.1	38.9 ± 4.3	31.4 ± 1.9	39.1 ± 1.5
Conj	28.7 ± 5.7	50.0 ± 0.1	26.3 ± 4.2	**23.6 ± 5.3**	62.2 ± 0.0	20.9 ± 16.1	**25.7 ± 3.3**	41.6 ± 1.4	37.7 ± 4.5	32.5 ± 2.6	38.7 ± 1.1

Table 28: Out-of-class transfer results from sequence labeling tasks to question answering tasks in the LIMITED → LIMITED regime.

Task	CCG	POS-PTB	POS-EWT	Parent	GParent	GGParent	ST	Chunk	NER	GED	Conj
Baseline	95.6	96.7	96.6	95.4	91.9	89.5	95.8	97.1	94.7	46.6	89.4
CoLA	95.5	96.7	**96.7**	95.2	91.8	89.4	95.8	97.0	94.6	46.6	89.8
SST-2	95.6	96.7	96.6	95.3	91.8	89.4	95.8	97.0	94.6	47.0	89.9
MRPC	95.6	96.6	96.6	95.2	91.9	89.4	95.8	97.0	94.5	47.0	90.3
STS-B	95.4	96.7	96.6	95.2	91.4	89.2	95.8	97.0	94.3	46.5	89.8
QQP	95.5	96.7	**96.7**	95.1	91.7	89.3	95.8	97.1	94.6	46.4	**90.4**
MNLI	95.4	96.7	96.6	95.1	91.9	89.0	95.7	97.1	94.6	46.6	**90.4**
QNLI	95.5	96.7	**96.7**	95.3	91.8	**89.6**	95.8	97.0	94.7	46.9	89.5
RTE	95.5	96.7	96.6	95.3	**92.0**	89.5	95.8	96.9	94.7	**47.4**	89.7
WNLI	95.5	96.7	96.5	95.4	91.8	89.5	95.8	97.0	94.5	46.3	89.4
SNLI	95.5	96.7	**96.7**	95.2	91.8	89.3	95.8	97.0	94.3	46.3	89.7
SciTail	95.5	96.7	**96.7**	95.2	**92.0**	89.4	95.8	97.0	94.5	46.2	89.5

Table 29: Out-of-class transfer results from classification/regression tasks to sequence labeling tasks in the FULL → FULL regime.

Task	CCG	POS-PTB	POS-EWT	Parent	GParent	GGParent	ST	Chunk	NER	GED	Conj
Baseline	53.2 ± 1.6	85.1 ± 0.9	89.3 ± 0.4	81.9 ± 0.9	62.8 ± 1.3	43.3 ± 1.7	76.7 ± 0.9	87.7 ± 0.5	77.4 ± 1.5	29.1 ± 1.3	73.3 ± 1.6
CoLA	48.9 ± 3.2	83.9 ± 0.9	88.8 ± 0.6	79.9 ± 0.9	63.2 ± 1.0	42.4 ± 1.5	75.2 ± 1.5	87.1 ± 0.6	77.3 ± 2.3	26.8 ± 2.1	71.6 ± 2.6
SST-2	50.0 ± 2.2	83.5 ± 1.1	88.4 ± 0.7	79.0 ± 1.0	61.8 ± 1.6	42.5 ± 2.4	75.7 ± 1.1	86.9 ± 0.7	78.6 ± 1.8	27.3 ± 0.7	73.3 ± 1.4
MRPC	53.0 ± 1.7	84.8 ± 0.9	89.3 ± 0.5	80.8 ± 1.3	62.3 ± 1.5	42.7 ± 1.7	76.8 ± 0.8	87.4 ± 0.5	77.2 ± 2.5	27.9 ± 2.6	72.7 ± 1.8
STS-B	**56.6 ± 1.5**	**86.6 ± 0.9**	**90.4 ± 0.4**	81.9 ± 1.2	61.4 ± 1.7	42.0 ± 2.8	**77.7 ± 0.9**	**87.9 ± 0.6**	72.1 ± 4.4	29.1 ± 3.2	72.1 ± 2.8
QQP	50.3 ± 3.3	83.6 ± 0.9	88.8 ± 0.5	80.3 ± 1.2	62.1 ± 1.2	43.2 ± 1.4	75.1 ± 1.2	86.7 ± 0.7	78.8 ± 1.4	26.5 ± 1.1	71.5 ± 1.6
MNLI	50.1 ± 1.8	82.5 ± 1.0	88.6 ± 0.5	79.6 ± 0.8	61.4 ± 1.2	41.9 ± 2.0	74.7 ± 1.4	86.5 ± 0.6	79.8 ± 1.6	27.0 ± 0.6	74.2 ± 1.4
QNLI	49.5 ± 3.2	83.5 ± 1.1	89.0 ± 0.4	80.3 ± 1.0	63.1 ± 1.3	41.9 ± 2.0	76.0 ± 0.8	87.1 ± 0.8	**80.7 ± 1.6**	27.0 ± 0.9	**75.2 ± 1.3**
RTE	51.3 ± 3.0	84.4 ± 1.0	88.9 ± 0.4	80.7 ± 1.2	62.8 ± 1.4	42.9 ± 1.8	76.1 ± 1.0	87.5 ± 0.5	77.3 ± 1.9	28.0 ± 1.8	73.7 ± 2.1
WNLI	53.3 ± 1.5	84.8 ± 1.0	89.3 ± 0.4	81.7 ± 1.2	62.5 ± 1.4	42.7 ± 1.9	76.4 ± 0.9	87.7 ± 0.5	76.5 ± 2.0	28.8 ± 1.4	73.2 ± 1.7
SNLI	52.0 ± 2.6	83.7 ± 1.0	88.6 ± 0.5	81.1 ± 1.0	62.7 ± 1.0	42.4 ± 2.1	76.1 ± 1.3	87.3 ± 0.6	79.1 ± 2.1	28.0 ± 0.8	73.8 ± 1.2
SciTail	51.8 ± 2.5	84.4 ± 1.0	89.1 ± 0.3	80.6 ± 1.1	**64.0 ± 1.3**	**43.5 ± 2.1**	76.3 ± 1.1	87.7 ± 0.5	77.6 ± 1.8	28.8 ± 1.4	**75.2 ± 1.7**

Table 30: Out-of-class transfer results from classification/regression tasks to sequence labeling tasks in the FULL → LIMITED regime.

Task	CCG	POS-PTB	POS-EWT	Parent	GParent	GGParent	ST	Chunk	NER	GED	Conj
Baseline	53.2 ± 1.6	85.1 ± 0.9	89.3 ± 0.4	81.9 ± 0.9	62.8 ± 1.3	43.3 ± 1.7	76.7 ± 0.9	87.7 ± 0.5	77.4 ± 1.5	29.1 ± 1.3	73.3 ± 1.6
CoLA	53.9 ± 1.4	85.3 ± 0.8	89.4 ± 0.5	**82.3 ± 1.0**	63.3 ± 1.5	43.4 ± 1.7	**77.8 ± 2.5**	**87.8 ± 0.4**	77.7 ± 2.6	**29.3 ± 1.4**	74.0 ± 1.4
SST-2	**54.0 ± 2.2**	85.2 ± 0.9	89.4 ± 0.4	82.1 ± 1.0	63.4 ± 2.0	43.8 ± 1.9	76.9 ± 0.8	**87.8 ± 0.7**	77.9 ± 1.9	28.9 ± 1.3	74.2 ± 1.1
MRPC	51.9 ± 2.2	84.8 ± 1.6	89.0 ± 0.5	81.0 ± 1.2	63.1 ± 1.4	43.3 ± 1.8	76.3 ± 1.0	87.6 ± 0.4	77.4 ± 2.2	28.6 ± 1.6	73.7 ± 2.0
STS-B	53.5 ± 2.7	85.1 ± 0.9	89.5 ± 0.4	81.3 ± 1.6	62.9 ± 2.0	43.4 ± 2.1	77.1 ± 0.8	87.6 ± 0.6	77.8 ± 1.7	28.8 ± 2.2	72.7 ± 2.2
QQP	52.5 ± 1.7	84.4 ± 1.0	88.8 ± 0.5	81.2 ± 1.1	**63.6 ± 2.0**	43.1 ± 1.8	76.2 ± 0.9	87.5 ± 0.8	77.7 ± 2.0	28.6 ± 1.5	**74.3 ± 1.4**
MNLI	53.3 ± 3.1	84.8 ± 1.0	89.4 ± 0.4	81.7 ± 1.1	62.8 ± 1.4	43.0 ± 1.8	77.1 ± 1.7	**87.8 ± 0.5**	77.4 ± 2.1	28.4 ± 1.6	73.4 ± 1.8
QNLI	53.4 ± 1.6	**85.7 ± 1.5**	**89.6 ± 0.3**	82.1 ± 1.0	63.2 ± 1.4	44.0 ± 2.6	77.1 ± 1.0	**87.8 ± 0.4**	**78.6 ± 2.9**	29.1 ± 1.4	73.6 ± 2.1
RTE	52.5 ± 1.6	84.5 ± 0.9	89.0 ± 0.5	81.2 ± 1.1	63.0 ± 1.4	43.4 ± 2.0	76.3 ± 0.9	87.4 ± 0.4	77.3 ± 2.0	28.7 ± 1.4	74.2 ± 2.1
WNLI	53.1 ± 1.7	84.7 ± 0.9	89.2 ± 0.5	81.5 ± 1.7	62.8 ± 1.4	**44.3 ± 2.4**	76.6 ± 0.9	87.7 ± 0.7	77.6 ± 2.4	29.1 ± 1.4	73.3 ± 1.9
SNLI	52.0 ± 2.1	84.4 ± 1.2	88.9 ± 0.4	80.9 ± 1.3	62.2 ± 1.4	42.7 ± 1.9	75.9 ± 1.1	87.6 ± 0.9	77.2 ± 2.2	28.7 ± 1.7	73.0 ± 2.0
SciTail	52.8 ± 1.5	84.8 ± 0.8	89.0 ± 0.4	81.4 ± 1.7	63.5 ± 1.3	43.9 ± 1.9	76.5 ± 0.9	87.4 ± 0.5	77.4 ± 2.1	28.8 ± 1.5	74.1 ± 1.6

Table 31: Out-of-class transfer results from classification/regression tasks to sequence labeling tasks in the LIMITED → LIMITED regime.

Task	CCG	POS-PTB	POS-EWT	Parent	GParent	GGParent	ST	Chunk	NER	GED	Conj
Baseline	95.6	96.7	96.6	95.4	91.9	89.5	95.8	97.1	94.7	46.6	89.4
SQuAD-1	95.4	96.7	**96.7**	95.3	91.8	89.5	95.8	97.1	94.8	46.7	**90.3**
SQuAD-2	95.4	96.7	96.6	95.3	91.8	89.4	95.8	97.1	94.5	46.4	89.9
NewsQA	95.5	96.7	96.4	95.3	91.6	89.2	95.8	97.0	94.4	45.6	90.0
HotpotQA	95.4	96.7	96.3	95.1	91.7	89.1	95.8	96.9	94.5	45.8	90.0
BoolQ	95.5	96.7	96.6	95.3	91.7	89.5	95.8	96.9	94.7	**47.2**	89.4
DROP	95.5	96.7	**96.7**	95.3	91.7	89.4	95.8	97.1	94.5	47.1	90.0
WikiHop	95.5	96.7	96.2	95.2	91.5	89.0	95.8	96.8	94.5	46.8	88.8
DuoRC-p	95.4	96.7	96.4	95.4	91.7	89.4	95.7	96.9	94.4	46.2	89.7
DuoRC-s	95.5	96.7	96.6	95.3	91.8	89.3	95.8	97.1	**94.9**	46.5	90.0
CQ	95.4	96.7	96.6	95.3	91.6	89.3	95.8	96.9	94.5	46.9	89.7
ComQA	95.5	96.7	96.5	95.1	91.7	89.3	95.7	96.8	94.1	46.6	89.2

Table 32: Out-of-class transfer results from question answering tasks to sequence labeling tasks in the FULL → FULL regime.

Task	CCG	POS-PTB	POS-EWT	Parent	GParent	GGParent	ST	Chunk	NER	GED	Conj
Baseline	53.2 ± 1.6	85.1 ± 0.9	89.3 ± 0.4	81.9 ± 0.9	62.8 ± 1.3	43.3 ± 1.7	76.7 ± 0.9	87.7 ± 0.5	77.4 ± 1.5	29.1 ± 1.3	73.3 ± 1.6
SQuAD-1	**57.5 ± 1.1**	86.7 ± 0.7	90.3 ± 0.3	**83.5 ± 0.7**	**67.2 ± 1.0**	**48.7 ± 1.5**	79.4 ± 0.8	**88.7 ± 0.3**	**84.2 ± 1.7**	27.9 ± 1.1	**77.6 ± 1.0**
SQuAD-2	56.8 ± 1.4	85.9 ± 0.8	89.9 ± 0.4	82.7 ± 0.7	66.3 ± 1.0	47.2 ± 1.3	78.7 ± 0.7	88.2 ± 0.5	83.7 ± 1.5	28.6 ± 1.2	75.6 ± 1.8
NewsQA	55.6 ± 2.2	85.2 ± 0.9	89.3 ± 0.4	81.3 ± 1.4	64.1 ± 1.2	46.3 ± 2.0	78.4 ± 0.7	87.5 ± 0.5	81.0 ± 1.7	27.0 ± 0.9	73.3 ± 1.0
HotpotQA	47.3 ± 4.1	81.9 ± 1.3	88.0 ± 0.6	77.5 ± 1.0	62.6 ± 1.1	41.7 ± 1.9	74.7 ± 1.4	86.1 ± 0.5	76.0 ± 2.9	26.7 ± 0.7	69.0 ± 2.1
BoolQ	50.8 ± 3.8	84.1 ± 1.4	88.6 ± 0.5	80.3 ± 1.4	60.8 ± 1.2	42.2 ± 2.2	75.6 ± 1.7	87.2 ± 0.6	74.1 ± 2.4	25.8 ± 2.8	73.8 ± 1.6
DROP	56.1 ± 1.2	**86.9 ± 1.1**	**90.6 ± 0.3**	82.6 ± 0.9	66.1 ± 0.8	47.3 ± 1.6	**80.2 ± 0.9**	88.4 ± 0.5	82.3 ± 1.5	29.7 ± 1.0	76.3 ± 1.1
WikiHop	53.3 ± 1.8	83.4 ± 1.1	88.6 ± 0.5	79.3 ± 0.9	60.5 ± 1.1	42.2 ± 2.2	77.2 ± 1.2	86.3 ± 1.1	77.5 ± 2.1	28.9 ± 1.4	66.3 ± 2.8
DuoRC-p	53.2 ± 2.4	84.0 ± 1.3	89.1 ± 0.7	80.1 ± 1.0	62.6 ± 1.2	43.0 ± 1.8	76.2 ± 1.1	87.0 ± 0.8	79.0 ± 2.5	26.4 ± 1.5	71.5 ± 2.1
DuoRC-s	55.4 ± 2.1	84.8 ± 0.9	89.5 ± 0.4	81.0 ± 0.9	64.3 ± 1.1	43.8 ± 1.9	77.3 ± 0.9	87.6 ± 0.5	81.9 ± 1.6	28.5 ± 0.9	72.9 ± 1.9
CQ	54.1 ± 1.4	85.4 ± 1.2	89.2 ± 0.3	80.6 ± 1.1	65.5 ± 0.9	47.2 ± 1.6	77.8 ± 1.1	87.5 ± 0.7	75.9 ± 1.7	**30.6 ± 1.1**	72.9 ± 1.2
ComQA	53.0 ± 2.1	81.9 ± 1.4	87.2 ± 1.0	79.0 ± 1.6	61.8 ± 1.0	44.3 ± 1.7	75.4 ± 1.5	86.6 ± 1.0	71.7 ± 2.8	27.2 ± 1.3	68.8 ± 1.9

Table 33: Out-of-class transfer results from question answering tasks to sequence labeling tasks in the FULL → LIMITED regime.

Task	CCG	POS-PTB	POS-EWT	Parent	GParent	GGParent	ST	Chunk	NER	GED	Conj
Baseline	53.2 ± 1.6	85.1 ± 0.9	89.3 ± 0.4	81.9 ± 0.9	62.8 ± 1.3	43.3 ± 1.7	76.7 ± 0.9	87.7 ± 0.5	77.4 ± 1.5	29.1 ± 1.3	73.3 ± 1.6
SQuAD-1	56.2 ± 1.4	86.4 ± 0.6	90.1 ± 0.4	83.0 ± 0.7	64.0 ± 2.1	45.7 ± 2.7	78.4 ± 0.6	**88.4 ± 0.5**	76.9 ± 3.4	30.3 ± 1.0	74.5 ± 1.5
SQuAD-2	56.4 ± 0.9	86.8 ± 0.6	90.3 ± 0.5	**83.1 ± 0.7**	63.7 ± 1.1	45.1 ± 2.3	78.7 ± 0.6	88.3 ± 0.4	77.0 ± 3.2	30.5 ± 0.9	**75.0 ± 2.0**
NewsQA	54.7 ± 1.1	86.2 ± 1.0	90.0 ± 0.4	82.4 ± 0.8	64.7 ± 1.0	46.2 ± 3.8	78.5 ± 0.6	88.2 ± 0.4	80.5 ± 2.7	30.9 ± 1.0	73.5 ± 2.1
HotpotQA	55.7 ± 3.9	85.7 ± 0.9	89.8 ± 0.4	81.3 ± 1.0	**65.1 ± 0.9**	46.4 ± 2.0	79.0 ± 1.6	88.1 ± 0.4	**82.0 ± 1.7**	31.6 ± 1.0	74.3 ± 1.5
BoolQ	53.4 ± 2.5	85.5 ± 0.8	89.5 ± 0.4	80.7 ± 1.1	63.2 ± 1.1	43.0 ± 3.1	76.5 ± 1.4	87.6 ± 0.4	71.7 ± 4.0	28.5 ± 1.3	74.6 ± 1.2
DROP	54.2 ± 2.4	85.3 ± 1.0	89.5 ± 0.5	82.5 ± 1.1	63.4 ± 1.2	44.2 ± 1.9	77.6 ± 0.9	88.0 ± 0.5	79.4 ± 2.9	29.0 ± 1.0	74.1 ± 1.2
WikiHop	55.6 ± 1.8	87.4 ± 0.8	**90.5 ± 0.3**	82.9 ± 1.2	64.8 ± 0.7	45.4 ± 2.2	**80.1 ± 0.9**	88.3 ± 0.6	81.3 ± 1.6	31.6 ± 0.9	73.4 ± 1.8
DuoRC-p	**56.5 ± 1.1**	**87.5 ± 1.0**	90.4 ± 0.7	82.9 ± 0.5	64.4 ± 0.9	46.1 ± 3.1	79.6 ± 0.6	**88.4 ± 0.3**	80.7 ± 1.5	**31.7 ± 0.7**	73.6 ± 1.4
DuoRC-s	55.7 ± 3.2	86.7 ± 0.7	90.0 ± 0.5	82.2 ± 0.8	64.4 ± 2.0	45.2 ± 1.7	78.5 ± 0.8	88.2 ± 0.5	80.4 ± 1.4	29.9 ± 1.4	73.4 ± 4.0
CQ	51.5 ± 2.5	84.6 ± 0.7	89.2 ± 0.6	81.4 ± 1.7	65.0 ± 1.3	45.7 ± 1.8	76.5 ± 1.1	87.3 ± 0.7	76.7 ± 1.9	30.8 ± 1.3	70.5 ± 2.4
ComQA	54.3 ± 1.5	85.4 ± 1.4	89.5 ± 0.7	81.8 ± 1.3	64.2 ± 1.5	**46.8 ± 2.1**	77.5 ± 1.4	**88.4 ± 0.4**	79.3 ± 2.5	29.1 ± 2.1	72.4 ± 2.3

Table 34: Out-of-class transfer results from question answering tasks to sequence labeling tasks in the LIMITED → LIMITED regime.