FETA: A Benchmark for Few-Sample Task Transfer in Open-Domain Dialogue

Alon Albalak[1] **Yi-Lin Tuan**[1] **Pegah Jandaghi**[2] **Connor Pryor**[3] **Luke Yoffe**[1]
Deepak Ramachandran[4] **Lise Getoor**[3] **Jay Pujara**[2] **William Yang Wang**[1]
[1]University of California, Santa Barbara [2]University of Southern California
[3]University of California, Santa Cruz [4]Google Research
alon_albalak@ucsb.edu

Abstract

Task transfer, transferring knowledge contained in related tasks, holds the promise of reducing the quantity of labeled data required to fine-tune language models. Dialogue understanding encompasses many diverse tasks, yet task transfer has not been thoroughly studied in conversational AI. This work explores conversational task transfer by introducing FETA: a benchmark for **FE**w-sample **TA**sk transfer in open-domain dialogue. FETA contains two underlying sets of conversations upon which there are 10 and 7 tasks annotated, enabling the study of intra-dataset task transfer; task transfer without domain adaptation. We utilize three popular language models and three learning algorithms to analyze the transferability between 132 source-target task pairs and create a baseline for future work. We run experiments in the single- and multi-source settings and report valuable findings, e.g., most performance trends are model-specific, and span extraction and multiple-choice tasks benefit the most from task transfer. In addition to task transfer, FETA can be a valuable resource for future research into the efficiency and generalizability of pre-training datasets and model architectures, as well as for learning settings such as continual and multitask learning. [1]

1 Introduction

Improving sample efficiency through transfer learning has been a long-standing challenge in the machine learning and natural language processing communities (Pratt et al., 1991; Ando and Zhang, 2005). Dialogue data requires multiple cohesive turns with consistent speaker personalities (Urbanek et al., 2019; Huang et al., 2020), creating a challenge for data collection and motivating the development of techniques that improve sample efficiency in conversational AI (Lin et al., 2020).

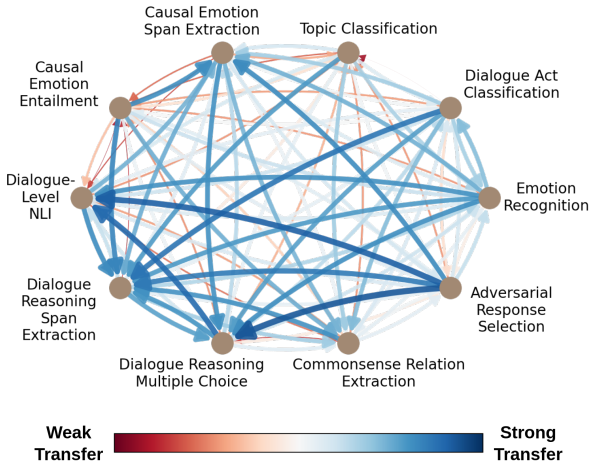

Figure 1: **Task Transfer Performance** on FETA-DailyDialog. Computed transfer performance is demonstrated by arrows leaving from source tasks and entering target tasks. Strength of the transfer is denoted by thickness and color of edges.

Furthermore, dialogue understanding tasks require a shared knowledge of semantics, pragmatics, human behavior, and commonsense, making dialogue an area of study that can benefit greatly from a deeper understanding of transfer learning.

Two essential transfer learning settings, namely domain adaptation and task transfer, have been studied on language tasks (Ruder et al., 2019). While domain adaptation has been studied in task-oriented dialogue (Mehri et al., 2020) , task transfer has been studied with less rigor in conversational AI. Prior studies of task transfer in dialogue consider only 2-4 tasks, focus on multitask learning, and do not compare learning algorithms (Hosseini-Asl et al., 2020; Peng et al., 2021b).

Prior studies have focused on cross-dataset task transfer, gathering tasks annotated on disjoint datasets (Vu et al., 2020; Ye et al., 2021), but this can lead to improvements in domain adaptation being confounded as improvements in task transfer. A precise study of task transfer should be on a sin-

[1]Benchmark available at alon-albalak.github.io/feta-website. We utilize the Transfer Learning in Dialogue Benchmarking Toolkit for all experiments (TLiDB python package).

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10936–10953
December 7-11, 2022

gle data source in an intra-dataset transfer setting, as in Zamir et al. (2018). Additionally, previous studies focus on learning algorithms and use only a single language model architecture (Pruksachatkun et al., 2020; Lourie et al., 2021; Aribandi et al., 2022), which may lead to a narrow understanding. To the best of our knowledge, this is the first rigorous study on task transfer in dialogue and the most extensive intra-dataset task transfer study in NLP.

In this work, we create FETA, a benchmark for few-sample task transfer for language understanding in open-domain dialogue with 17 total tasks. FETA datasets cover a variety of properties (dyadic vs. multi-party, anonymized vs. recurring speaker, varying dialogue lengths) and task types (utterance-level classification, dialogue-level classification, span extraction, multiple-choice), and maintain a wide variety of data quantities.

We study task transfer on FETA by comparing three task transfer algorithms and three commonly used language models in single-source and multi-source settings. Figure 1 illustrates some results in the single-source setting. For example, we find that Dialogue Reasoning Span Extraction benefits from nearly all source tasks. On the other hand, Adversarial Response Selection and Emotion Recognition improve the performance of many target tasks when utilized as a source task.

In this study, we find that: **(i)** Trends are largely model-dependent, a finding that previous works have not discussed. **(ii)** Out of all task types, span extraction tasks gain the most as a target, especially with few samples. **(iii)** Adding source tasks does not uniformly improve over a single source task, motivating a better understanding of the complex relationship between source and target tasks.

FETA provides a resource for various future studies, e.g., on the generalizability of model architectures, and pre-training datasets that enable efficient transfer. In addition to task transfer, FETA can also facilitate the study of continual and multitask learning.

In summary, our main contributions are:

- We create the first large-scale benchmark for task transfer in dialogue, with 132 source-target task pairs.
- Extensive experimentation on FETA in both the single-source and multi-source settings, and an in-depth analysis comparing models, learning algorithms, sample sizes, and task types, finding new and non-intuitive results.
- A readily extensible transfer learning framework[2] that allows for rapid experimentation and an online leaderboard[3] to encourage deeper research into task transfer.

2 Related Work

Transfer Learning in NLP Prior works on transfer learning in NLP have studied a wide variety of topics, including domain adaptation (Ben-David et al., 2010), multitask learning (Collobert and Weston, 2008; Bingel and Søgaard, 2017), and learning representations of words (Brown et al., 1992; Mikolov et al., 2013; Peters et al., 2017, 2018). More recently, DialoGLUE (Mehri et al., 2020) and RADDLE (Peng et al., 2021a) study domain adaptation for language understanding tasks in task-oriented dialogue. Shuster et al. (2020) focuses on multitasking in dialogue response generation across multiple datasets. Similar to this work, Pruksachatkun et al. (2020) study task transfer, although they study cross-dataset task transfer in general NLP tasks. Lourie et al. (2021) also study task transfer, but they focus on the T5 model and a suite of commonsenseQA datasets.

Task Transfer in Dialogue Task transfer has been applied in Task-Oriented Dialogue (TOD) settings but never rigorously studied. For example, Hosseini-Asl et al. (2020) and Lin et al. (2020) develop multitask models to perform 2-4 TOD tasks but do not aim to analyze the efficiency of models or learning algorithms for task transfer.

Intra-dataset Task Transfer Intra-dataset task transfer has been studied in computer vision applications (Zamir et al., 2018; Pal and Balasubramanian, 2019), but to our best knowledge it has never been studied in NLP.

3 FETA

In this section, we briefly define *intra-dataset task transfer*, the problem setting of FETA. Then, we introduce FETA, our benchmark for few-sample task transfer in open-domain dialogue. Finally, we define the metrics we use to evaluate models and learning algorithms on FETA.

3.1 Problem Definitions

Let a dataset be composed of the instance set, X, and n task-specific label sets $Y_1, Y_2, \ldots, Y_n$. In

[2] github.com/alon-albalak/TLiDB
[3] alon-albalak.github.io/feta-website/

	Task Name	Original Samples	FETA Samples Train	Dev	Test	Task Type	Metrics
DailyDialog	Emotion Recognition	102978	7230	1269	15885	Utt Cls	M/m-F1
	Dialogue Act Classification	102978	7230	1269	15885	Utt Cls	M/m-F1
	Topic Classification	13118	958	161	1919	Dial Cls	M/m-F1
	Causal Emotion Span Extraction	36324	2141	169	9133	Span Ex	T-F1,EM
	Causal Emotion Entailment	36324	2141	169	9133	Dial Cls	M-F1,Acc
	Dialogue-Level NLI	5817	569	52	1302	Dial Cls	M-F1,Acc
	Dialogue Reasoning Span Extraction	1098	123	13	244	Span Ex	T-F1,EM
	Dialogue Reasoning Multiple Choice	2165	224	26	496	Mult Ch	Acc
	Commonsense Relation Extraction	4009	350	38	851	Dial Cl.	M-F1,Acc
	Adversarial Response Selection	57145	3400	895	10750	Mult Ch	Acc
Friends	Emotion Recognition (EmoryNLP)	12606	844	207	1912	Utt Cls	m/W-F1
	Reading Comprehension	13865	912	181	2284	Mult Ch	Acc
	Character Identification	50247	3593	638	7803	Utt Cls	M/m-F1
	Question Answering	12257	819	191	1937	Span Ex	T-F1,EM
	Personality Detection	711	54	15	110	Dial Cls	Acc
	Relation Extraction	7636	519	121	1188	Dial Cls	m-F1
	Emotion Recognition (MELD)	9140	616	148	1247	Utt Cls	m/W-F1

Table 1: **Overview of FETA tasks**. Task types are abbreviated as follows: Utt Cls for utterance-level classification, Dial Cls for dialogue-level classification, Span Ex for span extraction, and Mult Ch for multiple choice. Metrics are abbreviated as follows: M-F1 for macro-F1, m-F1 for micro-F1, T-F1 for token-F1, W-F1 for weighted-F1, EM for exact match and Acc for accuracy.

FETA, each instance $x \in X$ is a dialogue.

Definition 1 (Domain and Task)**.** *A* domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ *consists of a feature space* $\mathcal{X}$ *and a marginal probability distribution* $P(X)$*. The marginal probabilities are over the instance set* $X = \{x_1, x_2, \ldots, x_n\} \in \mathcal{X}$*.*

A task $\mathcal{T} = \{\mathcal{Y}, f(X)\}$ *is composed of a label space* $\mathcal{Y}$ *and a predictive function,* $f : \mathcal{X} \rightarrow \mathcal{Y}$*.*

Definition 2 (Learning Algorithm)**.** *A* learning algorithm*,* $\mathcal{A}$*, is a protocol that determines the method by which the instance set* X *and task-specific label sets* $Y_1, Y_2, \ldots, Y_n$ *will be used to train a predictive function,* f*.*

Definition 3 (Task Transfer)**.** *Given a source task* $\mathcal{T}_S = \{\mathcal{Y}_S, f_S(X_S)\}$ *and target task* $\mathcal{T}_T = \{\mathcal{Y}_T, f_T(X_T)\}$*,* task transfer *is the use of a learning algorithm,* $\mathcal{A}$*, to improve the learning of* f_T *by using the knowledge in* $\mathcal{T}_S$*.*

In **cross-dataset task transfer**, when $X_S \neq X_T$, we also have $P(X_S) \neq P(X_T)$ and $\mathcal{D}_S \neq \mathcal{D}_T$; domain shift.

In **intra-dataset task transfer**, when $X_S = X_T$, there is no domain shift. This enables the study of the learning algorithm's performance on task transfer, isolated from domain adaptation.

We refer the reader to Pan and Yang (2010) and Zhuang et al. (2021) for expanded discussions on transfer learning definitions.

Few-Sample Due to the challenge and cost of collecting and annotating data, many real-world applications of NLP techniques are limited by data quantities. For this reason, we focus on the few-sample setting, defined in FETA as 10% of the original instance set. Out of 10%, 5%, and 1%, 10% was empirically determined to be the smallest percentage that retains labels from all label sets in both the train and development partitions. Given the recent attention focused on NLP applications in low-resource settings (Brown et al., 2020; Bansal et al., 2020; Mukherjee et al., 2021; Ye et al., 2021), we expect research done in such a low-data setting will lead to insights useful for many researchers and practitioners.

3.2 FETA Datasets

In this section, we describe the two dialogue sources we use, DailyDialog (Li et al., 2017) and Friends (Chen and Choi, 2016), and the tasks annotated on each source.

We select these datasets because they complement each other in desirable ways. DailyDialog contains 2-speaker dialogues where speakers are anonymized and averages 88 words per dialogue. In contrast, Friends consists of multiparty dialogues (3.6 speakers mean, 15 max) with recurring characters and averages 283 words per dialogue. These differences lead to each set of dialogue instances having different task annotations, giving FETA a wider variety of tasks. For example, DailyDialog tasks include understanding the causes of emotions and commonsense reasoning, while tasks annotated on Friends revolve more around recog-

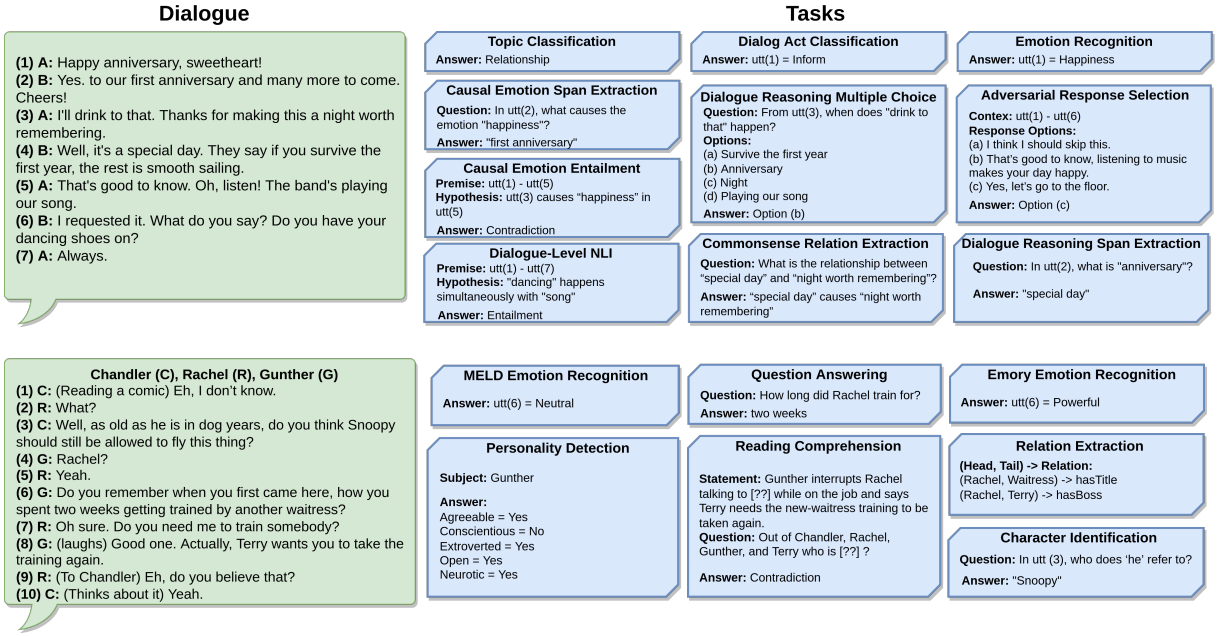

Figure 2: **Example dialogues and tasks** for FETA-DailyDialog (top) and FETA-Friends (bottom).

nizing entities and understanding personalities.

To create FETA versions of each dataset, we first partition the dialogues into 70/15/15% splits for training, validation, and test sets. After splitting, we randomly down-sample the train and development dialogues to 10% of the original quantities. Thus, FETA splits use 7/1.5/15% of the original dialogues. Not every dialogue is annotated for all tasks, allowing some tasks to have more samples than others. Crucially, the data splits are the same for all tasks, preventing data leakage. Table 1 shows an overview of the tasks, samples, and metrics used for each dataset.

FETA-DailyDialog Li et al. (2017) present the DailyDialog dataset, with chit-chat conversations covering 10 various topics including relationships, politics, and work.

Many works add annotations on top of these dialogues and FETA utilizes 10 of them. Figure 2 provides an overview of the tasks: *emotion recognition*, *dialogue act classification*, *topic classification* (from DailyDialog (Li et al., 2017)), *causal emotion span extraction*, *causal emotion entailment* (from RECCON (Poria et al., 2021)), *dialogue-level natural language inference*, *dialogue reasoning span extraction*, *dialogue reasoning multiple choice*, *commonsense relation extraction* (from CIDER (Ghosal et al., 2021)) *adversarial response selection* (from DailyDialog++ (Sai et al., 2020)). For further details of these tasks, we refer the reader to Appendix A and their original papers.

FETA-Friends The Friends dialogues come from transcripts of 10 seasons of the TV show by the same name (Chen and Choi, 2016). In addition to dialogue, the transcripts contain situational information such as behaviors and non-verbal information like scene information.

In total, FETA has 7 task annotations on top of the Friends scripts. As illustrated in Figure 2, the incorporated tasks include *Emory emotion recognition* (from (Zahiri and Choi, 2018)), *reading comprehension* (from (Ma et al., 2018)), *character identification* (from (Chen and Choi, 2016; Zhou and Choi, 2018)), *question answering* (from (Yang and Choi, 2019)), *personality detection* (from (Jiang et al., 2020)), and *relation extraction* (from DialogRE (Yu et al., 2020a)) and *MELD emotion recognition* (from MELD (Poria et al., 2019)). There are two emotion recognition label sets (Emory and MELD), but they have only 22% overlap in instance sets and have different label spaces. For further details of these tasks, we refer the reader to Appendix A and their original papers.

3.3 Evaluation Metrics

To define the metrics, we consider 4 variables: source task s, target task t, model f, and learning algorithm $\mathcal{A}$, and we abuse notation slightly to allow for $f_{\mathcal{A}}(s,t)$ to represent a model trained on the source and target tasks using the given learning

algorithm. In FETA, we evaluate the performance of a model and learning algorithm with multiple metrics: average and top-1 raw scores, as well as average and top-1 score Δs.

Average and Top-1 Scores First, we consider the two raw scores: average score and top-1 score. These metrics aim to answer the following questions: How well do a model and algorithm perform across all task pairs, and, how well do a model and algorithm perform supposing that we knew the best source task a priori.

We calculate an average score across all source-target task pairs to understand how each model and algorithm performs in the aggregate. Formally, let the score for a single task be computed as:

$$score(s, t, f, \mathcal{A}) = \frac{1}{|M_t|} \sum_{i=1}^{|M_t|} M_{t,i}(f_{\mathcal{A}}(s, t))$$

where M_t is the set of metrics associated with task t, found in Table 1, and $M_{t,i}(f)$ is the ith calculated metric of model f on task t. *All metrics range from 0 to 100.* Then, we calculate the average score as:

$$\text{Average Score}(f, \mathcal{A}) = \frac{\sum_{t \in \mathcal{T}} \sum_{s \neq t \in \mathcal{T}} score(s, t, f, \mathcal{A})}{|\mathcal{T}| \times (|\mathcal{T}| - 1)}$$

where $\mathcal{T}$ is the set of tasks.

Additionally, we calculate top-1 score to understand how models and algorithms perform if the best source task is known ahead of time. This score is calculated as the maximum score over source tasks averaged over target tasks. The top-1 score does not consider scores less than the baseline, which is a model trained directly on the target task. Denote the baseline algorithm by $\mathcal{A}_{\mathcal{B}}$ and the baseline score as $score(s, t, f, \mathcal{A}_{\mathcal{B}})$. Formally, the top-1 score is calculated as:

$$\text{Top-1}(f, \mathcal{A}) = \frac{\sum_{t \in \mathcal{T}} \max_{s \neq t \in \mathcal{T}} \Big(score(s, t, f, \mathcal{A}_{\mathcal{B}}), score(s, t, f, \mathcal{A}) \Big)}{|\mathcal{T}|}$$

Average and Top-1 Δs In addition to raw scores, we also calculate score differences to measure how much a source task benefits a target task. The average Δ describes how much benefit the model saw in the aggregate over all source tasks, while the top-1 Δ considers only the best source. Score Δs are calculated with respect to the baseline score as:

$$\Delta(s, t, f, \mathcal{A}) = score(s, t, f, \mathcal{A}) - score(s, t, f, \mathcal{A}_{\mathcal{B}})$$

and the average Δ is calculated as:

$$\text{Average } \Delta(f, \mathcal{A}) = \frac{\sum_{t \in \mathcal{T}} \sum_{s \neq t \in \mathcal{T}} \Delta(s, t, f, \mathcal{A})}{|\mathcal{T}| \times (|\mathcal{T}| - 1)}$$

Additionally, we calculate the top-1 Δ as the maximum positive score difference over source tasks averaged over target tasks:

$$\text{Top-1 } \Delta(f, \mathcal{A}) = \frac{\sum_{t \in \mathcal{T}} \max_{s \neq t \in \mathcal{T}} \Big(0, \Delta(s, t, f, \mathcal{A}) \Big)}{|\mathcal{T}|}$$

4 Task Transfer Algorithms

In this work, we consider three commonly used task transfer methods: Pre-train/Fine-tune, Multitask, Multitask/Fine-tune. We apply these methods with cross-entropy loss to further optimize pretrained language models on FETA.

Pre-train/Fine-tune Commonly used in NLP today, the pre-train/fine-tune algorithm consists of two stages of training (Pratt et al., 1991). First, the model is trained on the source task $\mathcal{T}_S$, optimizing Eq 1, followed by a separate stage of training on the target task $\mathcal{T}_T$, optimizing Eq 2:

$$\mathcal{L}_S = \underset{(x, y_s) \sim \{X, \mathcal{Y}_S\}}{-\mathbb{E}} \big[\log p(y_s | x) \big] \quad (1)$$

$$\mathcal{L}_T = \underset{(x, y_t) \sim \{X, \mathcal{Y}_T\}}{-\mathbb{E}} \big[\log p(y_t | x) \big] \quad (2)$$

Multitask In this algorithm, there is only a single stage of multitask training (Caruana, 1994). Formally, the training is conducted on both the source and target task by optimizing Eq 3:

$$\mathcal{L}_{S,T} = \underset{(x, y_s, y_t) \sim \{X, \mathcal{Y}_S, \mathcal{Y}_T\}}{-\mathbb{E}} \big[\log p(y_s | x) + \log p(y_t | x) \big] \quad (3)$$

Multitask/Fine-tune This algorithm combines the previous algorithms in two stages. In the first stage, the source and target task are optimized jointly, as in Eq 3. Then, the second stage trains using only the target task, as in Eq 2.

Even though model selection in multitasking is generally done w.r.t. multiple source and target tasks (Caruana, 1994), we modify the setting to validate a model on a single target task at a time. This allows hyperparameter search and early stopping to be controlled by the desired target task.

Model	Transfer Algorithm	DailyDialog Average Score (σ)	DailyDialog Average Δ	DailyDialog Top-1 Source Score	DailyDialog Top-1 Source Δ	Friends Average Score (σ)	Friends Average Δ	Friends Top-1 Source Score	Friends Top-1 Source Δ
BERT	Pre-train/Fine-tune	50.61 (0.24)	-0.93	52.22	+0.68	42.39 (0.30)	-0.89	44.36	+1.08
	Multitask	50.95 (0.24)	-0.59	52.40	+0.86	42.88 (0.29)	-0.40	45.14	+1.86
	Multitask/Fine-tune	**51.40** (0.25)	-0.15	52.76	+1.22	**44.69** (0.28)	**+1.41**	**46.00**	**+2.72**
GPT-2	Pre-train/Fine-tune	39.80 (0.25)	-1.28	42.19	+1.11	32.66 (0.18)	-0.64	34.34	+1.04
	Multitask	40.21 (0.24)	-0.86	41.77	+0.69	33.10 (0.16)	-0.20	34.83	+1.53
	Multitask/Fine-tune	41.15 (0.23)	+0.07	42.76	+1.68	34.62 (0.15)	+1.32	35.86	+2.56
T5	Pre-train/Fine-tune	49.92 (0.37)	+0.19	**53.04**	**+3.31**	41.73 (0.19)	-1.10	43.52	+0.69
	Multitask	49.49 (0.42)	-0.24	52.98	+3.25	40.42 (0.20)	-2.40	43.33	+0.51
	Multitask/Fine-tune	50.29 (0.36)	**+0.56**	52.85	+3.12	42.29 (0.17)	-0.53	43.87	+1.05

Table 2: **Average and Top-1 Source task transfer scores.** Average scores and Δs aggregate scores over all source tasks, compared with Top-1 scores and Δs which are calculated with scores from the highest performing source task. Δs are the difference from the baseline score without task transfer. Highest values for each model are underlined, highest values across all models are bolded.

5 Experiment Setup

To study task transfer on FETA, we run extensive experimentation. We utilize three task transfer algorithms: pre-train/fine-tune, multitask, and multitask/fine-tune, as described in Section 4. To draw broad conclusions about the performance of each learning algorithm, we utilize pretrained language models with three different architectures: encoder-only (BERT) (Devlin et al., 2019), decoder-only (GPT-2) (Radford et al., 2019), and encoder-decoder (T5) (Raffel et al., 2020). Implementation details, including hyperparameters and prompts, can be found in Appendix B.

A complete experiment for a single target task, $\mathcal{T}$, is as follows: First, we directly fine-tune on $\mathcal{T}$ to get the baseline score. Then, for each source task, $\mathcal{S}$, we take the model pre-trained on $\mathcal{S}$ and fine-tune on $\mathcal{T}$. Next, we jointly train on $\mathcal{S}$ and $\mathcal{T}$ together. Finally, we fine-tune the jointly trained model on $\mathcal{T}$.

FETA datasets have 10 and 7 tasks, giving $90 + 42 = 132$ unique source-target task pairs. Our experiments include three learning algorithms, three models, and we run each experiment with 5 random seeds. In total, we run $132 \times 3 \times 3 \times 5 = 5940$ transfer experiments, and $17 \times 3 \times 5 = 255$ baseline experiments leading to 6195 trained models.

In addition to the single-source setting described above, we also consider a subset of tasks to study in the multi-source setting, where multiple tasks are simultaneously used as source tasks to transfer to a single target task (6.2). For our experiments, we select two target tasks from each dataset that benefit the most from task transfer, and we use the three source tasks that transferred best onto those targets.

6 Results and Analysis

6.1 Single-Source Setting

Table 2 shows the results for all three models and algorithms, and we use this table to understand general trends. Figure 3 shows the relative improvement of a source task for each target task, demonstrating trends across tasks.

Aggregate Performance We find that, on average, Friends tasks get scores between 7-8 points less than DailyDialog, likely due to the greater number of speakers and utterance length of Friends. We find that GPT-2 lags behind the raw scores of BERT and T5 by ~10 points. This is expected as autoregressive decoder models are not designed with classification in mind. We find that the largest average Δ is 1.4, leaving room for improvement in task transfer on FETA.

Furthermore, we are interested in knowing: how much we would gain by using the best source task vs. a random source task. We calculate the differences between average Δ and top-1 Δ and find the mean difference to be ~1.6 and the largest difference to be ~3.5, motivating a further understanding of which source tasks transfer best to target tasks.

Performance Across Learning Algorithms We average scores across both datasets and find that pre-train/fine-tune gets an average score of 42.85, multitask 42.84, and multitask/fine-tune 44.07. Table 2 shows that multitask/fine-tune achieves the best average score for all models and datasets, and indeed its average score is a 2.8% improvement over the other algorithms. However, aggregate scores obscure some interesting nuances.

Do Trends Vary Across Models? Previous studies on task transfer have focused on a single model

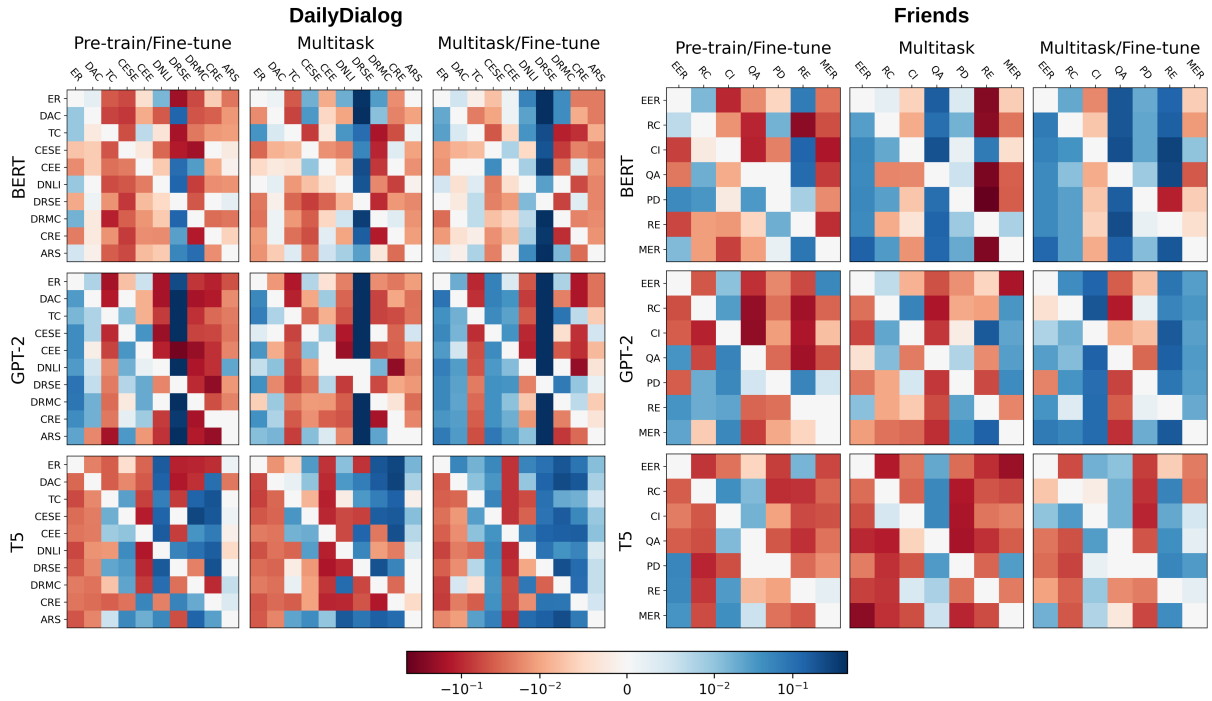

Figure 3: **Relative improvement of transfer over fine-tuned baselines**. Rows are source tasks and columns are target tasks. Diagonal cells are baseline scores. Looking at an individual column can demonstrate best source tasks for that target. Looking at rows can determine which source task works well across multiple targets.

(Pruksachatkun et al., 2020; Lourie et al., 2021; Aribandi et al., 2022), but we find that trends vary depending on the model. For example, we find results similar to Lourie et al. (2021), namely, that fine-tuning on the target task always benefits the T5 model. However, we discover that this does not hold for BERT and GPT-2, which achieve better scores from multitasking than pre-train/fine-tune.

Furthermore, Figure 3 shows that trends on individual tasks also vary depending on the model. For example, T5 positively transferred knowledge to question answering with all learning algorithms and from most source tasks, while GPT-2 had a negative transfer from all algorithms and sources.

For *nearly all* dimensions of analysis (e.g., sample sizes, learning algorithm), we find different trends between models. We *strongly suggest that future research be performed on multiple models* before attempting to draw broad conclusions on transfer learning.

Multitask/Fine-tune As Regularization We find that T5's top-1 score and Δ on DailyDialog are highest for pre-train/fine-tune, but the average score and Δ are highest for multitask/fine-tune. To understand why this occurred, we find the bottom-1 scores for T5 on DailyDialog: 46.78, 46.69, and 48.26 for pre-train/fine-tune, multitask, and multitask/fine-tune algorithms, confirming that multitask/fine-tune does achieve the best worst-case performance. Moreover, we find that for all datasets and models, multitask/fine-tune does achieve the best worst-case performance. In fact, for GPT-2 on Friends, utilizing the bottom-1 source tasks still lead to a 0.74% improvement over the baseline.

Do All Task Types Benefit Equally? We find that *span extraction tasks gain the most as target tasks*, shown in Figure 4 to benefit at all source-to-target sample ratios. Multiple choice tasks also stand to gain from task transfer, but we find that only occurs at a 10:1 ratio of source-target samples. This gain is likely due to the high-level language understanding required by both tasks.

Additionally, we find that utterance-level classification tasks decrease in score Δ at increasing source-to-target sample ratios. This is possibly due to models overfitting to specific tasks and a catastrophic forgetting of general skills learned during their large-scale pre-training.

Do All Task Types Give Equal Benefit? We find that *multiple-choice tasks give the greatest benefit as source tasks*, especially when the ratio of source-

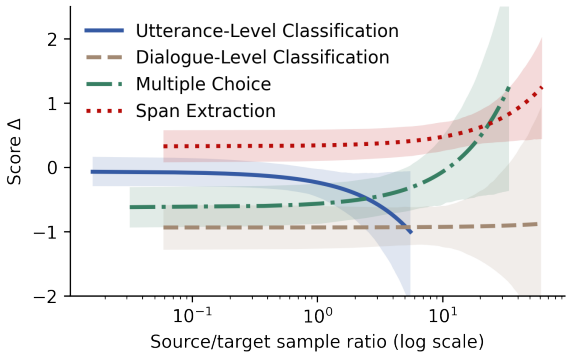

Figure 4: **Score Δ by target task type**. Lines show the average score Δ when the target task is of the specified task type, computed as a best-fit linear interpolation of the data with a 95% confidence interval. The number of samples for an individual task are fixed, but source/target ratios vary depending on which task pair is used.

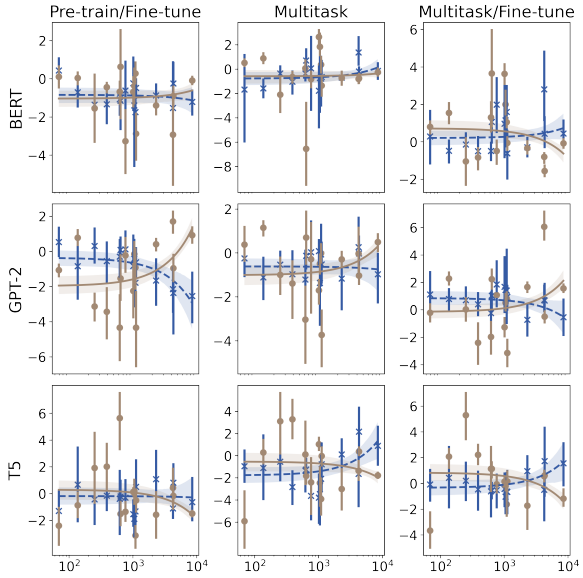

Figure 5: **Score Δ by sample count**. Sample count is on the x-axis (log scale) and score Δ is on the y-axis. The blue dotted line represents the average transfer Δ from a source task to all target tasks. The brown line represents the average transfer Δ to a target task from all sources. Trend lines are a linear best-fit on the data with a 95% confidence interval. The number of samples for an individual task are fixed, but source/target ratios vary depending on which task pair is used.

to-target samples is low, as shown in Figure 9 in the Appendix. Additionally, we find that at a ratio of 10:1 source-target samples, dialogue-level classification benefits downstream tasks, but utterance-level classification requires a ratio of 100:1.

How Do Sample Sizes Affect Transfer? Figure 5 shows that, interestingly, GPT-2 and T5 have opposite trends in relation to sample size. We find that Δs for GPT-2 increase with high target samples and decrease with high source samples. This suggests that GPT-2 may be overfitting to the source task and performs better with resource-rich target tasks. We find that T5 Δs decrease as target-task samples increase, *suggesting that T5 is more sample efficient* than both GPT-2 and BERT.

6.2 Multi-Source Setting

For multi-source transfer we select the two target tasks from each dataset with the best score differences from the single-source setting, shown in Figures 7 and 8 in the Appendix. We find those four tasks to be Dialogue Reasoning Span Extraction (DRSE), Dialogue-Level NLI (DNLI), Character Identification (CI), and Question Answering (QA). For each of these target tasks, we select the top-3 best source tasks, shown in Table 6 of the Appendix . Learning in this setting is similar to single-source, except we now simultaneously optimize the loss for multiple source tasks. Table 3 shows the multi-source results compared with the average score of the top-3 source tasks from the single-source setting. Full results, including score Δs from the single-source baselines, average top-3 score Δs, and multi-source score Δs are in Table 6 of the Appendix.

Does Multi-source Improve Over Single-source? We expect that by utilizing the top-3 source tasks from the single-source setting, the multi-source setting will improve performance for all models and algorithms, but find results to the contrary. We find that 6/9 multi-source algorithms outperform their average top-3 single-source counterparts in DRSE, 6/9 for DNLI, 3/9 for CI, and only 2/9 for QA, showing that naively combining source tasks is not always beneficial. The impressive result for DRSE follows our original intuition, given that there is an almost unanimous benefit from all source tasks, shown in Figure 3. Similarly, we find that *multi-source performance on CI also correlates with the performance of individual source tasks*. We find that in the single-source setting GPT-2 is the only model that improves with any source task, and indeed GPT-2 sees benefits from multi-source training on all algorithms.

Which Models Benefit From Multi-Source? Table 6 shows that GPT-2 improves in 8/12 experi-

	Target	DRSE	DNLI	CI	QA
BERT	P/F	-1.18	+1.37	-2.11	-0.99
	M	+2.77	+1.57	-0.54	-1.14
	M/F	+1.61	+2.28	-0.34	-0.55
GPT-2	P/F	+0.40	+0.16	+4.25	-3.90
	M	+0.78	+0.98	+1.28	-2.46
	M/F	+0.73	-0.09	+0.00	-0.95
T5	P/F	+0.60	+1.95	-0.79	+0.48
	M	-1.08	-0.96	-1.49	+0.08
	M/F	-1.22	-1.20	-0.24	-0.22

Table 3: **Multi-source score Δs from the average score of the top-3 source tasks.** Full results, including score Δs from the fine-tuned baseline are in Table 6.

ments over its average top-3 single-source counterparts, but BERT only 5/12 and T5 in only 4/12 experiments. It is counter-intuitive that T5 should perform the worst as we expect that it has a higher capacity for learning due to twice the model size. On the other hand, the additional parameters may be causing T5 to overfit on training data in the few-sample setting.

7 Conclusion

We introduce FETA, a comprehensive benchmark for evaluating language models and task transfer learning algorithms in open-domain dialogue with few samples. Through extensive experimentation, we find new and non-intuitive insights on the mechanisms of transfer learning. In particular, we find that most trends are model-specific, and we strongly encourage researchers to consider multiple model architectures before attempting to draw broad conclusions on transfer learning. It is our hope that FETA enables further research not only in task transfer, but also in other learning settings, and in the generalizability and efficiency of model architectures and pre-training datasets.

Limitations

A concern regarding any work that includes large-scale experiments with large language models is the energy consumption and environmental impact, the current work included. While there is a cost to running these experiments, the goal of this work is to improve sample efficiency in the future and we hope that the benefits in future energy saved will outweigh the up-front costs of discovering efficient methods.

Another concern of a large-scale benchmark is that of accessibility. A benchmark requiring too many resources will limit those who can reasonably compete. For this reason and others, in addition to our large-scale benchmark we also include a smaller multi-source setting which requires only 4 experiments to be run for a single model and algorithm, rather than 132 in the single-source setting. We believe this smaller setting will maintain the ability to extract high-quality insights on task transfer, yet allow for increased community access and reduce the carbon footprint of this benchmark.

While we do control for domain adaptation in our experiments on task transfer, there are some aspects that we cannot control. For example, each model has done language model pre-training with a different corpus. BERT was trained on English Wikipedia and BookCorpus (Zhu et al., 2015), GPT-2 was trained on a WebText (Radford et al., 2019), and T5 was trained on C4 (Raffel et al., 2020). This difference likely affects model performance on the dialogue tasks in FETA.

Additionally, we cannot exhaustively test every language model, but still try to provide enough variety in order to draw broad conclusions on task transfer. For example, we don't run any experiments on language models pre-trained in the dialogue domain or language models larger than base-sized. We expect that both of these changes would improve raw performance on FETA. More importantly though, it is unclear whether either of these changes would lead to improved task-transfer performance (average and top-1 Δs) and we leave this exploration for future work.

Furthermore, we cannot exhaustively test all learning algorithms. For example, Wang et al. (2020) propose a transfer learning method that minimizes negative task interference via meta-learning for multilingual models, Albalak et al. (2022) propose a policy-guided algorithm for task transfer in low-data settings, and Yu et al. (2020b) propose an optimization algorithm that mitigates gradient interference for reinforcement learning agents.

Finally, we stress the importance of *intra-dataset* task transfer in this work. However, this limits the number of pre-annotated tasks that are available, and there are certainly some tasks which we were not able to accomodate in FETA.

Acknowledgements

The authors would like to thank William Cohen and Tania Bedrax-Weiss for their valuable insights and constructive feedback about this work. This material is based on work that is partially funded by an unrestricted gift from Google. This work

was supported by the National Science Foundation award #2048122. The views expressed are those of the author and do not reflect the official policy or position of the US government. Finally, we thank the Robert N. Noyce Trust for their generous gift to the University of California via the Noyce Initiative.

References

Alon Albalak, Varun Embar, Yi-Lin Tuan, Lise Getoor, and William Yang Wang. 2022. D-REX: Dialogue relation extraction with explanations. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 34–46, Dublin, Ireland. Association for Computational Linguistics.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(61):1817–1853.

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

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *EMNLP*.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79:151–175.

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

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

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

Rich Caruana. 1994. Learning many related tasks at the same time with backpropagation. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA. Association for Computing Machinery.

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

Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. CIDER: Commonsense inference for dialogue explanation and reasoning. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 301–313, Singapore and Online. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38:1 – 32.

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13821–13822.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations,*

ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

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

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *AAAI*.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2039–2048, New Orleans, Louisiana. Association for Computational Linguistics.

S. Mehri, M. Eric, and D. Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *ArXiv*, abs/2009.13570.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Subhabrata (Subho) Mukherjee, Xiaodong Liu, Guoqing Zheng, Saghar Hosseini, Hao Cheng, Greg Yang, Chris Meek, Ahmed H. Awadallah, and Jianfeng Gao. 2021. Clues: Few-shot learning evaluation in natural language understanding. In *NeurIPS 2021*.

Arghya Pal and Vineeth N Balasubramanian. 2019. Zero-shot task transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2198.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

Baolin Peng, Chengkun Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021a. Raddle: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. *ArXiv*, abs/2012.14666.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021b. Soloist: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*.

Lorien Y. Pratt, Jack Mostow, and Candace A. Kamm. 1991. Direct transfer of learned information among neural networks. In *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, AAAI'91, page 584–589. AAAI Press.

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*, pages 5231–5247, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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*, 21(140):1–67.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *ACL*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across nlp tasks. In *EMNLP*.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Gloria Wilcox. 1982. The feeling wheel. *Transactional Analysis Journal*, 12:4:274–276.

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

Zhengzhe Yang and Jinho D. Choi. 2019. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197, Stockholm, Sweden. Association for Computational Linguistics.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. In *EMNLP*.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020a. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020b. Gradient surgery for multi-task learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

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)*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109:43–76.

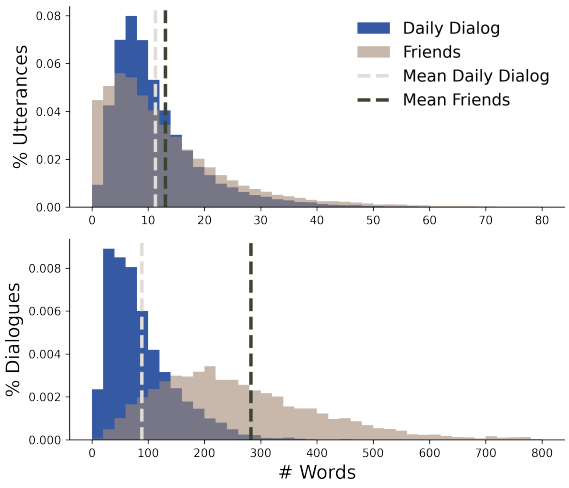

Figure 6: **Utterance and dialogue length distributions in FETA.**

A Dataset Details

A.1 DailyDialog

DailyDialog Along with the dialogues, Li et al. (2017) provide annotations for **emotion recognition**, **dialogue act classification**, and **topic classification**.

RECCON Poria et al. (2021) introduce the task of recognizing emotion causes in conversation and provide annotations for two subtasks: **causal emotion span extraction** and **causal emotion entailment**. Recognizing the cause behind emotions is an important aspect of developing conversational agents that can respond appropriately and these tasks test that ability. Both tasks assume that the emotion of an utterance is already known and require a model to identify the evidence or cause of the given emotion. In causal emotion span extraction, the model is given input as "The target utterance is <U_t>. The evidence utterance is <U_e>. What is the causal span from evidence in the context that is relevant to the target utterance's emotion <E_t>?". On the other hand, if the conversation history up to utterance U_t is $H(U_t)$, then the task of causal emotion entailment is to classify the triple (U_t,U_e,$H(U_t)$) as entailment or not entailment. In this case, entailment means that the emotion expressed in the target utterance, U_t, is caused by the evidence utterance, U_e.

CIDER Ghosal et al. (2021) provide annotations for four tasks designed to explore commonsense inference and reasoning in dialogue: **dialogue-level natural language inference** (DNLI), **dialogue reasoning span extraction**, **dialogue reasoning multiple choice**, and **commonsense relation extraction**. These tasks are created by annotating knowledge triplets on 31 relations that are either explicitly stated in the dialogue or that require commonsense reasoning using contextual information. In DNLI, the task is to determine whether a triplet is true or false given the dialogue. Given a knowledge triplet as <head, relation, tail>, the span extraction task is formulated as identifying the tail when given the head, relation, and dialogue for context. The multiple choice task is motivated by the SWAG commonsense inference task (Zellers et al., 2018), given a head, relation, and conversation as context, the goal is to predict the tail of the relation from 4 possible choices. Finally, commonsense relation extraction is formulated as usual relation extraction tasks; given the head, tail, and conversation as context, the goal is to predict the correct relation out of 31 options.

DailyDialog++ Sai et al. (2020) present the DailyDialog++ dataset, where they aim to improve evaluation of response generation. They do so by collecting five relevant responses and five adversarially crafted irrelevant responses for each dialogue in their dataset, and we recycle their data for a new task called **adversarial response selection**. Adversarial response selection is formulated as a multiple choice selection between a correct response, a randomly selected negative response, and an adversarial negative response.

A.2 Friends

EmoryNLP Chen and Choi (2016) and Zhou and Choi (2018) provide annotations for **character identification**, a subtask of entity linking, where entity mentions in an utterance need to be matched to their correct entity. For this task there are seven possible entities: the six main characters and an "other" entity.

Zahiri and Choi (2018) provide annotations on **emotion recognition**, with the 7 fine-grained emotions from the Feeling Wheel (Wilcox, 1982).

Ma et al. (2018) present annotations for a subtask of **reading comprehension**, called passage completion. In passage completion, given a dialogue and factual statement about the dialogue where character mentions are removed, the task is to fill in the blanks with the correct character from the dialogue. This task is similar to a multiple choice task because entity choices are presented to

the model, but because there are varying number of options in each dialogue, it is formulated as a span extraction that is evaluated based on accuracy.

Yang and Choi (2019) introduce annotations for **question answering**. The answers to question-answer pairs can either be a speaker name or exist as a span within the dialogue, and multiple spans may be correct.

Jiang et al. (2020) present the **personality detection** task by annotating speakers with five traits: agreeableness, conscientiousness, extraversion, openness, and neuroticism. The goal of the task is to correctly identify whether a given character from a dialogue either has or does not have each of the five traits.

DialogRE Yu et al. (2020a) introduce a **relation extraction** dataset annotated with 36 different relations. Their dataset anonymizes speakers which allows for an entity linking relation called "per:alternative_name". However, our version of the Friends dataset is named and so we remove this relation from our data. This task is similar to the relation extraction from DailyDialog, however the relations in DailyDialog are commonsense relations, and the relations in Friends are focused on information about entities.

MELD Poria et al. (2019) provide additional annotations for **emotion recognition**, with only 22.2% dialogue overlap with Zahiri and Choi (2018)'s dialogues. Additionally, while both use 7 total emotions, Poria et al. (2019) use 2 different emotions from Zahiri and Choi (2018).

B Implementation Details

For our experiments, we use the pretrained model implementations from the HuggingFace Transformers library (Wolf et al., 2020), where the bert-base-uncased model has 110M parameters, GPT-2 has 124M parameters, and T5-base has 223M parameters. We use the Adam optimizer (Kingma and Ba, 2015) with a batch size of 60 and run a learning rate sweep across $\{3\times10^{-6}, 1\times10^{-5}, 3\times10^{-5}, 1\times10^{-4}\}$ during the pre-training phase, finding that 3×10^{-5} worked well across all models. In all experiments we utilize validation-based best model selection, and train models for 30 epochs on DailyDialog tasks and 20 epochs on Friends tasks.

C Expanded Single-Source Results

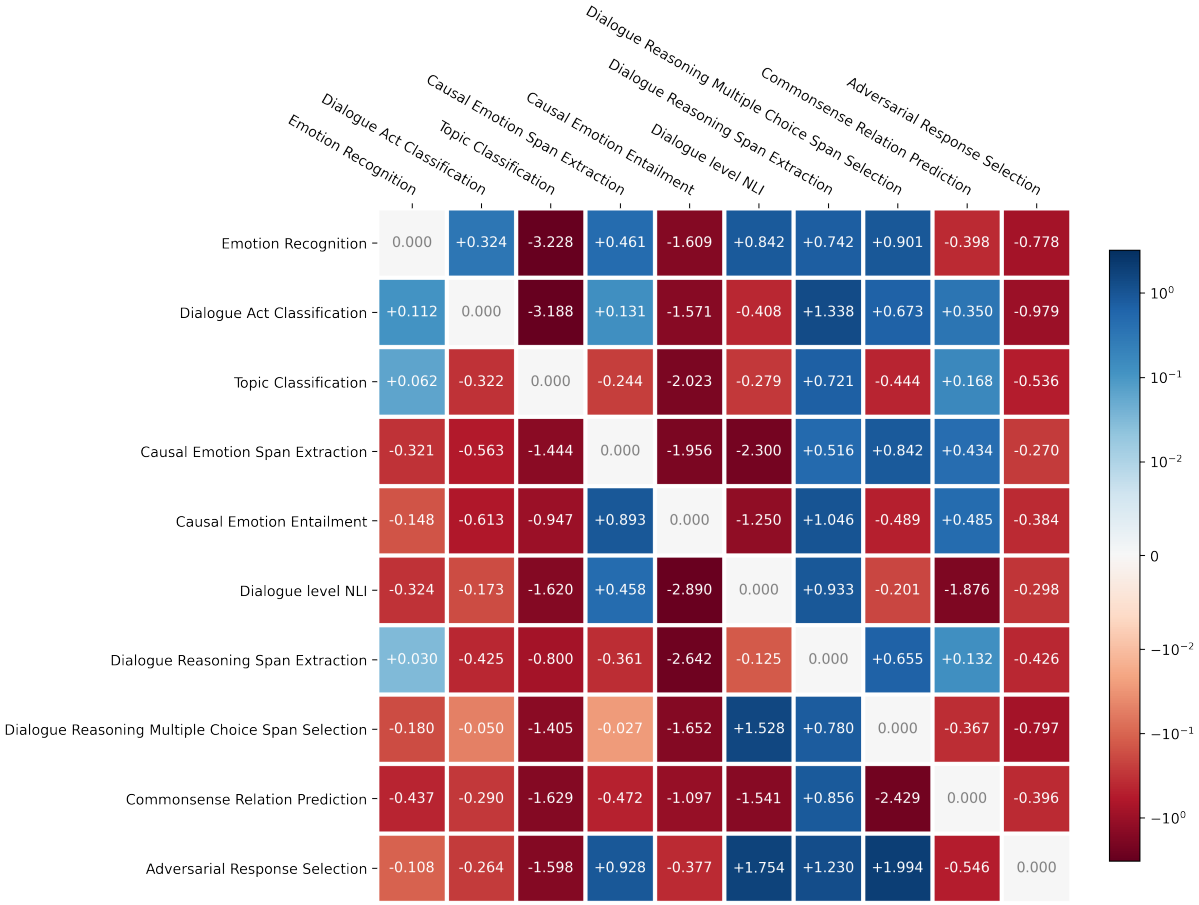

Figure 7: **Aggregate task transfer performance on DailyDialog.**

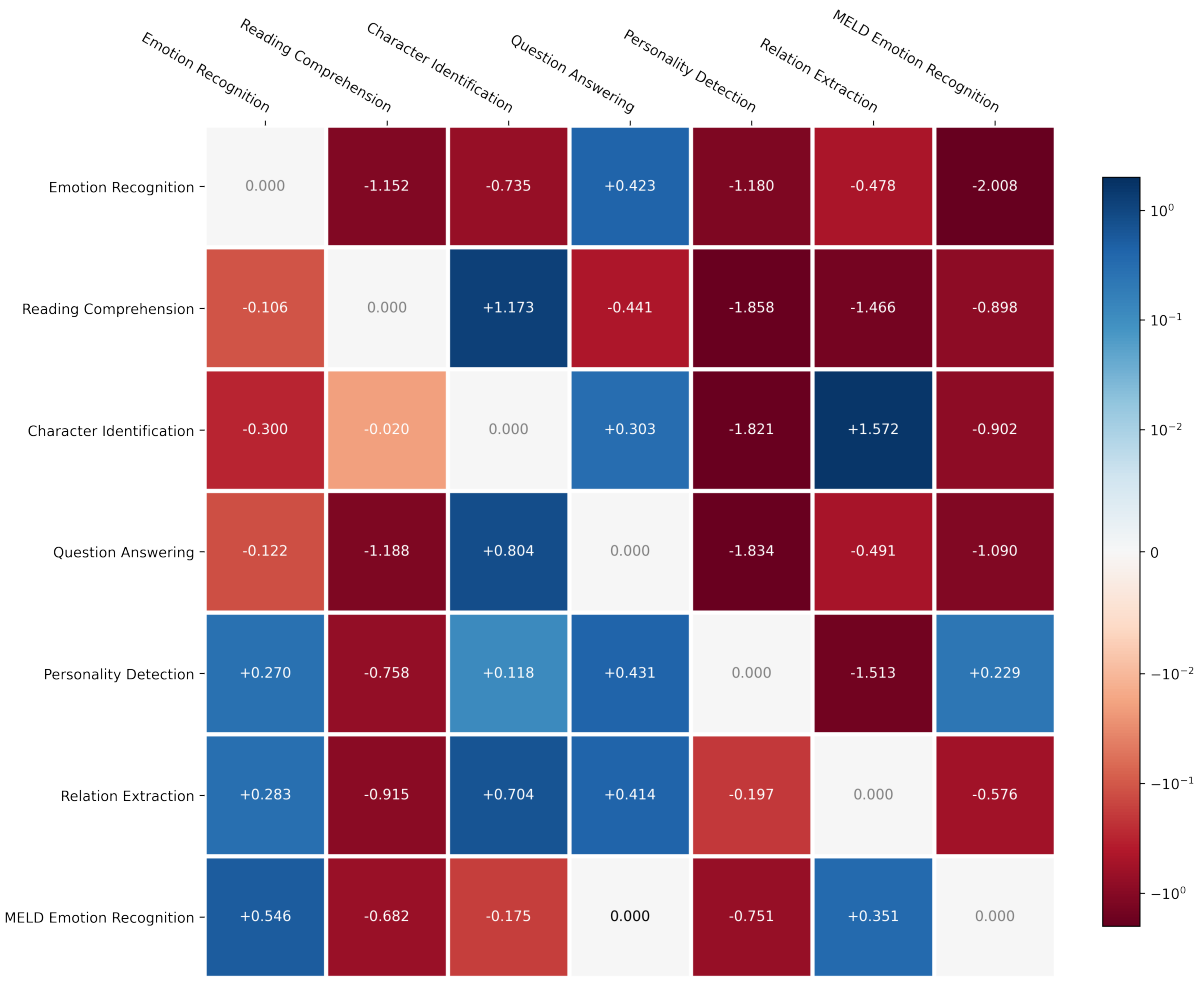

Figure 8: **Aggregate task transfer performance on Friends.**

Task	Prompt
Emotion Recognition	emotion:
Dialogue Act Classification	dialogue act:
Topic Classification	topic:
Causal Emotion Span Extraction	question: <question> answer:
Causal Emotion Entailment	context: <premise> causal emotion entailment: <hypothesis>
Dialogue-level NLI	context: <premise> entailment: <hypothesis>
Dialogue Reasoning Span Extraction	question: <question> answer:
Dialogue Reasoning Multiple Choice	question: <question> <options> The correct option is
Commonsense Relation Extraction	The relation between <head> and <tail> is
Adversarial Response Selection	question: <question> <options> The correct option is

Table 4: **Prompts for FETA-DailyDialog tasks**. All prompts start with "context: <context>", but we leave this out due to repetitiveness and space.

Task	Prompt
Emotion Recognition (Emory)	emotion:
Reading Comprehension	question: <question> out of <entities> [PLACEHOLDER] is
Character Identification	out of <options>, <mention> in the phrase <phrase> refers to
Question Answering	question: <question> answer:
Personality Detection	<entity> is <characteristic>
Relation Extraction	<head> has the following relations with <tail>
Emotion Recognition (MELD)	emotion:

Table 5: **Prompts for FETA-Friends tasks**. All prompts start with "context: <context>", but we leave this out due to repetitiveness and space.

D Expanded Multi-Source Results

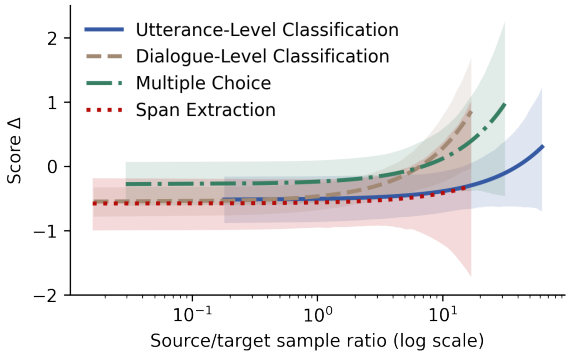

Figure 9: **Score Δ by source task type**. The number of samples for an individual task are fixed, but source/target ratios vary depending on which task pair is used..

Target		DRSE					DNLI					CI					QA				
		DAC	ARS	CEE	Top-3 Av.	Multi-Source	ARS	DRMC	ER	Top-3 Av.	Multi-Source	RC	QA	RE	Top-3 Av.	Multi-Source	PD	ER	RE	Top-3 Av.	Multi-Source
BERT	P/F	0.46	0.17	0.43	0.35	-0.83	-0.35	-0.39	0.88	0.05	1.48	-1.21	-0.76	-1.15	-0.16	-2.27	0.58	-0.28	-0.08	0.07	-0.92
	M	1.86	0.15	0.86	0.96	3.73	0.53	0.32	1.05	0.63	2.20	-0.77	-1.48	-0.27	-0.84	-1.38	1.89	2.98	2.62	2.50	1.36
	M/F	2.58	2.04	1.40	2.01	3.62	2.66	0.40	3.55	2.20	4.48	-0.58	-0.78	-0.48	-0.61	-0.95	3.04	3.64	4.49	3.72	3.17
GPT-2	P/F	0.93	1.14	-0.3	0.59	0.99	-3.65	0.00	-6.99	-3.55	-3.39	1.29	2.73	1.09	1.70	5.95	0.12	-1.76	-0.66	-0.77	-4.67
	M	1.30	1.59	0.89	1.26	2.04	-0.81	-1.73	-0.94	-1.16	-0.18	2.70	-1.03	-0.26	0.47	1.75	-1.59	-1.00	-1.14	-1.24	-3.70
	M/F	3.43	2.01	1.70	2.38	3.11	0.46	-0.32	-1.92	-0.59	-0.68	8.81	6.69	5.08	6.86	8.81	-1.31	-0.84	-0.83	-0.99	-1.94
T5	P/F	-3.08	-1.08	-1.48	-1.88	-1.28	2.52	5.53	8.60	5.55	7.50	2.22	0.70	1.59	1.50	0.71	0.03	-0.19	-0.31	-0.16	0.32
	M	1.54	1.77	2.93	2.08	1.00	8.83	5.83	0.55	5.07	4.11	-1.84	-0.30	0.22	-0.64	-2.13	1.10	0.82	0.27	0.73	0.81
	M/F	3.00	3.30	2.99	3.10	1.88	5.59	4.10	2.78	4.16	2.96	-0.06	1.46	0.52	0.64	0.40	0.02	0.42	-0.63	-0.06	-0.28

Table 6: **Results from the multi-source experiment**, where we use the top-3 source tasks in a multi-source task transfer setting. We include individual scores from all 3 top-3 source tasks and include their average score as a comparison. Multi-source experiments that improve over the top-3 average are underlined.