Improving Data Efficiency for Natural Language Processing

Alon Albalak

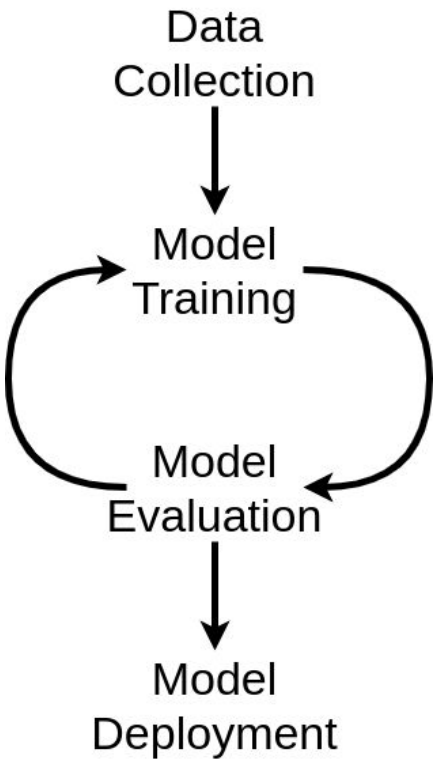
Ph.D. Proposal

02/03/2023

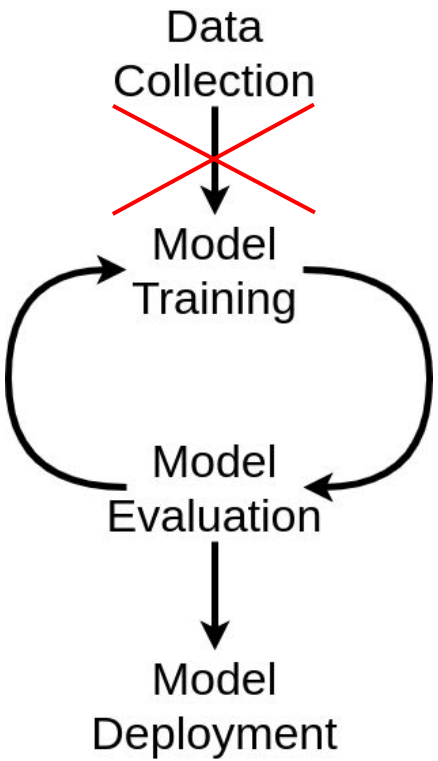
Department of Computer Science

UC SANTA BARBARA

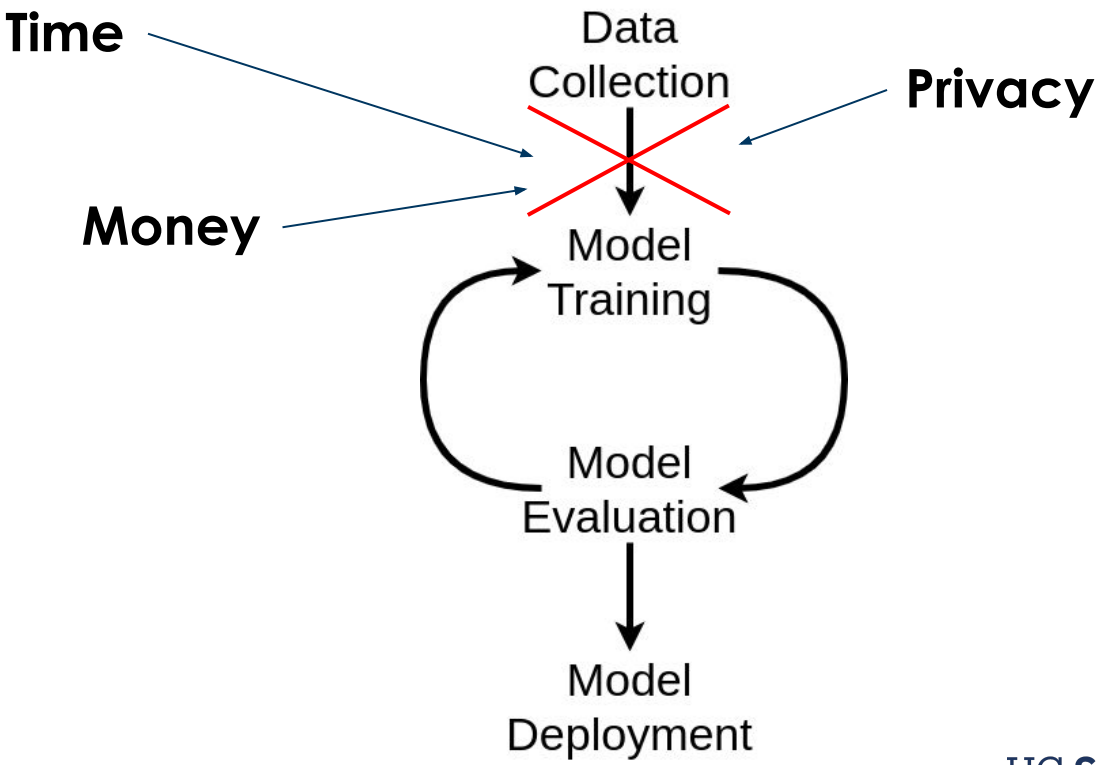
Machine Learning Pipeline



Machine Learning Pipeline



Machine Learning Pipeline



How can we utilize machine learning when DATA is a severe bottleneck?

Transferring External Knowledge

Transferring External Knowledge

Human	Data
• Model inductive bias • Domain knowledge • Relational information • Data augmentation	

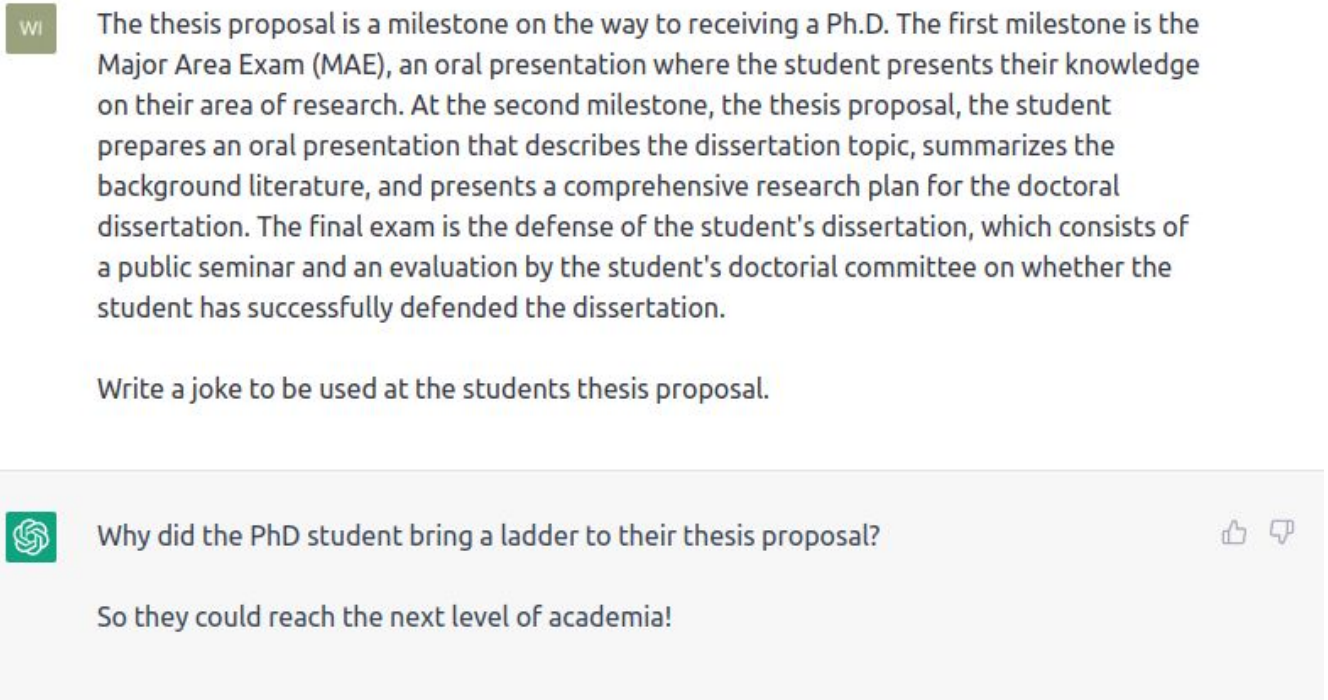
Transferring External Knowledge

Human	Data
• Model inductive bias • Domain knowledge • Relational information • Data augmentation	• Self-supervised domain data • Labeled out-of-domain data • Labeled related task data • Data augmentation

Self-supervised Training on Web Text

Self-supervised Training on Web Text

GPT-3



Self-supervised Training on Web Text

GPT-3

My Research on Data Efficiency

Human	Data
• Albalak et al. 2022 (preprint) • Augustine et al. 2022 (preprint) • Dickens et al. 2022 (TPM workshop) • Pryor et al. 2022 (preprint)	• Albalak et al. 2022 (EMNLP 2022) • Albalak et al. 2022 (ENSLP workshop 2022) • Li et al. 2022 (Alexa Prize 2022) • Albalak et al. 2022 (ConvAI workshop 2022)

My Research on Data Efficiency

Human	Data
• Albalak et al. 2022 (preprint) • Augustine et al. 2022 (preprint) • Dickens et al. 2022 (TPM workshop) • Pryor et al. 2022 (preprint)	• Albalak et al. 2022 (EMNLP 2022) • Albalak et al. 2022 (ENSLP workshop 2022) • Li et al. 2022 (Alexa Prize 2022) • Albalak et al. 2022 (ConvAI workshop 2022)

Zero-shot Transfer Methods

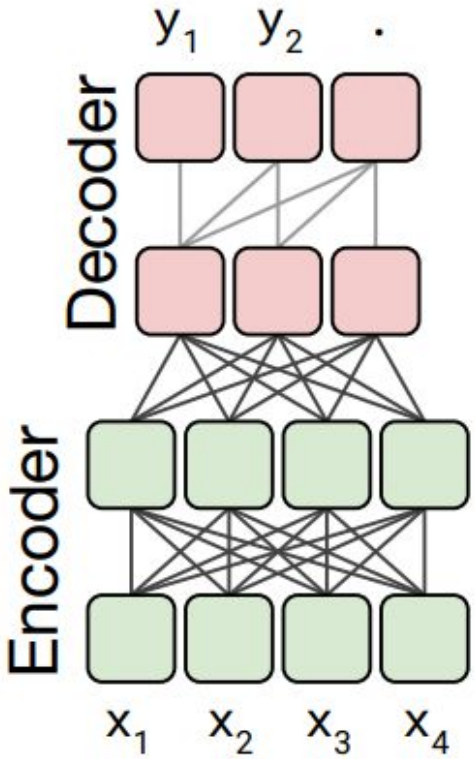
Problem: No target data samples

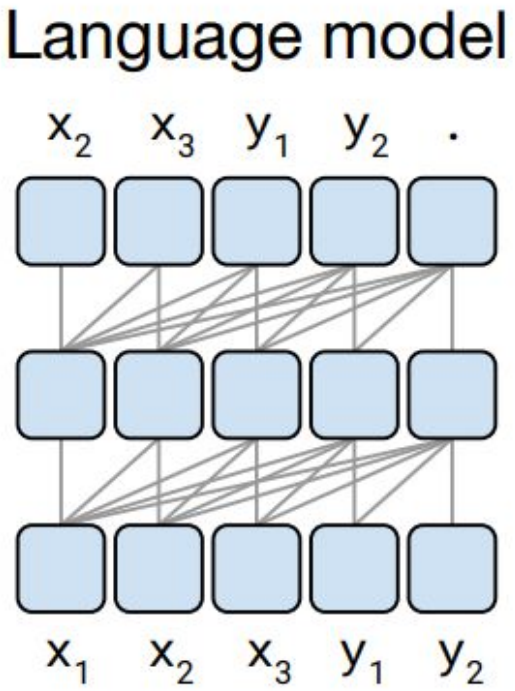
Idea: Convert many tasks into text-to-text format

Allows: 1 model can perform multiple tasks

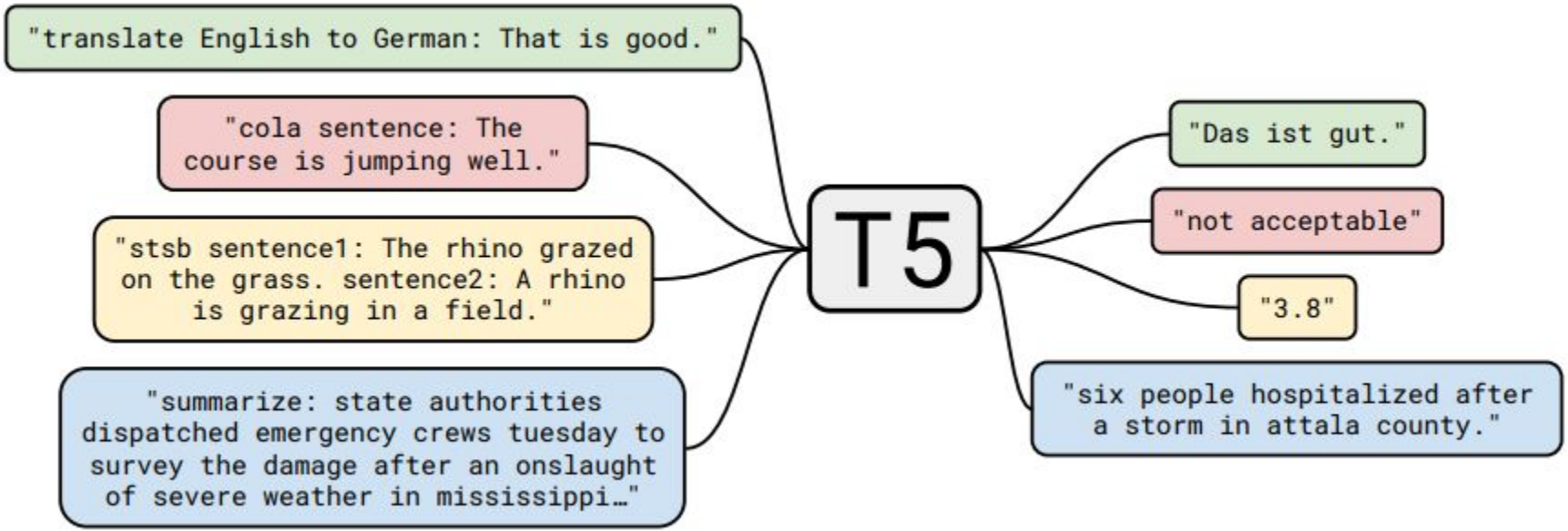
An Exploration of Methods for Zero-shot Transfer in Small Language Models.
Alon Albalak, Akshat Shrivastava, Chinnadhurai Sankar, Adithya Sagar, Mike Ross.
Efficient Natural Language and Speech Processing, 2022.

Generative Models

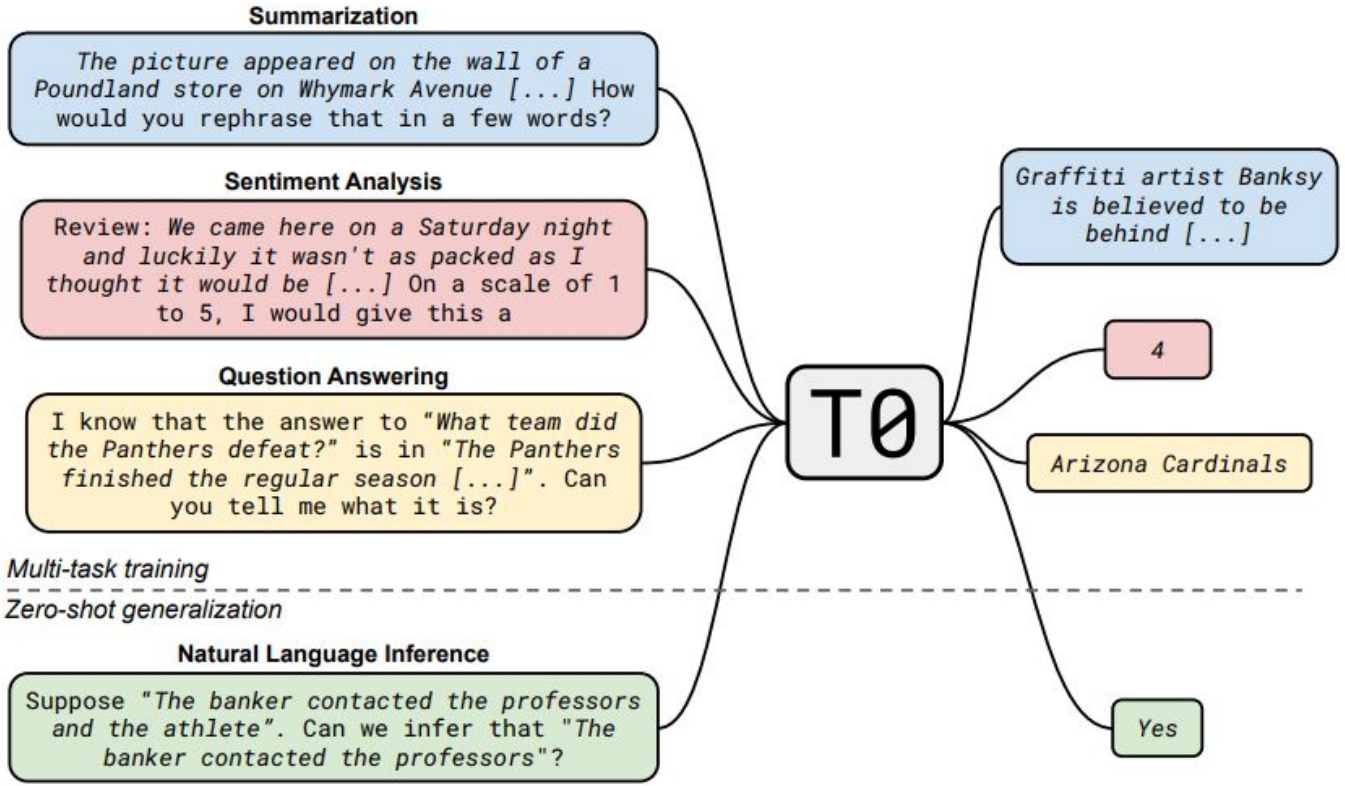




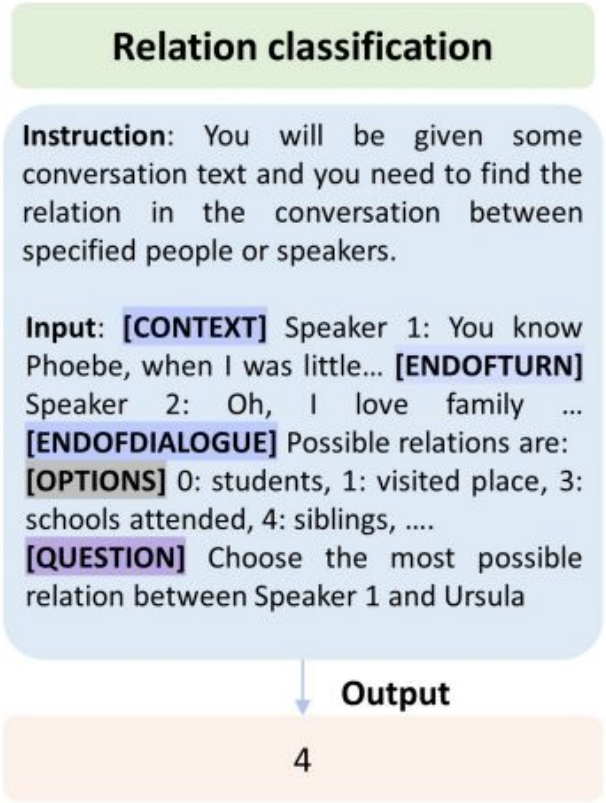
Multi-tasking with Text-to-text Transfer



Multi-tasking for Zero-shot Transfer



Instruction Tuning



Generalization methods have been well studied in large language models.

How do they interact within smaller language models?

Experiment Design

Data

- 46 Tasks from InstructDial[1]
- Tasks have between 3 and 10 instructions
- 3 splits of train/test tasks
  - 40 train tasks
  - 6 test tasks
- Tasks are divided into classification and generation

Models

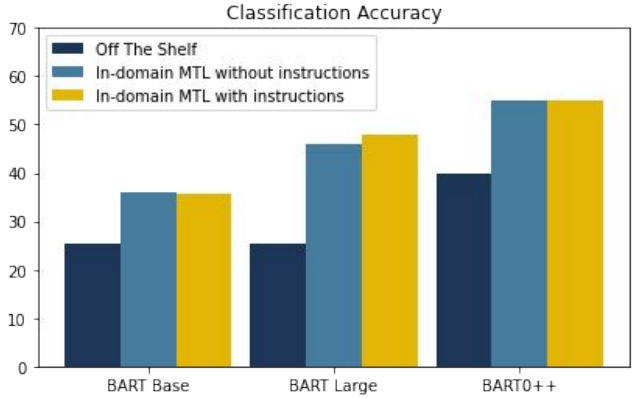
3 variants of BART

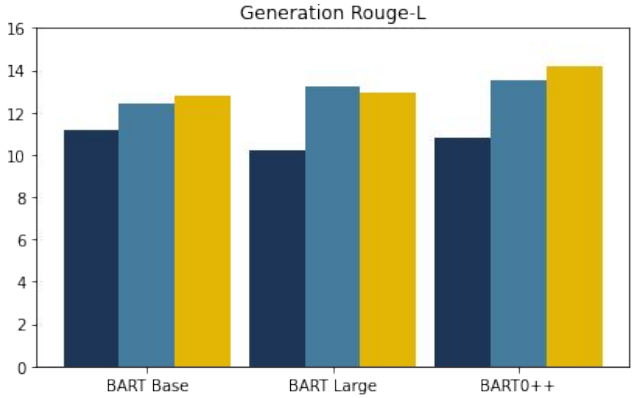
- BART-Base
- BART-Large
- BART0++

[1]Gupta et al. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. 2022

Zero-Shot Results

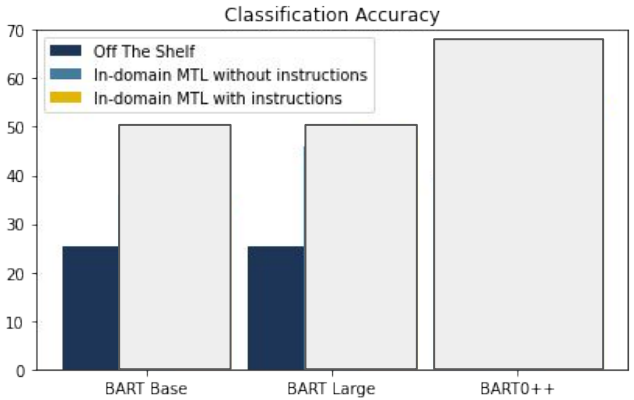
- Model Size
- Multi-Task Learning (MTL)
- In-Domain (Dialogue) MTL
- Instruction Tuning

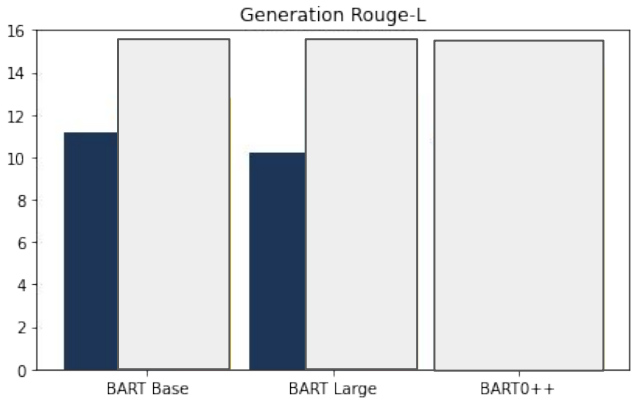




Effects of Model Size

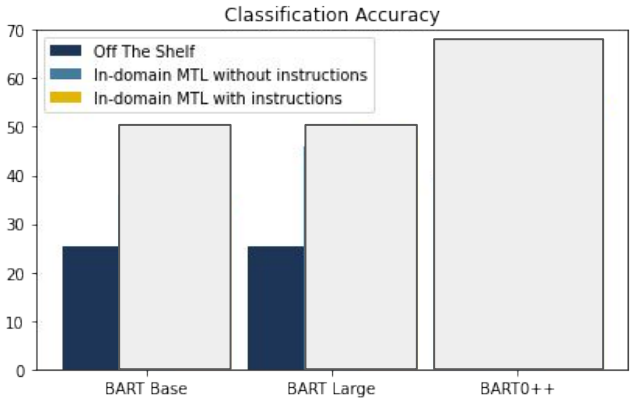
- Nearly identical performance on classification
- Slightly better BART-Base on generation

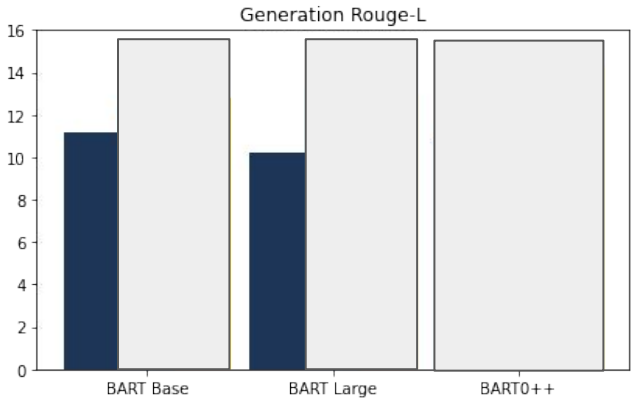




Effects of Model Size

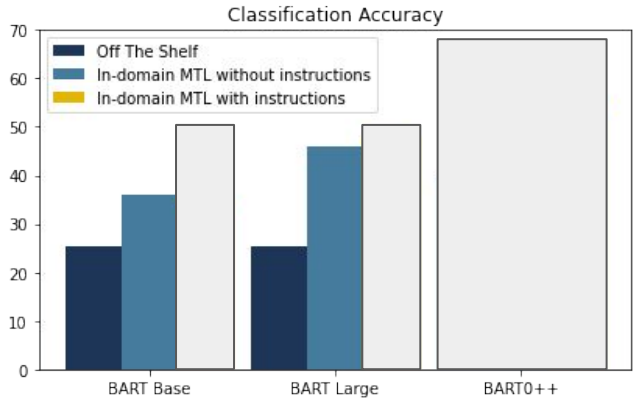
- Nearly identical performance on classification
- Slightly better BART-Base on generation
- ➢ Takeaway: With same data, larger model doesn't necessarily improve performance

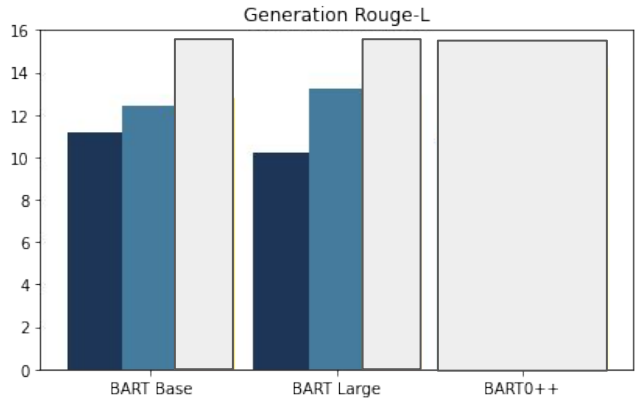




Effects of Model Size + In-domain MTL

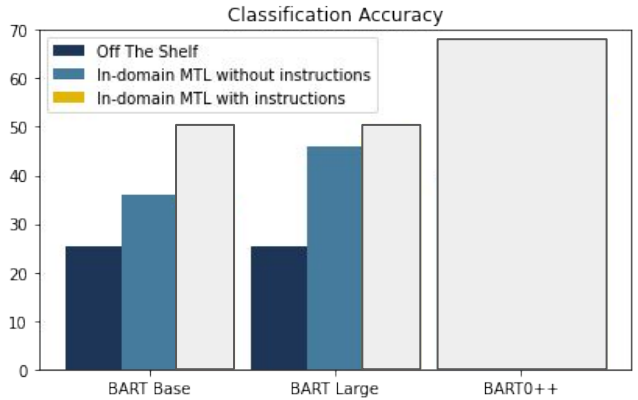
- BART-Base improves by 6.5 average
- BART-Large improves by 13.3 average

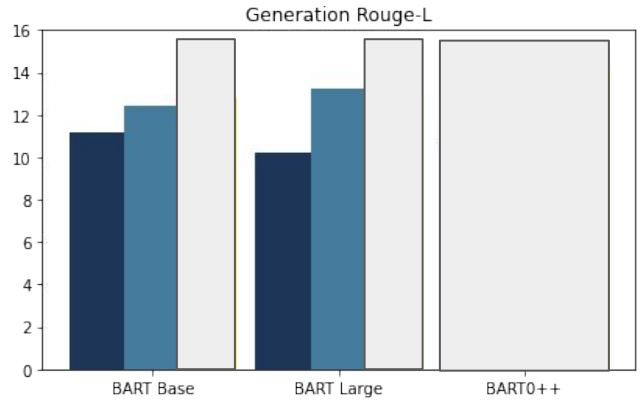




Effects of Model Size + In-domain MTL

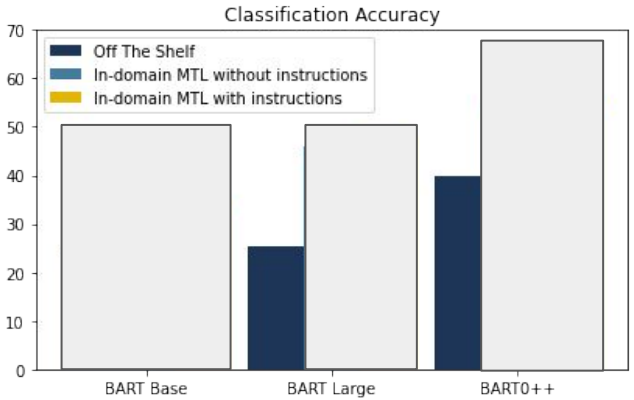
- BART-Base improves by 6.5 average
- BART-Large improves by 13.3 average
- ➢ Takeaway: Increasing model size AND training on in-domain data has better potential

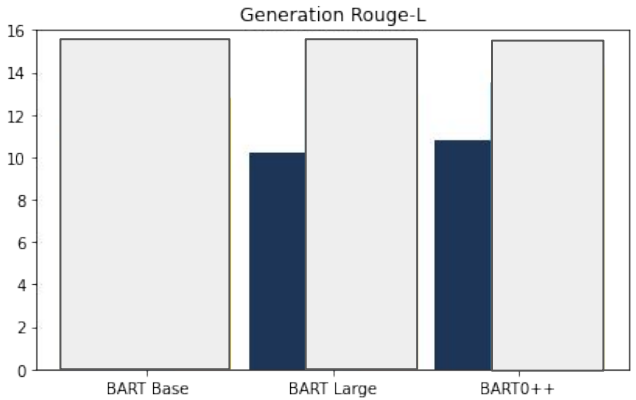




Effect of General Purpose Multi-Task Learning

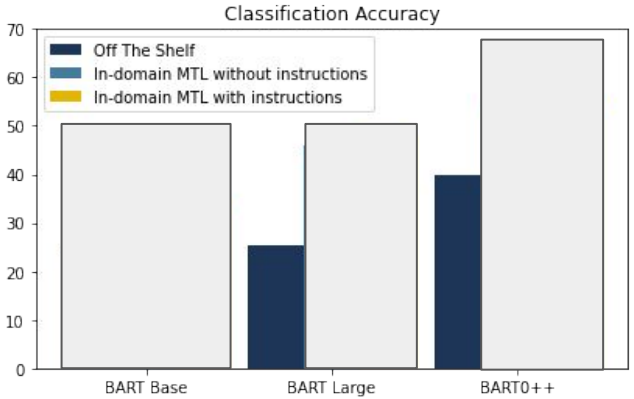
- 14.5 point (57.1% relative) improvement on classification
- 0.6 Rouge-L (5% relative) improvement on generation
- BART0++ MTL tasks are mainly classification

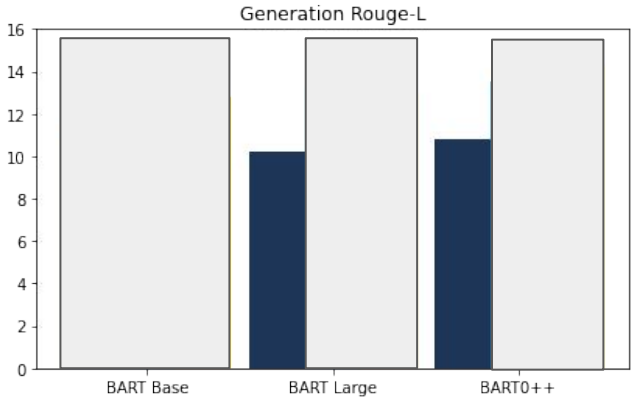




Effect of General Purpose Multi-Task Learning

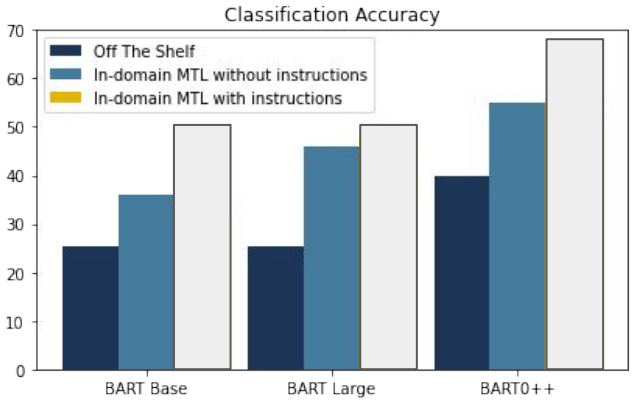
- 14.5 point (57.1% relative) improvement on classification
- 0.6 Rouge-L (5% relative) improvement on generation
- BART0++ MTL tasks are mainly classification
- ➢ Takeaway: General purpose MTL is incredibly beneficial when test tasks are in same distribution as MTL tasks

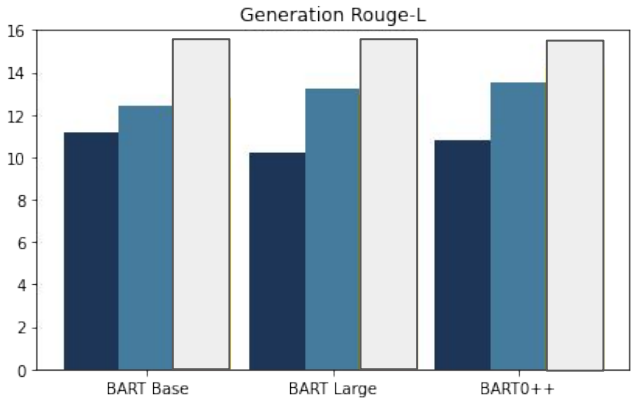




Effects of In-Domain Multi-Task Learning

Relative Improvement	BART-Base	BART-Large	BART0++
Classification	41.8%	**80%**	37.7%
Generation	11.5%	**29.3%**	25.3%

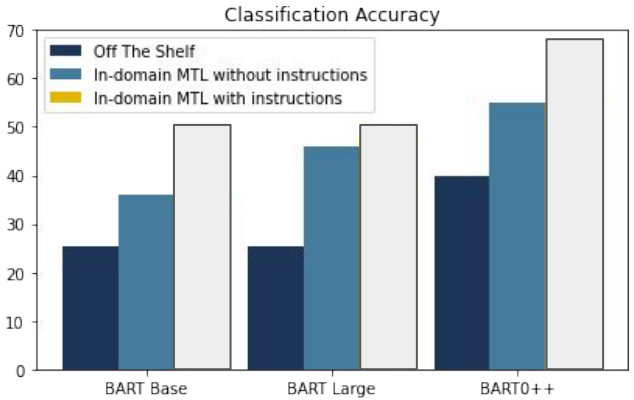


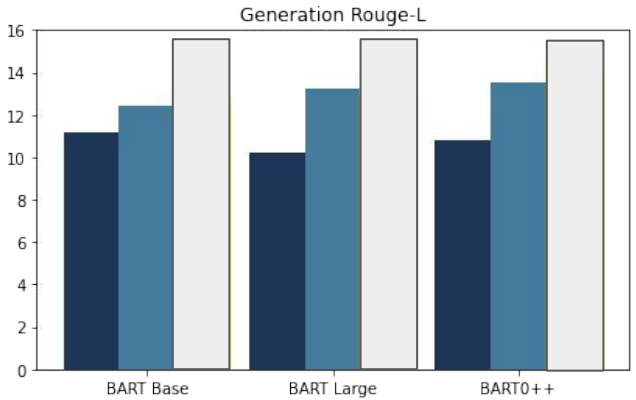


Effects of In-Domain Multi-Task Learning

Relative Improvement	BART-Base	BART-Large	BART0++
Classification	41.8%	**80%**	37.7%
Generation	11.5%	**29.3%**	25.3%

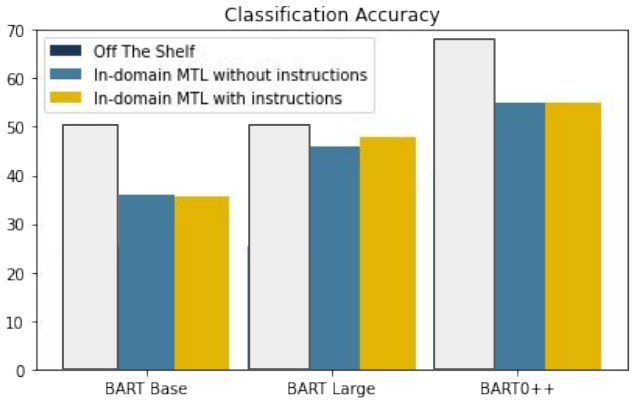
➢ Takeaway: In-domain MTL gives largest portion of generalization improvement

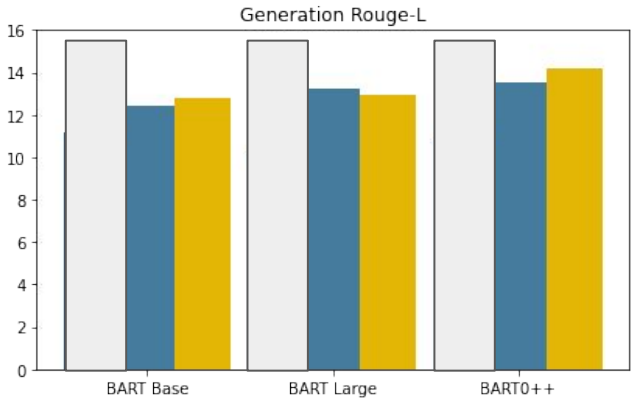




Effects of Instruction Tuning

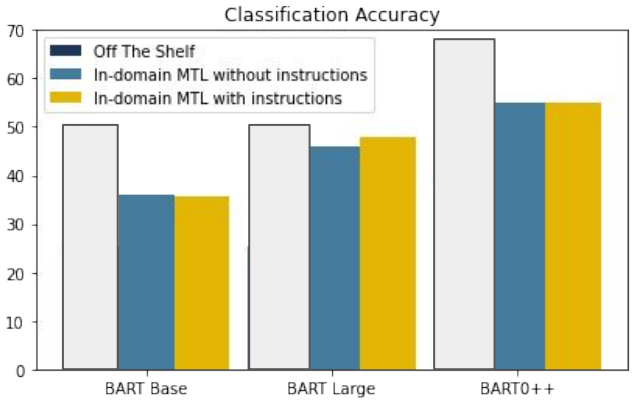
- Much smaller effect than previous variables
- 2% average improvement across models
- This finding runs counter to previous studies which found ~10% decrease in performance for models w/ <8B parameters

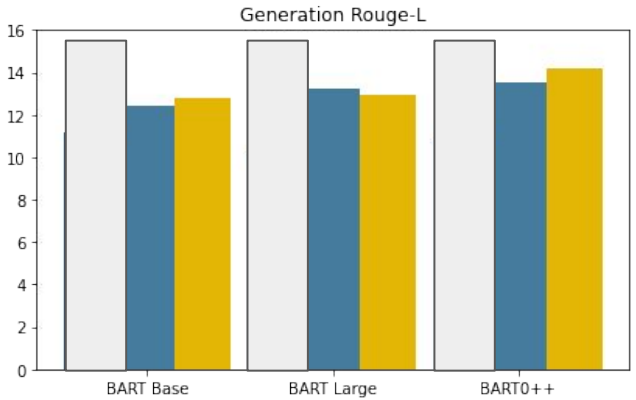




Effects of Instruction Tuning

- Much smaller effect than previous variables
- 2% average improvement across models
- This finding runs counter to previous studies which found ~10% decrease in performance for models w/ <8B parameters
- ➢ Takeaway: Instructions are beneficial, but gains are diminishing





Small models can benefit from multi-task training for zero-shot generalization

If we have a specific target task in mind, can we identify which tasks will transfer well?

Previous Studies on Task Transfer

Name	|Train|	|Dev|	task	metrics	genre/source
CommonsenseQA	9,741	1,221	question answering	acc.	ConceptNet
SciTail	23,596	1,304	natural language inference	acc.	science exams
Cosmos QA	25,588	3,000	question answering	acc.	blogs
SocialIQA	33,410	1,954	question answering	acc.	crowdsourcing
CCG	38,015	5,484	tagging	acc.	Wall Street Journal
HellaSwag	39,905	10,042	sentence completion	acc.	video captions & Wikihow
QA-SRL	44,837	7,895	question answering	F1/EM	Wikipedia
SST-2	67,349	872	sentiment classification	acc.	movie reviews
QAMR	73,561	27,535	question answering	F1/EM	Wikipedia
QQP	363,846	40,430	paraphrase detection	acc./F1	Quora questions
MNLI	392,702	20,000	natural language inference	acc.	fiction, letters, telephone speech
CB	250	57	natural language inference	acc./F1	Wall Street Journal, fiction, dialogue
COPA	400	100	question answering	acc.	blogs, photography encyclopedia
WSC	554	104	coreference resolution	acc.	hand-crafted
RTE	2,490	278	natural language inference	acc.	news, Wikipedia
MultiRC	5,100	953	question answering	$F1_\alpha$/EM	crowd-sourced
WiC	5,428	638	word sense disambiguation	acc.	WordNet, VerbNet, Wiktionary
BoolQ	9,427	3,270	question answering	acc.	Google queries, Wikipedia
CommonsenseQA	9,741	1,221	question answering	acc.	ConceptNet
Cosmos QA	25,588	3,000	question answering	acc.	blogs
ReCoRD	100,730	10,000	question answering	F1/EM	news (CNN, Daily Mail)

Classification

Sentiment Analysis
- Amazon_Polarity (McAuley et al. 2013)
- IMDB (Maas et al. 2011)
- Poem_Sentiment (Sheng et al. 2020) ...

Paraphrase Identification
- Quora Question Paraphrases (Quora)
- MRPC (Dolan et al. 2005)
- PAWS (Zhang et al. 2019) ...

Natural Language Inference
- MNLI (Williams et al. 2018)
- QNLI (Rajpurkar et al. 2016)
- SciTail (Knot et al. 2018) ...

Others (topic, hate speech, ...)

Question Answering

Reading Comprehension
- SQuAD (Rajpurkar et al. 2016)
- QuoRef (Dasigi et al. 2019)
- TweetQA (Xiong et al. 2019) ...

Multiple-Choice QA
- CommonsenseQA (Talmor et al. 2019)
- OpenbookQA (Mihaylov et al. 2018)
- AI2_ARC (Clark et al. 2018) ...

Closed-book QA
- WebQuestions (Berant et al. 2013)
- FreebaseQA (Jiang et al. 2019)
- KILT-NQ (Kwiatkowski et al. 2019) ...

Others (yes/no, long-form QA)

Conditional Generation

Summarization
- Gigaword (Napoles et al. 2012)
- XSum (Narayan et al. 2018) ...

Dialogue
- Empathetic Dialog (Rashkin et al. 2019)
- KILT-Wow (Dinan et al. 2019) ...

Others (text2SQL, table2text ...)

Others

Regression
- Mocha (Chen et al. 2020)
- Yelp Review Full (Yelp Open Dataset) ...

Others
- Acronym Identification
- Sign Language Translation
- Autoregressive Entity Linking
- Motion Recognition
- Pronoun Resolution ...

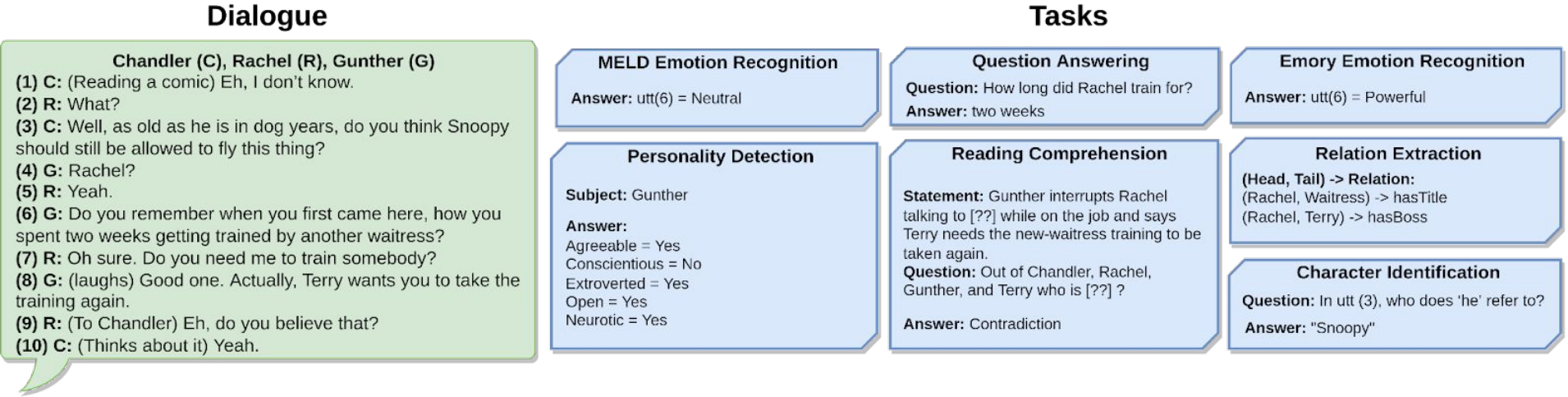
Intra-Dataset Task Transfer With FETA

Intra-Dataset Task Transfer:

Transferring knowledge from a source task to a target task, where both source and target are in the same distribution (domain)

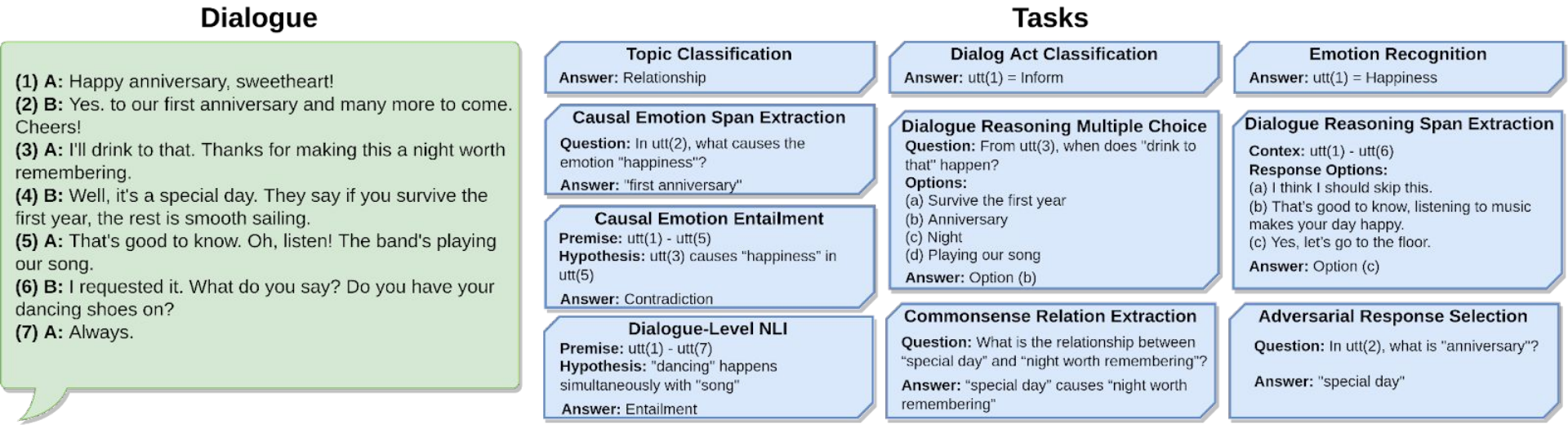
FETA: A Benchmark for Few-Sample Task Transfer in Open-Domain Dialogue
Alon Albalak, Yi-Lin Tuan, Pegah Jandaghi, Connor Pryor, Luke Yoffe, Deepak Ramachandran, Lise Getoor, Jay Pujara, William Yang Wang.
EMNLP 2022

Intra-Dataset Task Transfer With FETA-Friends



FETA: A Benchmark for Few-Sample Task Transfer in Open-Domain Dialogue
Alon Albalak, Yi-Lin Tuan, Pegah Jandaghi, Connor Pryor, Luke Yoffe, Deepak Ramachandran, Lise Getoor, Jay Pujara, William Yang Wang.
EMNLP 2022

Intra-Dataset Task Transfer With FETA-DailyDialog



FETA: A Benchmark for Few-Sample Task Transfer in Open-Domain Dialogue
Alon Albalak, Yi-Lin Tuan, Pegah Jandaghi, Connor Pryor, Luke Yoffe, Deepak Ramachandran, Lise Getoor, Jay Pujara, William Yang Wang.
EMNLP 2022

FETA Learning Setting

- Pairwise task transfer within dataset
- 2 sets with 10 and 7 tasks = 132 source-target pairs
- For each experiment:
  - Source task uses full data
  - Target task uses 10% of data

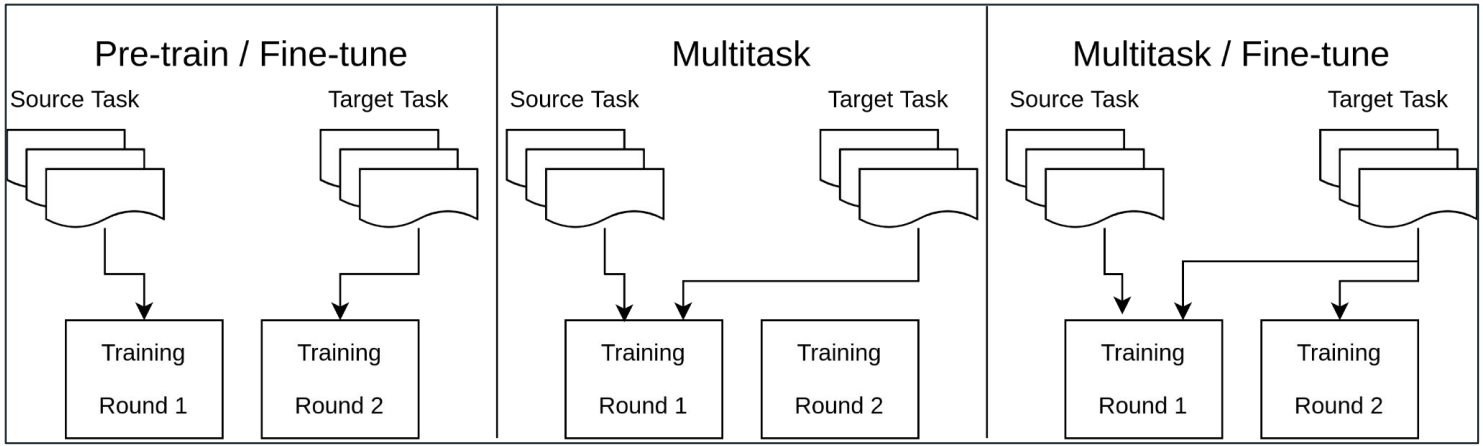
Experiments

3 Model Architectures

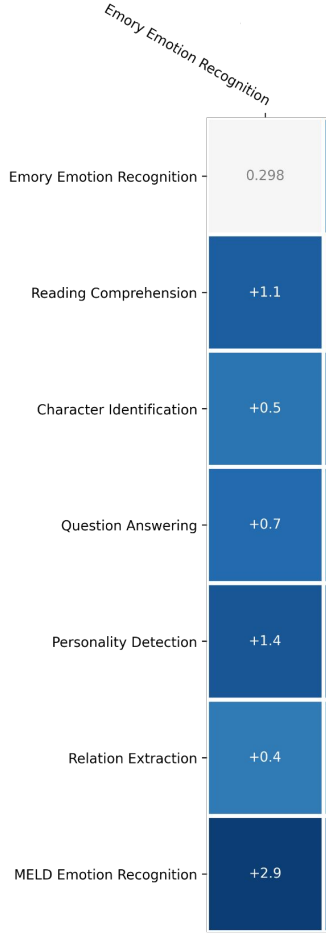
- Encoder - BERT
- Decoder - GPT
- Encoder-Decoder - T5

3 Transfer Algorithms

- Pre-train/Fine-tune
- Multitask
- Multitask/Fine-tune

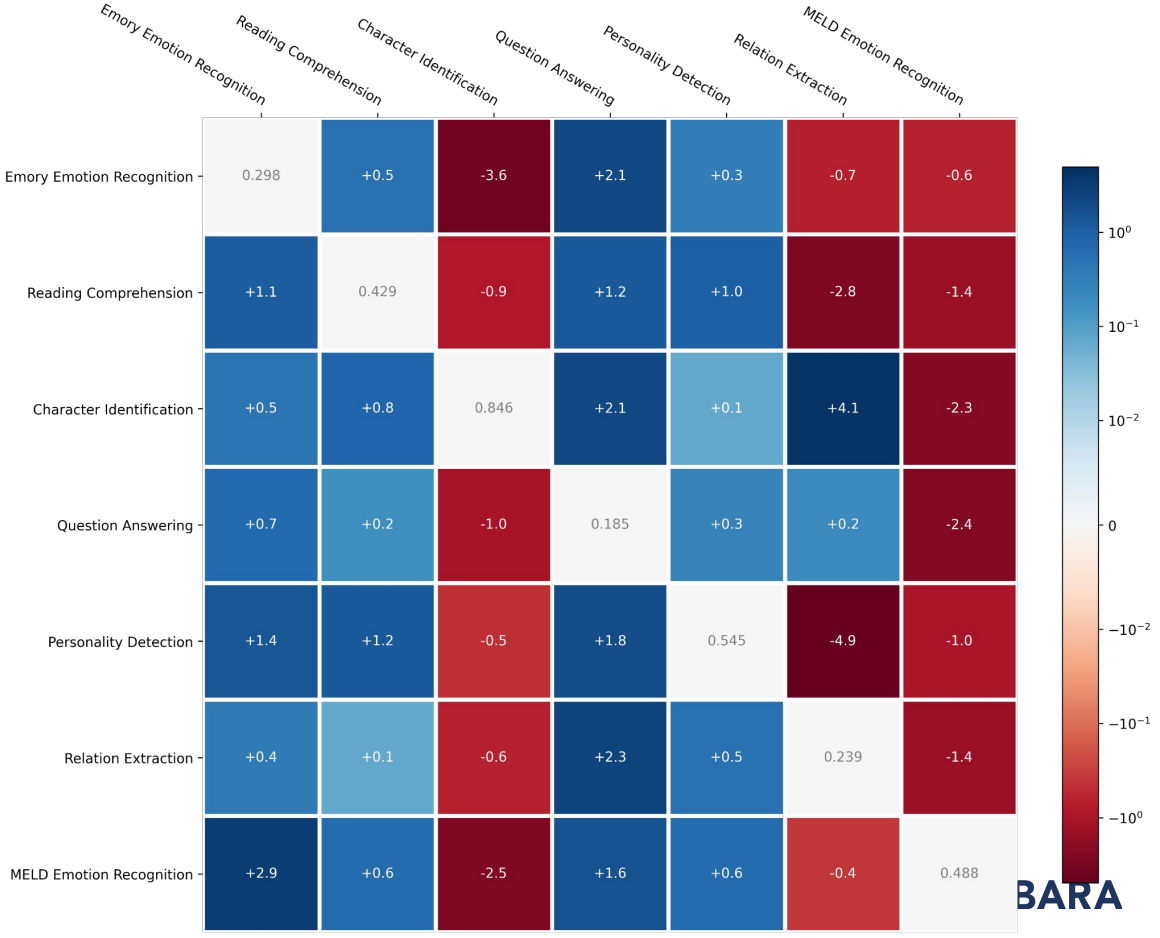


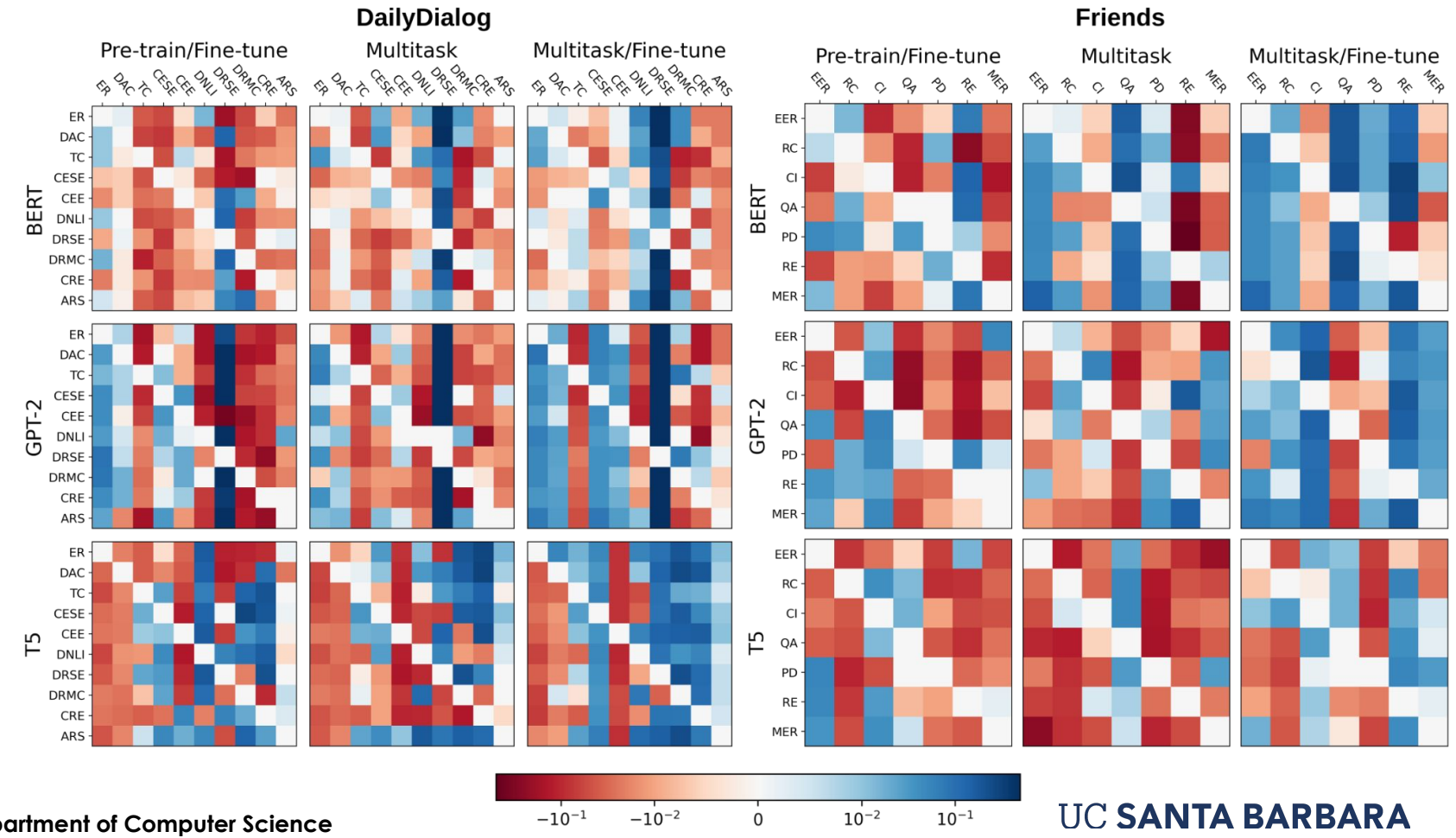
Sample Experiment



Sample Experiment

Sample Experiment




DailyDialog
Friends
Pre-train/Fine-tune
Multitask
Multitask/Fine-tune
BERT
GPT-2
T5
ER
DAC
TC
CESE
CEE
DNLI
DRSE
DRMC
CRE
ARS
EER
RC
CI
QA
PD
RE
MER
-10^{-1}
-10^{-2}
0
10^{-2}
10^{-1}

FETA Takeaways

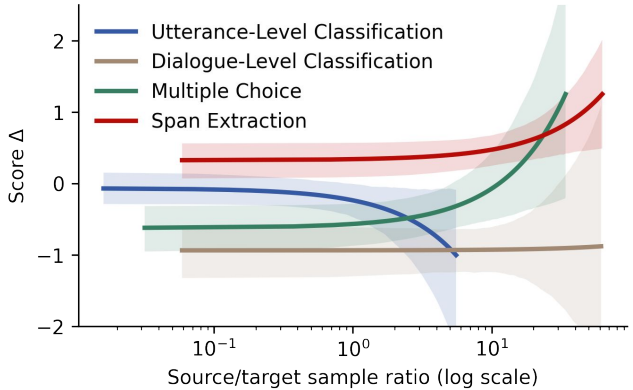
1. Finding the best source task can make a big difference
2. Label-space complexity affects transfer
3. Mitigating negative transfer with multitasking
4. Adding multiple sources can negatively impact transfer

Takeaway - Gap between average and best source

- We find the difference between using the best source task vs. average of all source tasks to be ~1.6 points, with the largest gap being 3.5 points
- **This strongly motivates the need for further understanding which source tasks will transfer best to specific target tasks**

Takeaway - Effect of label-space complexity

- We find that span extraction target tasks have positive transfer from all source task types
- Multiple choice target tasks also see positive transfer, but only when the ratio of source-to-target samples is large (>10-to-1 as shown in Fig. 2 below)
- Both classification tasks see increasingly negative transfer with increasing number of source task samples
- **Overall, as the label-space of a target task becomes more complex, the task benefits more from transfer**



Takeaway - Mitigate negative transfer with Multitask/Fine-tune

- We find that T5's best individual scores are with Pre-train/Fine-tune, but the best average scores are with Multitask/Fine-tune
- In fact, for GPT-2 on FETA-Friends, using the worst source task will still lead to a 0.74% improvement over the baseline
- **We find that across algorithms, Multitask/Fine-tune achieves the best worst-case performance**

Takeaway - Adding source tasks can negatively impact transfer

- Small scale experiment:
  - 4 target tasks with best source tasks
  - Train models using top-3 source tasks
- Results:
  - GPT-2 improves most (8/12 settings)
  - BERT improves on 5/12 settings
  - T5 only improves on 4/12 settings

➢ Takeaway - **Naively adding source tasks can actually *hurt* performance**

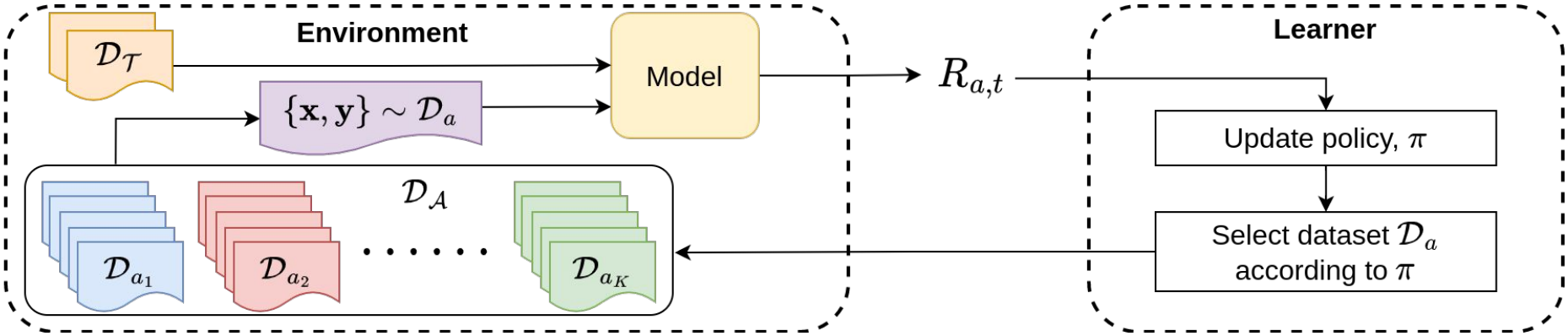
Not all source tasks are equal and adding more source tasks doesn't always improve performance

Moving forward, how can we mitigate negative transfer?

Current work

Few-shot Learning with Auxiliary Datasets - FLAD

Few-shot Learning with Multi-armed Bandits



- Explore (auxiliary dataset) arms to find which gives best reward
- Exploit knowledge of previous rewards

- Use gradient alignment as reward

Future Work

- Data pruning/selection
  - How can we determine which samples to best use, efficiently
- Demonstration selection for in-context learning (ICL)
  - How can we best select demonstrations for ICL based on a given task instance

Questions

@AlbalakAlon
alon_albalak@ucsb.edu
https://alon-albalak.github.io/

Department of Computer Science

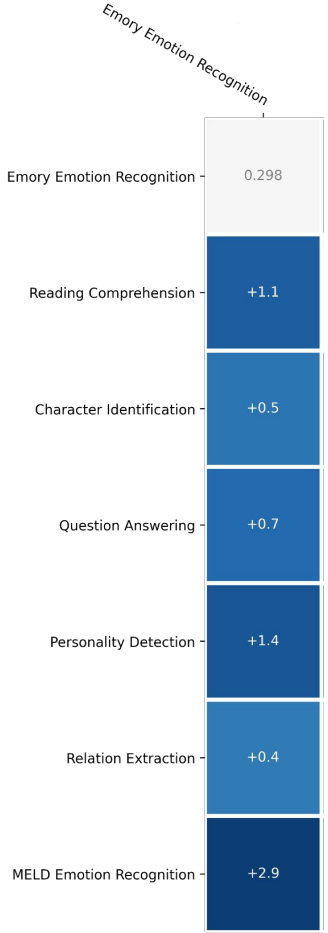
UC SANTA BARBARA

Supplementary Materials

FLAD Preliminary Results

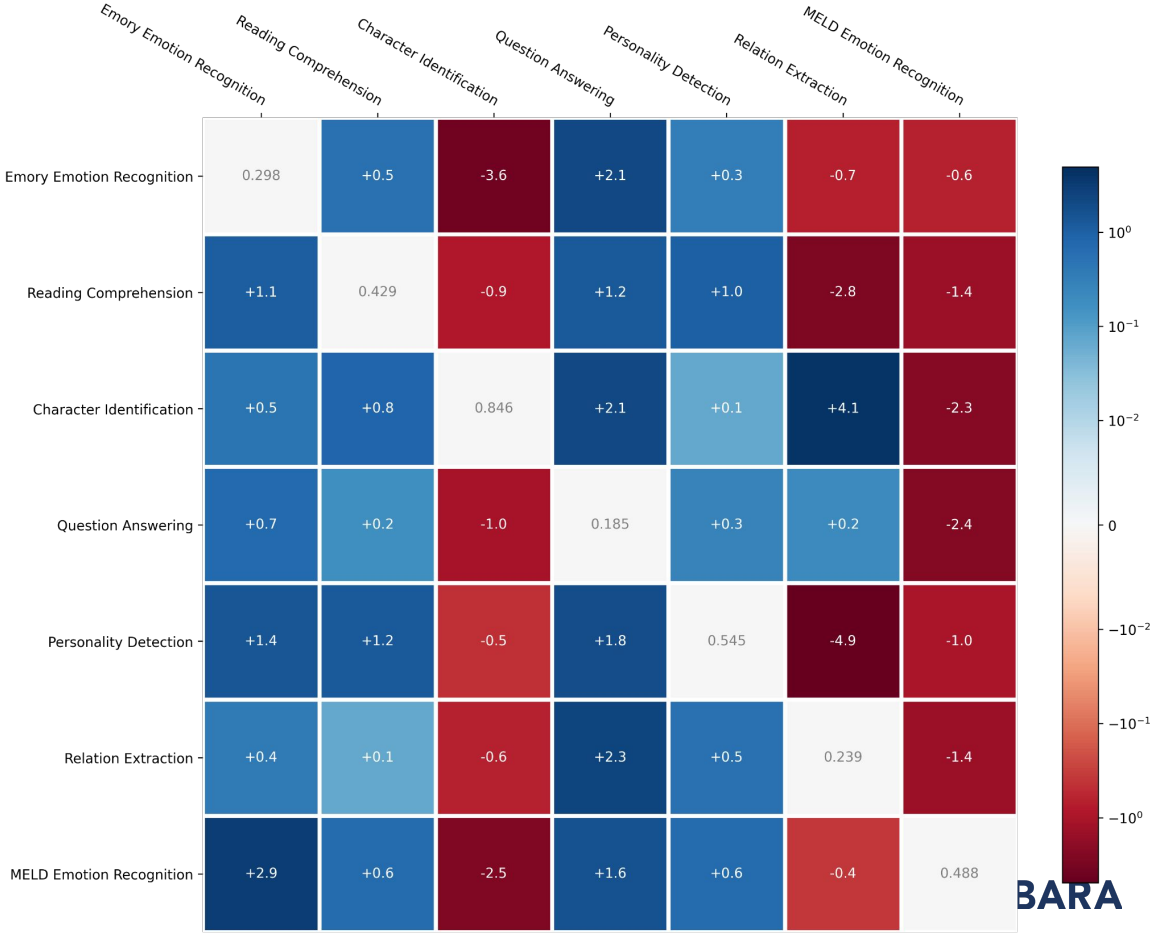
Base Model	T5-XL		T0-3B	
Aux. Data	T0Mix	P3	T0Mix	P3
Target-Only	$52.82_{3.34}$		$56.44_{4.70}$	
Explore-Only	$59.18_{5.52}$	$60.64_{4.92}$	$61.17_{3.30}$	$62.77_{4.83}$
Exploit-Only	$59.79_{5.63}$	$60.49_{5.01}$	$60.87_{3.35}$	$62.87_{3.69}$
EXP3-FLAD	$61.50_{4.25}$	$64.07_{4.81}$	$62.87_{3.85}$	$\underline{65.98}_{3.20}$
UCB1-FLAD	$62.01_{3.89}$	$\underline{65.52}_{3.86}$	$62.89_{3.68}$	$\mathbf{66.29}_{3.29}$

FETA Experiments

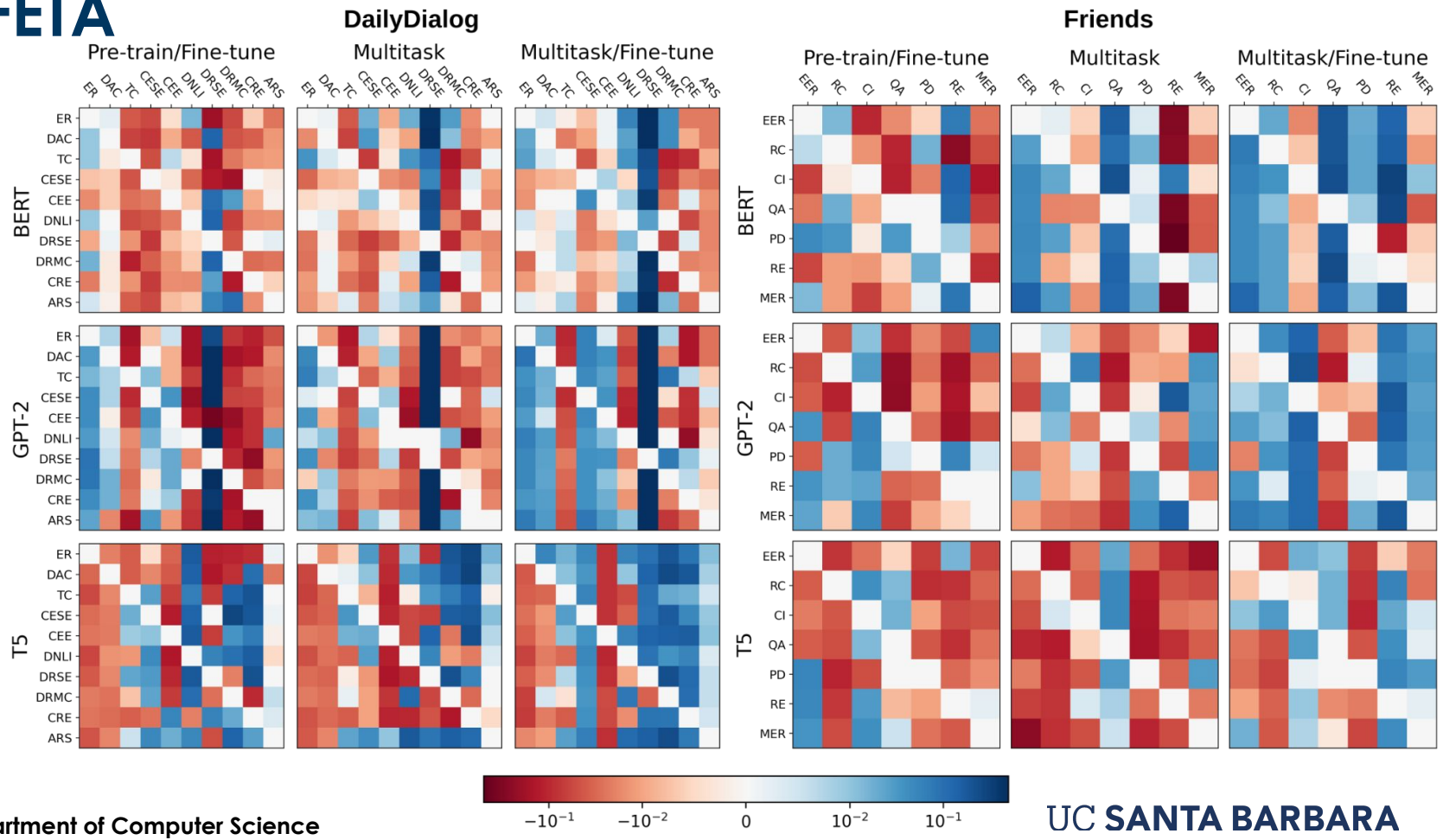


FETA Experiments

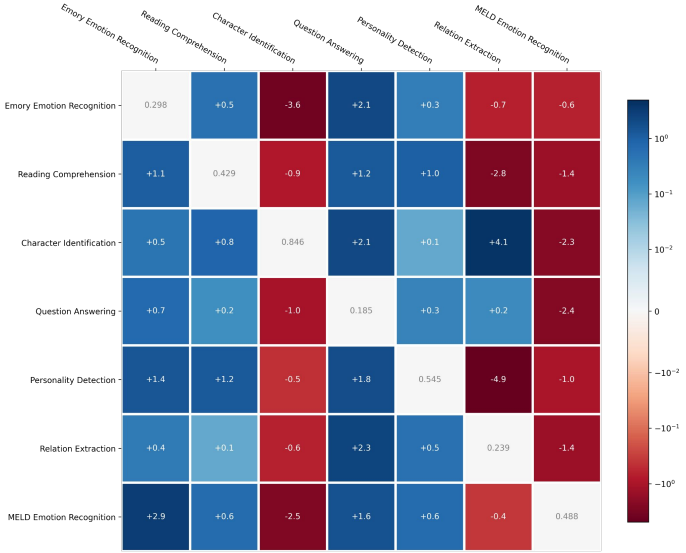
FETA Experiments



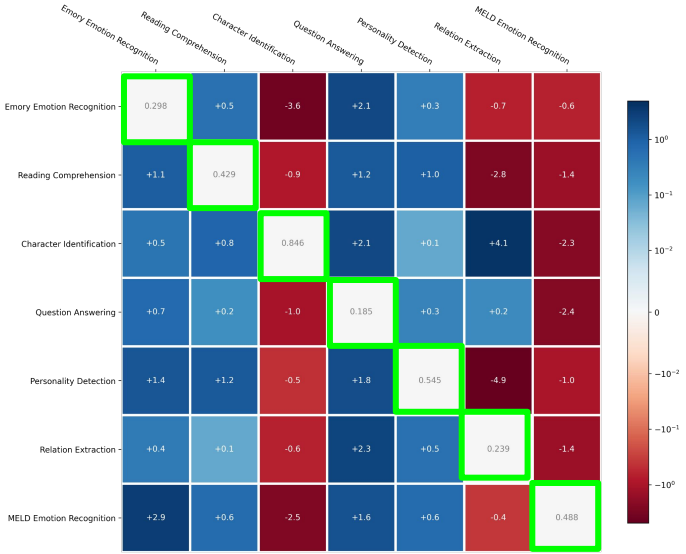
FETA


DailyDialog
Friends
Pre-train/Fine-tune
Multitask
Multitask/Fine-tune
BERT
GPT-2
T5
ER DAC TC CESE CEE DNLI DRSE DRMC CRE ARS
EER RC CI QA PD RE MER
-10^{-1} -10^{-2} 0 10^{-2} 10^{-1}

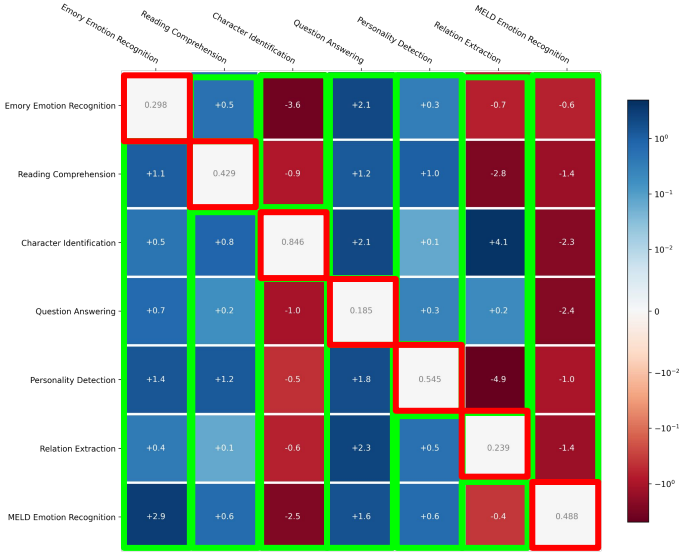
FETA Metrics



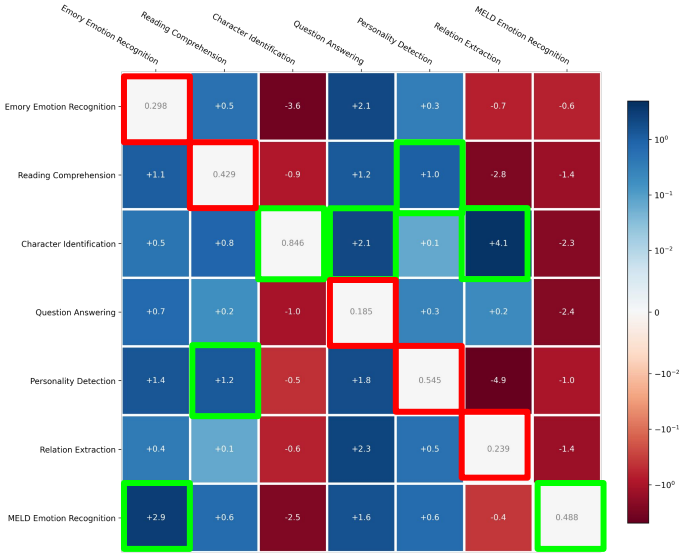
FETA Metrics - Baseline Score



FETA Metrics - Average Transfer Score



FETA Metrics - Top-1 Score



FETA Results - Aggregate Scores

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

Abstract

- Machine learning algorithms are data hungry
- Unlabeled data is easy to gather, but hard to utilize for specific tasks
- Labeled data can be time consuming and expensive to gather, and sometimes impossible due to privacy concerns or the nature of the problem
- For these reasons, few- and zero-shot settings (which require little to no labeled data) are attractive learning paradigms
- In this talk, I discuss methods of improving data efficiency in natural language processing inspired by transfer learning, reinforcement learning, and neuro-symbolic methods.
- In the end, I'll discuss current and future work for my PhD

Machine learning algorithms are data hungry, and although unlabeled data (e.g. web text) is easy to gather, it is difficult to utilize for specific tasks. Furthermore, labeled data can be time consuming and expensive to gather, and sometimes impossible due to privacy concerns or the nature of the problem.

This talk discusses methods of improving data efficiency in natural language processing inspired by transfer learning, reinforcement learning, and neuro-symbolic methods. The focus is on few- and zero-shot settings, which are attractive learning paradigms due to the challenges of gathering labeled data for specific tasks. The talk will also cover current and future work for the my PhD research in this area.

Outline

- ~5 introductory slides
- Outline of my research area
  - What is the broad area
  - What are the specific areas I have focused on
- In depth-ish on 1-2 works
  - Multi-task learning with prompts, instruction tuning, for zero- and few-shot
  - FETA as a good testbed
  - Finish with outcome that more source datasets didn't always help, how can we improve on this?
- Future work - addressing those weaknesses + other areas of interest
  - MAB for FLAD
  -
- Timeline

My research area

How do we improve data efficiency? Utilize additional information

We (roughly) categorize methods by the source of the information (2-sided figure with works on each side)

- Human knowledge (1-2 slides on how this works)
- Data related to our target domain/task

This presentation will focus on the second category

Specifically, we focus on methods for NLP

Zero-shot transfer methods

- Background:
  - How can multiple tasks be handled by one model?
  - Encoder-only + task-specific head (classification tasks only)
  - Generative models + text-to-text format
- Multi-task learning formulation
- Instruction tuning

Benefits have been demonstrated in large and massive language models, what about small models? (slide showing benefit from previous studies on large models)

Follow ENSLP slides to discuss data/models

Transition to FETA: Next, we'll take a closer look at the effects of individual source tasks on the target task

FETA

Transition to current/future work: So, we've seen the effects of negative transfer, not all source tasks are equal, and also that simply adding more source tasks doesn't always improve performance. So, moving forward, how can we mitigate negative transfer?