

# Large-scale Multi-task Pre-training

Zhuosheng Zhang

[zhangzs@sjtu.edu.cn](mailto:zhangzs@sjtu.edu.cn)

<https://bcmi.sjtu.edu.cn/~zhangzs>

# Introduction



## Zhuosheng Zhang

### ■ Education Background

- 2012-2016: B.S in Wuhan University
- 2016-2020: M.S. in Shanghai Jiao Tong University
- 2020-2023: Ph.D. in Shanghai Jiao Tong University, Advisor: Prof. Hai Zhao
- 2022: Internship at MSR from Feb. 2022 to Now, Mentor: Shuohang Wang

### ■ Research Interests

- Natural Language Processing
- Pre-trained Language Models
- Machine Reading Comprehension

# Overview: Large-scale Multi-task Pre-training

---

## ❑ Goals

- Bridge self-supervised pre-training with task requirements by leveraging **large-scale supervised tasks**
- Use a unified model to solve a wide range of tasks

## ❑ Benchmark Tasks

- Commonsense Reasoning (Rainbow)
- Legal Language Understanding (LexGLUE)

## ❑ Scientific Questions

- How to capture task relationships in large-scale multi-task pre-training

## ❑ Contributions

- A unified **encoder-only** multi-task pre-trained language model trained on 40 tasks
- A probing tool of using task prefix to explore the **task relationships** in large-scale MTL
- **Human-parity performance** on commonsense reasoning leaderboards.

# Benchmark Tasks

❑ **Rainbow:** develop models that use **commonsense knowledge** to answer multiple-choice questions.

Dataset	Goal
ANLI	Abductive reasoning in narratives. It asks models to identify the best explanation among several connecting a beginning and ending
COSMOSQA	asks commonsense reading comprehension questions about everyday narratives
HELLASWAG	requires models to choose the most plausible ending to a short context
PIQA	a multiple-choice question answering benchmark for physical commonsense reasoning
SOCIALIQA	evaluates commonsense reasoning about social situations and interactions.
WINOGRANDE	a large-scale collection of Winograd schema-inspired problems requiring reasoning about both social and physical interactions.

goal (string)	sol1 (string)	sol2 (string)	label (class label)
When boiling butter, when it's ready, you can	Pour it onto a plate	Pour it into a jar	1 (1)
To permanently attach metal legs to a chair, you can	Weld the metal together to get it to stay firmly in place	Nail the metal together to get it to stay firmly in place	0 (0)
how do you indent something?	leave a space before starting the writing	press the spacebar	0 (0)
how do you shake something?	move it up and down and side to side quickly.	stir it very quickly.	0 (0)
Clean tires	Pour water, cape off caked on dirt. Use speed wool to clean out crevices and sparrow spaces.	Pour water, scrape off caked on dirt. Use a steel wool to clean out crevices and narrow...	1 (1)
how do you taste something?	smell it enough to taste it.	place it in your mouth to taste.	1 (1)

# Benchmark Tasks

❑ **LexGLUE**: a benchmark dataset for **legal language understanding** in English

Dataset	Source	Sub-domain	Task Type	Training/Dev/Test Instances	Classes
ECtHR (Task A)	<a href="#">Chalkidis et al. (2019a)</a>	ECHR	Multi-label classification	9,000/1,000/1,000	10+1
ECtHR (Task B)	<a href="#">Chalkidis et al. (2021c)</a>	ECHR	Multi-label classification	9,000/1,000/1,000	10+1
SCOTUS	<a href="#">Spaeth et al. (2020)</a>	US Law	Multi-class classification	5,000/1,400/1,400	14
EUR-LEX	<a href="#">Chalkidis et al. (2021a)</a>	EU Law	Multi-label classification	55,000/5,000/5,000	100
LEDGAR	<a href="#">Tuggener et al. (2020)</a>	Contracts	Multi-class classification	60,000/10,000/10,000	100
UNFAIR-ToS	<a href="#">Lippi et al. (2019)</a>	Contracts	Multi-label classification	5,532/2,275/1,607	8+1
CaseHOLD	<a href="#">Zheng et al. (2021)</a>	US Law	Multiple choice QA	45,000/3,900/3,900	n/a

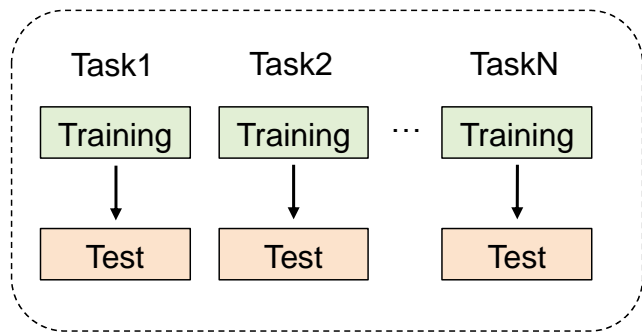
context (string)	endings (json)	label (class label)
Drapeau's cohorts, the cohort would be a "victim" of making the bomb. Further, firebombs are inherently...	[ "holding that possession of a pipe bomb is a crime of violence for purposes of 18 usc 3142f1", "holding...	0 (0)
Colameta used customer information that he took from Protégé. Additionally, Colameta admits to having take...	[ "recognizing that even if a plaintiff claims certain information constitutes trade secrets its claim may...	1 (1)
property tax sale. In reviewing section 6323(b)(6), this Court noted that it provides that a county's tax...	[ "holding that where there is a conflict between statutes the more recent statute is controlling and a...	4 (4)

# Language Understanding Needs Diverse Skills

- ❑ Different tasks may share common patterns (required skills)
- ❑ It is potential to build a unified foundation model and adapt it to different tasks

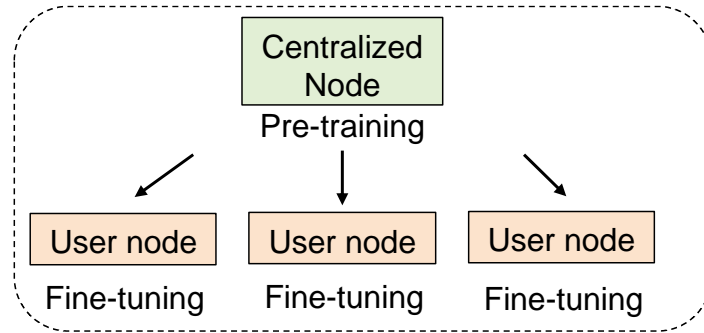
Skill	Description (Example)
Capable of	Whether an object is capable of performing an action (“A watch is capable of telling the past time”)
Long-tail knowledge	The question contains factual long-tail information (“Washington DC is located further south than Washington State”)
Plausibility	Quantifiers or always-never relations (“The peak of a mountain almost always reaches above the the tree line”)
Comparison	Comparison between two objects (“The end of a baseball bat is larger than the handle”)
Physical	Physical commonsense (“Do you build the walls on a house before putting on the roof?”)
Causality	Cause and effect relations (“If you get into an accident because you have been drinking alcohol you will be arrested?”)
Temporal	Temporal understanding (“None had ever reached the top of Mount Everest before 1977?”)
Negation	The question includes a negation phrase (“A mock trial is something with no legal consequence”)
Strategy	Reasoning steps are implicit and should be inferred using a strategy (“Blood banks almost never take cash or checks as deposits”)
Event chain	Question is about order of events (“Putting on shoes is done in this order normally: person ties shoelaces then slips shoes onto feet”)

# From Individual Task Modeling to Centralized Training



## Previous

Each user trains individual machine learning models for each task.



## Now

The central node pre-trains the generalized language model and provides the model to users for task-specific fine-tuning.

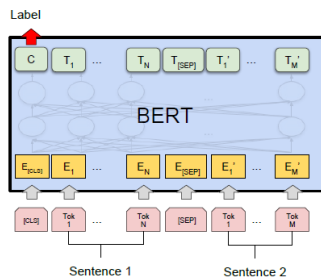
Individual  
training



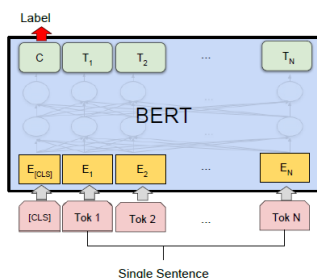
Centralized pre-training + individual fine-tuning

*\*Extreme case : GPT3 gives predictions directly, eliminating the fine-tuning process*

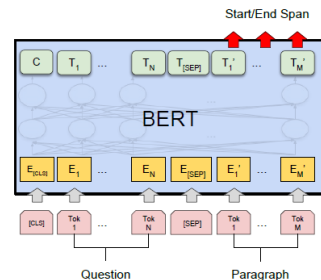
# Towards Multi-task Pre-training: Unified Modeling of Tasks



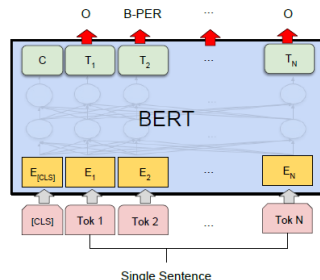
(a) Sentence Pair Classification Tasks:  
MNLi, QQP, QNLI, STS-B, MRPC,  
RTE, SWAG



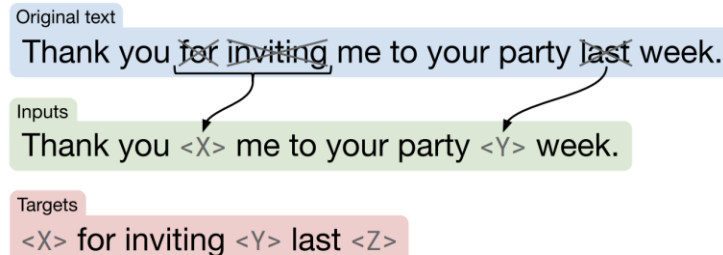
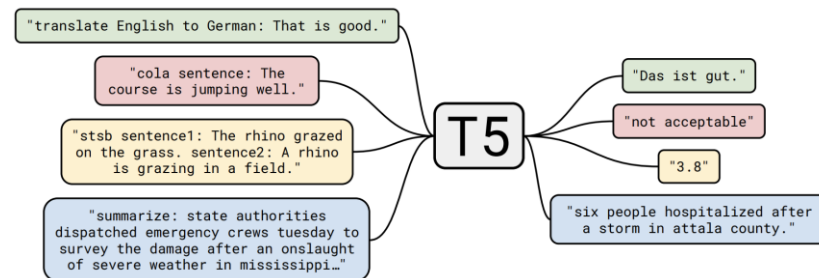
(b) Single Sentence Classification Tasks:  
SST-2, CoLA



(c) Question Answering Tasks:  
SQuAD v1.1



(d) Single Sentence Tagging Tasks:  
CoNLL-2003 NER



(a) Different formats of tasks

(b) Unified text-to-text format



# Large-scale Multi-task Pre-training

---

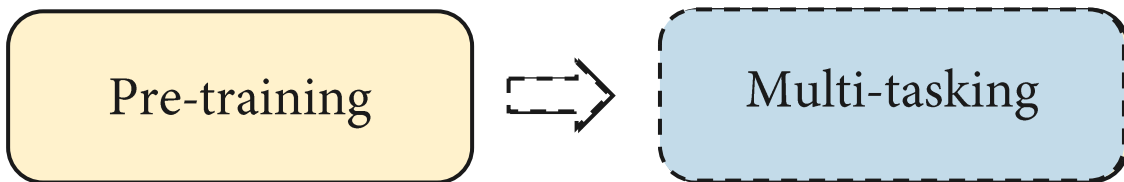
- ❑ **Theme:** Leveraging task-aware annotated data as supervised signals to assist with self-supervised learning on large-scale unlabeled data
- ❑ **Trend:** extreme scaling of task numbers, with little attention paid to the relationships between tasks
- ❑ **Challenges**
  - Catastrophic Forgetting
  - Negative Transfer

# Challenge: Catastrophic Forgetting

---

Additional large-scale learning stage between **pre-training** and **fine-tuning**

Also known as **multi-task pre-fine-tuning** or **sequential training**



# Challenge: Negative Transfer

**Observation:** tasks in different families may have side effects between each other.

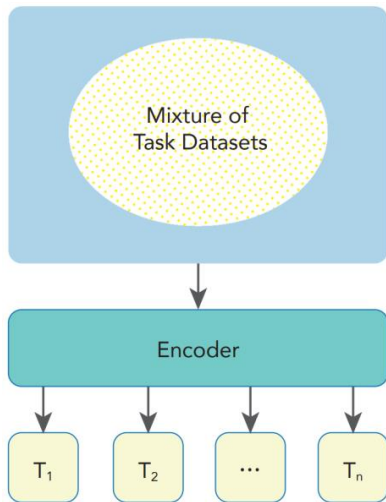
	SUM	DLG	NLI	CLS	SEM	CMNS	CBQA	RC	$\Delta_{\text{AVG}}$
SUM	27.89 29.36	37.81	60.45	77.10	78.25	61.92	7.84	65.37	-6.9%
DLG	29.05	38.56 39.76	63.62	77.10	75.55	64.05	13.39	64.75	+0.1%
NLI	28.61	40.60	64.91 67.23	77.29	77.72	67.60	15.24	66.40	+4.3%
CLS	29.52	40.16	66.69	77.14 77.47	76.05	65.29	12.93	65.20	+1.4%
SEM	29.30	38.86	62.46	76.83	72.09 72.79	57.84	12.44	63.52	-2.5%
CMNS	29.28	39.27	65.08	77.05	76.29	68.24 68.35	16.48	66.01	+4.7%
CBQA	29.75	39.29	64.96	77.66	75.21	66.84	14.68 19.98	66.37	+1.2%
RC	29.45	38.12	63.70	77.14	76.98	66.62	10.26	62.94 65.60	-2.4%
AVG <sub>\diag</sub>	29.28	39.16	63.77	77.17	76.43	64.31	12.65	65.37	

**Summarization tasks** generally seem to **hurt performance** on **dialogue system, natural language inference, and commonsense reasoning**

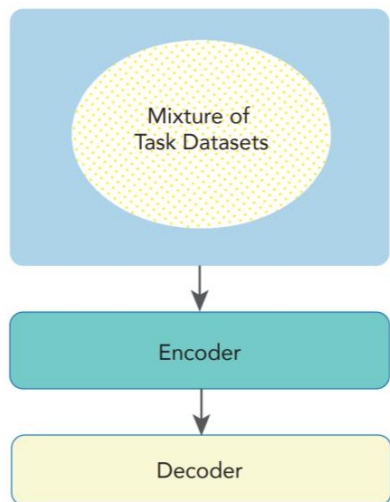
[6] Aribandi, Vamsi, et al. "ExtT5: Towards Extreme Multi-Task Scaling for Transfer Learning." International Conference on Learning Representations. 2021.

# Previous Multi-task Language Models

- a) Traditional methods: MT-DNN
- b) Unified Text-to-text Methods: T5, ExT5, FLAN, T0, etc.



a) Traditional Methods



b) Unified Text-to-text Methods

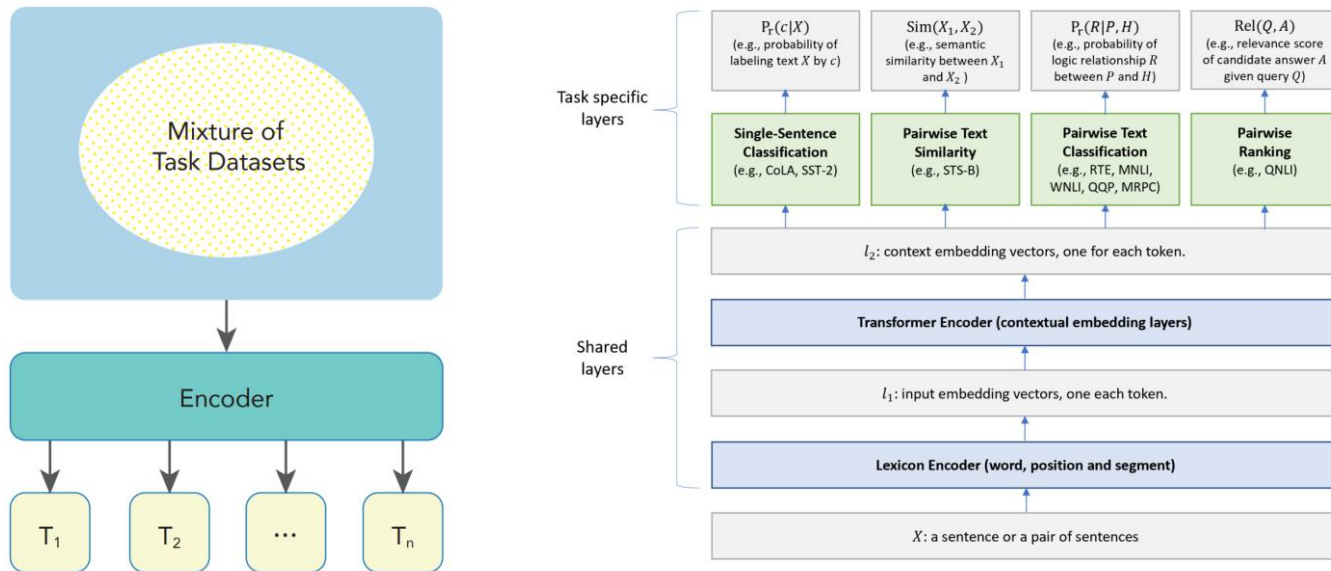
[7] Liu, Xiaodong, et al. "Multi-Task Deep Neural Networks for Natural Language Understanding." ACL. 2019.

[8] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." J. Mach. Learn. Res. 21.140 (2020): 1-67.

# Traditional Methods

## ❑ Traditional methods: MT-DNN

- Require additional modifications to model architecture and increase model complexity and computation cost
- Issue of catastrophic forgetting

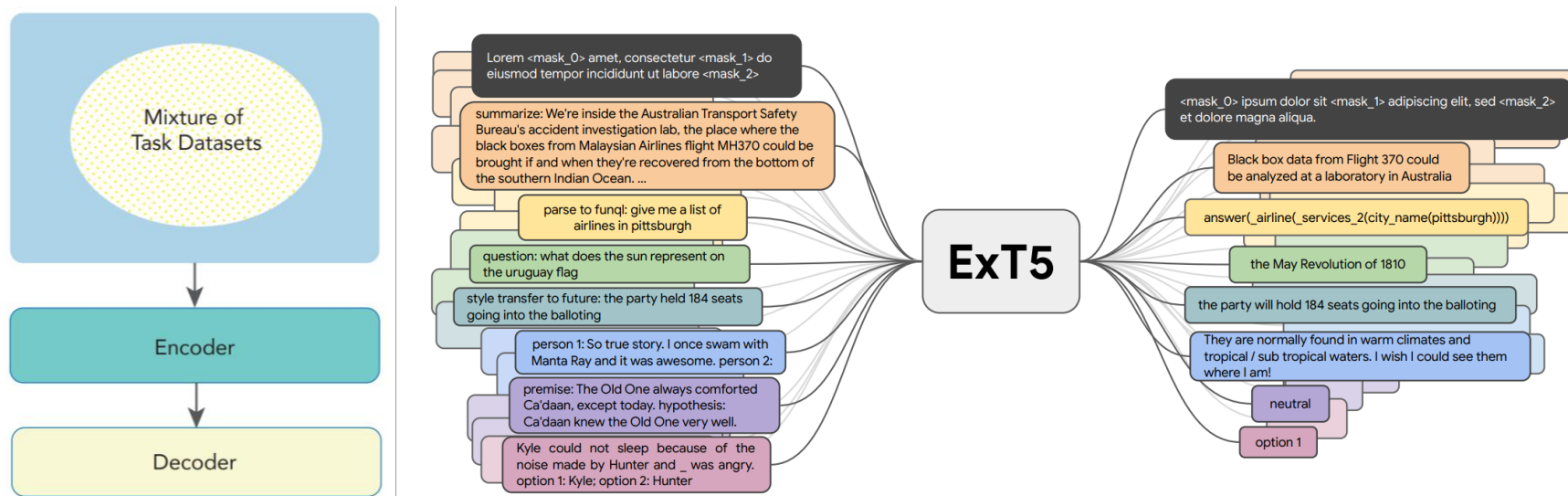


a) Traditional Methods

# Unified Text-to-text Methods

❑ Unified Text-to-text Methods: T5, ExT5, FLAN, T0, etc.

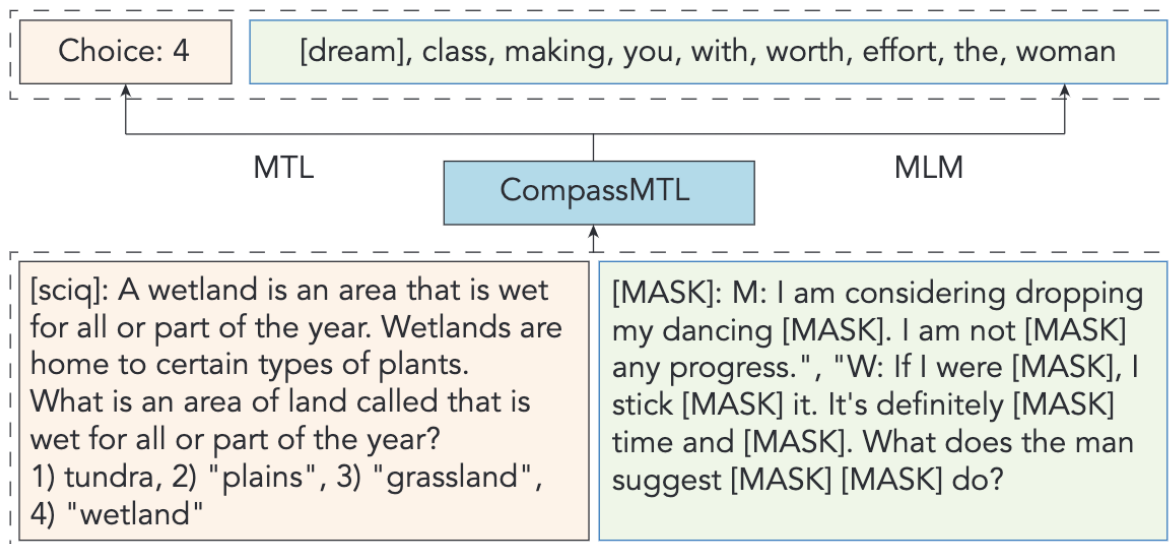
➤ Negative transfer between tasks



b) Unified Text-to-text Methods

# How to Capture Task Relationships: Our Solution

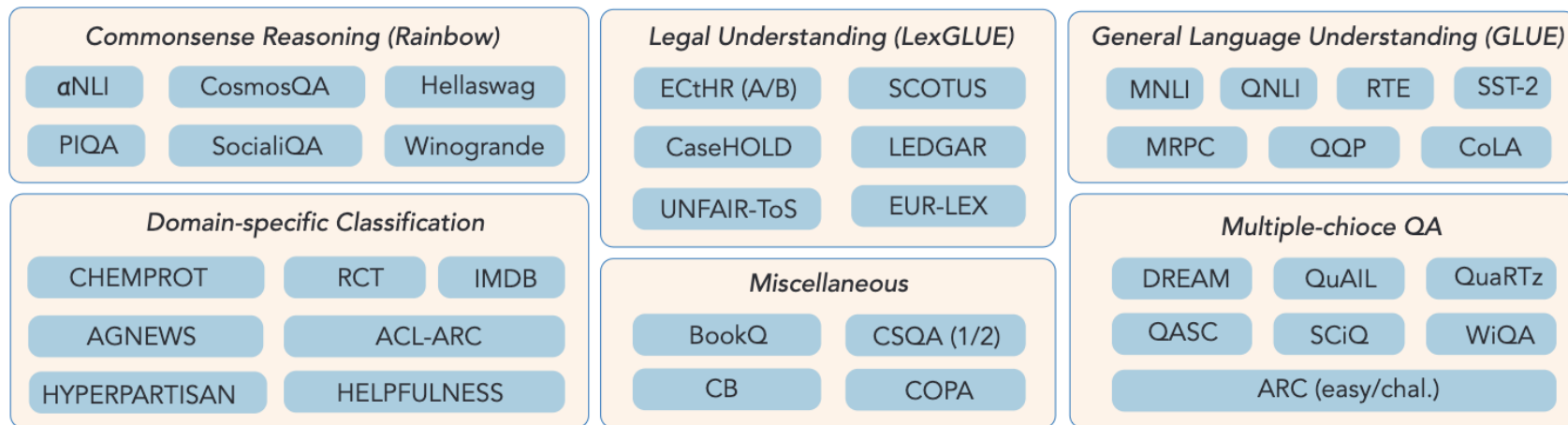
Ours: a task prefix guided multi-task pre-training framework



- 1) **Data:** Append a task prefix for each data sequence to capture common patterns from the task.
- 2) **Objective:** Require the model to predict some randomly masked prefixes to capture task differences.

# Task Taxonomy

There are 40 datasets used for training our multi-task model, some of which are collected from GLUE SuperGLUE, Rainbow, and LexGLUE





# Data Format (conversion)

**Basic:** Model tasks in a multiple-choice-like format to minimize the format transformation for NLU tasks

## Conversion Criteria:

- Ensure that each training data has a specific number of  $k$  candidate options
- Original pair-wise input texts are regarded as context and question in the view of multiple-choice problem

If the number of candidate options $> k$	the redundant options will be randomly discarded
If the number of candidate options $< k$	add "N/A" placeholder options
If the ground-truth is a list	randomly select a correct option from the gold list and randomly sample $k-1$ negative options from the held-out set
If ground-truth is a list and there is an empty choice	construct the truth option manually; the negative examples are constructed as the same as 3)

As a result, each training example will be formed as a sequence like

{ [Prefix]: context, question, option }

# Data Format (Examples)

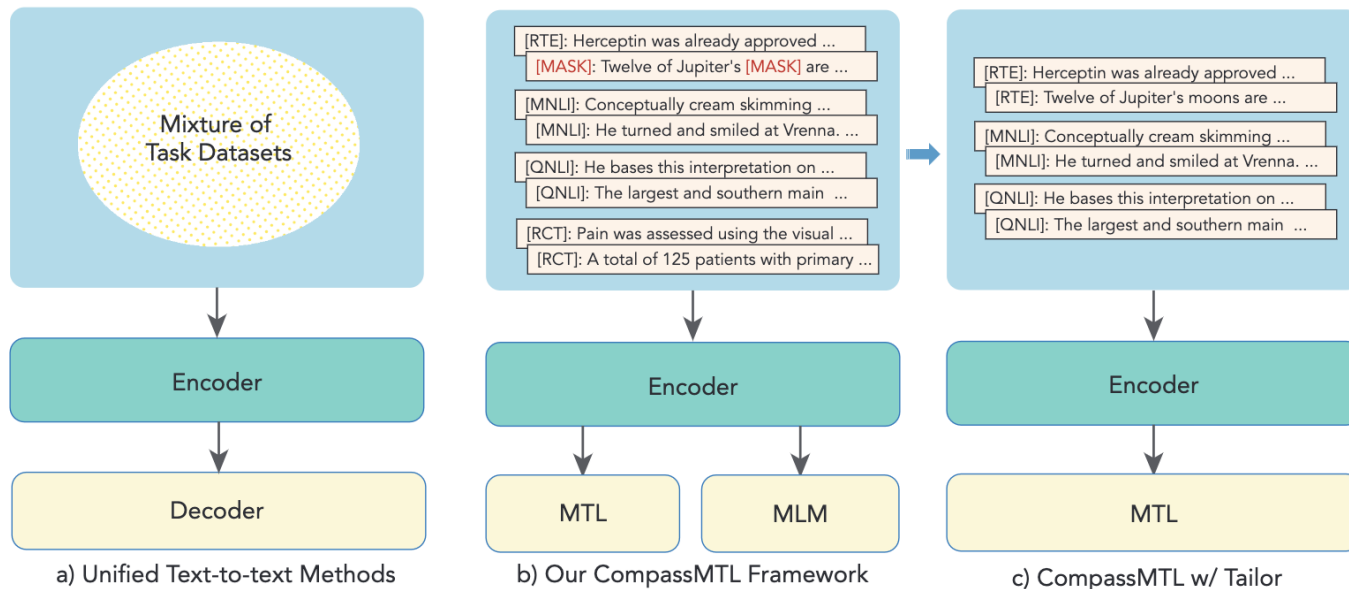
Context	Question	Option(s)
[sciq] A wetland is an area that is wet for all or part of the year. Wetlands are home to certain types of plants.	What is an area of land called that is wet for all or part of the year?	["tundra", "plains", "grassland", "wetland"]
[commonsense_qa] revolving door	A revolving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

# Model Architecture

**Backbone:** Encoder-only, based on the DeBERTa architecture

**Training Objectives:** Multi-task Learning (MTL) + Masked Language Modeling (MLM)

**Usages:** Unified Foundation Model + Probing Tool



# Model Architecture

**Data-centric:** without modification of model architecture. It can be regarded as an efficient implementation of the traditional MTL method composed of a **shared representation** module and **task-aware modules**.



the prefix is supposed to reflect the common patterns from the dataset

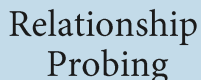


the model is required to predict randomly masked prefixes to capture task differences.

[sciq]: A wetland is an area that is wet for all or part of the year. Wetlands are home to certain types of plants. What is an area of land called that is wet for all or part of the year?  
1) tundra, 2) "plains", 3) "grassland", 4) "wetland"

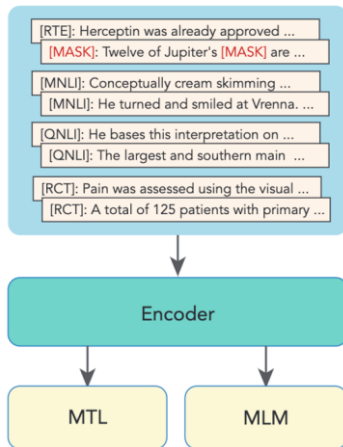
[MASK]: M: I am considering dropping my dancing [MASK]. I am not [MASK] any progress.", "W: If I were [MASK], I stick [MASK] it. It's definitely [MASK] time and [MASK]. What does the man suggest [MASK] [MASK] do?

# Model Evolution



CompassMTL  
w/ Tailor

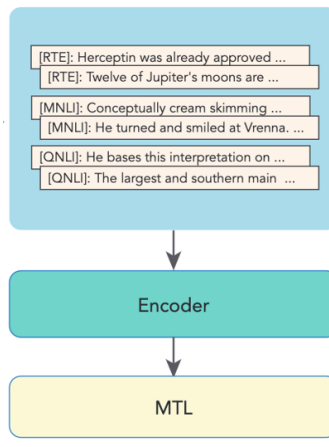
Trained on the 40 tasks



- 1) Only uses the MLM
- 2) Input data without options

[illegible]

### Trained with selected tasks



# Main Results

1. **CompassMTL** models outperform the related public models in general
2. Our **encoder-only models** yield better performance than the T5-based encoder-decoder models.
3. It is potential to achieve better performance by multi-task learning with **related tasks** (w/ Tailor)

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

Method	ECtHR (A)		ECtHR (B)		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ /m-F <sub>1</sub>
BERT	71.2	63.6	79.7	73.4	68.3	58.3	71.4	57.2	87.6	81.8	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	71.6	62.0	71.9	<b>57.9</b>	87.9	82.3	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	<b>72.1</b>	57.4	88.2	83.1	95.5	80.3	72.6
Longformer	69.9	<b>64.7</b>	79.4	71.7	72.9	64.0	71.6	57.7	88.2	83.0	95.5	80.9	71.9
BigBird	70.0	62.9	78.8	70.9	72.8	62.0	71.5	56.8	87.8	82.6	95.7	81.3	70.8
Legal-BERT	70.0	64.0	80.4	<b>74.7</b>	76.4	66.5	<b>72.1</b>	57.4	88.2	83.0	96.0	83.0	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	76.6	65.9	70.7	56.6	<b>88.3</b>	83.0	96.0	82.3	75.4
ExDeBERTa	-	-	-	-	-	-	-	-	-	-	-	-	74.8
CompassMTL	71.7	60.7	80.6	73.2	<b>77.7</b>	<b>68.9</b>	67.2	42.1	88.1	82.3	<b>96.3</b>	<b>84.3</b>	76.1
w/ Tailor	<b>73.0</b>	<b>64.7</b>	<b>80.7</b>	72.3	76.3	68.6	66.9	44.9	<b>88.3</b>	<b>83.2</b>	96.2	83.2	<b>78.1</b>

# Main Results

1. CompassMTL models outperform the related public models in general
2. Our **encoder-only models** yield better performance than the T5-based encoder-decoder models.
3. It is potential to achieve better performance by multi-task learning with **related tasks** (w/ Tailor)

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

Method	ECtHR (A)		ECtHR (B)		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ /m-F <sub>1</sub>
BERT	71.2	63.6	79.7	73.4	68.3	58.3	71.4	57.2	87.6	81.8	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	71.6	62.0	71.9	<b>57.9</b>	87.9	82.3	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	<b>72.1</b>	57.4	88.2	83.1	95.5	80.3	72.6
Longformer	69.9	<b>64.7</b>	79.4	71.7	72.9	64.0	71.6	57.7	88.2	83.0	95.5	80.9	71.9
BigBird	70.0	62.9	78.8	70.9	72.8	62.0	71.5	56.8	87.8	82.6	95.7	81.3	70.8
Legal-BERT	70.0	64.0	80.4	<b>74.7</b>	76.4	66.5	<b>72.1</b>	57.4	88.2	83.0	96.0	83.0	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	76.6	65.9	70.7	56.6	<b>88.3</b>	83.0	96.0	82.3	75.4
ExDeBERTa	-	-	-	-	-	-	-	-	-	-	-	-	74.8
CompassMTL	71.7	60.7	80.6	73.2	<b>77.7</b>	<b>68.9</b>	67.2	42.1	88.1	82.3	<b>96.3</b>	<b>84.3</b>	76.1
w/ Tailor	<b>73.0</b>	<b>64.7</b>	<b>80.7</b>	72.3	76.3	68.6	66.9	44.9	<b>88.3</b>	<b>83.2</b>	96.2	83.2	<b>78.1</b>

# Main Results

1. CompassMTL models outperform the related public models in general
2. Our **encoder-only models** yield better performance than the T5-based encoder-decoder models.
3. It is potential to achieve better performance by multi-task learning with **related tasks** (w/ Tailor)

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

Method	ECtHR (A)		ECtHR (B)		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ /m-F <sub>1</sub>
BERT	71.2	63.6	79.7	73.4	68.3	58.3	71.4	57.2	87.6	81.8	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	71.6	62.0	71.9	<b>57.9</b>	87.9	82.3	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	<b>72.1</b>	57.4	88.2	83.1	95.5	80.3	72.6
Longformer	69.9	<b>64.7</b>	79.4	71.7	72.9	64.0	71.6	57.7	88.2	83.0	95.5	80.9	71.9
BigBird	70.0	62.9	78.8	70.9	72.8	62.0	71.5	56.8	87.8	82.6	95.7	81.3	70.8
Legal-BERT	70.0	64.0	80.4	<b>74.7</b>	76.4	66.5	<b>72.1</b>	57.4	88.2	83.0	96.0	83.0	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	76.6	65.9	70.7	56.6	<b>88.3</b>	83.0	96.0	82.3	75.4
ExDeBERTa	-	-	-	-	-	-	-	-	-	-	-	-	74.8
CompassMTL	71.7	60.7	80.6	73.2	<b>77.7</b>	<b>68.9</b>	67.2	42.1	88.1	82.3	<b>96.3</b>	<b>84.3</b>	76.1
w/ Tailor	<b>73.0</b>	<b>64.7</b>	<b>80.7</b>	72.3	76.3	68.6	66.9	44.9	<b>88.3</b>	<b>83.2</b>	96.2	83.2	<b>78.1</b>



# Relationship Probing

**Probing Model:** only uses the MLM objective and is fed without options to alleviate possible shortcuts.

CosmosQA	αNLI	100	37	43	43	55	53	49	51	39	34	43	42	56	31	33	27	51	47	45	29	44	25	32	24	48	50	25	13	36	31	20	29	34	31	31	22	0	19	19	29	
	CosmosQA	35	100	33	36	44	39	28	40	54	42	44	26	36	15	29	18	31	27	30	26	24	6	29	19	37	39	20	35	26	17	32	33	30	19	20	8	0	12	31	25	
	Hellaswag	37	28	100	41	22	38	29	28	12	20	29	42	22	17	19	36	31	38	5	3	28	21	26	39	41	37	18	23	13	20	26	22	30	30	7	0	21	14	16		
	PIQA	38	32	42	100	37	50	54	39	23	48	49	48	28	11	45	37	32	47	30	2	37	32	34	31	45	47	33	31	15	10	25	38	27	16	16	1	0	22	23	23	
	SocialQA	53	43	27	40	100	53	44	40	33	37	44	32	35	22	45	30	39	40	38	24	36	17	42	34	42	45	26	29	29	22	31	32	39	16	16	0	0	17	30	25	
	Winogrande	51	38	42	52	53	100	61	34	33	38	29	37	35	13	30	37	37	42	29	15	35	29	40	30	42	44	30	23	18	29	32	23	21	22	7	0	17	23	24		
	CoLA	46	25	32	55	43	61	100	55	40	50	33	46	47	14	28	33	46	47	31	6	39	36	47	32	46	48	35	30	20	16	29	16	24	25	27	4	0	29	25	29	
	MNLI	51	41	35	44	42	36	57	100	54	49	49	48	69	34	34	34	58	45	28	18	43	32	40	35	49	51	40	25	30	28	37	26	36	32	33	25	0	36	31	29	
	MRPC	34	52	24	23	31	31	39	51	100	46	38	38	67	33	24	17	48	37	24	19	35	13	35	24	49	53	29	36	39	16	48	11	37	22	22	17	0	19	35	34	
	SST-2	32	42	24	51	37	29	52	48	48	100	30	31	38	24	39	39	41	39	26	15	34	16	45	28	39	41	29	34	28	32	48	30	38	21	24	9	0	27	45	23	
	QQP	41	43	26	51	45	29	34	47	41	30	100	54	26	34	21	40	39	30	9	34	20	42	23	50	52	33	36	24	14	33	30	31	25	25	13	0	28	32	28		
	QNLI	37	22	30	48	29	35	44	44	38	27	52	100	54	26	34	21	40	39	30	9	34	20	42	23	50	52	33	36	24	14	33	30	31	25	25	13	0	30	19	24	
	RTE	49	28	39	23	27	28	42	64	64	29	43	51	100	44	14	22	58	42	31	21	42	18	32	25	57	61	39	18	46	26	35	8	35	40	41	27	0	27	20	34	
	BookQ	19	3	18	4	13	3	5	23	28	14	17	29	44	100	41	24	53	37	33	33	19	0	21	33	43	45	39	3	40	27	39	8	30	35	37	28	15	0	27	39	14
	CommonsenseQA	18	15	8	38	35	18	16	19	13	27	22	18	9	38	100	34	35	47	33	18	49	4	43	37	36	38	21	21	16	11	33	28	26	12	12	2	0	17	35	17	
	CSQA 2.0	20	12	20	36	26	37	31	24	16	35	17	34	27	29	41	100	38	52	22	18	30	21	36	32	36	36	28	24	36	27	42	29	27	39	39	28	0	27	39	14	
CB	41	18	31	25	29	27	39	50	42	30	30	37	56	51	36	31	100	71	53	47	40	61	33	33	43	59	62	48	13	44	40	48	13	19	44	45	22	0	31	32	48	
COPA	41	22	32	46	36	39	45	40	35	34	35	43	45	40	52	52	74	100	49	37	59	48	54	50	66	67	45	29	40	32	43	22	24	44	43	22	0	33	43	46		
DREAM	30	13	30	19	25	15	17	10	11	10	15	12	25	27	31	10	51	42	100	28	45	7	17	30	45	47	26	6	28	25	20	11	5	28	27	11	0	11	14	18		
QuAIL	27	26	12	8	25	17	8	17	23	15	9	8	30	41	31	23	49	41	100	47	3	12	37	36	41	15	0	45	36	28	12	14	35	36	18	5	5	22	27			
QuARTz	33	10	27	30	26	25	30	31	27	22	23	27	39	47	49	22	61	55	47	38	100	29	37	52	58	61	41	16	27	16	36	11	21	31	31	4	0	26	21	28		
WiQA	29	13	39	40	25	36	43	36	23	22	27	39	33	19	26	32	47	55	30	11	44	100	43	45	46	47	34	25	14	0	16	13	15	24	25	4	2	31	19	27		
QASC	28	27	25	56	41	40	47	36	36	43	41	37	29	51	38	42	56	30	9	45	36	100	48	48	74	70	43	41	23	8	41	24	35	17	16	7	0	24	40	32		
SCIQ	13	10	24	28	28	23	27	26	20	15	20	26	35	41	29	46	48	37	30	54	34	44	100	50	51	41	32	26	6	43	17	22	23	23	0	4	24	33	23			
ARC-Easy	37	26	34	40	33	33	39	39	43	41	55	42	37	30	59	63	48	25	69	32	70	47	100	91	44	29	37	15	47	13	28	32	30	17	0	26	31	38				
ARC-Challenge	37	26	34	39	34	33	39	40	46	28	42	41	58	42	37	27	61	63	48	28	60	31	65	47	91	100	44	28	40	15	46	12	28	34	33	15	0	26	31	41		
CHEMPROT	9	6	31	25	14	19	26	28	21	16	22	22	37	37	22	20	48	39	29	0	41	16	34	38	43	45	100	37	32	22	53	24	30	37	36	12	14	32	33	26		
RCT	10	34	24	34	30	32	32	33	39	33	37	30	27	15	34	29	26	34	24	0	29	18	42	38	39	41	47	100	19	2	48	27	37	20	20	4	9	32	43	24		
HYPERPARTISAN	29	20	23	13	25	18	16	23	38	22	19	22	48	44	24	35	49	39	37	41	33	0	19	28	42	46	38	12	100	54	50	32	44	46	45	41	19	19	43	44		
IMDB	35	24	26	22	29	26	25	32	36	38	21	23	39	41	31	37	53	41	44	41	34	0	17	22	33	35	39	10	61	100	47	47	29	43	42	31	12	23	39	29		
ACL-ARC	10	26	19	23	26	23	24	30	46	43	27	23	37	42	39	41	52	42	28	21	0	0	38	43	50	51	56	43	50	37	100	27	31	37	35	23	12	22	56	37		
HELPPFULNESS	23	28	27	38	29	29	13	20	10	25	26	6	14	14	14	35	29	22	23	23	6	20	0	21	20	21	23	31	22	33	39	29	100	41	21	21	12	6	17	38	11	
AGNEWS	27	24	21	26	35	18	20	30	35	32	25	22	37	33	33	25	26	23	16	7	28	0	31	24	33	36	36	31	44	17	31	40	100	22	22	24	12	27	30	17		
ECCHR (A)	15	3	23	5	0	7	13	17	11	6	10	18	36	32	12	32	44	33	30	22	30	2	3	17	30	34	36	5	41	26	31	11	14	100	92	31	19	31	26	32		
ECCHR (B)	15	3	22	5	0	8	14	17	11	8	10	20	37	33	12	32	43	36	29	24	30	3	2	17	28	33	34	3	39	25	28	11	14	92	100	29	18	31	23	30		
CaseHold	22	11	16	10	20	12	10	26	24	12	16	20	38	39	20	35	36	30	20	21	21	0	12	13	32	32	28	7	48	28	32	20	32	45	43	100	30	25	24	34		
SCOTUS	3	6	13	11	7	8	9	4	11	6	7	11	17	30	21	12	20	12	23	11	20	0	9	19	20	22	31	15	29	11	25	17	24	36	36	32	100	19	23	22		
EUR-LEX	12	6	22	22	13	14	17	21	18	23	24	30	31	33	27	27	38	34	24	0	34	21	22	27	33	35	39	28	20	12	25	18	29	39	40	18	9	100	48	16		
LEDGAR	6	21	10	29	21	14	17	20	29	38	24	12	19	26	38	35	35	40	20	11	24	0	33	32	33	35	36	35	35	40	25	54	34	26	30	27	11	7	44	100	26	
UNFAIR-ToS	13	10	7	14	11	11	17	13	25	8	14	14	30	21	17	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	

CosmosQA	αNLI	100	37	43	43	55	53	49	51	39	34	43	42	56	31	33	27	51	47	45	29	44	25	32	24	48	50	25	13	36	31	20	29	34	31	31	22	0	19	19	29
	CosmosQA	35	100	33	36	44	39	28	40	54	42	44	26	36																											

## HowTo:

- 1) Fetch prefix embeddings
- 2) Calculate the Pearson correlation between each task pair

# Relationship Probing

1. The datasets inside the same task family (e.g., GLUE and Rainbow) correlate highly with each other.
2. The correlation scores also accord with the common practice of data augmentation.

CosmosQA	αNLI	100	37	43	43	55	53	49	51	39	34	43	42	56	31	33	27	51	47	45	29	44	25	32	24	48	50	25	13	36	31	20	29	34	31	31	22	0	19	19	29	
	CosmosQA	35	100	33	36	44	39	28	40	54	42	44	26	36	15	29	18	31	27	30	26	24	6	29	19	37	39	20	35	26	17	32	33	30	19	20	8	0	12	31	25	
	Hellaswag	37	28	100	41	22	38	29	28	22	18	20	29	42	22	17	19	36	31	38	5	33	28	21	26	39	41	37	18	23	13	20	26	22	30	30	7	0	21	14	16	
	PIQA	38	32	42	100	37	50	54	39	23	48	49	48	28	11	45	37	32	47	30	2	37	32	54	31	45	47	33	31	15	10	25	38	27	16	16	1	0	22	33	23	
	SocialQA	53	43	27	40	100	53	44	40	33	37	44	32	35	22	45	30	39	40	38	24	36	17	42	34	42	45	26	29	29	22	31	32	39	16	16	16	0	17	30	25	
	Winogrande	51	38	42	52	53	100	61	34	33	38	29	37	35	13	40	40	37	42	29	15	35	29	40	30	42	44	30	30	23	18	29	32	23	21	22	7	0	17	23	24	
	CoLA	46	25	32	55	43	61	100	55	40	50	33	46	47	14	28	33	46	47	31	6	39	36	47	32	46	48	35	30	20	16	29	16	24	25	27	4	0	29	25	29	
	MNLI	51	41	35	44	42	36	57	100	54	49	49	48	68	34	34	30	58	45	28	18	43	32	40	35	49	51	40	25	30	28	37	26	36	32	33	25	0	36	31	29	
	MRPC	34	52	24	23	31	31	39	51	100	46	38	38	67	33	24	17	48	37	24	19	35	13	35	24	49	53	29	36	39	16	48	11	37	22	22	17	0	19	35	34	
	SST-2	32	42	24	51	37	39	52	48	48	100	30	31	38	24	39	39	41	39	26	15	34	16	45	28	39	41	29	34	28	32	48	30	38	21	24	9	0	27	45	23	
QQP	41	43	26	51	45	29	34	47	41	30	100	54	45	26	34	21	40	39	30	9	34	20	42	23	50	52	33	36	24	14	33	30	31	25	25	13	0	28	32	28		
QNLI	37	22	30	48	29	35	44	44	38	27	52	100	54	34	28	34	44	44	24	3	35	31	40	24	47	48	30	26	23	12	26	7	24	28	30	13	0	30	19	24		
RTE	49	28	39	23	27	28	42	64	64	29	43	51	100	44	14	22	58	42	31	21	42	18	32	25	57	61	39	18	46	26	35	8	35	40	41	27	0	27	20	34		
CommonsenseQA	BookQ	19	3	18	4	13	3	5	23	28	14	17	29	44	100	41	24	53	37	33	33	49	0	21	33	43	45	39	3	40	27	39	8	30	35	37	28	15	28	26	24	
	CommonsenseQA	18	15	8	38	35	18	16	19	13	27	22	18	9	38	100	34	35	47	33	18	49	4	43	37	36	38	21	21	16	11	33	28	26	12	12	2	0	17	35	17	
	CSQA 2.0	20	12	20	36	26	37	31	24	16	35	17	34	27	29	41	100	38	52	22	18	30	21	36	32	36	36	28	24	36	27	42	29	27	39	39	28	0	27	39	14	
	CB	41	18	31	25	29	27	39	50	42	30	30	37	56	51	36	31	100	71	53	40	61	33	33	43	59	62	48	13	44	40	48	13	19	44	45	22	0	31	32	48	
	COPA	41	22	32	46	36	39	45	40	35	34	35	43	45	40	52	52	74	100	49	37	59	48	54	50	66	67	45	29	40	32	43	22	24	44	43	22	0	33	43	46	
	DREAM	30	13	30	19	25	15	17	10	11	10	15	12	25	27	31	10	51	42	100	28	45	7	17	30	45	47	26	6	28	25	20	11	5	28	27	11	0	11	14	18	
	QuAIL	27	26	12	8	25	17	8	17	23	15	9	8	30	41	31	23	49	41	41	100	47	3	12	37	36	41	15	0	45	36	28	12	14	35	36	18	5	5	22	27	
	QuARTz	33	10	27	30	26	25	30	31	27	22	23	27	39	47	49	22	61	55	47	38	100	29	37	52	58	61	41	16	27	16	36	11	21	31	31	4	0	26	21	28	
	WQIA	29	13	39	40	25	36	43	36	23	22	27	39	33	19	26	32	47	55	30	11	44	100	43	45	46	47	34	25	14	0	16	13	15	24	25	4	2	31	19	27	
	QASC	28	27	25	56	41	40	47	36	36	43	41	41	37	29	51	38	42	56	30	9	45	36	100	48	74	70	43	41	23	8	41	24	35	17	16	7	0	24	40	32	
SCIQ	13	10	24	28	28	23	27	26	20	15	20	26	35	41	29	46	48	37	30	54	34	44	100	50	51	41	32	26	6	43	17	22	23	23	0	4	24	33	23			
ARC-Challenge	ARC-Easy	37	26	34	40	33	33	39	39	43	28	42	41	55	42	37	30	59	63	48	25	59	32	70	47	100	91	44	29	37	15	47	13	28	32	30	17	0	26	31	38	
	ARC-Challenge	37	26	34	39	34	33	39	40	46	28	42	41	58	42	37	27	61	63	48	28	60	31	65	47	91	100	44	28	40	15	46	12	28	34	33	15	0	26	31	41	
	CHEMPROT	9	6	31	25	14	19	26	28	21	16	22	22	37	37	22	20	48	39	29	0	41	16	34	38	43	45	100	37	32	22	53	24	30	37	36	12	14	32	33	26	
	RCT	10	34	24	34	30	32	32	23	33	33	37	37	30	27	15	34	29	26	34	24	0	29	18	42	38	39	41	47	100	19	2	48	27	37	20	20	4	9	32	43	24
	HYPERPARTISAN	29	20	23	13	25	18	16	23	38	22	19	22	48	44	24	35	49	39	37	41	33	0	19	28	42	46	38	12	100	54	50	32	44	46	45	41	19	19	43	44	
	IMDB	35	24	26	22	29	26	25	32	26	38	21	23	39	41	31	37	53	41	44	41	34	0	17	22	33	35	39	10	61	100	47	47	29	43	42	31	12	23	39	29	
	ACL-ARC	10	26	19	23	26	23	24	30	46	43	27	23	37	42	39	41	52	42	28	21	40	0	38	43	50	51	56	43	50	37	100	27	31	37	35	23	12	22	56	37	
	HELPLEFULNESS	23	28	27	38	29	29	13	20	10	25	26	6	14	14	14	35	22	23	23	3	20	0	21	20	21	23	31	22	33	39	29	100	41	21	21	12	6	17	38	11	
	AGNEWS	27	24	21	26	35	18	20	30	35	32	25	22	37	37	33	33	25	26	23	16	7	28	0	31	24	33	36	36	31	44	17	31	40	100	22	22	24	12	27	30	17
	CaseHOLD	ECNHR (A)	15	3	23	5	0	7	13	17	11	6	10	18	36	32	12	32	44	37	30	22	30	2	3	17	30	34	36	5	41	26	31	11	14	100	92	31	19	31	26	32
ECNHR (B)		15	3	22	5	0	8	14	17	11	8	10	20	37	33	12	32	36	36	29	24	30	3	2	17	28	33	34	3	39	25	28	11	14	92	100	29	18	31	23	30	
CaseHOLD		22	11	16	10	20	12	10	26	24	12	16	20	38	39	20	35	36	30	21	21	0	0	12	13	32	32	28	7	48	28	32	20	32	45	43	100	30	25	24	34	
SCOTUS		3	6	13	11	7	8	9	4	11	6	7	11	17	30	21	12	20	12	23	11	20	0	9	19	20	22	31	15	29	11	25	17	24	36	36	32	100	19	23	22	
EUR-LEX		12	6	22	22	13	14	27	31	18	23	24	30	31	33	27	27	38	34	24	0	34	21	22	27	33	35	39	28	20	12	25	18	29	39	40	18	9	100	48	16	
LEDGAR		6	21	10	29	21	14	17	20	29	38	24	12	19	26	38	35	35	40	20	11	24	0	0	33	32	33	35	36	35	40	25	54	34	26	30	27	11	7	44	100	26
UNFAIR-ToS		13	10	7	14	11	11	17	13	25	8	14	14	14	4	47	40	20	20	13	27	5	21	17	36	42	25	9	38	8	31	0	8	33	31	18	1	5	23	100		
CosmosQA		αNLI	100	37	43	43	55	53	49	51	39	34	43	42	56	31	33	27	51	47	45	29	44	25	32	24	48	50	25	13	36	31	20	29	34	31	31	22	0	19	19	29
		CosmosQA	35	100	33																																					

# Relationship Probing

1. The datasets inside the same task family (e.g., GLUE and Rainbow) correlate highly with each other.
2. The correlation scores also accord with the common practice of data augmentation.

CommonsenseQA	anLI	100	37	43	55	49	54	39	34	43	42	56	31	33	27	51	47	45	29	44	25	32	24	48	50	25	13	36	31	20	29	34	31	31	22	0	19	19	29		
	CosmosQA	35	100	33	36	44	39	28	40	54	42	44	26	36	15	29	18	31	27	30	26	24	6	29	19	37	39	20	35	26	17	32	33	30	19	20	8	0	12	31	25
	Hellaswag	37	28	100	41	22	38	29	28	22	18	20	29	42	22	17	19	36	31	38	5	33	28	21	26	39	41	37	18	23	13	20	26	22	30	30	7	0	21	14	16
	PIQA	38	32	42	100	37	50	54	39	23	48	49	28	11	45	37	32	47	30	2	37	32	54	31	45	47	33	31	15	10	25	38	27	16	16	1	0	22	33	23	
	SocialQA	53	43	27	40	100	53	44	40	33	37	44	32	35	22	45	30	39	40	38	24	36	17	42	34	42	45	26	29	22	31	32	39	16	16	16	0	17	30	25	
	Winogrande	51	38	42	52	53	100	61	34	33	38	29	37	35	13	30	40	37	42	29	15	35	29	40	30	42	44	30	23	18	29	32	23	21	22	7	0	17	23	24	
	CoLA	46	25	32	55	43	61	100	55	40	50	33	46	47	14	28	33	46	47	31	6	39	36	47	32	46	48	35	20	16	29	16	24	25	27	4	0	29	25	29	
	MNLI	51	41	35	44	42	36	57	100	54	49	49	48	69	34	34	30	58	45	28	18	43	32	40	35	49	51	40	25	30	28	37	26	36	32	33	25	0	36	31	29
	MRPC	54	52	24	23	31	31	39	51	100	46	38	38	67	33	24	17	48	37	24	19	35	13	35	24	49	53	29	36	39	16	48	11	37	22	22	17	0	19	35	34
	SST-2	32	44	51	37	39	52	48	48	100	30	31	38	24	39	39	41	39	26	15	34	16	45	28	39	41	29	34	28	32	48	30	38	21	24	9	0	27	45	23	
QQP	41	45	46	51	45	29	34	47	41	30	100	54	49	26	34	21	40	39	30	9	34	20	42	23	50	52	33	36	24	14	33	30	31	25	25	13	0	28	32	28	
QNLI	37	22	30	48	29	35	44	44	38	27	52	100	54	34	28	34	44	24	3	35	31	40	24	47	48	30	26	23	12	26	7	24	28	30	13	0	30	19	24		
RTE	49	28	39	23	27	28	42	64	64	29	43	51	100	44	14	22	58	42	31	21	42	18	32	25	57	61	39	18	46	26	35	8	35	40	41	27	0	27	20	34	
BookQ	19	3	18	4	13	3	5	23	28	14	17	29	44	100	41	24	53	37	33	33	49	0	21	33	43	45	39	3	40	27	39	8	30	35	37	28	15	28	26	24	
CommonsenseQA 2.0	ARC-Easy	18	15	8	38	35	18	16	19	13	27	22	18	9	38	100	34	35	47	33	18	49	4	43	37	36	38	21	21	16	11	33	28	26	12	12	2	0	17	35	17
	CSQA 2.0	20	12	20	36	26	37	31	24	16	35	17	34	27	29	41	100	38	52	22	18	30	31	36	32	36	36	28	24	36	27	42	29	27	39	28	0	27	39	14	
	CB	41	18	31	25	29	27	39	50	42	30	30	37	56	51	36	31	100	71	53	40	61	33	33	43	59	62	43	13	44	40	46	13	19	44	45	22	0	31	32	48
	COPA	41	22	32	46	36	39	45	40	35	34	35	43	45	40	52	52	74	100	49	37	59	48	54	50	66	67	45	29	40	32	43	22	24	44	43	22	0	33	43	46
	DREAM	30	13	30	19	25	17	10	11	10	15	12	25	27	31	10	51	42	100	28	45	7	17	30	45	47	26	6	28	25	20	11	5	28	27	11	0	11	14	18	
	QUAIL	27	26	12	8	25	17	8	17	13	15	9	8	30	41	31	23	49	41	41	100	47	3	12	37	36	41	15	0	45	36	28	12	14	35	36	18	5	5	22	27
	Quartz	33	10	27	30	26	25	30	31	27	22	23	27	39	47	49	22	61	55	47	38	100	29	37	52	58	61	41	16	27	16	36	11	21	31	4	0	26	21	28	
	WIQA	29	13	39	40	25	36	43	36	23	22	27	39	33	19	26	32	47	55	30	11	44	100	43	45	46	47	34	25	14	0	16	13	15	24	25	4	2	31	19	27
	QASC	28	27	25	56	41	40	47	36	43	41	21	20	37	29	51	38	42	56	30	9	45	36	100	48	74	70	43	41	23	8	41	24	35	17	16	7	0	24	40	32
	SCIQ	13	10	24	28	28	23	27	26	20	15	20	26	35	41	29	46	48	37	30	54	34	44	100	50	51	41	32	26	6	43	17	22	23	23	0	4	24	33	23	
CommonsenseQA 2.0	ARC-Easy	37	26	34	40	33	39	39	43	28	42	41	55	42	37	30	59	63	48	25	59	32	70	47	100	91	44	29	37	15	47	12	28	32	30	17	0	26	31	38	
	ARC-Challenge	37	26	34	39	34	33	39	40	46	28	42	41	58	42	37	27	61	63	48	28	60	31	65	47	91	100	44	28	40	15	46	12	28	34	33	15	0	26	31	41
	CHEMPROT	9	6	31	25	14	19	26	28	21	16	22	37	37	22	20	48	39	29	0	41	16	34	38	43	45	100	37	32	22	53	24	30	37	36	12	14	32	33	26	
	RCT	10	34	24	34	30	32	32	33	39	33	30	27	15	34	29	26	34	24	0	29	18	42	38	39	41	47	100	32	2	48	27	37	20	20	4	9	32	43	24	
	HYPERPARTISAN	29	20	23	13	25	18	23	38	22	19	22	46	44	24	35	49	39	37	41	33	0	19	28	42	46	38	12	100	54	50	32	44	46	45	41	19	19	43	44	
	IMDB	35	24	26	22	29	26	25	32	26	38	21	23	39	41	31	37	53	41	44	41	34	0	17	22	33	35	39	10	61	100	47	47	29	43	42	31	12	23	39	29
	ACL-ARC	10	26	19	23	26	23	24	30	46	43	27	23	39	41	31	37	52	42	28	21	40	0	38	43	50	51	56	43	50	37	100	27	31	37	35	23	12	22	56	37
	HELPLESSNESS	23	28	27	38	29	13	20	10	25	26	6	14	14	35	29	22	23	23	6	20	0	21	20	21	23	31	22	33	39	29	100	41	21	21	12	6	17	38	11	
	AGNEWS	27	24	21	26	35	18	20	30	35	32	25	22	37	33	33	25	26	23	16	7	28	0	31	24	33	36	36	31	44	17	31	40	100	22	22	24	12	27	30	17
	CommonsenseQA 2.0	ECCHR (A)	15	3	23	5	0	7	13	17	11	6	10	18	36	32	12	32	44	37	30	22	30	2	3	17	30	34	36	5	41	26	31	11	14	100	92	31	19	31	26
ECCHR (B)		15	3	22	5	0	8	14	17	11	6	10	20	37	33	12	32	43	36	29	24	30	3	2	17	28	33	34	3	39	25	28	11	14	92	100	29	18	31	23	30
CaseHOLD		22	11	16	10	10	12	10	26	24	12	16	20	38	30	20	35	36	30	21	21	0	12	13	32	32	28	7	48	28	32	20	32	45	43	100	30	25	24	34	
SCOTUS		3	6	13	11	7	8	9	4	11	6	7	11	10	39	39	20	12	10	23	11	20	0	9	19	20	22	31	15	29	11	25	17	24	36	36	32	100	19	23	22
EUR-LEX		12	6	22	22	13	14	27	31	18	23	30	34	30	31	33	27	37	38	34	24	0	34	21	22	27	33	35	36	40	25	54	34	26	30	27	11	7	44	100	26
LEDGAR		6	21	10	29	21	14	17	20	29	38	24	12	19	26	38	35	35	40	20	11	24	0	33	32	33	35	36	40	25	54	34	26	30	27	11	7	44	100	26	
UNFAIR-TOS		13	10	7	14	11	11	17	13	25	8	14	14	30	21	17	47	40	20	13	27	0	5	21	17	36	42	25	9	38	8	31	1	5	23	23	100	0	0	0	0

- 1) the NLI datasets (MNLI, QNLI, RTE) share close relevance
- 2) helpful to initialize from an MNLI model to fine-tune RTE

# Relationship Probing

**Topic:** Whether the relationship scores coordinate with the model performance transferred between tasks

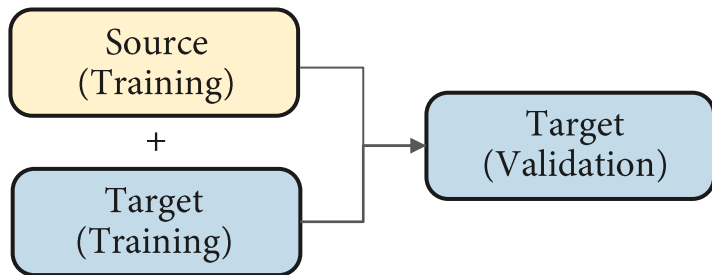
**Source tasks:** 13 source tasks from GLUE and Rainbow tasks

**Target Tasks:** 5 target tasks (ANLI, HellaSwag, MRPC, PIQA, QNLI, and RTE)

**Dual-task training setup:**

**Co-training:** train individual models using the mixture of training sets from each pair of source & target tasks

**Evaluation:** then evaluate the model on the validation set of the target dataset.



# Relationship Probing

Finally, we have 5 X 13 transfer results.

For each target dataset, we calculate Pearson correlation between relationship scores and transfer accuracy among the source datasets.

<b>Dataset</b>	<b>RTE</b>	<b>MRPC</b>	<b>QNLI</b>	<b>HellaSwag</b>	<b><i>α</i>NLI</b>
Correlation	0.19	0.22	0.38	0.12	0.51

Table 3: Pearson correlation between the relationship scores and the transfer accuracy.

**Result:** the relationship scores are positively bound up with the transfer performance

# Complementary Transfer

## Topic:

1. whether using more datasets always leads to better performance
2. whether using the most related datasets can lead to competitive results.

**Data Selection:** select a group of datasets to train an MTL model and fine-tuning the model on target datasets.

40-fullset	the same as our basic setting of CompassMTL
Top-5	Top-5 ranked dataset according to based on our probed relationship scores
Family	the datasets belonged to the same family with the target dataset
14-subset	the mixture of Rainbow and GLUE datasets

# Complementary Transfer

1. **Top-5** variant yields comparable, even better results than the others
2. **Small-scale datasets** (e.g., MRPC and RTE) are more likely to benefit from the complementary transfer

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

# Complementary Transfer

1. **Top-5** variant yields comparable, even better results than the others
2. **Small-scale datasets** (e.g., MRPC and RTE) are more likely to benefit from the complementary transfer

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



# Human-parity on Commonsense Reasoning Leaderboards

**Models:** The submissions are based on the ensemble of three models from complementary transfer.

**Results:** Compared with public methods that use much larger PrLMs, model ensemble, and knowledge graphs, our models establish new state-of-the-art results and reach **human-parity performance**.

Model	HellaSwag	$\alpha$ NLI
Human Performance	95.60	92.90
Previous SOTA	94.87	92.20
Our Results	95.94	92.80

<https://leaderboard.allenai.org/hellaswag/submissions/public>  
<https://leaderboard.allenai.org/anli/submissions/public>

Human Performance				Accuracy: 0.9560
				<a href="#">Download</a>
Rank	Submission	Created	Accuracy	
1	UniMTL Microsoft Azure Cognitive Ser...	05/11/2022	0.9594	
2	DeBERTa Large EMNLP Paper 3842 Authors	05/20/2022	0.9557	
3	CreAT Anonymous	05/03/2022	0.9487	
4	DeBERTa MCQ EMNLP Paper 3842 Authors	06/03/2022	0.9472	
5	DeBERTa Large Anonymous	04/14/2022	0.9437	
6	UL Test Google Research	03/14/2022	0.9413	
7	Deberta Testing Myself	05/06/2022	0.9394	
8	UNICORN Anonymous	07/24/2020	0.9385	

# Beyond The Unified Format

**Topic:** whether our model can be used for tasks that are unavailable to be transformed into our format

We evaluate the effectiveness by using the 1) reading comprehension datasets **SQuAD v1.1/2.0** and named entity recognition (NER) dataset **CoNLL 2003**.

Results show that our model is **generally effective across formats**

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

# Extension to T5

Our method is **generally applicable to other kinds of PrLMs**, such as encoder-decoder T5.

Model	$\alpha$ NLI	CosmosQA	HellaSwag	PIQA	SocialIQA	Winogrande	Average
T5	68.5	69.6	56.6	67.7	65.1	62.4	65.0
UNICORN	65.3	72.8	56.2	73.3	66.1	61.8	65.9
CompassMTL	<b>69.1</b>	<b>72.6</b>	<b>57.7</b>	<b>73.6</b>	<b>66.6</b>	<b>64.9</b>	<b>67.4</b>

Table 9: Results on the Rainbow validation sets by using T5-base as the backbone model.

# Conclusions

---

## ❑ A unified task prefix guided multi-task method

- Strong foundation backbone for a wide range of NLU tasks
- A probing tool for analyzing task relationships

## ❑ Effectiveness

- Generalizable advances over tasks in diverse formats
- Establishes human-parity results on commonsense reasoning tasks

## ❑ Findings

- prefixes reflect task relationships, which correlate with transfer learning performance between tasks
- suggest directions for data augmentation of complementary tasks

# Prospects for Future Studies

---

## 1) Collaborative multi-task learning of PrLMs

The recipe of using **task prefixes + prefix prediction in MLM** has shown effective for MTL pre-training.

## 2) Suggestive choice for data augmentation

The probed task relationships have shown informative in **finding complementary tasks**, which help obtain better performance for a target task, especially for small-scale datasets.

## 3) Guidance for skill-aware model evaluation

The discovery of task relationships may help determine redundant datasets that assess similar patterns of models to **avoid evaluation redundancy and save computation**.

# Thanks & QA

Zhuosheng Zhang

zhangzs@sjtu.edu.cn

<https://bcmi.sjtu.edu.cn/~zhangzs>