



**HAL**  
open science

# tasksource: A Dataset Harmonization Framework for Streamlined NLP Multi-Task Learning and Evaluation

Damien Sileo

► **To cite this version:**

Damien Sileo. tasksource: A Dataset Harmonization Framework for Streamlined NLP Multi-Task Learning and Evaluation. 2023. hal-04099649

**HAL Id: hal-04099649**

**<https://inria.hal.science/hal-04099649v1>**

Preprint submitted on 17 May 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# tasksource: A Dataset Harmonization Framework for Streamlined NLP Multi-Task Learning and Evaluation

Damien Sileo<sup>1</sup>

<sup>1</sup>Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France  
[damien.sileo@inria.fr](mailto:damien.sileo@inria.fr)

## Abstract

The HuggingFace Datasets Hub hosts thousands of datasets, offering exciting opportunities for language model training and evaluation. However, datasets for a specific task type often have different schemas, making harmonization challenging<sup>1</sup>. Multi-task training or evaluation necessitates manual work to fit data into task templates. Several initiatives independently tackle this issue by releasing harmonized datasets or providing harmonization codes to preprocess datasets into a consistent format. We identify patterns across previous preprocessing efforts, such as column name mapping and extracting specific sub-fields from structured data in a column. We then propose a structured annotation framework that ensures our annotations are fully exposed and not hidden within unstructured code. We release a dataset annotation framework and dataset annotations for more than 500 English tasks<sup>2</sup>. These annotations include metadata, such as the names of columns to be used as input or labels for all datasets, which can save time for future dataset preprocessing, regardless of whether our framework is utilized. We fine-tune a multi-task text encoder on all tasksource tasks, outperforming every publicly available text encoder of comparable size in an external evaluation<sup>3</sup>.

## 1 Introduction

Datasets are a key ingredient in modern artificial natural language processing (NLP).

Language understanding models trained on unannotated corpora need to be evaluated, and individual datasets or benchmarks with multiple datasets provide an objective measure of targeted model capabilities. Supervised fine-tuning on annotated datasets also leads to better evaluations, and multi-task learning (MTL) (Caruana, 1993) or extreme

MTL (Aribandi et al., 2022), i.e. MTL with many tasks, improves robustness.

The HuggingFace Datasets (Wolf et al., 2020) Hub hosts thousands of datasets. However, running evaluations or MTL on many datasets requires manual work because of a lack of standardization. Fine-tuning a model on multiple datasets requires alignment of datasets formats, even with a single task type (e.g. natural language inference). Because of that, various initiatives assemble datasets or preprocessing code to ease multi-task learning or benchmarking. However, they either distribute preprocessed copies of the datasets or preprocessing code associated with each dataset. Section 2 enumerates previous works enacting these two approaches.

The previous preprocessing codes implicitly use some metadata, such as mappings between column names and fields of a task, but extracting it is quite difficult. Code is not disentangled from metadata. We propose a very concise dataset annotation format by relying on patterns reoccurring across several preprocessings. Most annotations fit in a single line, e.g:

```
scitail = Classification(  
    'sentence1',  
    'sentence2',  
    'gold_label')
```

The SciTail (Khot et al., 2018) dataset on HF-Hub, noted `scitail_ds`<sup>4</sup> (Khot et al., 2018) can be standardized by calling the `tasksource.scitail` function, and associated metadata can be retrieved with `tasksource.scitail.dict()`.

We annotate 480 tasks, focusing on discriminative tasks to complement previous work better. We train a `deberta-base` text encoder on all of them simultaneously (Section 6) leading to unprecedented model performance (Section 7).

<sup>1</sup><https://xkcd.com/927/>

<sup>2</sup><https://github.com/sileod/tasksource>

<sup>3</sup>[hf.co/sileod/deberta-v3-base-tasksource-nli](https://hf.co/sileod/deberta-v3-base-tasksource-nli) <sup>4</sup><https://hf.co/datasets/scitail>

## 2 Related work

Various works harmonize existing datasets by either sharing preprocessed copies or preprocessings. Tasksource is a collection of preprocessings, and it is the largest for tasks excluding text generation tasks. Text generations tasks have a relatively simple format (optional input text, and output text), and previous work such as PromptSource (Bach et al., 2022) and SuperNatural Instructions (Wang et al., 2022) did not provide structured annotations, as defined in section 3, but these can still be combined with acceptable efforts.

**Preprocessed copies** BIG-Bench (Srivastava et al., 2022a), BigBio (Fries et al., 2022), Natural and SuperNatural Instructions (Mishra et al., 2022; Wang et al., 2022), PragmEval (Sileo et al., 2022a), UnifiedQA (Khashabi et al., 2020b), TweetEval (Barbieri et al., 2020), DiscoEval (Chen et al., 2019b), Silicone (Chapuis et al., 2020), LexGLUE (Chalkidis et al., 2022), SetFit (Tunstall et al., 2022) distribute preprocessed copies of the original data with standardized format.

**Collections of preprocessings** SentEval (Conneau et al., 2017), Jiant (Pruksachatkun et al., 2020), BLUE (Peng et al., 2019), MetaEval (Sileo and Moens, 2022a), CrossFit (Ye et al., 2021), PromptSource (Bach et al., 2022) distribute the code required to jointly use some datasets with initially distinct structures. ExMix (Aribandi et al., 2022) is not released to our knowledge. The Muppet (Aghajanyan et al., 2021) authors did not release their preprocessing either.

Our work also pertains to extreme MTL (Aribandi et al., 2022; Aghajanyan et al., 2021) and dataset count scaling.

## 3 Structured dataset annotation

We define *dataset parsing* as the mapping of a dataset into a task template.

A task template is a type of task, like paraphrase detection, associated with a predetermined set of fields. For example, Paraphrase detection can be mapped to a task template PARAPHRASEDETECTION(SENTENCE1, SENTENCE2, LABEL)

A dataset is a set of examples with named and typed columns. `quora` is an example of a dataset hosted on the HuggingFace Datasets Hub (Wolf et al., 2020), illustrated in Table 1.

A dataset parser for a specific dataset is a function that maps the whole dataset, or examples, to

questions (sequence)	is_duplicate (bool)
{ "id": [ 1, 2 ], "text": [ "What is the step by step guide to invest in share market in india?", "What is the step by step guide to invest in share market?" ] }	false
...	

Table 1: One row of the Quora dataset, as hosted on the HuggingFace Datasets Hub.

a task format, which can be PARAPHRASEDETECTION here.

As seen above, some benchmarks distribute harmonized datasets. This approach can save computations but can waste storage space, and make it harder to track all the design decisions that were applied to the original dataset. Users can also implement parsers themselves, or rely on external libraries to process examples on a restricted set of tasks. The previous preprocessings codes do not disentangle data and logic, and cannot be seen as semantic dataset annotations. This complicates the combinations of different preprocessing. Previous preprocessings also contain repetitive boilerplate code<sup>5</sup>.

We decompose dataset parsing logic from annotations with based on two observations:

(1) The fields of task type (e.g. SENTENCE1, SENTENCE2 for PARAPHRASEDETECTION) can often be independently mapped to functions of dataset examples. Therefore, we can annotate a dataset with a task type, then annotate each field of a task type with a function that extracts the desired information from a dataset example.

(2) The field mapping functions are often selecting a column from examples: in that case, they can be annotated with the name of the relevant columns. Sometimes, as in the Quora dataset in Table 1, they are selecting a path from a nested structure: in that case, they can be annotated with a path. Fields can also be mapped to a constant – some multiple-choice question-answering datasets always use the first choice as the correct choice and have an implicit constant label equal to 0. A field can also be mapped to a concatenation of the text of different columns, which can also be annotated

<sup>5</sup>i.e. [https://github.com/INK-USC/CrossFit/blob/master/tasks/aqua\\_rat.py](https://github.com/INK-USC/CrossFit/blob/master/tasks/aqua_rat.py)

with parameters.

#### 4 Tasksource dataset annotation format

Once a dataset is annotated with a task type, each field of the task type has to be annotated with a function that takes an example from the dataset and returns the intended part of the example. For brevity, we can annotate a field with a string `s` to denote the function `lambda x:x[s]`

The tasksource backend handles the annotations and turns them into harmonizing preprocessing. We consider 3 general task types:

`CLASSIFICATION(TEXT1, TEXT2, LABELS)` where `LABELS` has to be a function that takes an example and returns a class index. It can also return a float for regression tasks, or a fixed-size list for multi-label classification. `TEXT1` takes a dataset example as input and returns the text extracted from the example. `TEXT2` is optional and is here to leverage the fact that most text encoders process text pairs with special care.

`MULTIPLECHOICE(PROMPT, CHOICES, LABELS)`: `CHOICES` has to be a function that returns a list of text choices (the number of choices can differ across examples) extracted from an example. For concision, it can also be a list of column names to denote a list of textual choices already available in the example. `LABELS` has to return the index of the correct choice (most tasks have only one correct answer).

`TOKENCLASSIFICATION(TOKENS, LABELS)` where `TOKENS` takes an example as input and returns to a list of already split tokens, `LABELS` return a list of labels aligned to the tokens ( $i^{th}$  label annotates the  $i^{th}$  token).

We also provide 3 structured function factories to cover additional use cases while exposing their behavior with parameters.

`get` enables to access nested objects. `get.questions.text[0]` is equivalent to `lambda x:x['questions']['text'][0]`

`constant` provides constant functions. `constant(x)` is equivalent to `lambda _:x`.

`cat` concatenates multiple columns that contain strings. `cat(col1 col2)` is equivalent to `lambda x:x[col1]+x[col2]`.

An annotation to parse the Quora dataset in Table 1 can then be written as follows:

```
quora = Classification(  
    text1=get.questions.text[0],  
    text2=get.questions.text[1],  
    labels='is_duplicate')
```

For completeness, we also allow optional `preprocess` and `postprocess` arguments to a task type. They should be functions that take the full dataset as input and return a dataset. We found this feature to be necessary in a few cases where datasets had unusable labels (e.g. negative label indexes) that caused errors, or to edit the metadata of a dataset, like the name of the labels when it needs to be changed.

#### 5 Tasksource annotations

We select English datasets available on the HuggingFace Datasets Hub. We only consider discriminative tasks (Classification, Multiple-choice, Token Classification). We crawled all the tasks tagged with the English Language, and the Text-Classification task type<sup>6</sup> or Multiple Choice tag<sup>7</sup>, as of January 2023.

As many tags are missing, to increase the coverage, we crawled the 1000 most popular datasets and used heuristics to identify discriminative tasks with labels with their fields names. We then ran a `fasttext` (Joulin et al., 2016) `langid` classifier to filter out untagged datasets with non-English text.

We only annotate datasets that do not require the user to manually download data or sign an agreement. We exclude datasets that require a particular library, with the exception of BIG-bench. We also exclude tasks where high accuracy is not desirable, such as bias probing tasks (Nangia et al., 2020) where accuracy measures bias, and tasks with input length that mostly exceeds 256 tokens.

We manually deduplicate the datasets which can be available individually or in benchmarks.

We also annotate the mapping between split names and train/validation/test splits. When the test splits are obfuscated (labels unavailable), we split the validation set and use half of it as a test set. Our goal is to reduce friction and individually submitting model test predictions to data owners

<sup>6</sup>[https://hf.co/datasets?language=language:en&task\\_categories=task\\_categories:text-classification&sort=downloads](https://hf.co/datasets?language=language:en&task_categories=task_categories:text-classification&sort=downloads)

<sup>7</sup>[https://hf.co/datasets?task\\_categories=task\\_categories:multiple-choice&sort=downloads](https://hf.co/datasets?task_categories=task_categories:multiple-choice&sort=downloads)

can take a lot of time. When no split is available, we do a 80/10/10% split. We use a fixed 0 random seed. to help reproducibility.

Label handling was one of the pain points of the testing of the preprocessing functions. The tasksource backend preprocesses text labels to map them to integers.

The Table in Appendix A enumerates all datasets annotated in the current version of tasksource<sup>8</sup>

## 6 Pretraining a model on tasksource

To demonstrate the potential of tasksource, we fine-tune a single `deberta-base-v3` (He et al., 2021)<sup>9</sup> text encoder on all tasksource tasks.

Following BERT (Devlin et al., 2019) standard setup, for token-classification tasks, we use a softmax classifier on top of the last layer encoded tokens to predict the token classes. For classification tasks and multiple-choice tasks, we use a classifier on top of the [CLS] sentinel token last layer.

We assign each task a different classification layer, but we tie the label weights (not biases) to each other if they are all identical.

We oversample datasets by a factor of 2 if they have less than 64k examples then cap dataset size to 64k examples to foster dataset diversity. We randomly sample a task for each batch with a frequency proportional to the capped training dataset size and we add a learnable task-specific sentinel token to the shared sentinel token. We drop the task-specific token 10% of the time to teach the model to also work without these task embeddings, to reduce mismatch when using our model with the vanilla DeBERTa architecture. We also noticed that this tended to improve general accuracy, since this forces cooperation across tasks.

We limit the number of choices to 4 for multiple-choice tasks, to limit redundant computations, as some datasets have more than 100 choices.

We use a learning rate of  $3.10^{-5}$ , a sequence length of 256, and a batch size of 24, with 16 accumulation steps to stabilize the multi-task optimization (Yu et al., 2020a). We did not perform hyperparameter optimization.

We used the `tasknet`<sup>10</sup> library and a single

<sup>8</sup>Annotations: <https://github.com/sileood/tasksource/blob/main/src/tasksource/tasks.py>

<sup>9</sup>This is the best-performing unsupervisedly pretrained text encoder of this size according to the GLUE Benchmark (Wang et al., 2019a).

<sup>10</sup>Tasknet (Sileo, 2023) is interface Huggingface Datasets

RTX-6000 24GB GPU for 7 days (20k steps). Using tasksource with tasknet enables concise multi-task training<sup>11</sup>.

## 7 Results

As of January 2023, an early version of our model ranks first among 3574 base-sized<sup>12</sup> model on the *Model Recycling* (Choshen et al., 2022) external evaluation<sup>13</sup> This evaluation comprises 36 representative English NLP tasks (Consisting of sentiment, NLI, Twitter, topic classification, and other general classification tasks), over 5 random seeds. These results are competitive with `deberta-large` models on GLUE. We did not observe any sign of overfitting yet which suggests that the network might still be undertrained.

## 8 Conclusion

We described a semantic, structured, concise, expressive dataset preprocessing annotation framework, which is associated with a parser and annotations, that can greatly facilitate new experiments for multi-task learning and improve reproducibility. We only scratched the surface of the potential of this generated task collection due to computational limitations. For future work, we plan to use tasksource to fully automate dataset parsing on new datasets with machine learning techniques to learn the parsing process. We also plan to work on a multilingual extension of tasksource annotations.

## References

- 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems.
- 2017. Software applications user reviews.
- 2018. Sentimental analysis of tweets for detecting hate/racist speeches.
- 2019. Winogrande: An adversarial winograd schema challenge at scale.
- Zeinab Aghahadi and Alireza Talebpour. 2022. Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning.

with Huggingface Trainer.

<sup>11</sup>[https://colab.research.google.com/drive/1iB40x19\\_B5W3ZDzXoWJN-olUbqLBxgQS?usp=sharing](https://colab.research.google.com/drive/1iB40x19_B5W3ZDzXoWJN-olUbqLBxgQS?usp=sharing)

<sup>12</sup>This corresponds to 86M encoder parameters excluding embeddings.

<sup>13</sup><https://ibm.github.io/model-recycling>

- Journal of Applied Non-Classical Logics*, pages 1–17.
- Armen Aghajanyan, Ancht Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. [Muppet: Massive multi-task representations with pre-finetuning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tiago A. Almeida, Jose Maria Gomez Hidalgo, and Akebo Yamakami. 2011. Contributions to the study of sms spam filtering: New collection and results. In *Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11)*.
- 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*.
- Stephane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. [PROST: Physical reasoning about objects through space and time](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4597–4608, Online. Association for Computational Linguistics.
- Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. [Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).
- Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. [Promptsources: An integrated development environment and repository for natural language prompts](#).
- Qiming Bao, Alex Yuxuan Peng, Tim Hartill, Neseet Tan, Zhenyun Deng, Michael Witbrock, and Jiamou Liu. 2022. Multi-step deductive reasoning over natural language: An empirical study on out-of-distribution generalisation. The 2nd International Joint Conference on Learning and Reasoning and 16th International Workshop on Neural-Symbolic Learning and Reasoning (IJCLR-NeSy 2022).
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. [TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification](#). In *Proceedings of Findings of EMNLP*.
- Qiang Ning Ben Zhou, Daniel Khashabi and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. In *EMNLP*.
- Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. [Starc: Structured annotations for reading comprehension](#). In *ACL*. Association for Computational Linguistics.
- Sumithra Bhakthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. 2020. [Do dogs have whiskers? a new knowledge base of haspart relations](#).
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. [Piqa: Reasoning about physical commonsense in natural language](#). In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. [Nuanced metrics for measuring unintended bias with real data for text classification](#). *CoRR*, abs/1903.04561.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Sven Buechel and Udo Hahn. 2017. [EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.
- Shuyang Cao and Lu Wang. 2021. [Controllable open-ended question generation with a new question type ontology](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439, Online. Association for Computational Linguistics.

- Winston Carlike, Nishant Gurrupadi, Zixuan Ke, and Vincent Ng. 2018. [Give me more feedback: Annotating argument persuasiveness and related attributes in student essays](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.
- Rich Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *International Conference on Machine Learning*.
- Inigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. [Efficient intent detection with dual sentence encoders](#). In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*. Data available at <https://github.com/PolyAI-LDN/task-specific-datasets>.
- Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. [SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.
- Bharathi Raja Chakravarthi. 2020. [HopeEDI: A multilingual hope speech detection dataset for equality, diversity and inclusion](#). In *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality and Emotion’s in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.
- Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.
- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. [LexGLUE: A benchmark dataset for legal language understanding in English](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.
- Bhagavatula Chandra, Le Bras Ronan, Malaviya Chaitanya, Sakaguchi Keisuke, Holtzman Ari, Rashkin Hannah, Downey Doug, Wen-tau Yih Scott, and Choi Yejin. 2020. Abductive commonsense reasoning.
- Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloe Clavel. 2020. [Hierarchical pre-training for sequence labelling in spoken dialog](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.
- Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. [Semeval-2019 task 3: Emocontext contextual emotion detection in text](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Michael Chen, Mike Darcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019a. Codah: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69.
- Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019b. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proc. of EMNLP*.
- Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.
- Leshem Choshen, Elad Venezian, Shachar Don-Yehia, Noam Slonim, and Yoav Katz. 2022. Where to start? analyzing the potential value of intermediate models. *arXiv preprint arXiv:2211.00107*.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.
- Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *NAACL*.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. [Supervised learning of universal sentence representations from natural language inference data](#). In *Proceedings of EMNLP2017*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. 2018. [What you can cram into a single \\$&!#\\* vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.
- Alice Coucke, Alaa Saade, Adrien Ball, Theodore Bluche, Alexandre Caulier, David Leroy, Clement Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Mael Primet, and Joseph

- Dureau. 2018. [Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces](#). *CoRR*, abs/1805.10190.
- Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. [Mutual: A dataset for multi-turn dialogue reasoning](#). In *Proceedings of the 58th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. [Automated hate speech detection and the problem of offensive language](#). In *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17*, pages 512–515.
- Ona de Gibert, Naiara Perez, Aitor Garcia-Pablos, and Montse Cuadros. 2018. [Hate Speech Dataset from a White Supremacy Forum](#). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.
- Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. [The commitmentbank: Investigating projection in naturally occurring discourse](#). In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A Dataset of Fine-Grained Emotions](#). In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. [Results of the WNUT2017 shared task on novel and emerging entity recognition](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of NAACL2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. [Ncbi disease corpus: a resource for disease name recognition and concept normalization](#). *Journal of biomedical informatics*, 47:1–10.
- Mai ElSherief, Caleb Ziemis, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. [Moral stories: Situated reasoning about norms, intents, actions and their consequences](#). *ArXiv*, abs/2012.15738.
- Kawin Ethayarajh, Heidi Zhang, Yizhong Wang, and Dan Jurafsky. 2023. [Stanford human preferences dataset](#).
- Manaal Faruqui and Dipanjan Das. 2018. [Identifying well-formed natural language questions](#).
- Ziou Zheng Feng, Yufei, Quan Liu, Michael Greenspan, and Xiaodan Zhu. 2020. [Exploring end-to-end differentiable natural logic modeling](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1172–1185.
- William Ferreira and Andreas Vlachos. 2016. [Emergent: a novel data-set for stance classification](#). In *HLT-NAACL*.
- Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sanger, Bo Wang, Alison Callahan, Daniel Le3n Periann, Th3o Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pamies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. [Bigbio: A framework for data-centric biomedical natural language processing](#). In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Elizabeth Jasmi George and Radhika Mamidi. 2020. [Conversational implicatures in english dialogue: Annotated dataset](#). *Procedia Computer Science*, 171:2316–2323.
- Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. [Cicero: A dataset for contextualized commonsense inference in dialogues](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5010–5028.
- John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. [Switchboard: Telephone speech corpus for research and development](#). In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1, ICASSP'92*, pages 517–520, Washington, DC, USA. IEEE Computer Society.



- Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. [SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. [Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports](#). *Journal of Biomedical Informatics*, 45(5):885–892. Text Mining and Natural Language Processing in Pharmacogenomics.
- Ilya Gusev and Alexey Tikhonov. 2021. [Headlinecause: A dataset of news headlines for detecting casualties](#).
- Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. [The argument reasoning comprehension task: Identification and reconstruction of implicit warrants](#). 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 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.
- Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. [Folio: Natural language reasoning with first-order logic](#). *arXiv preprint arXiv:2209.00840*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *arXiv preprint arXiv:2111.09543*.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Seaghdha, Sebastian Pado, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. [SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals](#). In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. [Aligning ai with shared human values](#). *arXiv preprint arXiv:2008.02275*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Nabil Hossain, John Krumm, and Michael Gamon. 2019. ["president vows to cut taxes hair": Dataset and analysis of creative text editing for humorous headlines](#). *arXiv preprint arXiv:1906.00274*.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. [Cosmos QA: Machine reading comprehension with contextual commonsense reasoning](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.
- William Huang, Haokun Liu, and Samuel R. Bowman. 2020. [Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data](#). In *Proceedings of the 2020 EMNLP Workshop on Insights from Negative Results in NLP*. The Association for Computational Linguistics.
- Inc. huggingface. 2020. [A great new dataset](#).
- Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. [Are natural language inference models IMPPRESSive? Learning IMPLICature and PRESupposition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.
- Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. [Crowdsourcing multiple choice science questions](#).
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou, and Tomas Mikolov. 2016. [Fasttext.zip: Compressing text classification models](#). *arXiv preprint arXiv:1612.03651*.
- Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. [Learning the difference that makes a difference with counterfactually augmented data](#). *International Conference on Learning Representations (ICLR)*.
- Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reiser, and Kentaro Inui. 2019. [When choosing plausible alternatives, clever hans can be clever](#). In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.
- Mayank Kejriwal and Ke Shen. 2020. [Do fine-tuned commonsense language models really generalize?](#) *ArXiv*, abs/2011.09159.
- D. Khashabi, T. Khot, and A. Sabharwal. 2020a. [Natural perturbation for robust question answering](#). *arXiv preprint*.

- D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi. 2020b. Uni-fiedqa: Crossing format boundaries with a single qa system.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Tushar Khot, Peter Clark, Michal Guerin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*.
- Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.
- Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer.
- Neema Kotonya and Francesca Toni. 2020. [Explaining automated fact-checking for public health claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.
- Philippe Laban and Lucas Bandarkar. 2021. News headline grouping as a challenging nlu task. In *NAACL 2021*. Association for Computational Linguistics.
- Shibamouli Lahiri. 2015. [SQUINKY! A Corpus of Sentence-level Formality, Informativeness, and Implicature](#). *CoRR*, abs/1506.02306.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.
- Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. [Unsupervised prediction of acceptability judgements](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628, Beijing, China. Association for Computational Linguistics.
- Moritz Laurer, W v Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating, more classifying—addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli.
- Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues. In *Proceedings of the corpus linguistics 2003 conference*, volume 16, pages 441–446. Lancaster: Lancaster University.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Soren Auer, et al. 2015. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.
- Xin Li and Dan Roth. 2002. [Learning question classifiers](#). In *COLING 2002: The 19th International Conference on Computational Linguistics*.
- 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 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.
- Yichan Liang, Jianheng Li, and Jian Yin. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, pages 742–757.
- Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In *Proceedings of EMNLP*. To appear.
- Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021a. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2021b. [Truthfulqa: Measuring how models mimic human falsehoods](#).
- Marco Lippi, Przemyslaw Paka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. [CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service](#). *Artificial Intelligence and Law*, pages 117–139.
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. [Wanli: Worker and ai collaboration for natural language inference dataset creation](#).
- Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. [We’re afraid language models aren’t modeling ambiguity](#).
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

- Annie Louis, Dan Roth, and Filip Radlinski. 2020. I’d rather just go to bed: Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *CoRR*, abs/1902.01007.
- Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. Effective transfer learning for identifying similar questions: Matching user questions to covid-19 faqs.
- Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.
- Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.
- Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.
- Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. In *arXiv*.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- James O’Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. 2021. I wish i would have loved this one, but i didn’t – a multilingual dataset for counterfactual detection in product reviews.
- Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41. Association for Computational Linguistics.
- Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale

- multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.
- Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the first workshop on argumentation mining*, pages 29–38.
- Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*.
- Ethan Perez, Sam Ringer, Kamil Lukoic, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noem Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer E, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering language model behaviors with model-written evaluations.
- Thang M Pham, Seunghyun Yoon, Trung Bui, and Anh Nguyen. 2022. Pic: A phrase-in-context dataset for phrase understanding and semantic search. *arXiv preprint arXiv:2207.09068*.
- Mohammad Taher Pilehvar and osee Camacho-Collados. 2018. Wic: 10, 000 example pairs for evaluating context-sensitive representations. *CoRR*, abs/1808.09121.
- Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.
- Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. DynaSent: A dynamic benchmark for sentiment analysis. *arXiv preprint arXiv:2012.15349*.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 2020. jiant: A software toolkit for research on general-purpose text understanding models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–117, Online. Association for Computational Linguistics.
- Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789. Association for Computational Linguistics.
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics (ACL2019)*.
- Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. Condaqa: A contrastive reading comprehension dataset for reasoning about negation.
- Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference. *arXiv preprint arXiv:1901.03735*.
- Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.
- Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference*

- on Artificial Intelligence, *AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin V. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana. Association for Computational Linguistics.
- Mobashir Sadat and Cornelia Caragea. 2022. [SciNLI: A corpus for natural language inference on scientific text](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.
- Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *Int. J. Digit. Libr.*
- Viktor Schlegel, Kamen V. Pavlov, and Ian Pratt-Hartmann. 2022. [Can transformers reason in fragments of natural language?](#)
- Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.
- Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. [Get your vitamin C! robust fact verification with contrastive evidence](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.
- Emily Sheng and David Uthus. 2020. [Investigating societal biases in a poetry composition system](#).
- Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.
- Damien Sileo. 2023. [tasknet, multitask interface between Trainer and datasets](#).
- Damien Sileo and Antoine Lerno. 2023. Mindgames: Targeting theory of mind in large language models with dynamic epistemic modal logic. *arXiv preprint arXiv:2305.03353*.
- Damien Sileo and Marie-Francine Moens. 2022a. [Analysis and prediction of NLP models via task embeddings](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 633–647, Marseille, France. European Language Resources Association.
- Damien Sileo and Marie-Francine Moens. 2022b. Probing neural language models for understanding of words of estimative probability. *arXiv preprint arXiv:2211.03358*.
- Damien Sileo, Philippe Muller, Tim Van de Cruys, and Camille Pradel. 2022a. [A pragmatics-centered evaluation framework for natural language understanding](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2382–2394, Marseille, France. European Language Resources Association.
- Damien Sileo, Kanimozhi Uma, and Marie-Francine Moens. 2023. Generating multiple-choice questions for medical question answering with distractors and cue-masking. *arXiv preprint arXiv:2303.07069*.
- Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. [Mining discourse markers for unsupervised sentence representation learning](#). 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 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.
- Damien Sileo, Wout Vossen, and Robbe Raymaekers. 2022b. Zero-shot recommendation as language modeling. In *Advances in Information Retrieval*, pages 223–230, Cham. Springer International Publishing.
- Harold J. Spaeth, Lee Epstein, Jeffrey A. Segal Andrew D. Martin, Theodore J. Ruger, and Sara C. Benesh. 2020. [Supreme Court Database, Version 2020 Release 01](#). Washington University Law.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, et al. 2022a. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, et al. 2022b. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *NeurIPS*.
- Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. [DREAM: A challenge dataset and models for dialogue-based reading](#)

- comprehension. *Transactions of the Association for Computational Linguistics*.
- Roxana Szomiu and Adrian Groza. 2021. A puzzle-based dataset for natural language inference. *arXiv preprint arXiv:2112.05742*.
- Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. "quartz: An open-domain dataset of qualitative relationship questions".
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. **CommonsenseQA: A question answering challenge targeting commonsense knowledge**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Niket Tandon, Bhavana Dalvi Mishra, Keisuke Sakaguchi, Antoine Bosselut, and Peter Clark. 2019. Wiqa: A dataset for "what if..." reasoning over procedural text. *arXiv:1909.04739v1*.
- Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. 2021. Trusting roberta over bert: Insights from checklisting the natural language inference task. *ArXiv*, abs/2107.07229.
- Henry S Thompson, Anne H Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The hrc map task corpus: natural dialogue for speech recognition. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.
- Shivin Thukral, Kunal Kukreja, and Christian Kavouras. 2021. **Probing language models for understanding of temporal expressions**. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 396–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. **Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition**. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Don Tugener, Pius von Daniken, Thomas Peetz, and Mark Cieliebak. 2020. **LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts**. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France.
- Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.
- Chris Van Pelt and Alex Sorokin. 2012. Designing a scalable crowdsourcing platform. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 765–766.
- Amir Pouran Ben Veyseh, Franck Dernoncourt, Quan Hung Tran, and Thien Huu Nguyen. 2020. What Does This Acronym Mean? Introducing a New Dataset for Acronym Identification and Disambiguation. In *Proceedings of COLING*.
- Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682.
- David Vilares and Carlos Gomez-Rodriguez. 2019. **HEAD-QA: A healthcare dataset for complex reasoning**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.
- Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2022. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.
- Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2019b. **SphereRE: Distinguishing lexical relations with hyperspherical relation embeddings**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1727–1737, Florence, Italy. Association for Computational Linguistics.
- William Yang Wang. 2017. **"liar, liar pants on fire": A new benchmark dataset for fake news detection**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoor-molabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: generalization via declarative instructions on 1600+ tasks. In *EMNLP*.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2019. Blimp: A benchmark of linguistic minimal pairs for english. *arXiv preprint arXiv:1912.00582*.

- Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.
- Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. [Constructing datasets for multi-hop reading comprehension across documents](#).
- Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of NAACL2018*, pages 1112–1122. Association for Computational Linguistics.
- Thomas Wolf, Quentin Lhoest, Patrick von Platen, Yacine Jernite, Mariama Drame, Julien Plu, Julien Chaumond, Clement Delangue, Clara Ma, Abhishek Thakur, Suraj Patil, Joe Davison, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angie McMillan-Major, Simon Brandeis, Sylvain Gugger, François Lagunas, Lysandre Debut, Morgan Funtowicz, Anthony Moi, Sasha Rush, Philipp Schmid, Pierric Cistac, Victor Muštar, Jeff Boudier, and Anna Tordjmann. 2020. Datasets. *GitHub*. Note: <https://github.com/huggingface/datasets>, 1.
- Dustin Wright and Isabelle Augenstein. 2021. Semi-Supervised Exaggeration Detection of Health Science Press Releases. In *Proceedings of EMNLP*. Association for Computational Linguistics.
- Yang Y, Yih W, and C Meek. 2015. [WikiQA: A Challenge Dataset for Open-Domain Question Answering](#). page 2013–2018.
- Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40.
- Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. Help: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (\*SEM2019)*.
- Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Exploring transitivity in neural NLI models through veridicality. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 920–934.
- Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. [CrossFit: A few-shot learning challenge for cross-task generalization in NLP](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020a. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836.
- Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020b. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*.
- Amir Zeldes. 2017. [The GUM corpus: Creating multilayer resources in the classroom](#). *Language Resources and Evaluation*, 51(3):581–612.
- 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)*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noemi Aepli, Hamid Aghaei, veljko Agic, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabriele Aleksandraviciute, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, torunn Arnardottir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gozde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agne Bielinskiene, Kristin Bjarnadottir, Rogier Blokland, Victoria Bobicev, Loic Boizou, Emanuel Borges Volker, Carl Borstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaite, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gulsen Cebiroglu Eryigit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomir Ceplo, Savas Cetin, Ozlem Cetinoglu, Fabricio Chalub, Ethan Chi, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinkova, Aurelie Collomb, Cagri Coltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat,

Kira Drojanova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaz Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richard Farkas, Marilia Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Claudia Freitas, Kazunori Fujita, Katarina Gajdosova, Daniel Galbraith, Marcos Garcia, Moa Gardenfors, Sebastian Garza, Fabricio Ferraz Gerardi, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gokirmak, Yoav Goldberg, Xavier Gomez Guinovart, Berta Gonzalez Saavedra, Bernadeta Grieciute, Matias Grioni, Loic Grobol, Normunds Gruzitis, Bruno Guillaume, Celine Guillot-Barbance, Tunga Gungor, Nizar Habash, Hinrik Hafsteinsson, Jan Hajiv, Jan Hajiv jr., Mika Hamalainen, Linh Ha My, Na-Rae Han, Muhammad Yulistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladka, Jaroslava Hlavavova, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, dlajide Ishola, Tomav Jelinek, Anders Johannsen, Hildur Jonsdottir, Fredrik Jorgensen, Markus Juutinen, Sarveswaran K, Huner Kacikara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Vaclava Kettnerova, Jesse Kirchner, Elena Klementieva, Arne Kohn, Abdullatif Koksul, Kamil Kopacewicz, Timo Korhakangas, Natalia Kotsyba, Jolanta Kovalevskaite, Simon Krek, Parameswari Krishnamurthy, Sookyoung Kwak, Veronika Laipala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Le Hong, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, K Lim, Krister Linden, Nikola Ljubetic, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashenskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Catalina Maranduc, David Marcek, Katrin Marheinecke, Hector Martinez Alonso, Andre Martins, Jan Masek, Hiroshi Matsuda, Yuji Matsumoto, Ryan M, Sarah M, Gustavo Mendonca, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missila, Catalin Mititelu, Maria Mitrofan, Yusuke Miyao, A Mojiri Foroushani, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Muurisep, Pinkey Nainwani, Mariam Nakhle, Juan Ignacio Navarro Horniackek, Anna Nedoluzhko, Gunta Nevpore-Berzkalne, Lng Nguyen Thd, Huyen Nguyen Thd Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adedayd Oluokun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Ostling, Lilja Ovreid, caziye Betul Ozatec, Arzucan Ozgur, Balkiz Ozturk Bacaran, Niko Partanen, Elena Pascual,

Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Lapinska, Siyao Peng, Cenal-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalnina, Sophie Prevost, Prokopis Prokopidis, Adam Przepiorkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Raabis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Riesler, Erika Rimkute, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eirikur Rognvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roca, Davide Rovati, Olga Rudina, Jack Rueter, Kristjan Runarsson, Shoal Sadde, Pegah Safari, Benoit Sagot, Aleks Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardzic, Stephanie Samson, Manuela Sanguinetti, Dage Sarg, Baiba Saulite, Yanin Sawanakunanon, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djame Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Einar Freyr Sigursson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simko, Maria vimkova, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Steintor Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadova, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szanto, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Lisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Turk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdenka Uresova, Larraitz Uribe, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilian Wendt, Paul Widmer, Seyi Williams, Mats Wiren, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wroblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdenek Zabokrtsky, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. [Universal dependencies 2.7](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun.



- 2015b. Character-level convolutional networks for text classification. In *NIPS*.
- Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.
- Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pre-training help? assessing self-supervised learning for law and the casehold dataset. In *International Conference on Artificial Intelligence and Law*.
- Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed nli: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

## A Currently annotated preprocessings

	preprocessing	task type
0	glue/mnli (Williams et al., 2018)	Classification
1	glue/qnli (Williams et al., 2018)	Classification
2	glue/rte (Williams et al., 2018)	Classification
3	glue/wnli (Williams et al., 2018)	Classification
4	glue/mrpc (Williams et al., 2018)	Classification
5	glue/qqp (Williams et al., 2018)	Classification
6	glue/stsb (Williams et al., 2018)	Classification
7	super_glue/boolq (Clark et al., 2019)	Classification
8	super_glue/cb (De Marneffe et al., 2019)	Classification
9	super_glue/multirc (Khashabi et al., 2018)	Classification
10	super_glue/wic (Pilehvar and ose Camacho-Collados, 2018)	Classification
11	super_glue/axg (Rudinger et al., 2018)	Classification
12	anli/a1 (Nie et al., 2020)	Classification
13	anli/a2 (Nie et al., 2020)	Classification
14	anli/a3 (Nie et al., 2020)	Classification
15	babi_nli/lists-sets (Weston et al., 2015)	Classification
16	babi_nli/basic-deduction (Weston et al., 2015)	Classification
17	babi_nli/positional-reasoning (Weston et al., 2015)	Classification
18	babi_nli/basic-coreference (Weston et al., 2015)	Classification
19	babi_nli/three-supporting-facts (Weston et al., 2015)	Classification
20	babi_nli/path-finding (Weston et al., 2015)	Classification
21	babi_nli/three-arg-relations (Weston et al., 2015)	Classification
22	babi_nli/yes-no-questions (Weston et al., 2015)	Classification
23	babi_nli/time-reasoning (Weston et al., 2015)	Classification
24	babi_nli/indefinite-knowledge (Weston et al., 2015)	Classification
25	babi_nli/counting (Weston et al., 2015)	Classification
26	babi_nli/size-reasoning (Weston et al., 2015)	Classification
27	babi_nli/compound-coreference (Weston et al., 2015)	Classification
28	babi_nli/basic-induction (Weston et al., 2015)	Classification
29	babi_nli/single-supporting-fact (Weston et al., 2015)	Classification
30	babi_nli/simple-negation (Weston et al., 2015)	Classification
31	babi_nli/two-arg-relations (Weston et al., 2015)	Classification
32	babi_nli/two-supporting-facts (Weston et al., 2015)	Classification
33	babi_nli/conjunction (Weston et al., 2015)	Classification
34	sick/label (Marelli et al., 2014)	Classification
35	sick/relatedness (Marelli et al., 2014)	Classification
36	sick/entailment_AB (Marelli et al., 2014)	Classification
37	snli (Bowman et al., 2015)	Classification
38	scitail/snli_format (Khot et al., 2018)	Classification
39	hans (McCoy et al., 2019)	Classification
40	WANLI (Liu et al., 2022)	Classification
41	recast/recast_verbcorner (Poliak et al., 2018)	Classification
42	recast/recast_megaveridicality (Poliak et al., 2018)	Classification
43	recast/recast_sentiment (Poliak et al., 2018)	Classification
44	recast/recast_ner (Poliak et al., 2018)	Classification
45	recast/recast_kg_relations (Poliak et al., 2018)	Classification

Continued on next page

	preprocessing	task type
46	recast/recast_factuality (Poliak et al., 2018)	Classification
47	recast/recast_puns (Poliak et al., 2018)	Classification
48	recast/recast_verbnet (Poliak et al., 2018)	Classification
49	probability_words_nli/reasoning_1hop (Sileo and Moens, 2022b)	Classification
50	probability_words_nli/usnli (Sileo and Moens, 2022b)	Classification
51	probability_words_nli/reasoning_2hop (Sileo and Moens, 2022b)	Classification
52	nan-nli/joey234–nan-nli	Classification
53	nli_fever	Classification
54	breaking_nli	Classification
55	conj_nli	Classification
56	fracas	Classification
57	dialogue_nli	Classification
58	mpe	Classification
59	dnc	Classification
60	recast_white/fnplus	Classification
61	recast_white/spri	Classification
62	recast_white/dpr	Classification
63	joci	Classification
64	robust_nli/IS_CS	Classification
65	robust_nli/LI_LI	Classification
66	robust_nli/ST_WO	Classification
67	robust_nli/PI_SP	Classification
68	robust_nli/PI_CD	Classification
69	robust_nli/ST_SE	Classification
70	robust_nli/ST_NE	Classification
71	robust_nli/ST_LM	Classification
72	robust_nli_is_sd	Classification
73	robust_nli_li_ts	Classification
74	gen_debiased_nli/snli_seq_z	Classification
75	gen_debiased_nli/snli_z_aug	Classification
76	gen_debiased_nli/snli_par_z	Classification
77	gen_debiased_nli/mnli_par_z	Classification
78	gen_debiased_nli/mnli_z_aug	Classification
79	gen_debiased_nli/mnli_seq_z	Classification
80	add_one_rte	Classification
81	imppres/presupposition_all_n_presupposition (Jereti c et al., 2020)	Classification
82	imppres/presupposition_possessed_definites_existence (Jereti c et al., 2020)	Classification
83	imppres/presupposition_cleft_uniqueness(Jereti c et al., 2020)	Classification
84	imppres/presupposition_question_presupposition(Jereti c et al., 2020)	Classification
85	imppres/presupposition_possessed_definites_uniqueness(Jereti c et al., 2020)	Classification
86	imppres/presupposition_only_presupposition(Jereti c et al., 2020)	Classification
87	imppres/presupposition_both_presupposition(Jereti c et al., 2020)	Classification
88	imppres/presupposition_change_of_state(Jereti c et al., 2020)	Classification
89	imppres/presupposition_cleft_existence(Jereti c et al., 2020)	Classification
90	imppres/implicature_quantifiers/prag (Jereti c et al., 2020)	Classification
91	imppres/implicature_numerals_2_3/prag (Jereti c et al., 2020)	Classification
92	imppres/implicature_numerals_10_100/prag (Jereti c et al., 2020)	Classification
93	imppres/implicature_modals/prag (Jereti c et al., 2020)	Classification

Continued on next page

	preprocessing	task type
94	imppres/implicature_connectives/prag (Jereti c et al., 2020)	Classification
95	imppres/implicature_gradable_verb/prag (Jereti c et al., 2020)	Classification
96	imppres/implicature_gradable_adjective/prag (Jereti c et al., 2020)	Classification
97	imppres/implicature_quantifiers/log (Jereti c et al., 2020)	Classification
98	imppres/implicature_numerals_2_3/log (Jereti c et al., 2020)	Classification
99	imppres/implicature_numerals_10_100/log (Jereti c et al., 2020)	Classification
100	imppres/implicature_gradable_adjective/log (Jereti c et al., 2020)	Classification
101	imppres/implicature_connectives/log (Jereti c et al., 2020)	Classification
102	imppres/implicature_modals/log (Jereti c et al., 2020)	Classification
103	imppres/implicature_gradable_verb/log (Jereti c et al., 2020)	Classification
104	glue_diagnostics/diagnostics	Classification
105	hlgd (Laban and Bandarkar, 2021)	Classification
106	paws/labeled_final (Zhang et al., 2019)	Classification
107	paws/labeled_swap (Zhang et al., 2019)	Classification
108	quora	Classification
109	medical_questions_pairs (McCreery et al., 2020)	Classification
110	glue/cola (Williams et al., 2018)	Classification
111	glue/sst2 (Williams et al., 2018)	Classification
112	utilitarianism (Hendrycks et al., 2020)	Classification
113	amazon_counterfactual/en (ONeill et al., 2021)	Classification
114	insincere-questions	Classification
115	toxic_conversations	Classification
116	TuringBench (huggingface, 2020)	Classification
117	trec (Li and Roth, 2002)	Classification
118	vitaminc/tals-vitaminc (Schuster et al., 2021)	Classification
119	hope_edi/english (Chakravarthi, 2020)	Classification
120	rumoureval_2019/RumourEval2019 (Gorrell et al., 2019)	Classification
121	ethos/binary (Mollas et al., 2020)	Classification
122	ethos/multilabel (Mollas et al., 2020)	Classification
123	tweet_eval/emotion (Barbieri et al., 2020)	Classification
124	tweet_eval/irony (Barbieri et al., 2020)	Classification
125	tweet_eval/offensive (Barbieri et al., 2020)	Classification
126	tweet_eval/sentiment (Barbieri et al., 2020)	Classification
127	tweet_eval/stance_abortion (Barbieri et al., 2020)	Classification
128	tweet_eval/stance_atheism (Barbieri et al., 2020)	Classification
129	tweet_eval/stance_climate (Barbieri et al., 2020)	Classification
130	tweet_eval/stance_feminist (Barbieri et al., 2020)	Classification
131	tweet_eval/stance_hillary (Barbieri et al., 2020)	Classification
132	tweet_eval/emoji (Barbieri et al., 2020)	Classification
133	tweet_eval/hate (Barbieri et al., 2020)	Classification
134	discovery/discovery (Sileo et al., 2019)	Classification
135	pragmeval/squinky-informativeness (Lahiri, 2015)	Classification
136	pragmeval/squinky-implicature (Lahiri, 2015)	Classification
137	pragmeval/verifiability (Park and Cardie, 2014)	Classification
138	pragmeval/squinky-formality (Lahiri, 2015)	Classification
139	pragmeval/emobank-valence (Buechel and Hahn, 2017)	Classification
140	pragmeval/emobank-dominance (Buechel and Hahn, 2017)	Classification
141	pragmeval/emobank-arousal (Buechel and Hahn, 2017)	Classification

Continued on next page

	preprocessing	task type
142	pragmeval/switchboard (Godfrey et al., 1992)	Classification
143	pragmeval/mrda (Shriberg et al., 2004)	Classification
144	pragmeval/sarcasm (Oraby et al., 2016)	Classification
145	pragmeval/persuasiveness-premisetype (Carlile et al., 2018)	Classification
146	pragmeval/persuasiveness-eloquence (Carlile et al., 2018)	Classification
147	pragmeval/persuasiveness-claimtype (Carlile et al., 2018)	Classification
148	pragmeval/persuasiveness-specificity (Carlile et al., 2018)	Classification
149	pragmeval/gum (Zeldes, 2017)	Classification
150	pragmeval/emergent (Ferreira and Vlachos, 2016)	Classification
151	pragmeval/persuasiveness-strength (Carlile et al., 2018)	Classification
152	pragmeval/stac (Asher et al., 2016)	Classification
153	pragmeval/pdtb (Prasad et al., 2008)	Classification
154	pragmeval/persuasiveness-relevance (Carlile et al., 2018)	Classification
155	silicone/meld_s (Chen et al., 2018)	Classification
156	silicone/sem (McKeown et al., 2011)	Classification
157	silicone/oasis (Leech and Weisser, 2003)	Classification
158	silicone/meld_e (Chen et al., 2018)	Classification
159	silicone/maptask (Thompson et al., 1993)	Classification
160	silicone/iemocap (Busso et al., 2008)	Classification
161	silicone/dyda_e (Li et al., 2017)	Classification
162	silicone/dyda_da (Li et al., 2017)	Classification
163	lex_glue/eurlex (Chalkidis et al., 2021)	Classification
164	lex_glue/scotus (Spaeth et al., 2020)	Classification
165	lex_glue/ledgar (Tuggener et al., 2020)	Classification
166	lex_glue/unfair_tos (Lippi et al., 2019)	Classification
167	language-identification	Classification
168	imdb (Maas et al., 2011)	Classification
169	rotten_tomatoes (Pang and Lee, 2005)	Classification
170	ag_news (Zhang et al., 2015b)	Classification
171	yelp_review_full/yelp_review_full (Zhang et al., 2015a)	Classification
172	financial_phrasebank/sentences_allagree (Malo et al., 2014)	Classification
173	poem_sentiment (Sheng and Uthus, 2020)	Classification
174	dbpedia_14/dbpedia_14 (Lehmann et al., 2015)	Classification
175	amazon_polarity/amazon_polarity (McAuley and Leskovec, 2013)	Classification
176	app_reviews (Zur, 2017)	Classification
177	hate_speech18 (de Gibert et al., 2018)	Classification
178	sms_spam (Almeida et al., 2011)	Classification
179	humicroedit/subtask-1 (Hossain et al., 2019)	Classification
180	humicroedit/subtask-2 (Hossain et al., 2019)	Classification
181	snips_built_in_intents (Coucke et al., 2018)	Classification
182	banking77 (Casanueva et al., 2020)	Classification
183	hate_speech_offensive (Davidson et al., 2017)	Classification
184	yahoo_answers_topics	Classification
185	stackoverflow-questions	Classification
186	hyperpartisan_news	Classification
187	sciie	Classification
188	citation_intent	Classification
189	go_emotions/simplified (Demszky et al., 2020)	Classification

Continued on next page

	preprocessing	task type
190	scicite (Cohan et al., 2019)	Classification
191	liar (Wang, 2017)	Classification
192	lexical_relation_classification/K&H+N (Wang et al., 2019b)	Classification
193	lexical_relation_classification/CogALexV (Wang et al., 2019b)	Classification
194	lexical_relation_classification/BLESS (Wang et al., 2019b)	Classification
195	lexical_relation_classification/EVALution (Wang et al., 2019b)	Classification
196	lexical_relation_classification/ROOT09 (Wang et al., 2019b)	Classification
197	linguisticprobing/subj_number (Conneau et al., 2018)	Classification
198	linguisticprobing/bigram_shift (Conneau et al., 2018)	Classification
199	linguisticprobing/top_constituents (Conneau et al., 2018)	Classification
200	linguisticprobing/odd_man_out (Conneau et al., 2018)	Classification
201	linguisticprobing/past_present (Conneau et al., 2018)	Classification
202	linguisticprobing/coordination_inversion (Conneau et al., 2018)	Classification
203	linguisticprobing/tree_depth (Conneau et al., 2018)	Classification
204	linguisticprobing/obj_number (Conneau et al., 2018)	Classification
205	linguisticprobing/sentence_length (Conneau et al., 2018)	Classification
206	crowdfower/sentiment_nuclear_power (Van Pelt and Sorokin, 2012)	Classification
207	crowdfower/tweet_global_warming (Van Pelt and Sorokin, 2012)	Classification
208	crowdfower/corporate-messaging (Van Pelt and Sorokin, 2012)	Classification
209	crowdfower/economic-news (Van Pelt and Sorokin, 2012)	Classification
210	crowdfower/airline-sentiment (Van Pelt and Sorokin, 2012)	Classification
211	crowdfower/political-media-bias (Van Pelt and Sorokin, 2012)	Classification
212	crowdfower/text_emotion (Van Pelt and Sorokin, 2012)	Classification
213	crowdfower/political-media-audience (Van Pelt and Sorokin, 2012)	Classification
214	crowdfower/political-media-message (Van Pelt and Sorokin, 2012)	Classification
215	ethics/commonsense (Hendrycks et al., 2020)	Classification
216	ethics/deontology (Hendrycks et al., 2020)	Classification
217	ethics/justice (Hendrycks et al., 2020)	Classification
218	ethics/virtue (Hendrycks et al., 2020)	Classification
219	emo/emo2019 (Chatterjee et al., 2019)	Classification
220	google_wellformed_query (Faruqui and Das, 2018)	Classification
221	tweets_hate_speech_detection (ZRo, 2018)	Classification
222	has_part (Bhakhavatsalam et al., 2020)	Classification
223	blog_authorship_corpus/gender (Schler et al., 2006)	Classification
224	blog_authorship_corpus/age (Schler et al., 2006)	Classification
225	blog_authorship_corpus/horoscope (Schler et al., 2006)	Classification
226	blog_authorship_corpus/job (Schler et al., 2006)	Classification
227	open_question_type (Cao and Wang, 2021)	Classification
228	health_fact (Kotonya and Toni, 2020)	Classification
229	mc_taco (Ben Zhou and Roth, 2019)	Classification
230	ade_corpus_v2/Ade_corpus_v2_classification (Gurulingappa et al., 2012)	Classification
231	circa (Louis et al., 2020)	Classification
232	EffectiveFeedbackStudentWriting	Classification
233	promptSentiment (McAuley and Leskovec, 2013)	Classification
234	promptNLI (Nie et al., 2020)	Classification
235	promptSpoke	Classification
236	promptProficiency	Classification
237	promptGrammar (Warstadt et al., 2018)	Classification

Continued on next page

	preprocessing	task type
238	promptCoherence	Classification
239	phrase_similarity (Pham et al., 2022)	Classification
240	scientific-exaggeration-detection (Wright and Augenstein, 2021)	Classification
241	quarel	Classification
242	fever-evidence-related/mwong-fever-related	Classification
243	numer_sense (Lin et al., 2020)	Classification
244	dynasent/dynabench.dynasent.r1.all/r1 (Potts et al., 2020)	Classification
245	dynasent/dynabench.dynasent.r2.all/r2 (Potts et al., 2020)	Classification
246	Sarcasm_News_Headline	Classification
247	sem_eval_2010_task_8 (Hendrickx et al., 2010)	Classification
248	auditor_review/demo-org-auditor_review	Classification
249	Dynasent_Disagreement	Classification
250	Politeness_Disagreement	Classification
251	SBIC_Disagreement	Classification
252	SChem_Disagreement	Classification
253	Dilemmas_Disagreement	Classification
254	wiki_qa (Y et al., 2015)	Classification
255	cyclic_classification (Kejriwal and Shen, 2020)	Classification
256	sts-companion (Cer et al., 2017)	Classification
257	commonsense_qa_2.0	Classification
258	lingnli (Parrish et al., 2021)	Classification
259	monotonicity-entailment (Yanaka et al., 2019a)	Classification
260	scinli (Sadat and Caragea, 2022)	Classification
261	naturallogic (Feng et al., 2020)	Classification
262	dynahate (Vidgen et al., 2021)	Classification
263	syntactic-augmentation-nli (Min et al., 2020)	Classification
264	autotnli	Classification
265	CONDAQA (Ravichander et al., 2022)	Classification
266	scruples	Classification
267	attempto-nli	Classification
268	defeasible-nli/atomic	Classification
269	defeasible-nli/snli	Classification
270	help-nli (Yanaka et al., 2019b)	Classification
271	nli-veridicality-transitivity (Yanaka et al., 2021)	Classification
272	natural-language-satisfiability (Schlegel et al., 2022)	Classification
273	lonli (Tarunesh et al., 2021)	Classification
274	dadc-limit-nli (Wallace et al., 2022)	Classification
275	FLUTE	Classification
276	strategy-qa	Classification
277	folio (Han et al., 2022)	Classification
278	tomi-nli	Classification
279	avicenna (Aghahadi and Talebpour, 2022)	Classification
280	CREAK	Classification
281	puzzte (Szomiu and Groza, 2021)	Classification
282	spartqa-yn (Mirzaee et al., 2021)	Classification
283	temporal-nli (Thukral et al., 2021)	Classification
284	clcd-english	Classification
285	twentyquestions	Classification

Continued on next page

	preprocessing	task type
286	counterfactually-augmented-imdb (Kaushik et al., 2020)	Classification
287	counterfactually-augmented-snli (Kaushik et al., 2020)	Classification
288	cnli (Huang et al., 2020)	Classification
289	boolq-natural-perturbations (Khashabi et al., 2020a)	Classification
290	acceptability-prediction (Lau et al., 2015)	Classification
291	equate (Ravichander et al., 2019)	Classification
292	implicit-hate-stg1 (ElSherief et al., 2021)	Classification
293	chaos-mnli-ambiguity (Zhou et al., 2022)	Classification
294	headline_cause/en_simple (Gusev and Tikhonov, 2021)	Classification
295	logiqa-2.0-nli	Classification
296	oasst1_dense_flat/quality	Classification
297	oasst1_dense_flat/toxicity	Classification
298	oasst1_dense_flat/helpfulness	Classification
299	PARARULE-Plus (Bao et al., 2022)	Classification
300	mindgames (Sileo and Lernould, 2023)	Classification
301	ambient (Liu et al., 2023)	Classification
302	civil_comments/toxicity (Borkan et al., 2019)	Classification
303	civil_comments/severe_toxicity (Borkan et al., 2019)	Classification
304	civil_comments/obscene (Borkan et al., 2019)	Classification
305	civil_comments/threat (Borkan et al., 2019)	Classification
306	civil_comments/insult (Borkan et al., 2019)	Classification
307	civil_comments/identity_attack (Borkan et al., 2019)	Classification
308	civil_comments/sexual_explicit (Borkan et al., 2019)	Classification
309	I2D2	Classification
310	hh-rlhf	MultipleChoice
311	model-written-evals (Perez et al., 2022)	MultipleChoice
312	truthful_qa/multiple_choice (Lin et al., 2021b)	MultipleChoice
313	fig-qa	MultipleChoice
314	bigbench/strange_stories (Srivastava et al., 2022b)	MultipleChoice
315	bigbench/arithmetic (Srivastava et al., 2022b)	MultipleChoice
316	bigbench/formal_fallacies_syllogisms_negation (Srivastava et al., 2022b)	MultipleChoice
317	bigbench/implicatures (Srivastava et al., 2022b)	MultipleChoice
318	bigbench/salient_translation_error_detection (Srivastava et al., 2022b)	MultipleChoice
319	bigbench/causal_judgment (Srivastava et al., 2022b)	MultipleChoice
320	bigbench/discourse_marker_prediction (Srivastava et al., 2022b)	MultipleChoice
321	bigbench/timedial (Srivastava et al., 2022b)	MultipleChoice
322	bigbench/general_knowledge (Srivastava et al., 2022b)	MultipleChoice
323	bigbench/evaluating_information_essentiality (Srivastava et al., 2022b)	MultipleChoice
324	bigbench/cause_and_effect (Srivastava et al., 2022b)	MultipleChoice
325	bigbench/hyperbaton (Srivastava et al., 2022b)	MultipleChoice
326	bigbench/hindu_knowledge (Srivastava et al., 2022b)	MultipleChoice
327	bigbench/crass_ai (Srivastava et al., 2022b)	MultipleChoice
328	bigbench/movie_recommendation (Srivastava et al., 2022b)	MultipleChoice
329	bigbench/cifar10_classification (Srivastava et al., 2022b)	MultipleChoice
330	bigbench/logic_grid_puzzle (Srivastava et al., 2022b)	MultipleChoice
331	bigbench/sentence_ambiguity (Srivastava et al., 2022b)	MultipleChoice
332	bigbench/fact_checker (Srivastava et al., 2022b)	MultipleChoice
333	bigbench/strategyqa (Srivastava et al., 2022b)	MultipleChoice

Continued on next page



	preprocessing	task type
334	bigbench/elementary_math_qa (Srivastava et al., 2022b)	MultipleChoice
335	bigbench/temporal_sequences (Srivastava et al., 2022b)	MultipleChoice
336	bigbench/penguins_in_a_table (Srivastava et al., 2022b)	MultipleChoice
337	bigbench/goal_step_wikihow (Srivastava et al., 2022b)	MultipleChoice
338	bigbench/dark_humor_detection (Srivastava et al., 2022b)	MultipleChoice
339	bigbench/logical_fallacy_detection (Srivastava et al., 2022b)	MultipleChoice
340	bigbench/irony_identification (Srivastava et al., 2022b)	MultipleChoice
341	bigbench/emojis_emotion_prediction (Srivastava et al., 2022b)	MultipleChoice
342	bigbench/sports_understanding (Srivastava et al., 2022b)	MultipleChoice
343	bigbench/contextual_parametric_knowledge_conflicts (Srivastava et al., 2022b)	MultipleChoice
344	bigbench/intent_recognition (Srivastava et al., 2022b)	MultipleChoice
345	bigbench/crash_blossom (Srivastava et al., 2022b)	MultipleChoice
346	bigbench/real_or_fake_text (Srivastava et al., 2022b)	MultipleChoice
347	bigbench/ruin_names (Srivastava et al., 2022b)	MultipleChoice
348	bigbench/logical_deduction (Srivastava et al., 2022b)	MultipleChoice
349	bigbench/identify_math_theorems (Srivastava et al., 2022b)	MultipleChoice
350	bigbench/vitamins_fact_verification (Srivastava et al., 2022b)	MultipleChoice
351	bigbench/hhh_alignment (Srivastava et al., 2022b)	MultipleChoice
352	bigbench/simple_ethical_questions (Srivastava et al., 2022b)	MultipleChoice
353	bigbench/checkmate_in_one (Srivastava et al., 2022b)	MultipleChoice
354	bigbench/similarities_abstraction (Srivastava et al., 2022b)	MultipleChoice
355	bigbench/novel_concepts (Srivastava et al., 2022b)	MultipleChoice
356	bigbench/snarks (Srivastava et al., 2022b)	MultipleChoice
357	bigbench/abstract_narrative_understanding (Srivastava et al., 2022b)	MultipleChoice
358	bigbench/social_iqa (Srivastava et al., 2022b)	MultipleChoice
359	bigbench/phrase_relatedness (Srivastava et al., 2022b)	MultipleChoice
360	bigbench/physics (Srivastava et al., 2022b)	MultipleChoice
361	bigbench/gr_reading_comprehension (Srivastava et al., 2022b)	MultipleChoice
362	bigbench/logical_sequence (Srivastava et al., 2022b)	MultipleChoice
363	bigbench/winowhy (Srivastava et al., 2022b)	MultipleChoice
364	bigbench/movie_dialog_same_or_different (Srivastava et al., 2022b)	MultipleChoice
365	bigbench/riddle_sense (Srivastava et al., 2022b)	MultipleChoice
366	bigbench/metaphor_understanding (Srivastava et al., 2022b)	MultipleChoice
367	bigbench/moral_permissibility (Srivastava et al., 2022b)	MultipleChoice
368	bigbench/nonsense_words_grammar (Srivastava et al., 2022b)	MultipleChoice
369	bigbench/bbq_lite_json (Srivastava et al., 2022b)	MultipleChoice
370	bigbench/physical_intuition (Srivastava et al., 2022b)	MultipleChoice
371	bigbench/navigate (Srivastava et al., 2022b)	MultipleChoice
372	bigbench/reasoning_about_colored_objects (Srivastava et al., 2022b)	MultipleChoice
373	bigbench/metaphor_boolean (Srivastava et al., 2022b)	MultipleChoice
374	bigbench/analytic_entailment (Srivastava et al., 2022b)	MultipleChoice
375	bigbench/mnist_ascii (Srivastava et al., 2022b)	MultipleChoice
376	bigbench/misconceptions (Srivastava et al., 2022b)	MultipleChoice
377	bigbench/authorship_verification (Srivastava et al., 2022b)	MultipleChoice
378	bigbench/social_support (Srivastava et al., 2022b)	MultipleChoice
379	bigbench/tracking_shuffled_objects (Srivastava et al., 2022b)	MultipleChoice
380	bigbench/analogical_similarity (Srivastava et al., 2022b)	MultipleChoice
381	bigbench/figure_of_speech_detection (Srivastava et al., 2022b)	MultipleChoice

Continued on next page

	preprocessing	task type
382	bigbench/understanding_fables (Srivastava et al., 2022b)	MultipleChoice
383	bigbench/question_selection (Srivastava et al., 2022b)	MultipleChoice
384	bigbench/undo_permutation (Srivastava et al., 2022b)	MultipleChoice
385	bigbench/conceptual_combinations (Srivastava et al., 2022b)	MultipleChoice
386	bigbench/unit_interpretation (Srivastava et al., 2022b)	MultipleChoice
387	bigbench/logical_args (Srivastava et al., 2022b)	MultipleChoice
388	bigbench/geometric_shapes (Srivastava et al., 2022b)	MultipleChoice
389	bigbench/code_line_description (Srivastava et al., 2022b)	MultipleChoice
390	bigbench/fantasy_reasoning (Srivastava et al., 2022b)	MultipleChoice
391	bigbench/identify_odd_metaphor (Srivastava et al., 2022b)	MultipleChoice
392	bigbench/empirical_judgments (Srivastava et al., 2022b)	MultipleChoice
393	bigbench/color (Srivastava et al., 2022b)	MultipleChoice
394	bigbench/symbol_interpretation (Srivastava et al., 2022b)	MultipleChoice
395	bigbench/suicide_risk (Srivastava et al., 2022b)	MultipleChoice
396	bigbench/date_understanding (Srivastava et al., 2022b)	MultipleChoice
397	bigbench/cs_algorithms (Srivastava et al., 2022b)	MultipleChoice
398	bigbench/play_dialog_same_or_different (Srivastava et al., 2022b)	MultipleChoice
399	bigbench/international_phonetic_alphabet_nli (Srivastava et al., 2022b)	MultipleChoice
400	bigbench/emoji_movie (Srivastava et al., 2022b)	MultipleChoice
401	bigbench/mathematical_induction (Srivastava et al., 2022b)	MultipleChoice
402	bigbench/implicit_relations (Srivastava et al., 2022b)	MultipleChoice
403	bigbench/anachronisms (Srivastava et al., 2022b)	MultipleChoice
404	bigbench/odd_one_out (Srivastava et al., 2022b)	MultipleChoice
405	bigbench/human_organs_senses (Srivastava et al., 2022b)	MultipleChoice
406	bigbench/english_proverbs (Srivastava et al., 2022b)	MultipleChoice
407	bigbench/key_value_maps (Srivastava et al., 2022b)	MultipleChoice
408	bigbench/dyck_languages (Srivastava et al., 2022b)	MultipleChoice
409	bigbench/known_unknowns (Srivastava et al., 2022b)	MultipleChoice
410	bigbench/disambiguation_qa (Srivastava et al., 2022b)	MultipleChoice
411	bigbench/entailed_polarity (Srivastava et al., 2022b)	MultipleChoice
412	bigbench/epistemic_reasoning (Srivastava et al., 2022b)	MultipleChoice
413	bigbench/presuppositions_as_nli (Srivastava et al., 2022b)	MultipleChoice
414	blimp/sentential_negation_npi_scope (Warstadt et al., 2019)	MultipleChoice
415	blimp/left_branch_island_echo_question (Warstadt et al., 2019)	MultipleChoice
416	blimp/inchoative (Warstadt et al., 2019)	MultipleChoice
417	blimp/principle_A_reconstruction (Warstadt et al., 2019)	MultipleChoice
418	blimp/complex_NP_island (Warstadt et al., 2019)	MultipleChoice
419	blimp/npi_present_2 (Warstadt et al., 2019)	MultipleChoice
420	blimp/existential_there_quantifiers_2 (Warstadt et al., 2019)	MultipleChoice
421	blimp/wh_vs_that_with_gap (Warstadt et al., 2019)	MultipleChoice
422	blimp/superlative_quantifiers_1 (Warstadt et al., 2019)	MultipleChoice
423	blimp/coordinate_structure_constraint_complex_left_branch (Warstadt et al., 2019)	MultipleChoice
424	blimp/matrix_question_npi_licensor_present (Warstadt et al., 2019)	MultipleChoice
425	blimp/principle_A_c_command (Warstadt et al., 2019)	MultipleChoice
426	blimp/drop_argument (Warstadt et al., 2019)	MultipleChoice
427	blimp/tough_vs_raising_1 (Warstadt et al., 2019)	MultipleChoice
428	blimp/npi_present_1 (Warstadt et al., 2019)	MultipleChoice
429	blimp/coordinate_structure_constraint_object_extraction (Warstadt et al., 2019)	MultipleChoice

Continued on next page

	preprocessing	task type
430	blimp/animate_subject_passive (Warstadt et al., 2019)	MultipleChoice
431	blimp/wh_vs_that_with_gap_long_distance (Warstadt et al., 2019)	MultipleChoice
432	blimp/wh_questions_subject_gap_long_distance (Warstadt et al., 2019)	MultipleChoice
433	blimp/sentential_subject_island (Warstadt et al., 2019)	MultipleChoice
434	blimp/wh_questions_object_gap (Warstadt et al., 2019)	MultipleChoice
435	blimp/principle_A_domain_2 (Warstadt et al., 2019)	MultipleChoice
436	cos_e/v1.0 (Rajani et al., 2019)	MultipleChoice
437	cosmos_qa (Huang et al., 2019)	MultipleChoice
438	dream (Sun et al., 2019)	MultipleChoice
439	openbookqa (Mihaylov et al., 2018)	MultipleChoice
440	qasc (Khot et al., 2020)	MultipleChoice
441	quartz (Tafjord et al., "2019")	MultipleChoice
442	quail (Rogers et al., 2020)	MultipleChoice
443	head_qa/en (Vilares and Gomez-Rodriguez, 2019)	MultipleChoice
444	sciq (Johannes Welbl, 2017)	MultipleChoice
445	social_i_qa	MultipleChoice
446	wiki_hop/original (Welbl et al., 2018)	MultipleChoice
447	wiqa (Tandon et al., 2019)	MultipleChoice
448	piqa (Bisk et al., 2020)	MultipleChoice
449	hellaswag (Zellers et al., 2019)	MultipleChoice
450	super_glue/copa (Roemmele et al., 2011)	MultipleChoice
451	balanced-copa (Kavumba et al., 2019)	MultipleChoice
452	e-CARE	MultipleChoice
453	art (Chandra et al., 2020)	MultipleChoice
454	mmlu/nutrition (Hendrycks et al., 2021)	MultipleChoice
455	mmlu/college_medicine (Hendrycks et al., 2021)	MultipleChoice
456	mmlu/philosophy (Hendrycks et al., 2021)	MultipleChoice
457	mmlu/global_facts (Hendrycks et al., 2021)	MultipleChoice
458	mmlu/college_mathematics (Hendrycks et al., 2021)	MultipleChoice
459	mmlu/college_computer_science (Hendrycks et al., 2021)	MultipleChoice
460	mmlu/college_chemistry (Hendrycks et al., 2021)	MultipleChoice
461	mmlu/college_biology (Hendrycks et al., 2021)	MultipleChoice
462	mmlu/clinical_knowledge (Hendrycks et al., 2021)	MultipleChoice
463	mmlu/business_ethics (Hendrycks et al., 2021)	MultipleChoice
464	mmlu/astronomy (Hendrycks et al., 2021)	MultipleChoice
465	mmlu/machine_learning (Hendrycks et al., 2021)	MultipleChoice
466	mmlu/moral_scenarios (Hendrycks et al., 2021)	MultipleChoice
467	mmlu/sociology (Hendrycks et al., 2021)	MultipleChoice
468	mmlu/us_foreign_policy (Hendrycks et al., 2021)	MultipleChoice
469	mmlu/virology (Hendrycks et al., 2021)	MultipleChoice
470	mmlu/world_religions (Hendrycks et al., 2021)	MultipleChoice
471	mmlu/prehistory (Hendrycks et al., 2021)	MultipleChoice
472	mmlu/professional_accounting (Hendrycks et al., 2021)	MultipleChoice
473	mmlu/professional_law (Hendrycks et al., 2021)	MultipleChoice
474	mmlu/professional_medicine (Hendrycks et al., 2021)	MultipleChoice
475	mmlu/professional_psychology (Hendrycks et al., 2021)	MultipleChoice
476	mmlu/electrical_engineering (Hendrycks et al., 2021)	MultipleChoice
477	mmlu/elementary_mathematics (Hendrycks et al., 2021)	MultipleChoice

Continued on next page

	preprocessing	task type
478	mmlu/anatomy (Hendrycks et al., 2021)	MultipleChoice
479	mmlu/abstract_algebra (Hendrycks et al., 2021)	MultipleChoice
480	mmlu/medical_genetics (Hendrycks et al., 2021)	MultipleChoice
481	mmlu/miscellaneous (Hendrycks et al., 2021)	MultipleChoice
482	mmlu/logical_fallacies (Hendrycks et al., 2021)	MultipleChoice
483	mmlu/jurisprudence (Hendrycks et al., 2021)	MultipleChoice
484	mmlu/computer_security (Hendrycks et al., 2021)	MultipleChoice
485	mmlu/international_law (Hendrycks et al., 2021)	MultipleChoice
486	mmlu/human_sexuality (Hendrycks et al., 2021)	MultipleChoice
487	mmlu/human_aging (Hendrycks et al., 2021)	MultipleChoice
488	mmlu/high_school_world_history (Hendrycks et al., 2021)	MultipleChoice
489	mmlu/college_physics (Hendrycks et al., 2021)	MultipleChoice
490	mmlu/high_school_us_history (Hendrycks et al., 2021)	MultipleChoice
491	mmlu/high_school_statistics (Hendrycks et al., 2021)	MultipleChoice
492	mmlu/conceptual_physics (Hendrycks et al., 2021)	MultipleChoice
493	mmlu/high_school_psychology (Hendrycks et al., 2021)	MultipleChoice
494	mmlu/high_school_physics (Hendrycks et al., 2021)	MultipleChoice
495	mmlu/high_school_microeconomics (Hendrycks et al., 2021)	MultipleChoice
496	mmlu/high_school_mathematics (Hendrycks et al., 2021)	MultipleChoice
497	mmlu/econometrics (Hendrycks et al., 2021)	MultipleChoice
498	mmlu/high_school_macroconomics (Hendrycks et al., 2021)	MultipleChoice
499	mmlu/high_school_government_and_politics (Hendrycks et al., 2021)	MultipleChoice
500	mmlu/high_school_geography (Hendrycks et al., 2021)	MultipleChoice
501	mmlu/high_school_european_history (Hendrycks et al., 2021)	MultipleChoice
502	mmlu/high_school_computer_science (Hendrycks et al., 2021)	MultipleChoice
503	mmlu/high_school_chemistry (Hendrycks et al., 2021)	MultipleChoice
504	mmlu/high_school_biology (Hendrycks et al., 2021)	MultipleChoice
505	mmlu/marketing (Hendrycks et al., 2021)	MultipleChoice
506	mmlu/management (Hendrycks et al., 2021)	MultipleChoice
507	mmlu/moral_disputes (Hendrycks et al., 2021)	MultipleChoice
508	mmlu/formal_logic (Hendrycks et al., 2021)	MultipleChoice
509	mmlu/security_studies (Hendrycks et al., 2021)	MultipleChoice
510	mmlu/public_relations (Hendrycks et al., 2021)	MultipleChoice
511	winogrande/winogrande_xl (ai2, 2019)	MultipleChoice
512	codah/codah (Chen et al., 2019a)	MultipleChoice
513	ai2_arc/ARC-Challenge/challenge (Clark et al., 2018)	MultipleChoice
514	ai2_arc/ARC-Easy/challenge (Clark et al., 2018)	MultipleChoice
515	definite_pronoun_resolution (Rahman and Ng, 2012)	MultipleChoice
516	swag/regular (Zellers et al., 2018)	MultipleChoice
517	math_qa	MultipleChoice
518	lex_glue/case_hold (Zheng et al., 2021)	MultipleChoice
519	commonsense_qa (Talmor et al., 2019)	MultipleChoice
520	discosense	MultipleChoice
521	medmcqa (Pal et al., 2022)	MultipleChoice
522	aqua_rat/tokenized (ACL, 2017)	MultipleChoice
523	logiqa (Liu et al., 2020)	MultipleChoice
524	cyclic_multiplechoice (Kejriwal and Shen, 2020)	MultipleChoice
525	arct (Habernal et al., 2018)	MultipleChoice

Continued on next page

preprocessing	task type
526 onestop_qa (Berzak et al., 2020)	MultipleChoice
527 moral_stories/full (Emelin et al., 2021)	MultipleChoice
528 prost (Aroca-Ouellette et al., 2021)	MultipleChoice
529 webgpt_comparisons (Nakano et al., 2021)	MultipleChoice
530 synthetic-instruct-gptj-pairwise	MultipleChoice
531 wouldyourather	MultipleChoice
532 summarize_from_feedback/comparisons (Stiennon et al., 2020)	MultipleChoice
533 SHP (Ethayarajh et al., 2023)	MultipleChoice
534 MedQA-USMLE-4-options-hf	MultipleChoice
535 wikimedqa/medwiki (Sileo et al., 2023)	MultipleChoice
536 cicero (Ghosal et al., 2022)	MultipleChoice
537 mutual (Cui et al., 2020)	MultipleChoice
538 NeQA	MultipleChoice
539 quote-repetition	MultipleChoice
540 redefine-math	MultipleChoice
541 implicatures (George and Mamidi, 2020)	MultipleChoice
542 race/high (Lai et al., 2017)	MultipleChoice
543 race/middle (Lai et al., 2017)	MultipleChoice
544 race-c (Liang et al., 2019)	MultipleChoice
545 spartqa-mchoice (Mirzaee et al., 2021)	MultipleChoice
546 riddle_sense (Lin et al., 2021a)	MultipleChoice
547 reaclor (Yu et al., 2020b)	MultipleChoice
548 ScienceQA_text_only (Saikh et al., 2022)	MultipleChoice
549 ekar_english	MultipleChoice
550 path-naturalness-prediction	MultipleChoice
551 cloth	MultipleChoice
552 dgen	MultipleChoice
553 oasst1_pairwise_rlhf_reward	MultipleChoice
554 conll2003/pos_tags (Tjong Kim Sang and De Meulder, 2003)	TokenClassification
555 conll2003/chunk_tags (Tjong Kim Sang and De Meulder, 2003)	TokenClassification
556 conll2003/ner_tags (Tjong Kim Sang and De Meulder, 2003)	TokenClassification
557 wnut_17/wnut_17 (Derczynski et al., 2017)	TokenClassification
558 ncbi_disease/ncbi_disease (Dogan et al., 2014)	TokenClassification
559 acronym_identification (Veyseh et al., 2020)	TokenClassification
560 jnlpba/jnlpba (Kim et al., 2004)	TokenClassification
561 species_800/species_800 (Pafilis et al., 2013)	TokenClassification
562 ontonotes_english (Tjong Kim Sang and De Meulder, 2003)	TokenClassification
563 universal_dependencies/en_partut/depel (Zeman et al., 2020)	TokenClassification
564 universal_dependencies/en_lines/depel (Zeman et al., 2020)	TokenClassification
565 universal_dependencies/en_gumreddit/depel (Zeman et al., 2020)	TokenClassification
566 universal_dependencies/en_esl/depel (Zeman et al., 2020)	TokenClassification
567 universal_dependencies/en_ewt/depel (Zeman et al., 2020)	TokenClassification
568 universal_dependencies/en_gum/depel (Zeman et al., 2020)	TokenClassification

## B Model Recycling results

model_name	deberta-v3-base	+tasksource
avg	79.04	<b>80.73</b>
mnli (linear probe)	-	<b>93.73</b>
20_newsgroup	86.41	<b>86.46</b>
ag_news	90.44	<b>90.67</b>
amazon_reviews_multi	66.86	<b>66.90</b>
anli	58.78	<b>60.38</b>
boolq	82.99	<b>85.66</b>
cb	75.00	<b>82.14</b>
cola	86.57	<b>87.15</b>
copa	58.40	<b>81.00</b>
dbpedia	<b>79.43</b>	79.20
esnli	91.93	<b>91.54</b>
financial_phrasebank	84.48	<b>85.20</b>
imdb	94.49	<b>94.67</b>
isear	71.86	<b>71.90</b>
mnli_mismatched	89.78	<b>91.14</b>
mrpc	<b>89.20</b>	88.73
multirc	62.26	<b>63.82</b>
poem_sentiment	86.73	<b>92.31</b>
qnli	93.51	<b>93.72</b>
qqp	91.79	<b>91.92</b>
rotten_tomatoes	90.42	<b>90.99</b>
rte	82.35	<b>90.61</b>
sst2	95.06	<b>95.41</b>
sst_5bins	56.98	<b>58.60</b>
stsb	90.28	<b>91.81</b>
trec_coarse	97.76	<b>96.80</b>
trec_fine	<b>91.02</b>	90.80
tweet_ev_emoji	46.19	<b>47.82</b>
tweet_ev_emotion	83.95	<b>85.71</b>
tweet_ev_hate	56.21	<b>57.47</b>
tweet_ev_irony	79.82	<b>83.04</b>
tweet_ev_offensive	85.06	<b>85.23</b>
tweet_ev_sentiment	71.80	<b>72.01</b>
wic	<b>71.21</b>	69.44
wnli	<b>70.21</b>	67.61
wsc	64.09	<b>66.35</b>
yahoo_answers	72.03	<b>72.07</b>