RAFT: A Real-World Few-Shot Text Classification Benchmark

Neel Alex[*]
Ought

Eli Lifland[*]
Ought

Lewis Tunstall
Hugging Face

Abhishek Thakur
Hugging Face

Pegah Maham
Stiftung Neue Verantwortung

C. Jess Riedel
NTT Research

Emmie Hine
Oxford Internet Institute

Carolyn Ashurst
Alan Turing Institute

Paul Sedille
Harvard Kennedy School

Alexis Carlier
University of Oxford

Michael Noetel
Australian Catholic University

Andreas Stuhlmüller
Ought

Abstract

Large pre-trained language models have shown promise for few-shot learning, completing text-based tasks given only a few task-specific examples. Will models soon solve classification tasks that have so far been reserved for human research assistants? Existing benchmarks are not designed to measure progress in applied settings, and so don't directly answer this question. The RAFT benchmark (**R**eal-world **A**nnotated **F**ew-shot **T**asks) focuses on naturally occurring tasks and uses an evaluation setup that mirrors deployment. Baseline evaluations on RAFT reveal areas current techniques struggle with: reasoning over long texts and tasks with many classes. Human baselines show that some classification tasks are difficult for non-expert humans, reflecting that real-world value sometimes depends on domain expertise. Yet even non-expert human baseline F1 scores exceed GPT-3 by an average of 0.11. The RAFT datasets and leaderboard will track which model improvements translate into real-world benefits at https://raft.elicit.org.

1 Introduction

Few-shot learning, the capacity to complete a task given a small number of demonstrations [11], is one of the hallmarks of human intelligence [30, 17]. As researchers, we leverage this capacity when we delegate work on crowdsourcing platforms or give a task with examples to a human research assistant.

Brown et al. [6] show that large pre-trained language models exhibit few-shot learning capabilities for a wide range of natural language tasks. If those capabilities were comparable to people on economically relevant tasks, this would be important to know: a single model could be used across multiple real-world tasks, with low per-task data labeling cost. However, these models have also been shown to have inconsistent few-shot performance depending on the exact setup and task being solved [e.g. 21, 24]. The mixed evidence suggests that it would be valuable to measure and track few-shot performance on a set of tasks that is representative of what appears in practice.

---

[*]Equal contribution. Correspondence to eli.d.lifland@gmail.com and salexucb@berkeley.edu. NA contributed during an internship at Ought.

Task Specification in Natural Language

Organization	Paper Title	Label
North Carolina State Univ.,Raleigh,NC,USA	3Gb/s AC-coupled chip-to-chip communication using a low-swing pulse receiver	university
Advanced LCD Technology Development Center...	Sub-Micron CMOS / MOS-Bipolar Hybrid TFTs for System Displays	company
imec,Heverlee,Belgium	24.4 A 680nA fully integrated implantable ECG-acquisition IC with analog...	research institute
...	...	...

Partially-labeled Dataset

Label the sentence based on whether it is related to an adverse drug effect (ADE)

The following is a banking customer service query. Classify it as one of the following 77 classes.

Label the impact statement as "mentions a harmful application" or "doesn't mention a harmful application" based on whether it mentions a harmful application of the research done in the paper.

Identify whether this paper should be included in a meta-review which includes the findings of systematic reviews on interventions designed to promote charitable donations.
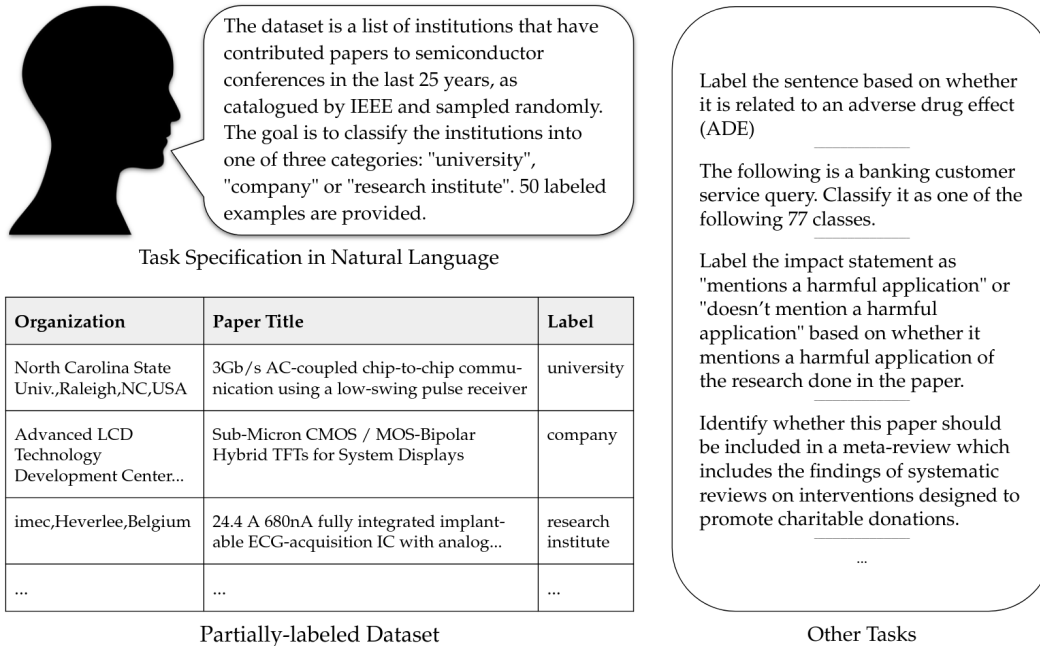
...

Other Tasks

Figure 1: RAFT includes naturally occurring classification datasets, mimicking work that is usually given to human research assistants. Each task comes with natural language instructions and labels in addition to 50 training examples.

Natural language tasks coarsely split into generation, classification, and retrieval. We focus on classification tasks because they support high-quality automated evaluation, cover a wide range of economically valuable tasks, and yet don't have existing real-world benchmarks.

Existing few-shot classification benchmarks are typically designed to highlight areas where models fall short [29] or to study particular model abilities [5, 37, 21]. The tasks and evaluation setup aren't optimized to measure progress in applied settings:

- *Tasks* that are generated or chosen specifically to test language models may not represent some of the challenges found when applying these models in real-world settings. For example, SuperGLUE [32] and the few-shot equivalent FewGLUE [29] mainly include short texts. Doing well on applied tasks sometimes requires reasoning over long texts. Existing systems struggle with long texts due to a limited context window, especially in the few-shot setting where some systems learn from examples presented in context.

- The *evaluation* does not closely mirror deployment, and may both under- and overestimate models' capabilities. It may underestimate model capability by restricting models to the closed-book setting (e.g., no retrieval from online sources) and using uninformative labels (e.g., 0/1 instead of "about literature" vs. "about movies"). It may overestimate model capability by using many more than a few examples for setting hyperparameters during validation [24].

RAFT is a real-world few-shot text classification benchmark designed to measure how much recent and upcoming NLP advances benefit applications:

- The *tasks* are naturally occurring tasks. Their labeling is inherently valuable to someone, and they may have challenges that are not reflected in synthetic tasks. Inherent value means that, if it were sufficiently fast and cheap, it would be desirable to outsource the task to human research assistants or crowd workers. Challenges refers to the need for information retrieval, domain expertise, parsing long documents, and making use of instructions. Table 1 shows the real-world challenges presented by RAFT, including 4 datasets with long input texts.

- The *evaluation* closely mirrors deployment. For each task, we release a public training set with 50 examples and a larger unlabeled test set[2]. We encourage unsupervised pre-training on the unlabelled examples and open-domain information retrieval. We keep the test-set labels private and provide automated evaluation through a Hugging Face leaderboard[3].

In addition to the gold-standard human labels, we collect automatic and crowdsourced baselines. The automatic baselines reveal areas where current techniques struggle, such as reasoning over long texts and tasks with many classes. The crowdsourced baseline reveals that RAFT includes a mix of moderate to difficult tasks. We also observe difficulties in collecting human crowdsourced baselines on some datasets, particularly when domain expertise is important, which suggests that real-world value often depends on domain knowledge.

The RAFT datasets and leaderboard can be viewed and submitted to at https://raft.elicit.org.

2   Related Work

We briefly review few-shot learning in NLP, then the benchmarks that are most similar to RAFT.

2.1   Few-shot learning in NLP

Pre-trained language models (PLMs) such as BERT [10] and GPT-3 [6] can learn to do some NLP tasks when prompted with a few demonstrations, including some classification tasks. The two primary approaches to few-shot classification using PLMs are in-context learning and prompt-based fine-tuning.

In-context learning. A PLM is primed with labeled examples in its prompt. It classifies the example included at the end of the prompt by predicting the classification conditioned on the priming. GPT-3 [6] used in-context learning to achieve promising results on a variety of classification tasks. UniFew [5] similarly achieved strong results on classification tasks via in-context learning, converting classification tasks into a multiple-choice question answer format for prompting.

Prompt-based fine-tuning. A PLM is fine-tuned with masked-language modeling objectives to learn from few examples. This is also known as Pattern-exploiting training (PET) [28]. While PET requires task-specific prompts, it achieves better performance than GPT-3 in-context with smaller models [29]. LM-BFF [13] improves prompt-based fine-tuning by dynamically constructing prompts.

2.2   Few-shot NLP benchmarks

The most closely related few-shot NLP benchmarks are FLEX [5], FewGLUE [29], CrossFit [37], and NaturalInstructions [21]. Each of these benchmarks includes at least some classification tasks with meaningful textual labels.

These benchmarks are designed to study transfer between tasks [5, 37], pinpoint where NLP models fall short [29], and evaluate ability of models to follow instructions [21], whereas RAFT is designed to be representative of real-world classification tasks. This difference in goals is reflected in selection of tasks and evaluation:

Tasks. FLEX, FewGLUE and NaturalInstructions test on traditional NLP tasks. CrossFit tests on 160 tasks from the Hugging Face Datasets[4] and includes some naturally occurring datasets, including the TweetEval dataset [2] that RAFT uses as well. CrossFit excludes tasks that leverage external sources or information retrieval techniques, need domain knowledge (e.g. COVID-19 datasets), and long documents (e.g. scientific papers). Like RAFT, FLEX includes tasks with strong class imbalance.

Evaluation. None of the existing benchmarks allow open-domain information retrieval. Like FLEX, RAFT provides no extra validation data beyond the training examples. Perez et al. [24] argue that the performance of state-of-the-art few-shot methods has been overestimated by most other existing benchmarks because they use labeled examples beyond the few training instances provided for model and parameter selection.

---

[2]Datasets are at https://raft.elicit.org/datasets
[3]Instructions for submission are at https://raft.elicit.org/submit
[4]https://huggingface.co/datasets

3 Benchmark Description

RAFT is a few-shot *classification* benchmark. We focus on classification primarily because automatic evaluation is more reliable than for generation tasks. We believe (as our results will later confirm) that there still is a substantial gap between even non-expert humans and automated systems in the few-shot classification setting.

Both tasks (datasets and metadata) and evaluation (rules for submission, metrics) are chosen to mirror real-world classification problems.

3.1 Tasks

A classification task is a dataset with labeled natural language entries. Each label corresponds one-to-one with a natural language class name. Each task has instructions for labeling.

3.1.1 Dataset selection criteria

We selected datasets based on the following criteria ("non-trivial real world tasks"):

Naturally occurring. We focus on data that are naturally occurring, rather than being synthetically generated to test and improve language models.

Intrinsic value. We select datasets for which the correct labeling inherently provides real-world value. RAFT includes tasks like hate-speech detection, medical case report parsing, and literature review automation, where better performance translates into practical benefits. This criterion involves subjectivity, but we aimed to select tasks that approximate the distribution of valuable classification tasks well.

Realistic class distribution. We did not exclude datasets with heavily imbalanced classes.

Open-domain feasibility. As we provide an open-domain setting where information retrieved from the web may be used to augment predictions, we excluded tasks for which the correct label is extremely easily discoverable through a Google search. For example, we considered including the LIAR [35] dataset which includes Politifact statements and their veracity. We decided against including it since it would be trivial to get 100% accuracy by running a site search on https://www.politifact.com/.

In order to gather datasets meeting the above requirements, we put out a collaboration request. We also reached out to users of classification on Elicit [23]. Lastly, we conducted a search of existing datasets on the Hugging Face Hub[5] and PapersWithCode [6].

3.1.2 Dataset preparation

In cases where the test set was over 5,000 data points, we randomly selected 5,000 to serve as a test set in order to keep the test set sizes manageable. When the dataset didn't already have textual labels, we added textual labels according to our best understanding of the task.

3.1.3 Selected RAFT datasets

We selected 11 datasets, accessible at https://raft.elicit.org/datasets. Table 1 presents an overview of the datasets. More details are available in the Appendix.

ADE Corpus V2 (*ADE*). The ADE corpus V2 [15] contains sentences from medical case reports annotated for relation to adverse drug effects. We focus on the binary classification task of whether a sentence is related to an adverse drug effect (ADE).

Banking77 (*B77*). Banking77 [7] contains online banking customer service queries annotated with their intents.

NeurIPS impact statement risks (*NIS*). We include the broader impact statements from NeurIPS 2020 papers collected in the dataset from Ashurst et al. [1]. We annotate these based on whether they mention possibly harmful applications of the research done in the paper .[7]

---

[5]https://huggingface.co/datasets

[6]https://paperswithcode.com

[7]The raw scraped NeurIPS impact statements can be found at https://raft.elicit.org/neurips-impact.

Dataset Name	Long inputs	Domain expertise	Detailed instructions	Number of classes	Test set size
ADE Corpus V2 (*ADE*)	–	✓	–	2	5,000
Banking77 (*B77*)	–	–	–	77	5,000
NeurIPS impact statement risks (*NIS*)	✓	–	–	2	150
OneStopEnglish (*OSE*)	✓	–	–	3	516
Overruling (*Over*)	–	✓	–	2	2,350
Semiconductor org types (*SOT*)	–	–	–	3	449
Systematic review inclusion (*SRI*)	✓	–	✓	2	2,243
TAI safety research (*TAI*)	✓	✓	✓	2	1,639
Terms of Service (*ToS*)	–	✓	✓	2	5,000
TweetEval Hate (*TEH*)	–	–	✓	2	2,966
Twitter complaints (*TC*)	–	–	–	2	3,399

Table 1: Overview of the tasks in RAFT. *Long inputs*, *Domain expertise*, and *Detailed instructions* are some of the real-world challenges posed by RAFT.

OneStopEnglish (*OSE*). OneStopEnglish [31] contains articles sourced from The Guardian newspaper and rewritten by teachers to suit three levels of adult English as a Second Language (ESL) learners.

Overruling (*Over*). Overruling [39] contains statements from a law corpus annotated based on whether they are overruling, defined as nullifying a previous case decision as a precedent.

Semiconductor org types (*SOT*). We collect a dataset of institutions that have contributed to semiconductor conferences in the last 25 years, then classify these institutions into organization types: "university", "company", and "research institute".

Systematic review inclusion (*SRI*). We use data from a systematic meta-review studying interventions to increase charitable donations [22]. The task is to predict whether a paper advances past the screening stage.

TAI safety research (*TAI*). We include data from the formation of a bibliographic database for research on the safety of transformative artificial intelligence (TAI) [27]. We choose the binary task of predicting whether a work is classified as TAI safety research.

Terms of Service (*ToS*). The Terms of Service dataset [19] contains clauses from Terms of Services, annotated by whether they are potentially unfair to consumers.

TweetEval Hate (*TEH*). We include the hate-speech detection task from the TweetEval dataset [2], which was curated from Basile et al. [3].

Twitter complaints (*TC*). We include a dataset of tweets annotated by whether they contain a complaint [25].

3.2 Evaluation

3.2.1 Setting and rules

The RAFT evaluation replicates real-world few-shot classification problems by restricting to 50 labeled examples without validation set, providing meaningful instructions and labels, and using a no-holds-barred setting:

50 labeled examples. We provide 50 labeled examples per task (not per class). In the authors' experience with users of the classification tool Elicit [23], this is approximately the number of examples people are willing to label for a task with a few thousand unlabeled examples. The 50

examples are chosen randomly, mirroring the applied setting in which one can't easily choose a balanced set. No examples beyond the chosen 50 are available for validation.[8]

Task-specific instructions. As an important replacement for large amounts of labeled data, instructions can specify how a task should be done. Therefore, we provide the instructions we give to human labelers so that they can be used in instructing automatic systems. The level of detail of the instructions varies. We write the instructions based on information from publications (for datasets published elsewhere) or in consultation with the dataset creator (for new datasets).

Meaningful label names. Similar to instructions, textual labels are an important aspect of few-shot and especially zero-shot learning. We create default textual labels for each dataset as recommended by FLEX [5].

Transfer learning permitted. Transfer and meta-learning using other datasets is permitted, including further pre-training on other corpora.

Unlabeled data permitted. Use of the unlabeled RAFT test sets is permitted, as unlabeled data are usually available in the applied setting.

Open-domain retrieval permitted. Models may be augmented with information retrieved from the internet, e.g. via automated web searches.[9]

Submission requires only labels. Submission on the test set is open to the public and only requires upload of test set labels. This is in line with benchmarks like GLUE [33] and SuperGLUE [32], but is in contrast to the few-shot benchmark FLEX [5]. By only requiring labels, we give submission creators maximal flexibility in what models to set up.

Weekly evaluation. Evaluation is run on a weekly basis to minimize information gained from frequent repeated submissions.

Baseline	Avg	ADE	B77	NIS	OSE	Over	SOT	SRI	TAI	ToS	TEH	TC
Human (crowdsourced)	**.735**	**.830**	**.607**	**.857**	**.646**	.917	**.908**	.468	.609	**.627**	**.722**	**.897**
GPT-3 (175B)	.627	.686	.299	.679	.431	**.937**	.769	**.516**	**.656**	.574	.526	.821
AdaBoost	.514	.543	.023	.626	.475	.838	.455	.506	.556	.560	.443	.625
GPT-Neo (2.7B)	.481	.452	.149	.408	.343	.681	.406	.493	.605	.565	.554	.636
GPT-2 (1.6B)	.458	.600	.121	.561	.245	.498	.380	.492	.612	.498	.311	.723
BART MNLI Zero-shot	.382	.234	.332	.615	.360	.462	.644	.026	.469	.122	.543	.400
Plurality class	.331	.446	.000	.353	.164	.337	.271	.493	.344	.471	.366	.391
GPT-3 Zero-shot	.292	.163	.000	.572	.323	.378	.628	.027	.362	.164	.303	.290

Table 2: Performance of RAFT baselines (F1)

3.3 Metrics

Since some RAFT datasets have substantial class imbalances, we use F1 as our evaluation metric. We compute macro-averaged F1 scores, even for binary datasets. To get an overall score, we average across all datasets.

---

[8]By only releasing 50 labelled examples, we make it difficult to cheat by using more than 50 examples for validation. For the NeurIPS impact statement risks and Semiconductor org type datasets, the test set labels aren't available publicly. For other datasets, the test set labels are available publicly but it is non-trivial and discouraged to seek them out.

[9]Ideally, we'd only allow information from before the time each dataset needed to be labeled. This isn't currently feasible, so we settle for a fully open-domain setting.

4 Baselines

The code for all automatic baselines is open-sourced at https://raft.elicit.org/baselines.

4.1 GPT-3 baseline

We provide a simple automatic baseline using GPT-3 [6], accessed through the OpenAI API[10]. As in Brown et al. [6], we use in-context learning, adding labeled examples to the prompt to prime GPT-3. We also run a zero-shot version with no training examples included in the prompt.

4.1.1 Prompt construction

We build a prompt consisting of:

1. Task-specific instructions
2. N labeled examples, with N selected on a per-task basis
3. The target example to classify

Example prompts for all datasets are available at https://raft.elicit.org/baseline-prompts.

Truncation. GPT-3's context window support up to $2,048$ tokens. The experiments in Brown et al. [6] include as many complete labeled examples as would fit in the context, reporting that typically 10 to 100 examples fit. However, datasets such as *OSE* have very long inputs so that only 1-2 complete labeled examples would fit, suggesting that another approach may be better.

We select N training examples to include in a given prompt and then truncate the examples. In a given task, the instructions take up I tokens. Separators between the instructions and each example take up S tokens. No more than $T = 2,048$ tokens can be used.

1. $E = T - I - S$ tokens are allotted for the training examples and classification target.
2. The classification target is truncated to $\frac{1}{4}E$ tokens.
3. Each of the N remaining training examples is truncated to $\frac{3}{4}\frac{E}{N}$ tokens.

We truncate from a training example's data fields first, leaving the label intact.

Field selection and sorting. We exclude data fields that are unlikely to contribute substantially to GPT-3's performance. These fields either deal with the authors of the textual example or are URLs. Additionally, we sort the order in which the text fields occur to put the most important fields first. When examples are truncated, the most important information is preserved.

Semantic selection. To select training examples to include in the prompt for a given test example, we selected the most similar training examples as in Liu et al. [20]. To perform semantic search, we use the OpenAI API search endpoint with the `ada` engine.

4.1.2 Classification

With the prompt formed, we retrieve GPT-3's 100 most likely next tokens using the `davinci` engine. For each class, we assign the probability that its first token is generated. We then normalize the probabilities to sum to 1. For the *B77* dataset, multiple labels share the same first token so we prepend a numerical prefix such as "1. " to each class.

4.1.3 Parameter selection

We tune the GPT-3 baseline on the training set using *leave-one-out cross validation* (LOOCV): k-fold cross validation with $k = n$ so that only one test example is used at a time for validation. While LOOCV isn't robust with as few as 50 examples as discussed in Perez et al. [24], it is one of the best options for parameter selection in the few-shot setting. Detailed LOOCV results are in Section A.5.

Instructions. We test two modes of instruction: (a) a generic classification prompt: "Possible labels:" followed by a list of textual labels. (b) instructions similar to the ones given to human labelers, plus

---

[10]https://beta.openai.com

the list of textual labels. The instructions are taken whole when possible, and otherwise shortened and summarized manually to limit usage of the GPT-3 context window. Task-specific instructions outperform generic instructions by an .04 on averaged F1 score, thus we include task-specific instructions in the baseline.

Semantic training example selection. To select training examples for inclusion in the prompt from a larger set, we consider (a) selecting examples randomly and (b) using semantic search to identify the training examples most similar to the test example. Semantic selection outperforms random selection by 0.03 on averaged F1, thus we include semantic selection in the baseline.

Number of examples in the prompt. We select the number of examples to include in the prompt on a per-dataset basis, as our truncation strategy induces a quality-quantity trade-off. For each dataset, we test performance with 5, 10, 25, and 50[11] training examples and choose the number that performs best by F1. For datasets with long inputs, smaller numbers of more detailed samples often produce better performance, while datasets with smaller inputs can fit more complete labeled examples in the prompt.

4.2 Other automatic baselines

In-context baselines. We run further in-context baselines GPT-Neo [4] and GPT-2 [26]. We provide code[12] for generating predictions on RAFT using these models and any other causal language model available on the HuggingFace Hub. For semantic search, we use a MiniLM [34] fine-tuned on sentence pairs via the sentence-transformers package[13].

Zero-shot baselines. We run two transformers in the zero-shot setting:

- GPT-3, to judge to what extent training examples in the prompt aid performance
- BART [18] trained on MNLI [36], as suggested by Yin et al. [38] and Davison [9] as an effective zero-shot classification approach

Non-neural baselines. We run AdaBoost [12] to establish a strong non-neural baseline. We construct feature vectors for each example based on the counts of n-grams of 1-5 words as the input to a weighted ensemble of 100 depth-3 decision trees. These decision trees and weights are trained with AdaBoost with learning rate 1, and evaluated through weighted voting. We also include a plurality (most frequent) class baseline.

4.3 Human baseline

To collect human baselines, we use the Surge[14] crowdsourcing platform. Following Wang et al. [32], we randomly select 100 data points from each test set and use a 2-step labeling process: qualification then annotation. The crowdsourced label is the plurality vote of 5 labelers.

We put crowd workers in a similar situation to automated systems. We link to a sheet with the same 50 labeled examples, use the same textual labels, and give the same task-specific instructions that we are providing to practitioners to adapt for instructing language models.[15]

4.4 Analysis

Humans generally outperform GPT-3. Humans outperform GPT-3 on 8 out of 11 tasks, demonstrating room for improvement for models on real-world few-shot tasks. We expect that exceeding the crowdsourced baseline will require substantial advances in model performance, and even more so for a future expert human baseline.

Weaknesses of GPT-3 include:

- **Many classes**: Humans most outperform GPT-3 on *B77*, which has by far the most classes in RAFT. With 77 classes and 50 labeled examples, many classes have no corresponding

---

[11]49 rather than 50 training examples for LOO experiments

[12]https://raft.elicit.org/baselines

[13]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[14]https://www.surgehq.ai/

[15]For details on the human baseline gathering process, see Section A.7.

labeled examples. Additionally, just listing out the possible classes takes up a large portion of GPT-3's context window.

- **Long inputs**: GPT-3 performs poorly on some tasks requiring reasoning over long inputs, such as *NIS* and *OSE*. GPT-3's context window may be a contributing factor.

Crowd-sourced baselines struggle on domain-specific tasks. Crowd-sourced humans substantially outperform GPT-3 on only 1 of 4 tasks we identified as requiring domain expertise:

- Humans substantially outperform GPT-3 on *ADE*, which requires medical expertise.
- Humans outperformed GPT-3 by just .053 on *ToS*, which requires parsing legal language.
- GPT-3 outperforms humans on *Over*, which requires greater legal expertise than *ToS* [39], and *TAI*, which requires expertise in AI safety research.

Zero-shot performance is weak. GPT-3 zero-shot does poorly on RAFT, performing worse than the plurality class baseline. BART zero-shot exceeds the plurality class baseline but does not do so in every dataset, and it is not competitive with few-shot language models. We encourage future research on improving performance in the zero-shot setting, perhaps through improved prompt construction and transfer learning.

Neural baselines besides few-shot GPT-3 perform worse than AdaBoost. Generative language models smaller than GPT-3 comfortably outperform the plurality class baseline but remain below AdaBoost. We use the same amount of labelled examples in the prompt as with GPT-3 despite the context window being smaller; performance may improve with fewer (but longer) examples.

5 Discussion

5.1 Limitations

Linguistic diversity. The benchmark only includes English tasks. Dealing with multilingual corpora is a real-world challenge for many NLP systems, especially for those deployed in countries where there are multiple national languages. To fully capture the distribution of real-world tasks, additional languages will be needed.

Possible biases in data collection. While we attempted to execute our dataset selection process as described in Section 3.1.3 in an unbiased manner, the datasets we ended up selecting are part of a subjective human process that may be subject to biases. For example, the organizations we work with are disproportionately in technology and policy.

5.2 Impact

Offensive content. By including a hate-speech detection dataset, we include offensive content and may harm readers of the dataset. We believe the advantages from studying hate-speech detection are likely greater than the disadvantages of publicizing hate-speech datasets.

Prohibitive costs. The models best equipped to perform well on RAFT will often be the massive transformer models trained by private corporations. In advancing this benchmark as a means of evaluating models, we risk further widening the gap between what a dedicated individual or team can do, and what can only be done by industry research labs with sufficient funding.

5.3 Future Work

Stronger human baselines. Human baselines are intended to tell us how well the dataset would be labeled in the absence of automated systems. For many RAFT datasets, this process would involve a stronger baseline than is easily available via a crowd-worker platform: for example, the *Over* dataset would be labeled by someone with law expertise. In addition to ML submissions, we welcome efforts to collect stronger human baselines for RAFT.

Additional automatic baselines. We expect that systems that use prompt-based fine-tuning rather than in-context learning may provide an even stronger automatic baseline. We further expect that models that leverage the open-domain information retrieval option can surpass models that don't.

Application-specific metrics. Different applications care about different metrics; e.g., in some applications it is more important to minimize false positives, whereas in others the focus is on false negatives. An ideal measure of real-world value would take that into account.

Learning from natural language In this work, we focused on instructions as a supplement to labeled examples. Similarly to Mishra et al. [21], we found that including task-specific instructions improved performance. Like humans, NLP systems could also learn from other types of natural language. For example, could including explanations with each labeled example be used to further improve few-shot performance?

6 Conclusion

RAFT is a benchmark that tests language models across multiple domains on economically valuable classification tasks in the true few-shot setting. To our knowledge, this is the first multi-task benchmark designed to closely mirror how models are applied in both the task distribution and the evaluation setup. By complementing existing synthetic benchmarks designed to highlight where models fall short, it helps measure the gap between research and practice, incentivizes work that is valuable for deployed systems, and provides a template for future benchmarks that mirror deployment.

Acknowledgments and Disclosure of Funding

References

[1] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. Ai ethics statements - analysis and lessons learnt from neurips broader impact statements. *arXiv preprint arXiv:2111.01705*, 2021.

[2] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. Tweet-Eval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://aclanthology.org/2020.findings-emnlp.148.

[3] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://aclanthology.org/S19-2007.

[4] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata.

[5] Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. Flex: Unifying evaluation for few-shot nlp. *arXiv: Computation and Language*, 2021. doi: null.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh,

Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

[7] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders, 2020.

[8] Dataset of affiliation types, 2021. URL `https://docs.google.com/spreadsheets/d/1CT3hCvbKxyJeS1FdrZtlK5MTuvWsfXR_`.

[9] Joe Davison. Zero-shot learning in modern nlp. URL `https://joeddav.github.io/blog/2020/05/29/ZSL.html`.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[11] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:594–611, 2006.

[12] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1997.1504. URL `https://www.sciencedirect.com/science/article/pii/S002200009791504X`.

[13] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners, 2021.

[14] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2020.

[15] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 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, 2012. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2012.04.008. URL `https://www.sciencedirect.com/science/article/pii/S1532046412000615`. Text Mining and Natural Language Processing in Pharmacogenomics.

[16] Sergey Ivanov. Neurips affiliation locations, 2020. URL `https://github.com/nd7141/icml2020/blob/master/university2.csv`.

[17] B. Lake, R. Salakhutdinov, Jason Gross, and J. Tenenbaum. One shot learning of simple visual concepts. *Cognitive Science*, 33, 2011.

[18] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019. URL `http://arxiv.org/abs/1910.13461`.

[19] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27 (2):117–139, Feb 2019. ISSN 1572-8382. doi: 10.1007/s10506-019-09243-2. URL `http://dx.doi.org/10.1007/s10506-019-09243-2`.

[20] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3?, 2021.

[21] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Natural instructions: Benchmarking generalization to new tasks from natural language instructions, 2021.

[22] Michael Noetel, Peter Slattery, Alexander K Saeri, Joannie Lee, Thomas Houlden, Neil Farr, Romy Gelber, Jake Stone, Lee Huuskes, Shane Timmons, and et al. What works to increase charitable donations? a meta-review with meta-meta-analysis, Feb 2020. URL `psyarxiv.com/yxmva`.

[23] Ought. Elicit: The ai research assistant, 2021. URL `https://elicit.org`.

[24] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models, 2021.

[25] Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1495. URL `https://aclanthology.org/P19-1495`.

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

[27] Jess Riedel and Angelica Deibel. Tai safety bibliographic database, Dec 2020. URL `https://www.lesswrong.com/posts/4DegbDJJiMX2b3EKm/tai-safety-bibliographic-database#Inclusion___categorization1`.

[28] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference, 2021.

[29] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners, 2021.

[30] J. Tenenbaum, Charles Kemp, T. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331:1279 – 1285, 2011.

[31] Sowmya Vajjala and Ivana Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0535. URL `https://aclanthology.org/W18-0535`.

[32] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*, 2019.

[33] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[34] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

[35] William Yang Wang. "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, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL `https://aclanthology.org/P17-2067`.

[36] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426, 2017. URL `http://arxiv.org/abs/1704.05426`.

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

[38] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR*, abs/1909.00161, 2019. URL `http://arxiv.org/abs/1909.00161`.

[39] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset, 2021.

A Appendix

A.1 Dataset licensing

The curating authors (NA, EL, and AS) bear all responsibility in the case of violation of rights. Below we provide information on the license for each dataset:

ADE Corpus V2 (*ADE*). Unlicensed.

Banking77 (*B77*). Creative Commons Attribution 4.0 International.

NeurIPS impact statement risks (*NIS*). The NeurIPS impact statement dataset has an MIT License. We license the derivative NeurIPS impact statement risks dataset under Creative Commons Attribution 4.0 International.

OneStopEnglish (*OSE*). Creative Commons Attribution-ShareAlike 4.0 International.

Overruling (*Over*). Unlicensed.

Semiconductor org types (*SOT*). We license it under Creative Commons Attribution-NonCommercial 4.0 International.

Systematic review inclusion (*SRI*). Creative Commons Attribution 4.0 International

TAI safety research (*TAI*). Creative Commons Attribution-ShareAlike 4.0 International

Terms of Service (*ToS*). Unlicensed.

TweetEval Hate (*TEH*). Unlicensed.

Twitter complaints (*TC*). Unlicensed.

A.2 Dataset examples

See Table 3 for one training example from each dataset.

A.3 Task-specific instructions

Table 4 contains an excerpt from the instructions for each dataset.

Below we provide the full instructions given to human annotators and adapted for automatic baselines for each RAFT task.

ADE Corpus V2 (*ADE*)

Label the sentence based on whether it is related to an adverse drug effect (ADE). Details are described below:

Drugs: Names of drugs and chemicals that include brand names, trivial names, abbreviations and systematic names were annotated. Mentions of drugs or chemicals should strictly be in a therapeutic context. This category does not include the names of metabolites, reaction byproducts, or hospital chemicals (e.g. surgical equipment disinfectants).

Adverse effect: Mentions of adverse effects include signs, symptoms, diseases, disorders, acquired abnormalities, deficiencies, organ damage or death that strictly occur as a consequence of drug intake.

Banking77 (*B77*)

The following is a banking customer service query. Classify the query into one of the 77 categories available.

NeurIPS impact statement risks (*NIS*)

Label the impact statement based on whether it mentions a harmful application of the research done in the paper. Make sure the statement is sufficient to conclude there are harmful applications of the research being done, not a past risk that this research is solving.

Table 3: A training example from every dataset, with textual label

Dataset	Training Sample
ADE Corpus V2 (*ADE*)	`{'Sentence': 'No regional side effects were noted.', 'Label': 'not ADE-related'}`
Banking77 (*B77*)	`{'Query': 'Is it possible for me to change my P...', 'Label': 'change_pin'}`
NeurIPS impact statement risks (*NIS*)	`{'Paper title': 'Auto-Panoptic: Cooperative Mul...', 'Paper link': 'https://proceedings.neurips.cc/...', 'Impact statement': 'This work makes the first...', 'ID': '0', 'Label': "doesn't mention a harmful application"}`
OneStopEnglish (*OSE*)	`{'Article': 'For 85 years, it was just a grey b...', 'Label': 'intermediate'}`
Overruling (*Over*)	`{'Sentence': 'in light of both our holding toda...', 'Label': 'overruling'}`
Semiconductor org types (*SOT*)	`{'Paper title': '3Gb/s AC-coupled chip-to-chip ...', 'Organization name': 'North Carolina State Uni...', 'Label': 'university'}`
Systematic review inclusion (*SRI*)	`{'Title': 'Prototyping and transforming facial ...', 'Abstract': 'Wavelet based methods for prototy...', 'Authors': 'Tiddeman, B.; Burt, M.; Perrett, D.', 'Journal': 'IEEE Comput Graphics Appl', 'Label': 'not included'}`
TAI safety research (*TAI*)	`{'Title': 'Malign generalization without intern...', 'Abstract Note': "In my last post, I challenge...", 'Url': 'https://www.alignmentforum.org/posts/y...', 'Publication Year': '2020', 'Item Type': 'blogPost', 'Author': 'Barnett, Matthew', 'Publication Title': 'AI Alignment Forum', 'Label': 'TAI safety research'}`
Terms of Service (*ToS*)	`{'Sentence': 'Crowdtangle may change these term...', 'Label': 'potentially unfair'}`
TweetEval Hate (*TEH*)	`{'Tweet': 'New to Twitter-- any men on here kno...', 'Label': 'not hate speech'}`
Twitter complaints (*TC*)	`{'Tweet text': '@HMRCcustomers No this is my fi...', 'Label': 'no complaint'}`

OneStopEnglish (*OSE*)

The following is an article sourced from The Guardian newspaper, and rewritten by teachers to suit three levels of adult English as Second Language (ESL) learners: elementary, intermediate, and advanced. Predict the level of the article.

Overruling (*Over*)

In law, an overruling sentence is a statement that nullifies a previous case decision as a precedent, by a constitutionally valid statute or a decision by the same or higher ranking court which establishes a different rule on the point of law involved. Label the sentence based on whether it is overruling or not.

Dataset Name	Instructions excerpt
ADE Corpus V2 (*ADE*)	Label the sentence based on whether it is related to an adverse drug effect (ADE).
Banking77 (*B77*)	The following is a banking customer service query.
NeurIPS impact statement risks (*NIS*)	Label the impact statement as "mentions a harmful application" or "doesn't mention a harmful application" based on whether it mentions a harmful application of the research done in the paper.
OneStopEnglish (*OSE*)	The following is an article sourced from The Guardian newspaper, and rewritten by teachers to suit three levels of adult English as Second Language (ESL) learners: elementary, intermediate, and advanced.
Overruling (*Over*)	In law, an overruling sentence is a statement that nullifies a previous case decision as a precedent, by a constitutionally valid statute or a decision by the same or higher ranking court which establishes a different rule on the point of law involved.
Semiconductor org types (*SOT*)	The dataset is a list of institutions that have contributed papers to semiconductor conferences in the last 25 years, as catalogued by IEEE and sampled randomly.
Systematic review inclusion (*SRI*)	Identify whether this paper should be included in a meta-review which includes the findings of systematic reviews on interventions designed to promote charitable donations.
TAI safety research (*TAI*)	The contents of the paper are directly motivated by, and substantively inform, the challenge of ensuring good outcomes for Transformative AI.
Terms of Service (*ToS*)	According to art. 3 of the Directive 93/13 on Unfair Terms in Consumer Contracts, a contractual term is unfair if: 1) it has not been individually negotiated; and 2) contrary to the requirement of good faith, it causes a significant imbalance in the parties rights and obligations, to the detriment of the consumer.
TweetEval Hate (*TEH*)	Label whether the following tweet contains hate speech against either immigrants or women.
Twitter complaints (*TC*)	A complaint presents a state of affairs which breaches the writer's favorable expectation.

Table 4: Instructions excerpt for each dataset.

Semiconductor org types (*SOT*)

The dataset is a list of institutions that have contributed papers to semiconductor conferences in the last 25 years, as catalogued by IEEE and sampled randomly. The goal is to classify the institutions into one of three categories: "university", "company" or "research institute".

Systematic review inclusion (*SRI*)

Identify whether this paper should be included in a meta-review which includes the findings of systematic reviews on interventions designed to promote charitable donations.

Papers should be included if they meet **all** of these criteria:

1. systematic reviews, scoping reviews, or similar reproducible reviews;
2. reviews describing monetary charitable donations;
3. reviews assessing any population of participants in any context; and
4. peer reviewed and written in English (due to logistical constraints).

15

They shouldn't be included if they meet **any** of these criteria:

1. primary research reporting new data (e.g., randomised experiments);
2. non-systematic reviews, theory papers, or narrative reviews;
3. reviews on cause-related marketing; and
4. reviews of other kinds of prosocial behaviour (e.g., honesty, non-financial donations like volunteering, blood, or organ donations).

TAI safety research (*TAI*)

Transformative AI (TAI) is defined as AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution.

Label a paper as "TAI safety research" if:

1. **The contents of the paper are directly motivated by, and substantively inform, the challenge of ensuring good outcomes for TAI.** The paper need not mention TAI explicitly, but it must be motivated by it, since there are far too many papers that are merely relevant to safety. Judging motivation is, unfortunately, inherently subjective, but this is necessary to avoid penalizing papers that do not explicitly mention TAI for appearance reasons, while also not including every paper on, e.g., adversarial examples (which are motivated by capabilities and near-term safety). If the paper would likely have been written even in the absence of TAI-safety concerns, it is excluded. Ultimately, we want to support researchers who are motivated by TAI safety and allow them to find each other's work

2. **There is substantive content on AI safety, not just AI capabilities.** That said, for more speculative papers it is harder to distinguish between safety vs. not safety, and between technical vs. meta, and we err on the side of inclusion. Articles on the safety of autonomous vehicles are generally excluded, but articles on the foundations of decision theory for AGI are generally included.

3. **The intended audience is the community of researchers.** Popular articles and books are excluded. Papers that are widely released but nevertheless have substantial research content (e.g., Bostrom's Superintelligence) are included, but papers that merely try to recruit researchers are excluded.

4. **It meets a subjective threshold of seriousness/quality.** This is intended to be a very low threshold, and would, for instance, include anything that was accepted to be placed on the ArXiv. Web content not intended for review (e.g., blog posts) is only accepted if it has reached some (inevitably subjective) threshold of notability in the community. It is of course infeasible for us to document all blog posts that are about TAI safety, but we do not want to exclude some posts that have been influential but have never been published formally.

5. **Peer review is not required. White papers, preprints, and book chapters are all included.**

Otherwise, label it as "not TAI safety research".

Terms of Service (*ToS*)

Label the sentence from a Terms of Service based on whether it is potentially unfair. If it seems clearly unfair, mark it as potentially unfair.

According to art. 3 of the Directive 93/13 on Unfair Terms in Consumer Contracts, a contractual term is unfair if: 1) it has not been individually negotiated; and 2) contrary to the requirement of good faith, it causes a significant imbalance in the parties rights and obligations, to the detriment of the consumer.

Details on types of potentially unfair clauses are found below:

The **jurisdiction** clause stipulates what courts will have the competence to adjudicate disputes under the contract. Jurisdiction clauses giving consumers a right to bring disputes in their place of residence were marked as clearly fair, whereas clauses stating that any judicial proceeding takes a residence away (i.e. in a different city, different country) were marked as clearly unfair.

The **choice of law** clause specifies what law will govern the contract, meaning also what law will be applied in potential adjudication of a dispute arising under the contract. Clauses defining the applicable law as the law of the consumer's country of residence were marked as clearly fair. In every other case, the choice of law clause was considered as potentially unfair.

The **limitation of liability** clause stipulates that the duty to pay damages is limited or excluded, for certain kind of losses, under certain conditions. Clauses that explicitly affirm non-excludable providers' liabilities were marked as clearly fair. Clauses that reduce, limit, or exclude the liability of the service provider were marked as potentially unfair when concerning broad categories of losses or causes of them, such as any harm to the computer system because of malware or loss of data or the suspension, modification, discontinuance or lack of the availability of the service. Also those liability limitation clauses containing a blanket phrase like "to the fullest extent permissible by law", were considered potentially unfair. Clause meant to reduce, limit, or exclude the liability of the service provider for physical injuries, intentional damages as well as in case of gross negligence were marked as clearly unfair.

The **unilateral change** clause specifies the conditions under which the service provider could amend and modify the terms of service and/or the service itself. Such clause was always considered as potentially unfair.

The **unilateral termination** clause gives provider the right to suspend and/or terminate the service and/or the contract, and sometimes details the circumstances under which the provider claims to have a right to do so.

The **contract by using** clause stipulates that the consumer is bound by the terms of use of a specific service, simply by using the service, without even being required to mark that he or she has read and accepted them. We always marked such clauses as potentially unfair.

The **content removal** gives the provider a right to modify/delete user's content, including in-app purchases, and sometimes specifies the conditions under which the service provider may do so.

The **arbitration** clause requires or allows the parties to resolve their disputes through an arbitration process, before the case could go to court. It is therefore considered a kind of forum selection clause. However, such a clause may or may not specify that arbitration should occur within a specific jurisdiction. Clauses stipulating that the arbitration should (1) take place in a state other then the state of consumer's residence and/or (2) be based not on law but on arbiter's discretion were marked as clearly unfair. Clauses defining arbitration as fully optional would have to be marked as clearly fair.

TweetEval Hate (*TEH*)

WARNING: This task involves labeling offensive and hateful content, particularly toward immigrants and women.

Label whether the following tweet contains hate speech against either immigrants or women.

Label whether the following tweet contains hate speech against either immigrants or women. Hate Speech (HS) is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics. Detailed guidelines are provided below, please read before labeling.

Hate speech against immigrants

More specifically, HS against immigrants may include:

- insults, threats, denigrating or hateful expressions

- incitement to hatred, violence or violation of rights to individuals or groups perceived as different for somatic traits (e.g. skin color), origin, cultural traits, language, etc.

- presumed association of origin/ethnicity with cognitive abilities, propensity to crime, laziness or other vices

- references to the alleged inferiority (or superiority) of some ethnic groups with respect to others

- delegitimization of social position or credibility based on origin/ethnicity

- references to certain backgrounds/ethnicities as a threat to the national security or welfare or as competitors in the distribution of government resources
- dehumanization or association with animals or entities considered inferior

While answering the question "Is this tweet hateful?", you must take into account the following aspects:

1. the tweet content MUST have IMMIGRANTS/REFUGEES as main TARGET, or even a single individual, but considered for his/her membership in that category (and NOT for the individual characteristics)

2. we must deal with a message that spreads, incites, promotes or justifies HATRED OR VIOLENCE TOWARDS THE TARGET, or a message that aims at dehumanizing, hurting or intimidating the target

The joint presence of both elements in a tweet is considered essential to determine whether the tweet has hateful contents, therefore if both of them occur, your answer will be 'Yes'.

In case even just one of these conditions is not detected, HS (at least against immigrants) is assumed not to occur, then your answer will be 'No'.

Here a list of other aspects that are NOT considered hate speech for our purposes:

- HATE SPEECH AGAINST OTHER TARGETS
- offensive language
- blasphemy
- historical denial
- overt incitement to terrorism
- offense towards public servants and police officers
- defamation

Hate speech against women

Label the tweet as hate speech if it is misogynous against women. A tweet is misogynous if it expresses hating towards women in particular (in the form of insulting, sexual harassment, threats of violence, stereotype, objectification and negation of male responsibility).

Twitter complaints (*TC*)

A complaint presents a state of affairs which breaches the writer's favorable expectation. Label the tweet text based on whether it contains a complaint.

A.4 Dataset documentation

We provide documentation using applicable questions from the datasheets framework [14] for the *NIS*, *SOT*, and *TAI* datasets. For documentation on other datasets we refer readers to the works in which the datasets were originally introduced as cited in Section 3.1.3.

A.4.1 NeurIPS impact statement risks

The labeling section of this documentation contains information on how the impact statements were annotated based on whether they mention a harmful application. The other sections largely contain information on how the original dataset of NeurIPS impact statements [1] was collected.

Motivation

- **For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.** The original dataset was created to evaluate the then new requirement for authors to include an "impact statement" in their 2020 NeurIPS papers. Had it been successful? What kind of things did

authors mention the most? How long were impact statements on average? See [1] for more details.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** The original dataset was created as part of a project based at the Centre for the Governance of AI, which involved individual researchers and developers from the University of Oxford, Oxford Internet Institute, Harvard Kennedy School and the Alan Turing Institute.

- **Who funded the creation of dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.** The project was based at the Centre for the Governance of AI. There was no grant associated with the project. Individuals were funded by their respective organisations, or as contractors.

Composition

- **Is any information missing from individual instances in the dataset? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.** No.

- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** This dataset has limitations that should be taken into consideration when using it. In particular, the method used to collect broader impact statements involved automated downloads, conversions and scraping and was not error-proof (see https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements/blob/main/main-dataset/notes-on-data.md for details). Although care has been taken to identify and correct as many errors as possible, not all texts have been reviewed by a human. This means it is possible some of the broader impact statements contained in the dataset are truncated or otherwise incorrectly extracted from their original article. The original dataset also contains labels describing whether authors chose to effectively "opt-out" of the requirement (for example by stating that a broader impact section is "Not Applicable"). Several statements were ambiguous in this respect, and so this label represents a subjective judgement on what constituted an opt-out. The labeling performed for this paper (whether a harmful application is mentioned) also constitutes a subjective judgment, and will contain human biases. Please see the section on Preprocessing, Cleaning, Labeling for more details.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.** The dataset contains authors' names. These were scraped from publicly available scientific papers submitted to NeurIPS 2020.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.** No.

- **Does the dataset relate to people?** The dataset does not relate to people directly, although it does contain authors' names. These were scraped from publicly available scientific papers submitted to NeurIPS 2020.

Collection

- **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.** The data was directly observable (raw text scraped) for the most part; although some data was taken from previous datasets (which themselves had taken it from raw text). The data was validated, but only in part, by human reviewers. Further details can be found here: https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements/blob/main/main-dataset/notes-on-data.md

19

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?** The main dataset was collected using software, and a combination of code iteration and human review was used to validate the results. Further details may be found here: https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements/blob/main/main-dataset/notes-on-data.md.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** The subset annotated based on harmful applications was sampled randomly.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** The original dataset was created as part of a project based at the Centre for the Governance of AI, which involved individual researchers and developers as described above. The labeling for this paper (whether a harmful application is mentioned) was performed by Ought contractors.

- **Does the dataset relate to people?** The dataset does not relate to people directly, although it does contain authors' names. These were scraped from publicly available scientific papers submitted to NeurIPS 2020.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?** The impact statements were collected from the NeurIPS websites. Metadata included in the original dataset was collected from the NeurIPS chairs, and websites (for example where affiliated institutions are geographically based). See [1] for further details. The labeling for this paper (whether a harmful application is mentioned) was collected from the contractors directly.

Preprocessing, Cleaning, Labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** For the original dataset [1], the manuscript pdfs for accepted papers were obtained from the NeurIPS 2020 proceedings website. The pdfs were converted to XML, and the title and impact statement section were extracted. The dataset was appended with information about paper subject area, author names, affiliations, affiliation type and affiliation institution locations, as follows. Primary and secondary subject area, as selected by authors on submission, were supplied to us by the NeurIPS programme chairs. Author names and affiliations were obtained from separate scrapes of the NeurIPS papers. Each affiliation was tagged with a location and type (industry or academia) based on [16] and [8] respectively. Further details on the generation of the original dataset, and its assumptions and limitations, can be found at https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements/blob/main/main-dataset/notes-on-data.md. Contractors paid by Ought performed the labeling of whether impact statements mention harmful applications. A majority vote was taken from three annotators.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** The original NeurIPS impact statements data is available at https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements. The accepted papers containing the statements can also be found at https://proceedings.neurips.cc/paper/2020.

Uses

- **Has the dataset been used for any tasks already? If so, please provide a description.** An analysis of the original dataset has been prepared by the dataset authors, which can be found in Ashurst et al. [1].

- **What (other) tasks could the dataset be used for?** Other researchers are encouraged to use the dataset to provide further analysis on the outcomes of the NeurIPS broader impact requirement. The dataset could also be used for additional meta-analysis of NeurIPS 2020 accepted papers.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?** This dataset has limitations that should be taken into consideration when using it. In particular, the method used to collect broader impact statements involved automated downloads, conversions and scraping and was not error-proof. Although care has been taken to identify and correct as many errors as possible, not all texts have been reviewed by a human. This means it is possible some of the broader impact statements contained in the dataset are truncated or otherwise incorrectly extracted from their original article. More details may be found at https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements/blob/main/main-dataset/notes-on-data.md. For this paper, individual labelers were asked whether harmful applications were mentioned in the statement, but what constitutes a harmful application is of course highly subjective, and will depend on the particular views and experiences of the labeler. For example, many applications will provide some benefits to some individuals and groups, while creating risks and harms to others. The intention was to capture a rough measure of whether the authors had intended to point out potential negative effects that could arise from the use of their work, or whether they chose to limit to potential positive impacts only. This will likely exclude applications that are typically viewed as beneficial or neutral, despite the fact that such applications can cause harm to individuals or subgroups in society. We therefore urge caution in how such labels are interpreted for future tasks.

A.4.2 Semiconductor org types

This Labeling section of this documentation contains information on how the semiconductor organizations were annotated by type. The other sections mainly contain information describing how the unlabeled dataset of semiconductor organizations was collected.

Motivation

- **For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.** The data set was originally created to understand better which countries' organisations have contributed most to semiconductor R&D over the past 25 years using three main conferences. Moreover, to estimate the share of academic and private sector contributions, the organisations were classified as "university", "research institute" or "company".

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** The data science unit of Stiftung Neue Verantwortung (Berlin).

- **Who funded the creation of dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.** The Stiftung Mercator is funding the data science unit in general

Composition

- **Is any information missing from individual instances in the dataset? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.** This data set is a sample of 500 out of many more organisations. Examples where the institution names contain "universit" were deleted because all language models can classify this as "university" and no discrimination is gained.

- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** The human-created labels could be wrong.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the**

content of individuals' non-public communications)? If so, please provide a description. No.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.** No.

- **Does the dataset relate to people?** No.

Collection

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?** We used the IEEE API to obtain institutions that contributed papers to semiconductor conferences in the last 25 years. This is a random sample of 500 of them with a corresponding conference paper title. The three conferences were the International Solid-State Circuits Conference (ISSCC), the Symposia on VLSI Technology and Circuits (VLSI) and the International Electron Devices Meeting (IEDM).

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** It was probabilistic. Duplicate entries (by organisation name) were deleted.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** A student was involved and paid according to German law.

- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.** March 2021

Preprocessing, Cleaning, Labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** Yes. Contractors paid by Ought performed the labeling of organization types. A majority vote was taken from 3 annotators.

Distribution

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.** It can only be used for non-commercial research purposes. See here and here. The annotated data is licensed under Creative Commons Attribution-NonCommercial 4.0 International.

A.4.3 TAI Safety Research

Motivation

- **For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.** The primary motivations for assembling this database were to: (1) Aid potential donors in assessing organizations focusing on TAI safety by collecting and analyzing their research output. (2) Assemble a comprehensive bibliographic database that can be used as a base for future projects, such as a living review of the field.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** Angelica Deibel and myself (Jess Riedel). We did not do it on behalf of any entity.

- **Who funded the creation of dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.** I volunteered my own time and paid Angelica Deibel for her time from my personal funds.

Composition

- **Is any information missing from individual instances in the dataset? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.** Not really sure what this means for our case. There's no redacted information, but there are undoubtedly tons of papers we failed to find in our literature search. Also, we kept/excluded articles based on a set of subjective criteria we invented, and we undoubtedly made mistakes applying this criteria.

- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** See above. No redundancies that I know of.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.** No.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.** No.

- **Does the dataset relate to people? If not, you may skip the remaining questions in this section.** Sort of. It's a database of papers, and those papers have authors.

- **Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.** It's a database of papers, and those papers have authors. This information is already public.

- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.** No.

Collection

- **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.** We asked TAI safety organizations for what their employees had written, emailed some individual authors, and searched Google Scholar. See the LessWrong post for more details: https://www.lesswrong.com/posts/4DegbDJJiMX2b3EKm/tai-safety-bibliographic-database

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?** Mostly be hand. We collected citation information using an automated API call to Google Scholar. See the LessWrong post for more details: https://www.lesswrong.com/posts/4DegbDJJiMX2b3EKm/tai-safety-bibliographic-database

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** It was Angelica Deibel and me. I volunteered and paid Angelica $20/hour.

- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.** It was collected haphazardly between in 2019 and 2020

- **Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.** No.

- **Does the dataset relate to people?** It's a database of papers, which have authors.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?** Both.

- **Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.** We asked authors to suggest papers that should be included in the database.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.** No.

Preprocessing, Cleaning, Labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** Yes. See the LessWrong post for more details on our labels, which was done largely by hand, on citation numbers, collected from Google Scholar by automated API call, and on the basic bibliographic information, which was collected with the automated tools from Zotero: https://www.lesswrong.com/posts/4DegbDJJiMX2b3EKm/tai-safety-bibliographic-database

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** No. There was no clean distinction between raw and processed data. We used several automated tools that interacted, plus corrections and additions by hand.

- **Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.** See link to the Citation numbers API called for Google Scholar in the the LessWrong post for more details: https://www.lesswrong.com/posts/4DegbDJJiMX2b3EKm/tai-safety-bibliographic-database

Uses

- **Has the dataset been used for any tasks already? If so, please provide a description.** Yes, for the report we posted on LessWrong here. It was also used by "Larks" in his review.

- **Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.** No, this hasn't been used in any academic papers yet.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?** No.

- **Are there tasks for which the dataset should not be used? If so, please provide a description.** Don't use it to create a dangerous AI that could bring the end of days.

Distribution

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license**

and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. As mentioned in the LessWrong post: "We release the Zotero database under the Creative Commons Attribution-ShareAlike 4.0 International License. In short, the means you are free to use, modify, and reproduce the database for anything so long as you cite us and release any derivative works under the same license."

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.** No. The CC-SA-BY license is the only restriction

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.** No.

A.5 GPT-3 baseline details

The code for the GPT-3 baseline is available at https://raft.elicit.org/baselines under an MIT license. Running the automatic baseline of GPT-3 davinci on the test sets cost approximately $2,600.

A.5.1 Parameter selection

Tables 5, 6 and 7 detail the results of parameter selection runs. All runs were done using GPT-3.

We mistakenly use 50 rather than 25 training examples in the prompt for *TEH* when running in-context baselines, despite 25 performing better in LOOCV.

When running in-context baselines besides GPT-3, we use the same number of training examples in the prompt. Note that this may be suboptimal due to other models having smaller context windows; we leave improving upon these baselines to future work.

Instructions	Avg	ADE	B77	NIS	OSE	Over	SOT	SRI	TAI	ToS	TEH	TC
Task-Specific	**0.593**	0.752	0.081	0.566	0.231	0.940	0.777	0.495	0.480	0.691	0.731	0.780
Generic	**0.551**	0.797	0.052	0.476	0.184	0.899	0.717	0.495	0.462	0.438	0.708	0.830

Table 5: LOO Cross Validation performance for task-specific versus generic instructions, F1 scores. The experiment was run with 20 training examples for all datasets and no semantic selection.

Selection	Avg	ADE	B77	NIS	OSE	Over	SOT	SRI	TAI	ToS	TEH	TC
Semantic	**0.622**	0.696	0.098	0.635	0.454	0.940	0.716	0.419	0.696	0.578	0.778	0.836
Random	**0.593**	0.752	0.081	0.566	0.231	0.94	0.777	0.495	0.480	0.691	0.731	0.780

Table 6: LOO Cross Validation performance for semantic versus random training example selection, F1 scores. The experiment was run with 20 training examples for all datasets and task-specific instructions.

A.6 AdaBoost baseline details

We concatenated all non-label data in every training example into a single string, separated by periods, then constructed n-grams from all words and adjacent sets of n words in the dataset for $n \in [1, 5]$ after removing letter cases and certain special symbols. Each training or test example was vectorized as the count of each n-gram in the example. For the base estimator, we used decision trees with a maximum depth of 3. We ensembled 100 estimators with a learning rate of 1.0.

Examples	Avg	ADE	B77	NIS	OSE	Over	SOT	SRI	TAI	ToS	TEH	TC
5	0.611	0.696	0.076	0.571	**0.528**	0.860	**0.745**	0.474	**0.672**	**0.789**	0.618	0.688
10	0.593	0.667	**0.096**	0.559	0.456	0.920	0.623	**0.479**	0.642	0.610	0.735	0.733
25	0.617	**0.714**	0.090	**0.740**	0.445	**0.960**	0.591	0.412	0.643	0.548	**0.778**	**0.862**
49	0.598	0.653	0.074	0.643	0.394	0.960	0.692	0.375	0.586	0.643	0.718	0.842

Table 7: LOO Cross Validation performance for number of training examples, F1 scores. The experiment was run with task-specific instructions and semantic selection of training examples.

We tuned several hyperparameters in our AdaBoost implementation. First, we tested the learning rate of AdaBoost, the rate at which the weights of the ensembled classifiers are changed, finding that it didn't change results substantially from within a reasonable range. We then tested a number of different depths of decision trees in the ensemble, finding that low depths were ideal. Finally, we tested the number of trees to ensemble, finding that around 50 to 100 trees perform the best. All hyperparameters were tuned with leave-one-out cross validation.

Learn Rate	Avg	ADE	B77	NIS	OSE	Over	SOT	SRI	TAI	ToS	TEH	TC
0.03	**0.547**	0.587	0.012	0.760	0.344	0.900	0.538	0.495	0.720	0.621	0.475	0.561
0.1	**0.544**	0.587	0.004	0.760	0.344	0.900	0.515	0.495	0.720	0.621	0.475	0.561
0.3	**0.536**	0.714	0.009	0.667	0.352	0.919	0.422	0.495	0.660	0.642	0.451	0.561
1.0	**0.548**	0.636	0.000	0.718	0.444	0.919	0.385	0.495	0.699	0.621	0.538	0.569
3.0	**0.410**	0.643	0.000	0.392	0.432	0.597	0.284	0.495	0.653	0.368	0.333	0.319

Table 8: LOO Cross Validation performance for learning rate, F1 scores from an AdaBoost ensemble classifier of 50 depth-1 decision trees trained on n-grams of the dataset for $n \in [1, 5]$.

Depth	Avg	ADE	B77	NIS	OSE	Over	SOT	SRI	TAI	ToS	TEH	TC
1	**0.546**	0.636	0.000	0.718	0.466	0.919	0.385	0.495	0.699	0.621	0.494	0.569
2	**0.511**	0.592	0.000	0.716	0.405	0.900	0.366	0.495	0.466	0.527	0.626	0.524
3	**0.549**	0.721	0.004	0.735	0.463	0.919	0.366	0.495	0.507	0.621	0.626	0.586
4	**0.531**	0.556	0.000	0.672	0.410	0.880	0.438	0.495	0.607	0.602	0.524	0.653
5	**0.506**	0.619	0.004	0.583	0.318	0.860	0.360	0.495	0.451	0.602	0.592	0.684

Table 9: LOO Cross Validation performance for depth of trees, F1 scores from an AdaBoost ensemble classifier of 50 decision trees with learning rate 1.0 trained on n-grams of the dataset for $n \in [1, 5]$.

A.7 Human baseline details

A.7.1 Labeling process

For each dataset, we first conduct a qualification phase with 20 data points from the training set, showing labelers the other 30 as reference examples. Labelers who label at least 10 data points and achieved at least median accuracy advance to the annotation phase. In the annotation round, we

# Trees	Avg	ADE	B77	NIS	OSE	Over	SOT	SRI	TAI	ToS	TEH	TC
10	**0.547**	0.649	0.000	0.616	0.394	0.860	0.516	0.495	0.660	0.691	0.582	0.554
50	**0.544**	0.636	0.000	0.697	0.482	0.919	0.381	0.495	0.699	0.621	0.491	0.569
100	**0.554**	0.667	0.000	0.756	0.318	0.919	0.390	0.495	0.697	0.642	0.592	0.616
500	**0.537**	0.649	0.000	0.814	0.305	0.919	0.385	0.495	0.557	0.642	0.592	0.548

Table 10: LOO Cross Validation performance for number of trees, F1 scores from an AdaBoost ensemble classifier with learning rate 1.0 trained on n-grams of the dataset for $n \in [1, 5]$.

collect 5 labels for each of the 100 data points. We then take the plurality vote for each data point, breaking ties randomly.

Due to extreme class imbalance, we conduct only an annotation phase of 200 data points for the *SRI* dataset.

A.7.2   Instructions

We attempted to mimic annotation instructions reported by the works introducing datasets whenever possible. The instructions we gave to annotators was as follows (parts enclosed in brackets denote variations in the instructions depending on the task or phase):

> [If qualification phase: This task will serve as a qualification stage for annotation on a larger set. Label at least 10 examples to be considered for qualification for the annotation task. Please only complete this qualification if you're available to label 100 more data points in the next day.]
>
> [Task-specific instructions]
>
> There are 50 [30 if qualification phase] labeled examples here to help you. If it seems that the instructions and examples are in conflict, use the examples as a guide.
>
> You may use info on the internet (e.g. Google searches) to help you.
>
> We know that labeling accuracy will (a) vary some based on level of background knowledge and (b) have some inherent subjectivity. Please select your best guess for each data point.

Task-specific instructions are detailed in Section A.3.

A.7.3   Costs

We spent $2,030 compensating crowdworkers for human baselines. We conservatively estimate that workers were paid $15/hr.