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).

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

- **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.