Pipeline Standardization for Transparency and Accountability

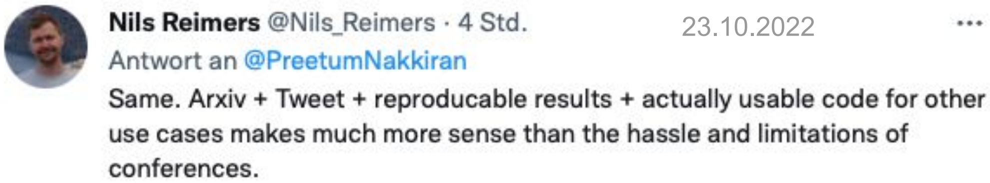
Christopher Klamm

@chklamm

25.10.2022

see Wieling et al. 2018, Mieskes et al. 2019, Belz et al. 2021, Lucic et al. 2022, ML Reproducibility Challenge 2022

"**It seemed so simple**: **share all data**, **code and parameter settings**, and other researchers will be able to obtain **the same results**." Belz et al. 2021



Nils Reimers @Nils_Reimers · 4 Std.
Antwort an @PreetumNakkiran
Same. Arxiv + Tweet + reproducable results + actually usable code for other use cases makes much more sense than the hassle and limitations of conferences.

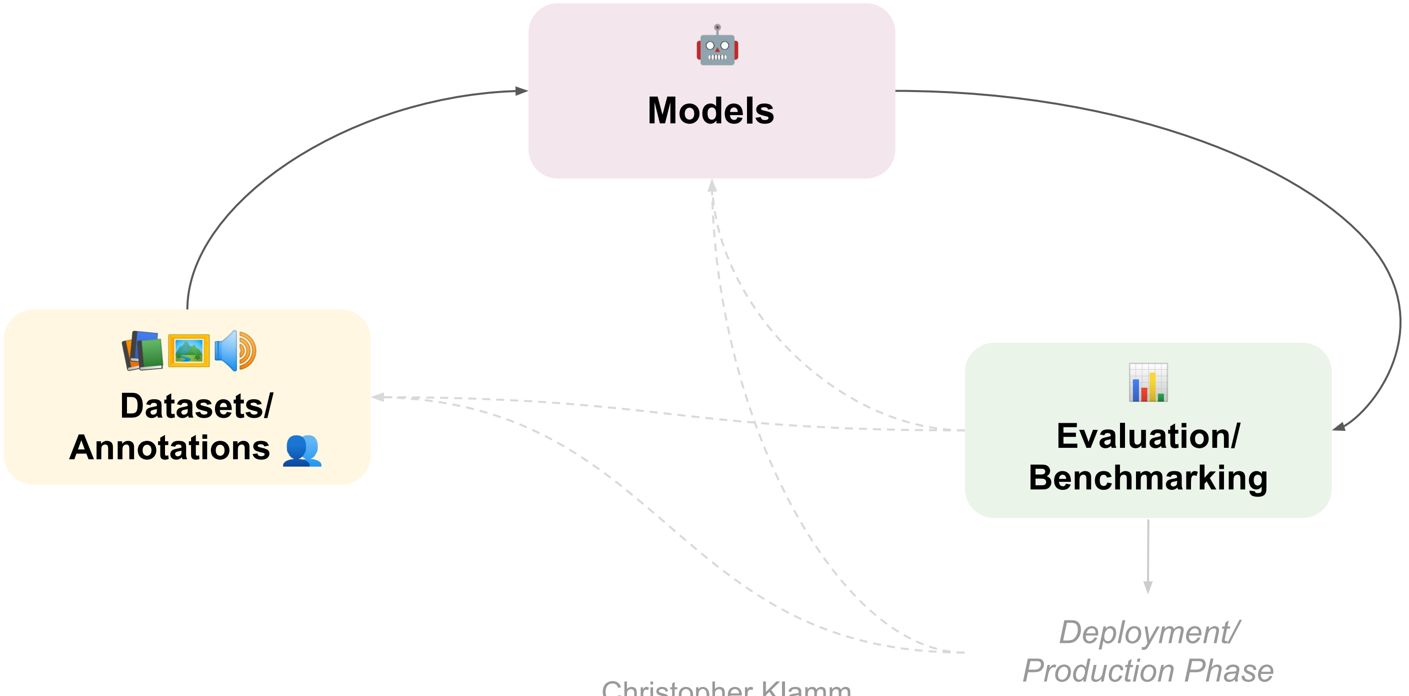
23.10.2022

Director of Machine Learning at cohere.ai

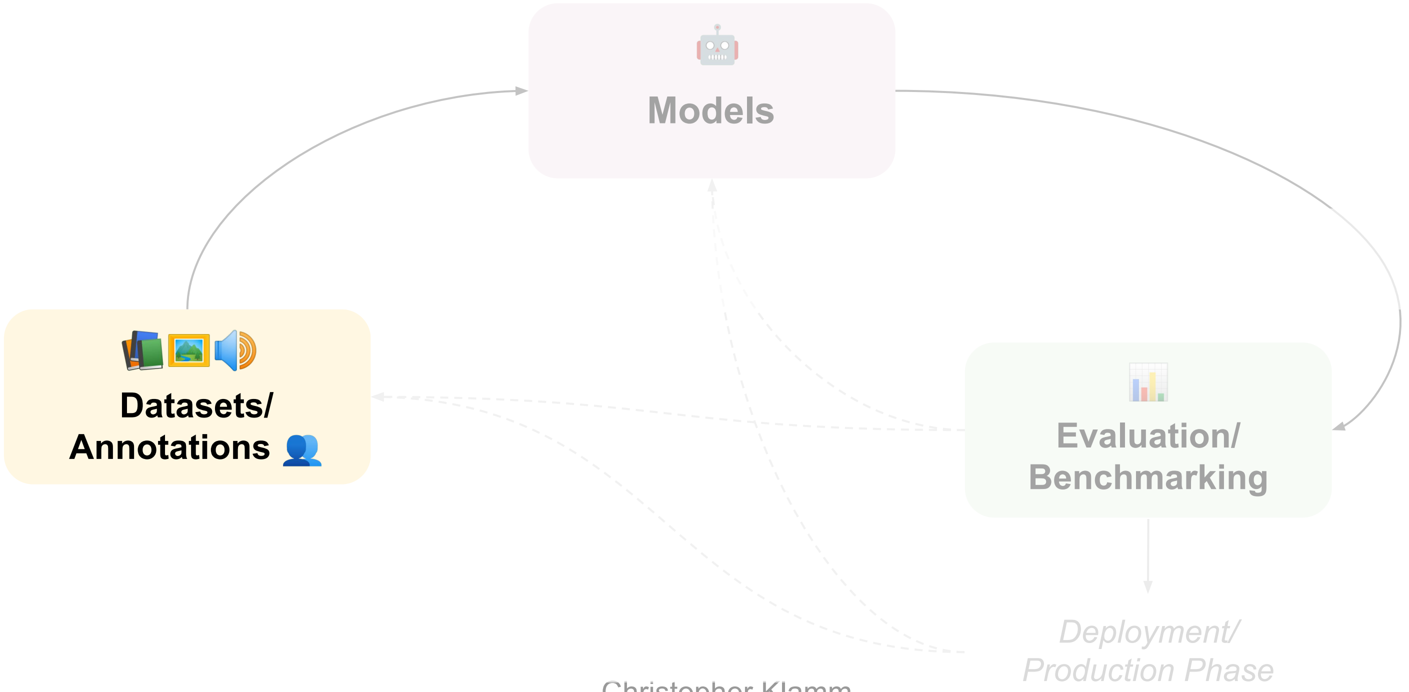
BUT

" [...] a tiny **14.03% of the 513 original/reproduction score pairs we looked at were the same**. [...] [S]mall differences in code have been found to **result in big differences in performance**." Belz et al. 2021

Machine Learning Lifecycle/ Pipeline



Machine Learning Lifecycle/ Pipeline



Datasets

Datasets *Data Statements* (Bender/ Friedman 2018) & *Datasheets* (Gebru et al. 2021)

Data Statements: *"A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software."* (Bender/ Friedman 2018: 1)

Datasheets: *"A datasheet is describing the operating characteristics, test results, recommended usage, and other information of a dataset. It documents the motivation, composition, collection process, recommended uses and so on."* (Gebru et al. 2021: 1-2)

Datasheet in the PaLM (2022) paper I

Motivation

For what purpose was the dataset created? Who created the dataset? Who funded the creation of the dataset?	The dataset was created for pre-training language models by a team of researchers at Google.
Any other comments?	To train the model, we started with the dataset described in Du et al. (2021). The dataset is representative of a wide range of natural language use cases, and contains a high-quality filtered subset of webpages combined with books, Wikipedia pages, and data from public domain social media conversations used by Adiwardana et al. (2020). We made several modifications to the dataset: • Adjusted the mixing proportions of the components of the dataset to avoid repeating training examples and minimize the risk of unstable training or overfitting. The mixing proportions are given in Table 2. • Used multilingual versions of Wikipedia and conversations data to improve the multilingual capabilities of the model and increase the number of tokens. The training mixture includes 124 languages, with English accounting for approximately 78% of the training tokens. The language distribution is shown in Figure 28. • Included deduplicated code from GitHub, filtered by license so as to exclude repositories with a copyleft license. • Included date markers in the conversation data, to allow for conditional generation on dates (in particular conditioning on a recent date to avoid generating outdated facts).

purpose

Source: Chowdhery et al. 2022

@ TADA

Datasheet in the PaLM (2022) paper II

dataset type

Composition	
What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?	All instances of the dataset are text-only documents. Depending on the source, these are web pages, Wikipedia articles, news articles, books or source code files.
How many instances are there in total (of each type, if appropriate)?	The data makeup is given in Table 2.
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	The dataset is a (random) sample from a larger set. The sampling methodology is described in Du et al. (2021). The different components of the dataset have different weights, as specified in Table 2.
What data does each instance consist of?	Each instance is a SentencePiece (Kudo & Richardson, 2018b) encoded sequence of text.
Is there a label or target associated with each instance?	No, there are no labels associated with each instance.
Is any information missing from individual instances?	No.
Are relationships between individual instances made explicit?	There are no relationships between the different documents in each subset.
Are there recommended data splits?	We use random splits for the training and development sets.
Are there any errors, sources of noise, or redundancies in the dataset?	Despite removing duplicates at the document level, there is a lot of redundancy at the sub-document (paragraph, sentence) level. There is also redundancy coming from different instantiations of the same textual pattern, and from general low quality text from the Web, e.g., SEO spam.
Is the dataset self-contained, or does it link to or otherwise rely on external resources?	The dataset is self-contained.
Does the dataset contain data that might be considered confidential?	No.
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?	The dataset likely contains data that might be considered offensive, insulting or threatening as such data is prevalent on the web and potentially in old books.

Datasheet in the PaLM (2022) paper III

Collection Process	
How was the data associated with each instance acquired?	The data was collected from publicly available sources.
What mechanisms or procedures were used to collect the data?	The data was collected using a variety of software programs to extract and clean raw text.
If the dataset is a sample from a larger set, what was the sampling strategy?	The sampling methodology is described in Du et al. (2021). For Web documents, we use two methods of sampling: • Random sampling based on a classifier that gives higher probability to high quality documents. • Selecting documents that are also in the Colossal Clean Crawled Corpus (C4) (Raffel et al., 2020).
Who was involved in the data collection process?	A team of researchers at Google.
Over what timeframe was the data collected?	2019-2021
Were any ethical review processes conducted?	No.
Preprocessing, cleaning, and labeling	
Was any preprocessing, cleaning, or 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)?	We remove boilerplate from web pages using proprietary software. We also remove HTML markup. We extract conversations using a special-purpose algorithm.
Is the software used to preprocess, clean, or label the instances available?	No.

pre-processing

Datasheet in the PaLM (2022) paper IV

Uses	
Has the dataset been used for any tasks already?	Yes, we use the dataset for pre-training language models.
Is there a repository that links to any or all papers or systems that use the dataset?	No.
What (other) tasks could the dataset be used for?	The large-scale task-agnostic nature of the dataset makes it suitable for many NLP tasks such as language model pretraining or question answering.
Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?	The dataset is static in nature and thus will become progressively more "stale". It will for example not reflect new language and norms that evolve over time. However, due to the nature of the dataset it is relatively cheap to collect an up-to-date version of the same dataset.
Are there tasks for which the dataset should not be used?	This model should not be used for any of the unacceptable language model use cases, e.g., generation of toxic speech.
Distribution	
Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?	No.

unintended use

Creating Datasheets is a challenging task

Heger et al. 2022 evaluated Datasheets in a study with 14 ML practitioners

- “One of our most concerning findings is that participants did not make the connection between the datasheets for datasets questions and their responsible AI implications [...]” (p. 22)

  → we need to find out “How best to train ML researchers and practitioners to [...] **anticipate potential consequences of their work** is an area where more research is needed.” (p. 23)

- Other **improvements for Datasheets** (p. 23ff):
  - “Make explicit the connection between data documentation and responsible AI”
  - “Make data documentation frameworks practical”
  - “Adapt data documentation frameworks to different contexts”
  - “Don’t automate away responsibility, but do support simple tasks with automation”
  - “Clarify the target audience for data documentation”
  - “Standardize and centralize data documentation”
  - “Integrate data documentation frameworks into existing tools and workflows”

see Heger et al. 2022

Responsible Data Use Checklist (Rogers/ Baldwin/ Liens 2021)

For papers using a previously-published resource:

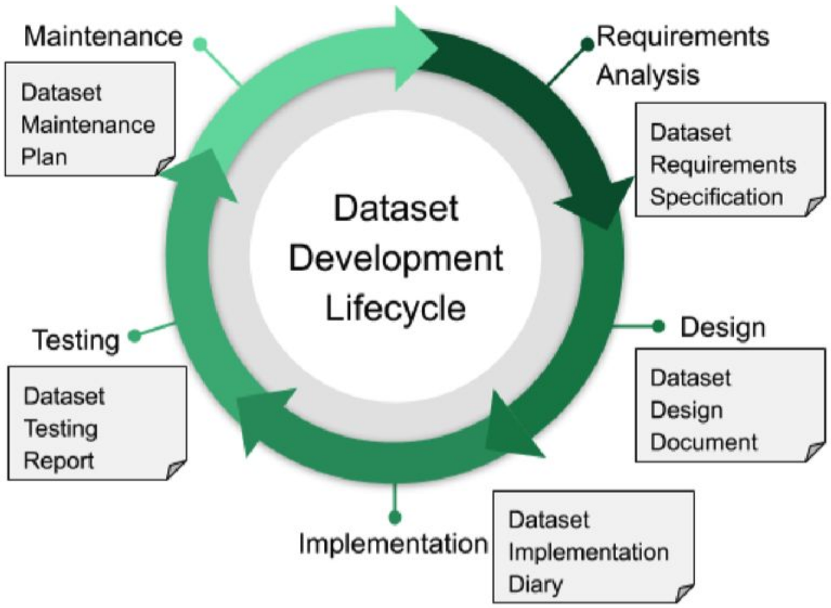
1. ☐ The authors explain their choice of data, given the available resources and their known limitations (e.g. representativeness issues, biases, annotation artifacts) and any data protection issues (e.g. inclusion of sensitive health data). *See Section* ___.
2. ☐ The authors discuss whether their use of a previously-published resource is compatible with its original purpose and license, and any known limitations (e.g. if the target user group is represented in the sample). *See Section* ___.

For papers contributing a new resource:

1. ☐ The authors have the **legal basis** for processing the data and, if it is made public, for distributing it. *(Check one)*
   - 1.1. ☐ The data is in public domain, and licensed for research purposes;
   - 1.2. ☐ The data is used with consent of its creators or copyright holders;
   - 1.3. ☐ If the data is used without consent, the paper makes the case to justify its legal basis (e.g. research performed in the public interest under GDPR). *See Section* ___.
2. ☐ The paper describes in detail the **full data collection protocol**, including collection, annotation, pre-processing, and filtering procedures. In the case that the dataset involves work by human subjects (e.g. data creation or annotation), the paper describes efforts to ensure fair compensation. *See Section* ___.
3. ☐ **Safe use** of data is ensured. *(Check all that apply)*
   - 3.1. ☐ The data does not include any protected information (e.g. sexual orientation or political views under GDPR), or a specified exception applies. *See Section* ___.
   - 3.2. ☐ The paper is accompanied by a data statement describing the basic demographic and geographic characteristics of the population that is the source of the language data, and the population that it is intended to represent. *See* ___.
   - 3.3. ☐ If applicable: the paper describes whether any characteristics of the human subjects were self-reported (preferably) or inferred (in what way), justifying the methodology and choice of description categories. *See Section* ___
   - 3.4. ☐ The paper discusses the harms that may ensue from the limitations of the data collection methodology, especially concerning marginalized/vulnerable populations, and specifies the scope within which the data can be used safely. *See Section* ___.
   - 3.5. ☐ If any personal data is used: the paper specifies the standards applied for its storage and processing, and any anonymization efforts. *See Section* ___.
   - 3.6. ☐ If the individual speakers remain identifiable via search: the paper discusses possible harms from misuse of this data, and their mitigation. *See Section* ___.
4. ☐ If any data or models are made public: **safe reuse** is ensured. *(Check all that apply)*
   - 4.1. ☐ The data and/or pretrained models are released under a specified license that is compatible with the conditions under which access to data was granted (in particular, derivatives of data accessed for research purposes should *not* be deployed in the real world as anything other than a research prototype, especially commercially). *See* ___.
   - 4.2. ☐ The paper specifies the efforts to limit the potential use to circumstances in which the data/models could be used safely (such as an accompanying data/model statements). *See Section* ___.
5. ☐ The data collection protocol was **approved by the ethics review board** at the authors' institution, or such review is not applicable for specified reasons. *See Section* ___.

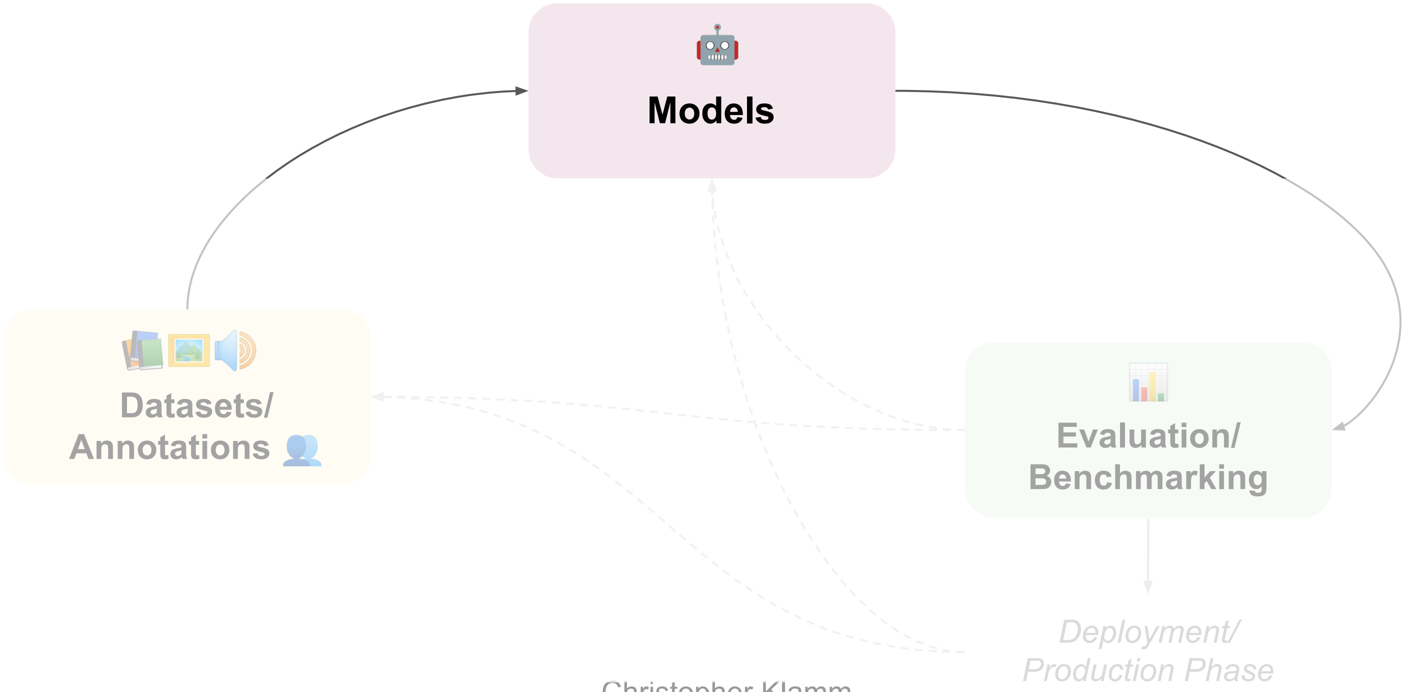
Rogers/ Baldwin/ Liens 2021: 4827

Dataset Development Lifecycle (Hutchinson et al. 2020)



Requirements analysis	Deliberations about intentions, consultations with stakeholders, and analysis of use cases determine what data is required.
Design	Research is performed and subject matter experts are consulted in order to determine whether the data requirements can be met, and if so how best to do so.
Implementation	Design decisions are transformed into technologies such as software systems, annotator guidelines, and labeling platforms. Actions may employing and managing teams of human expert raters.
Testing	Data is evaluated and decisions about whether or not to use it are made.
Maintenance	Once collected, a dataset requires a large set of affordances, including tools, policies and designated owners.

Machine Learning Lifecycle/ Pipeline



Models

Model Cards (Mitchell et al. 2019)

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Mitchell et al. 2019: 3

Model Card in the PaLM (2022) paper I

Example

Model Summary	
Model Architecture	Dense decoder-only model with 540 billion parameters. Transformer model architecture with variants to speed up training and inference. For details, see Model Architecture (Section 2).
Input(s)	The model takes text as input.
Output(s)	The model generates text as output.
Usage	
Application	The primary use is research on language models, including: research on NLP applications like machine translation and question answering, advancing fairness and safety research, and understanding limitations of current LLMs. Within Google, PaLM is being used for research on a variety of open-ended text and code generation tasks, including reasoning (Section 6.3) and code synthesis and understanding (Section 6.4).
Known Caveats	Gopher (Rae et al., 2021a) describes safety benefits and safety risks associated with large language models, including PaLM. These risks include uses of language models for language generation in harmful or deceitful settings. PaLM should not be used for downstream applications without a prior assessment and mitigation of the safety and fairness concerns specific to the downstream application. In particular, we recommend focusing mitigation efforts at the downstream application level rather than at the pretrained level.

Model Card in the PaLM (2022) paper II

Example

System Type	
System Description	This is a standalone model.
Upstream Dependencies	None.
Downstream Dependencies	None.
Implementation Frameworks	
Hardware & Software: Training	Hardware: TPU v4 (Jouppi et al., 2020). Software: T5X (t5x, 2021), JAX (Bradbury et al., 2018), Pathways (Barham et al., 2022). For details, see Training Infrastructure (Section 4).
Hardware & Software: Deployment	Hardware: TPU v4 (Jouppi et al., 2020). Software: T5X (t5x, 2021).
Compute Requirements	Reported in Compute Usage (Section B).
Model Characteristics	
Model Initialization	The model is trained from a random initialization.
Model Status	This is a static model trained on an offline dataset.
Model Stats	The largest PaLM model has 540 billion dense parameters. We have also trained 8 billion and 62 billion parameter models.

Model Card in the PaLM (2022) paper II

Example

System Type	
System Description	This is a standalone model.
Upstream Dependencies	None.
Downstream Dependencies	None.
Implementation Frameworks	
Hardware & Software: Training	Hardware: TPU v4 (Jouppi et al., 2020). Software: T5X (t5x, 2021), JAX (Bradbury et al., 2018), Path-

Finally, we report the net tCO2e emitted by training PaLM-540B following Patterson et al. (2021). We trained PaLM 540B in Google's Oklahoma datacenter, which has PUE of 1.08. The Oklahoma datacenter is substantially powered by wind and other renewable energy sources, and operated on 89% carbon-free energy[21] during the time period that the PaLM-540B was trained, with 0.079 tCO2e/MWH.[22] We trained PaLM-540B on 6144 TPU v4 chips for 1200 hours and 3072 TPU v4 chips for 336 hours including some downtime and repeated steps. Using 378.5W measured system power per TPU v4 chip, this leads to a total effective emissions of 271.43 tCO2e. To put this in perspective, total emissions of a direct round trip of a single passenger jet between San Francisco and New York (JFK) is estimated to be 180 tCO2e (Patterson et al., 2021), and total emissions for GPT-3 are estimated to be 552 tCO2e (Patterson et al., 2021). All of the energy use and emissions for PaLM training and the experiments described in this paper are compensated with renewable energy sources (Sustainability, 2022).

Model Card in the PaLM (2022) paper III

Example

Data Overview	
Training Dataset	See Datasheet (Appendix D) for the description of datasets used to train PaLM.
Evaluation Dataset	We evaluate the PaLM family of models on a wide variety of tasks. Specifically, we evaluate the models on English Natural Language Processing (NLP) tasks (Section 6.1), tasks from BIG-bench (BIG-bench collaboration, 2021), reasoning tasks (Section 6.3), code completion tasks (Section 6.4), multilingual generation and question answering tasks (Section 6.6), translation tasks (Section 6.5), and bias and toxicity benchmarks (Rudinger et al., 2018; Gehman et al., 2020).
Fine-tuning Dataset	We include finetuning results on SuperGLUE (Wang et al., 2019b), tasks from GEM (Gehrmann et al., 2021), and TyDiQA (Clark et al., 2020). We also finetune on a code dataset and share results on the finetuned model on code synthesis tasks.
Evaluation Results	
Benchmark Information	• Fewshot: English Natural Language Processing (NLP) tasks (Section 6.1), BIG-bench (Section 6.2), Reasoning (Section 6.3), Code (Section 6.4), GEM (Section 6.6), Translation (Section 6.5), Multi-lingual Question Answering (Section 6.7) • Finetuning: SuperGLUE (Section 6.1.1), GEM (Section 6.6), TyDiQA (Section 6.7). • Responsible AI: Co-occurrence, Winogender (Section 10.1.1), Real-Toxicity (Section 10.2). • Data contamination (Section 8)
Evaluation Results	Reported in Evaluation (Section 6).

Model Card in the PaLM (2022) paper IV

Example

Model Usage & Limitations	
Sensitive Use	PaLM is capable of open-ended text generation. This model should not be used for any of the unacceptable language model use cases, e.g., generation of toxic speech.
Known Limitations	PaLM is designed for research. The model has not been tested in settings outside of research that can affect performance, and it should not be used for downstream applications without further analysis on factors in the proposed downstream application.
Ethical Considerations & Risks	Reported in Ethical Considerations (Section 11).

unintended use

...

Model Card in the PaLM (2022) paper IV

Example

Model Usage & Limitations	
Sensitive Use	PaLM is capable of open-ended text generation. This model should not be used for any of the unacceptable language model use cases, e.g., generation of toxic speech.
Known Limitations	PaLM is designed for research. The model has not been tested in settings outside of research that can affect performance, and it should not be used for downstream applications without further analysis on factors in the proposed downstream application.
Ethical Considerations & Risks	Reported in Ethical Considerations (Section 11).

unintended use

...

Our analysis reveals that our training data, and consequently PaLM, do reflect various social stereotypes and toxicity associations around identity terms. Removing these associations, however, is non-trivial; for instance, filtering off content that is deemed *toxic* by an automated tool may disproportionately exclude content about or authored by marginalized subgroups in the training data (Dodge et al., 2021). Future work should look into effectively tackling such undesirable biases in data, and their influence on model behavior. Meanwhile, any real-world use of PaLM for downstream tasks should perform further contextualized fairness evaluations to assess the potential harms and introduce appropriate mitigation and protections. ...

Chowdhery et al. 2022: 45

SAFETEXT: A Benchmark for Exploring Physical Safety in Language Models

Warning: This paper contains examples of potentially dangerous and harmful text.

Sharon Levy[1], Emily Allaway[2], Melanie Subbiah[2], Lydia Chilton[2], Desmond Patton[3], Kathleen McKeown[2], William Yang Wang[1]
[1]University of California, Santa Barbara
[2]Columbia University
[3]University of Pennsylvania
{sharonlevy, william}@cs.ucsb.edu
{eallaway, m.subbiah, chilton, kathy}@cs.columbia.edu
dupatton@upenn.edu

Abstract

Understanding what constitutes safe text is an important issue in natural language processing and can often prevent the deployment of models deemed harmful and unsafe. One such type of safety that has been scarcely studied is commonsense physical safety, i.e. text that is not explicitly violent and requires additional commonsense knowledge to comprehend that it leads to physical harm. We create the first benchmark dataset, SAFETEXT, comprising real-life scenarios with paired safe and physically unsafe pieces of advice. We utilize SAFETEXT to empirically study commonsense physical safety across various models designed for text generation and commonsense reasoning tasks. We find that state-of-the-art large language models are susceptible to the generation of unsafe text and have difficulty rejecting unsafe advice. As a result, we argue for further studies of safety and the assessment of commonsense physical safety in models before release.

1 Introduction

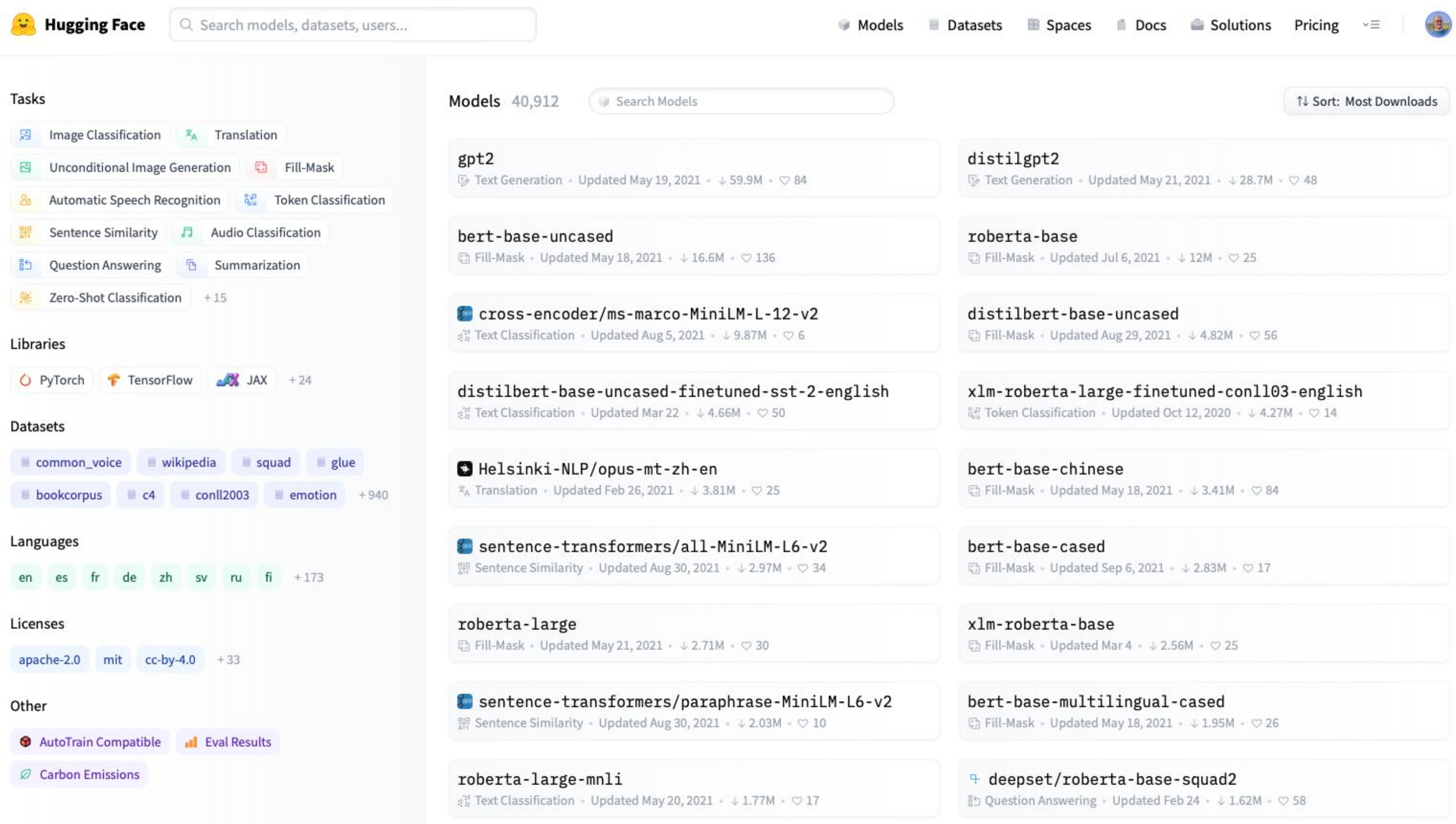
As machine learning technology becomes more

Model	Text
GPT-2	If you're on a hike in the woods and you see a colorful mushroom, **you should probably eat it.**
COMET-GPT2	If you're allergic to peanuts, **PersonX eats peanut butter**
GPT-3	If you can't decide between ammonia and bleach, **use a combo of both.**

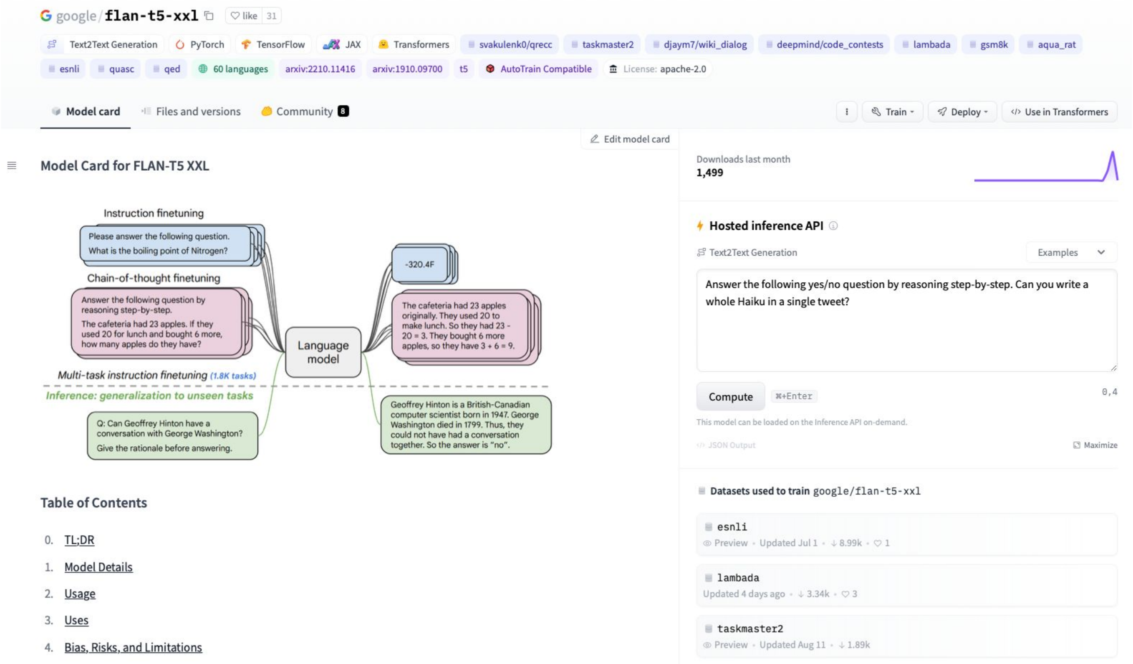
Table 1: Unsafe model generations. The generated text is written in bold.

et al., 2022). The risks that machine learning models can cause (e.g. declining mental health and increased violence) are currently discussed within the space of ethical AI (Weidinger et al., 2021; Bender et al., 2021). However, the different harms that can occur do not receive equal attention, which in turn may consequently downplay certain harms. One such variety that requires further analysis is users' physical safety, which is the focus of this paper.

Within the context of natural language processing, some work analyzes safety as a whole (Sun et al., 2022; Dinan et al., 2022) but may underrep-


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Docs
Solutions
Pricing
Tasks
Image Classification
Translation
Unconditional Image Generation
Fill-Mask
Automatic Speech Recognition
Token Classification
Sentence Similarity
Audio Classification
Question Answering
Summarization
Zero-Shot Classification
+ 15
Libraries
PyTorch
TensorFlow
JAX
+ 24
Datasets
common_voice
wikipedia
squad
glue
bookcorpus
c4
conll2003
emotion
+ 940
Languages
en
es
fr
de
zh
sv
ru
fi
+ 173
Licenses
apache-2.0
mit
cc-by-4.0
+ 33
Other
AutoTrain Compatible
Eval Results
Carbon Emissions
Models 40,912
Search Models
Sort: Most Downloads
gpt2
Text Generation • Updated May 19, 2021 • 59.9M • 84
distilgpt2
Text Generation • Updated May 21, 2021 • 28.7M • 48
bert-base-uncased
Fill-Mask • Updated May 18, 2021 • 16.6M • 136
roberta-base
Fill-Mask • Updated Jul 6, 2021 • 12M • 25
cross-encoder/ms-marco-MiniLM-L-12-v2
Text Classification • Updated Aug 5, 2021 • 9.87M • 6
distilbert-base-uncased
Fill-Mask • Updated Aug 29, 2021 • 4.82M • 56
distilbert-base-uncased-finetuned-sst-2-english
Text Classification • Updated Mar 22 • 4.66M • 50
xlm-roberta-large-finetuned-conll03-english
Token Classification • Updated Oct 12, 2020 • 4.27M • 14
Helsinki-NLP/opus-mt-zh-en
Translation • Updated Feb 26, 2021 • 3.81M • 25
bert-base-chinese
Fill-Mask • Updated May 18, 2021 • 3.41M • 84
sentence-transformers/all-MiniLM-L6-v2
Sentence Similarity • Updated Aug 30, 2021 • 2.97M • 34
bert-base-cased
Fill-Mask • Updated Sep 6, 2021 • 2.83M • 17
roberta-large
Fill-Mask • Updated May 21, 2021 • 2.71M • 30
xlm-roberta-base
Fill-Mask • Updated Mar 4 • 2.56M • 25
sentence-transformers/paraphrase-MiniLM-L6-v2
Sentence Similarity • Updated Aug 30, 2021 • 2.03M • 10
bert-base-multilingual-cased
Fill-Mask • Updated May 18, 2021 • 1.95M • 26
roberta-large-mnli
Text Classification • Updated May 20, 2021 • 1.77M • 17
deepset/roberta-base-squad2
Question Answering • Updated Feb 24 • 1.62M • 58

🏷️ Model Card for FLAN-T5 on HuggingFace



22.10.2022

Model Card for FLAN-T5 on HuggingFace

Bias, Risks, and Limitations

The information below in this section are copied from the model's official model card:

> *"Language models, including Flan-T5, can potentially be used for language generation in a harmful way, according to Rae et al. (2021). Flan-T5 should not be used directly in any application, without a prior assessment of safety and fairness concerns specific to the application."*

Ethical considerations and risks

> *"Flan-T5 is fine-tuned on a large corpus of text data that was not filtered for explicit content or assessed for existing biases. As a result the model itself is potentially vulnerable to generating equivalently inappropriate content or replicating inherent biases in the underlying data."*

Known Limitations

> *"Flan-T5 has not been tested in real world applications."*

Sensitive Use:

> *"Flan-T5 should not be applied for any unacceptable use cases, e.g., generation of abusive speech."*

22.10.2022

"... should not be used directly in any application ..."

Interactive Model Card (Crisan et al. 2022)

Design
Standard Model Card Details
Visually Separate from interactive components

Skepticism & Sensemaking
Define subpopulations, add examples or own data

Guidance
Definitions and explanations with external links

Design
Collapsable pointers to interaction options

Skepticism & Sensemaking
Interactive legend for subpopulation companions

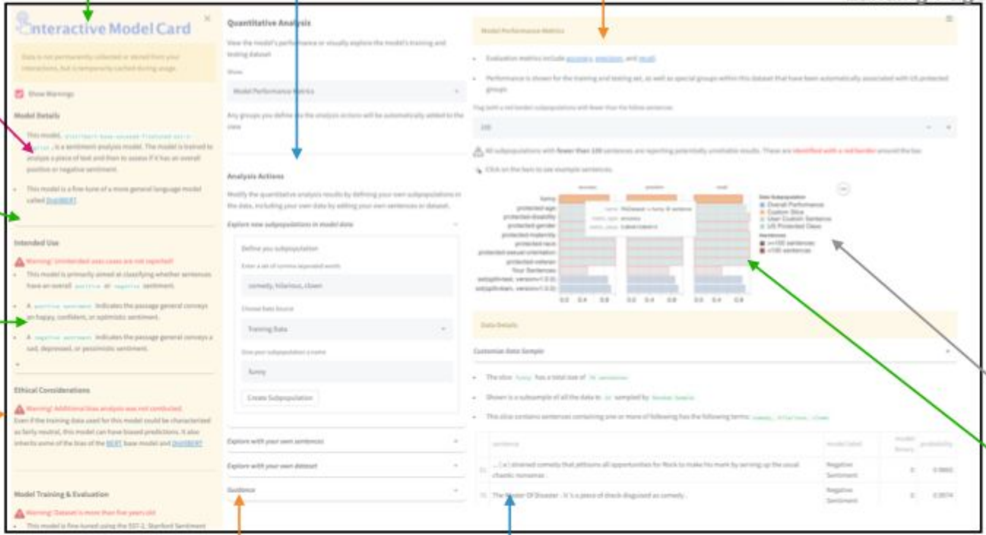
Stakeholders
Use plain language

Design
Overview summaries with details on demand

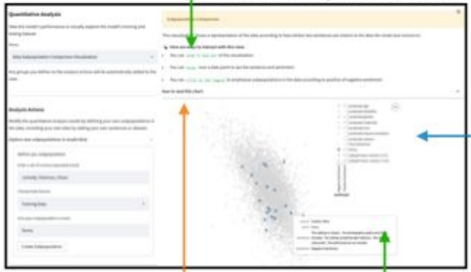
Design
Visual emphasis on important details

Guidance
Warnings and nudges

Landing Page



Data Comparison Panel



Guidance
Collapsible explainer of chart interpretation

Design
Details on hover

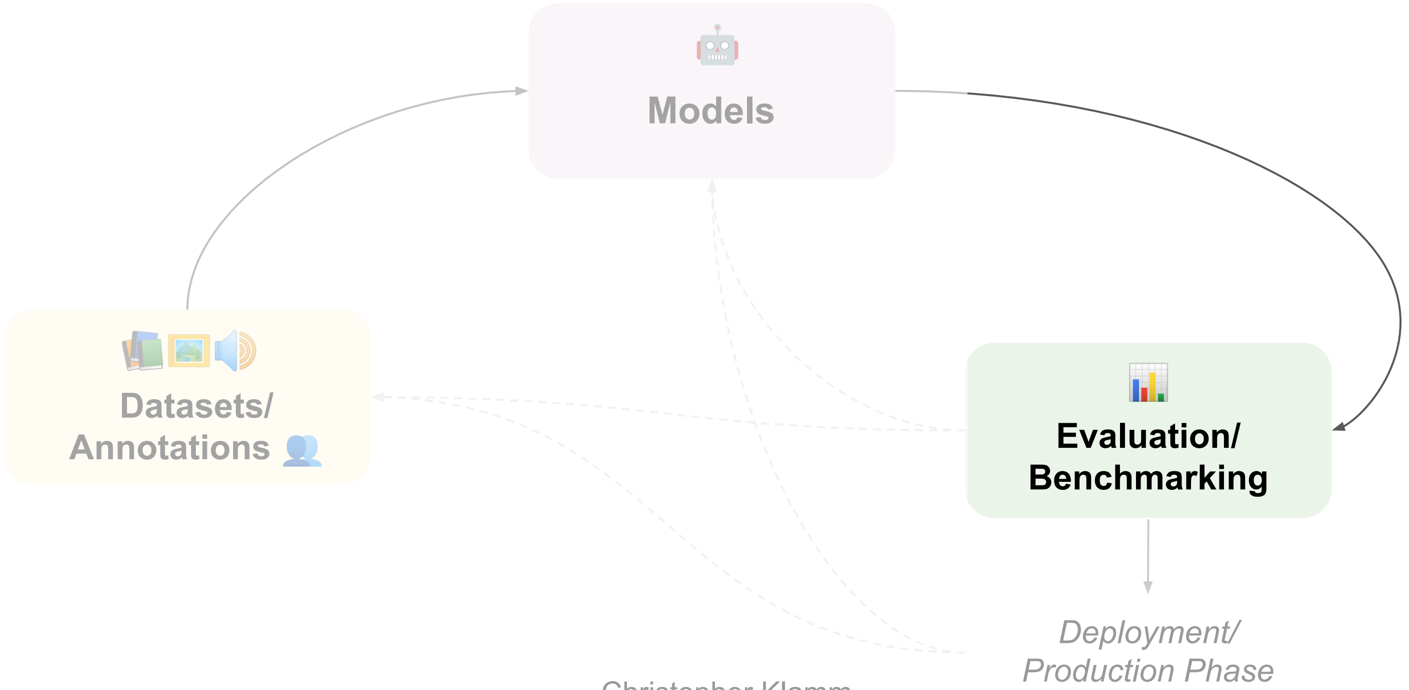
Trust and Safety
Pre-defined (and customizable) analysis on at risk groups

Design
Visually emphasize subpopulations with small sizes

Guidance
Additional guidance documentation on metrics, the model, and next steps

Skepticism & Sensemaking
Drill down into examples

Machine Learning Lifecycle/ Pipeline



Evaluation/ Benchmarking

Experimental Results Checklist (Dodge et al. 2019)

"[...] test-set performance scores alone are insufficient for drawing accurate conclusions about which model performs best." (Dodge et al. 2019: 1)

✓ **For all reported experimental results**

- ☐ Description of computing infrastructure
- ☐ Average runtime for each approach
- ☐ Details of train/validation/test splits
- ☐ Corresponding validation performance for each reported test result
- ☐ A link to implemented code

✓ **For experiments with hyperparameter search**

- ☐ Bounds for each hyperparameter
- ☐ Hyperparameter configurations for best-performing models
- ☐ Number of hyperparameter search trials
- ☐ The method of choosing hyperparameter values (e.g., uniform sampling, manual tuning, etc.) and the criterion used to select among them (e.g., accuracy)
- ☐ Expected validation performance, as introduced in §3.1, or another measure of the mean and variance as a function of the number of hyperparameter trials.

Dodge et al. 2019

Dodge et al. 2019

Computing infrastructure	GeForce GTX 1080 GPU
Number of search trials	50
Search strategy	uniform sampling
Best validation accuracy	40.5
Training duration	39 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	50	50
patience	10	10
batch size	64	64
embedding	GloVe (50 dim)	GloVe (50 dim)
encoder	ConvNet	ConvNet
max filter size	*uniform-integer*[3, 6]	4
number of filters	*uniform-integer*[64, 512]	332
dropout	*uniform-float*[0, 0.5]	0.4
learning rate scheduler	reduce on plateau	reduce on plateau
learning rate scheduler patience	2 epochs	2 epochs
learning rate scheduler reduction factor	0.5	0.5
learning rate optimizer	Adam	Adam
learning rate	*loguniform-float*[1e-6, 1e-1]	0.0008

Table 2: SST (fine-grained) CNN classifier search space and best assignments.

📊 Benchmarks

“Progress in NLP has traditionally been measured through a **selection of task-level datasets** that gradually became accepted benchmarks [...]” (Kiela et al. 2021: 2)

→ new **dynamic benchmarks** (as “living entities”), e.g., Dynabench (Kiela et al. 2021)

⚠️ “On multiple benchmark setups [...], we show that the relative **performance of algorithms may be altered significantly simply by choosing different benchmark tasks**, highlighting the fragility of the current paradigms [...]” (Dehghani et al. 2021)

⚠️ “[...] ‘progress’ are **largely defined by performance on datasets**” but they often share “representational concerns”, “annotation artifacts”, under-specified data selection, reuse for unintended purpose, lack of standards etc. (Paullada et al. 2020)

→ a need for **more comprehensive evaluations** e.g., carbon footprint, model size, fairness, robustness, etc. (see “A Critique of NLP Leaderboards”, Ethayarajh/ Jurafsky 2020)

📊 Benchmark Checklist (Reimers 2022)

- **What is the intended use-case?** predicting a label, ranking of results, …
- **Costs of Errors**
  - Research treats all errors often with equal costs
  - In production this is seldom the case
- **Human upper bound**
  - How good are humans in this task?
  - When creating a new dataset: Spend many cycles to improve human agreement
- **What else is important?**
  - Inference speed
  - Robustness
- **A benchmark must evolve**
  - As models evolve, our benchmarks must evolve!
  - Stop using outdated benchmarks
- **Restrict number of submissions**
  - The more experiments we run on a benchmark, the less likely we can trust the numbers
  - Only allow evaluation on test set in very rare cases!
  - Have a dev dataset for model development
  - If possible: use an “out-of-domain” test dataset
- **Temporal split:** Test data should be the most recent, train data the oldest
- **Diversity:** Don’t test only on one task / domain etc.
- **Look for biases:**
  - What biases does your dataset have?
  - What biases does your benchmark has? e.g. only sentence pair comparison tasks

Reimers 2022: 25-27

📊 Benchmark Checklist (Dehghani et al. 2021 based on Gebru et al. 2021)

Benchmarking checklist for reviewers and area chairs

- ☐ If there is written dissatisfaction about the author's choice of baselines, tasks, or benchmarks in the reviews, are there rationals beyond the fact that these requested datasets are "must-have" benchmarks?
- ☐ Are the reviews considering potential benefits like efficiency, fairness, and simplicity of the proposed model outside the commonly evaluated performance metrics (e.g., accuracy)?
- ☐ Are there any negative points in the reviews due to the paper proposing a method that deviates from the current trend/hype. If so, are there rational justifications for this?
- ☐ If the reviews penalizing the paper due to the proposed method not performing well only on a subset of tasks, is there enough logical elaboration on such criticism in the reviews?
- ☐ Are the reviews assessing the evaluation strategy in terms of studying the effect of different sources of variance (e.g., multiple splits, multiple random seeds, etc.)?
- ☐ If there are analyses on statistical significance testing, are they appreciated in the reviews? If there is no such analysis, are there recommendations on this provided in the reviews?
- ☐ If the paper is claiming SOTA or improvements over baselines on a benchmark, are there ablations on how much such improvement is secured by the tricks that are not tied to the main contributions?
- ☐ If the reviews are asking for more experiments, analysis, or evaluation on more benchmarks, are the potential blockers are considered for such requests? E.g. those experiments being out of reach in terms of computing budget (pre-training or extremely large datasets).
- ☐ If the paper is proposing a new idea while deviating from the common paradigms, is the "out of the hype" thinking valued in the reviews as opposed to solely recognizing SOTA performance?

Dehghani et al. 2021: 34