

Dataset Card for Paloma

Evaluations of language models (LMs) commonly report perplexity on monolithic data held out from training. Implicitly or explicitly, this data is composed of domains—varying distributions of language. We introduce Perplexity Analysis for Language Model Assessment (Paloma), a benchmark to measure LM fit to 546 English and code domains, instead of assuming perplexity on one distribution extrapolates to others. Among 16 source curated in Paloma, we include two new datasets of the top 100 subreddits (e.g., r/depression on Reddit) and programming languages (e.g., Java on GitHub), both sources common in contemporary LMs.

Dataset Details

Evaluating with Paloma

In addition to the dataset hosted here, Paloma introduces guidelines for making perplexity results comparable across models and code that implements these guidelines with specific experimental controls.

Whether you are just evaluating an off-the-shelf model or preparing to conduct your own pretraining experiment from scratch, we recommend that you employ as much of our standardized code as possible to ensure the greatest level comparability with existing results.

How to conduct fully comparable pretraining experiments with Paloma

Dataset Description

Paloma aims to enable research on differences in LM fit over hundreds of domains by curating and standardizing the text datasets with the most fine-grained domains readily available from existing metadata.

We define two terms: *Sources* are as existing datasets (or curated subsets thereof) in use for research. *Domains* are fine-grained partitions of sources based on available metadata that attempt to surface a distinct and intuitive distribution of language (e.g., Wikipedia articles about visual arts or a subreddit for advice on PC builds). Paloma is derived from 16 sources. Where we curate previous fine-grained corpora, we inherit their operationalization of domains, ranging from the community-driven Wikipedia ontology to expert curation and automatic classification. Where we build our own fine-grained domains from Reddit and GitHub, we make similar use of metadata about subreddits and file extensions.

Curated by: Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge

Languages: We elect to focus just on the language modeling of English and code data.

License: The data subsets are licensed under the AI2 ImpACT License - Low Risk Artifacts, except as listed below. - Wikitext-103 - CC BY-SA - TwitterAAE - for research purposes only - Red Pajama - see license details - M2D2 - CC BY-NC

Paper: <https://arxiv.org/abs/2312.10523>

Dataset Sources

- Code
- Paloma 1B Baseline Models: Dolma, Pile, RedPajama, C4, mC4-en, Falcon-RefinedWeb

Uses

This benchmark is intended for use in evaluating language model fit to fine-grained domains.

Direct Use

This dataset should be used for evaluating the likelihood of text from a given domain by a language model.

Out-of-Scope Use

Note that the sources contained in this benchmark include varying licenses with differing restrictions (see License)

Dataset Structure

The sources in this dataset are each organized into their own subcorpus. This consists of a `val` and `test` split. Data within this is organized as files with lines separated JSON data where each line represents a document and its associated metadata. The type of metadata available varies from source to source, but each line contains at least a field `'text'` which contains the text of the document.

Dataset Creation

Curation Rationale

Perplexity is conventionally reported on held out data from a model’s training distribution or a small number of traditional test sets. Such monolithic evaluation ignores potential variation of model fit across different domains that LMs implicitly learn to model. We curate sources of fine-grained textual domains in Paloma to enable evaluation of language model fit to specific domains of text.

Source Data

Standard language modeling benchmarks Though it is common practice to evaluate on held out data from the pretraining corpus of a given model, we evaluate *across* several major pretraining corpora and standard language modeling benchmarks. We also break down performance per domain within the datasets that have multiple domains.

Source	Citation	Description
c4-en	Raffel et al (2019) via Dodge et al (2021)	Standard contemporary LM pretraining corpus automatically filtered from the April 2019 Common Crawl scrape
mc4-en	Xue et al (2021)	The English language portion of a pretraining corpus automatically filtered from 71 Common Crawl scrapes
Wikitext-103	Merity et al (2016)	A standard collection of verified “Good” and “Featured” articles on Wikipedia
Penn Tree Bank	Marcus et al (1999) via Nunes, Davide. (2020)	Classic Wall Street Journal benchmark with linguistic structure annotations omitted
RedPajama	Together Computer (2023)	A publicly available reproduction of the LLaMA (Touvron et al., 2023) pretraining source mixture, combining large amounts of webscraped text with smaller curated sources
Falcon-RefinedWeb	Penedo et al (2023)	A corpus of English sampled from all Common Crawl scrapes until June 2023, more aggressively filtered and deduplicated than c4 and mc4-en
Dolma v1.5	Soldaini et al. (2023)	A three trillion token corpus that samples sources commonly used to train LMs in order to enable open research on pretraining data

Fine-grained domain benchmarks Where typical pretraining corpora offer at most tens of labeled domains usually based on where the data is sourced, we examine datasets with up to an order of magnitude more domains. Existing datasets (M2D2 and c4 100 Domains) and datasets we curate from Dolma v1.5 use metadata to define hundreds of domains over Wikipedia, Semantic Scholar, Common Crawl, Reddit, and Github data. These include diverse domains from *Culture and the arts: Performing arts*, a topic on Wikipedia, to *r/depression*, a forum on Reddit for mental health support.

Source	Citation	Description
M2D2 S2ORC	Reid et al (2022)	Papers from Semantic Scholar grouped by hierarchical academic field categories
M2D2 Wiki	Reid et al (2022)	Wikipedia articles grouped by hierarchical categories in the Wikipedia ontology
c4 100 Domains	Chronopoulou et al (2021)	Balanced samples of the top 100 URL domains in C4
Dolma 100 Subreddits	Soldaini et al. (2023)	Balanced samples of the top 100 Subreddits from the Dolma Reddit subset
Dolma 100 Programming Languages	Kocetkov et al. (2022) via Soldaini et al. (2023)	Balanced samples of the top 100 programming languages from the Dolma Stack subset

Disparities between speech communities LMs today primarily process dominant dialects in countries, such as the US, where they are most often trained and deployed. Even within English, hundreds of millions of people around the world speak other dialects that have been shown to be underserved by existing models. As a starting point for measuring disparities between dialects, we include TwitterAAE two corpora representing African-American and White-aligned English, automatically classified via geolocation information and demographic census statistics.

Source	Citation	Description
TwitterAAE	Balch et al. (2016) via Liang et al (2022)	Balanced sets of tweets classified as African American or White aligned English

Fringe sources previously studied for problematic discourse Text from some fringe online communities has been shown to contain larger proportions of hate speech and toxicity than more mainstream sources. Measuring perplexity on Manosphere, Gab, and 4chan characterises model exposure to distinct social contexts in which toxic language arises.

Source	Citation	Description
Manosphere Corpus	Ribeiro et al (2020)	9 forums where a set of related masculinist ideologies developed over the 2000s and 2010s

Source	Citation	Description
Gab Corpus	Zannettou et al (2018)	Data from 2016-18 from an alt-right, free-speech-oriented social media platform shown to contain more hate speech than mainstream platforms
4chan Corpus	Papasavva et al (2020)	Data from 2016-19 from a politics subforum of an anonymity-focused forum found to contain among the highest rates of toxic content

Data Collection and Processing The data in Paloma are sampled from existing sources. Most often perplexity evaluation data is subsampled uniformly over the original distribution of domains in a source, resulting in more or less tokens from each domain in the evaluation data based on how well represented they are in the corpus. We instead employ stratified sampling, in which all sources with marked domains are partitioned by domain and a uniform sample of the same size is taken from each partition. Specifically, documents are sampled from each domain until a target number of tokens is reached. This helps ensure that no domains are lost or very small after subsampling.

In social media domains with additional metadata that is typically displayed along with posts, we format metadata such as timestamps into the document 'text' field. Where information is available about how threads of posts are connected, documents in that domain contain all posts in a given thread.

Additional details on source specific processing are available in our paper.

Who are the source data producers? Text data from each of the sources curated in Paloma is created by varying sets of original authors. Some sources are collected from users of specific internet fora such as specific subreddits. Other data is collected on the basis of expert or automated classification of demographic groups. Other data is collected from authors of archival material including scientific preprints, Wikipedia, and code repositories. Lastly, data sampled from standard pretraining corpora comes from authors collected through automatic webscraping and large scale sampling of archival sources, making it difficult to recover much specific information about these authors.

Annotation process No annotation is done on this data.

Who are the annotators? No annotation is done on this data.

Personal and Sensitive Information Sources in Paloma may contain personally identifiable information (PII). No attempt is made to measure or remove this information for the following reason: Paloma provides a small subsample of already publicly available data. The small size of this subsample renders this data less useful for aggregation of PII information than the already available public sources which we subsample.

Bias, Risks, and Limitations

It is beyond the scope of any one group of researchers to prescribe an exhaustive set of domains that should be examined for a LM. Rather Paloma brings together a substantial selection of domains that are identifiable from already available metadata to demonstrate the kinds of analyses possible with hundreds of domains and rigorous experimental controls. Different research goals will motivate different definitions and selections of domains, but other researchers can apply the guidelines we detail in our paper to novel fine-grained domains suitable for their research questions. One of the key advantages of evaluating a model by its fit to a collection of text representing a domain is that such domains can be identified not just by researchers who study LMs. We hope future work will identify many more domains that no one discipline would think to look at.

Interpreting language model fit to domains also poses challenges. Instead of relying on LM fit to represent alignment to a domain’s human salient features, we examine anomalies in domain fit to deepen understanding of language modeling dynamics and illuminate gaps in existing approaches to evaluation.

Also, some domains in Paloma appear in multiple sources, such as academic papers. Though Dolma and RedPajama process academic papers differently, the subcorpora on academic papers in each source represent different approximations of the same or very similar domains. However for the sake of simplicity, we make the reductive assumption of counting all 546 domains in Paloma as fully distinct.

Recommendations

In our paper we outline guidelines for evaluating language model fit. We encourage users of Paloma to adopt these experimental controls for metric variance when subsampling, benchmark contamination, differing tokenization, training data order, and evaluation data format.

Citation

BibTeX:

```
@article{Magnusson2023PalomaAB,  
  title={Paloma: A Benchmark for Evaluating Language Model Fit},  
  author={Ian Magnusson and Akshita Bhagia and Valentin Hofmann and Luca Soldaini and A. Jha},  
  journal={ArXiv},  
  year={2023},  
  volume={abs/2312.10523},  
  url={https://api.semanticscholar.org/CorpusID:266348815}  
}
```

Dataset Card Contact

{ianm,jessed}@allenai.org